

Fast, "robust", and approximately correct: estimating mixed demand systems

Bernard Salanié
Frank A. Wolak

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP64/18

Fast, “Robust”, and Approximately Correct: Estimating Mixed Demand Systems*

Bernard Salanié[†] Frank A. Wolak[‡]

October 9, 2018

Abstract

Many econometric models used in applied work integrate over unobserved heterogeneity. We show that a class of these models that includes many random coefficients demand systems can be approximated by a “small- σ ” expansion that yields a straightforward 2SLS estimator. We study in detail the models of market shares popular in empirical IO (“macro BLP”). Our estimator is only approximately correct, but it performs very well in practice. It is extremely fast and easy to implement, and it accommodates to misspecifications in the higher moments of the distribution of the random coefficients. At the very least, it provides excellent starting values for more commonly used estimators of these models.

*We are grateful to Dan Akerberg, Steve Berry, Xiaohong Chen, Chris Conlon, Pierre Dubois, Jeremy Fox, Han Hong, Guy Laroque, Simon Lee, Arthur Lewbel, Thierry Magnac, Lars Nesheim, Ariel Pakes, Mathias Reynaert, Tobias Salz, Richard Smith, Pedro Souza, Frank Verboven, and Martin Weidner for their useful comments, as well as to seminar audiences at NYU, UCL, and Stanford Institute for Theoretical Economics (SITE). We also thank Zeyu Wang for excellent research assistance.

[†]Department of Economics, Columbia University, 1022 International Affairs Building, 420 West 118th Street, New York, NY 10027, bsalanie@columbia.edu.

[‡]Department of Economics and Program on Energy and Sustainable Development, Stanford University, Stanford CA 94305-6072, wolak@zia.stanford.edu.

Introduction

Many econometric models are estimated from conditional moment conditions that express the mean independence of random unobservable terms η and instruments Z :

$$E(\eta|Z) = 0.$$

In structural models, the unobservable term is usually obtained by solving a set of equations—often a set of first-order conditions—that define the observed endogenous variables as functions of the observed exogenous variables and unobservables. That is, we start from

$$G(y, \eta, \theta_0) = 0 \tag{1}$$

where y stands for the observed data and θ_0 for the unknown parameters, while the function G is to be known and can depend on observed exogenous variables. Then (assuming that the solution exists and is unique) we invert this system into

$$\eta = F(y, \theta_0)$$

and we seek an estimator of θ_0 by minimizing an empirical analog of a norm

$$\|E(F(y, \theta)Z)\|.$$

Inversion often is a step fraught with difficulties. Even when a simple algorithm exists, inversion is still costly: it must be done with a high degree of numerical precision, as errors may jeopardize the “outer” minimization problem. One alternative is to minimize an empirical analog of the norm

$$\|E(\eta Z)\|$$

subject to the structural constraints (1). This “MPEC approach” has met with some success in dynamic programming and empirical industrial organization (Sujd 2012, Dubé et al 2012.) It still requires solving a nonlinearly constrained, nonlinear minimization problem; convergence to a solution can be a challenging task in the absence of very good initial values.

We propose an alternative that derives a linear model from a very simple series expansion. To fix ideas, suppose that θ_0 can be decomposed into a pair (β_0, σ_0) , where

σ_0 is a scalar that we have reasons to think is not too far from zero. We rewrite (1) as

$$G(y, F(y, \beta_0, \sigma_0), \beta_0, \sigma_0) = 0.$$

We expand $\sigma \rightarrow F(y, \beta_0, \sigma)$ in a Taylor series around 0 and re-write $F(y, \beta_0, \sigma_0)$ as:

$$F(y, \beta_0, \sigma_0) = F(y, \beta_0, 0) + F_\sigma(y, \beta_0, 0)\sigma_0 + \dots + F_{\sigma\sigma\dots\sigma}(y, \beta_0, 0)\frac{\sigma_0^L}{L!} + O(\sigma_0^{L+1}),$$

where the subscript σ denotes a partial derivative with respect to the argument σ .

This suggests a sequence of “approximate estimators” that minimize the analogs of the following norms

$$\begin{aligned} & \|E(F(y, \beta, 0)Z)\| \\ & \|E((F(y, \beta, 0) + F_\sigma(y, \beta, 0)s)Z)\| \\ & \|E\left(\left(F(y, \beta, 0) + F_\sigma(y, \beta, 0)s + F_{\sigma\sigma}(y, \beta, 0)\frac{\sigma^2}{2}\right)Z\right)\| \\ & \dots \end{aligned}$$

If the true value σ_0 is not too large, one may hope to obtain a satisfactory estimator with the third of these “approximate estimators.” In general, this still requires solving a nonlinear minimization problem. However, suppose that the function F satisfies the following three conditions:

C1: $F_\sigma(y, \beta_0, 0) \equiv 0$

C2: $F(y, \beta, 0) \equiv f_0(y) - f_1(y)\beta$ is affine in β

C3: the second derivative $F_{\sigma\sigma}(y, \beta, 0)$ does not depend on β .

Denote $f_2(y) = -F_{\sigma\sigma}(y, \beta, 0)$. Then we would minimize

$$\left\| E\left(\left(f_0(y) - f_1(y)\beta - f_2(y)\frac{\sigma^2}{2}\right)Z\right)\right\|.$$

Taking the parameters of interest to be (β_0, σ_0^2) , this is simply a two-stage least squares regression of $f_0(y)$ on $f_1(y)$ and $f_2(y)$ with instruments that are functions of Z . As this is a linear problem, the optimal¹ instruments Z^* are simply

$$Z^* = (E(f_1(y)|Z), E(f_2(y)|Z)).$$

¹In the sense of Amemiya (1975).

They could be estimated directly from the data using nonparametric regressions. Or more simply, we can include flexible functions of the columns of Z in the instruments used to compute the 2SLS estimates.

The resulting estimators of β_0 and σ_0^2 are only approximately correct, because they consistently estimate an approximation of the original model. On the other hand, they can be estimated in closed form using linear 2SLS. Moreover, because they only rely on limited features of the data generating process, they are “robust” in interesting and useful ways that we will explore later.

Conditions **C1–C3** extend directly to a multivariate parameter σ_0 . They may seem very demanding. Yet as we will show, under very weak conditions the Berry, Levinsohn, and Pakes (1975) (macro-BLP) model that is the workhorse of empirical IO satisfies all three. In this application, σ_0 is taken to be the square root of the variance–covariance matrix Σ of the random coefficients in the mixed demand model. More generally, we will characterize in Section 6.4 a general class of models with unobserved heterogeneity to which conditions **C1–C3** apply.

Our approach builds on “small- Σ ” approximations to construct successive approximations to the inverse mapping (from market shares to product effects). Kadane (1971) pioneered the “small- σ ” method. He applied it to a linear, normal simultaneous equation system and studied the asymptotics of k -class estimators² when the number of observations n is fixed and σ goes to zero. He showed that when the number of observations is large, under these “small- σ asymptotics” the k -class estimators have biases in σ^2 , and that their mean-squared errors differ by terms of order σ^4 . Kadane argued that small σ , fixed n asymptotics are often a good approximation to finite-sample distributions when the estimation sample is large enough.

The small- σ approach was used by Chesher (1991) in models with measurement error. Most directly related to us, Chesher and Santos-Silva (2002) used a second-order approximation argument to reduce a mixed multinomial logit model to a “heterogeneity adjusted” unmixed multinomial logit model in which mean utilities have additional terms. They suggested estimating the unmixed logit and using a score statistic based on these additional covariates to test for the null of no random variation in preferences. Like them, we introduce additional covariates. Unlike them, we develop a method to

²Which include OLS and 2SLS.

estimate jointly the mean preference coefficients and parameters characterizing their random variation; and we only use linear instrumental variables estimators. To some degree, our method is also related to that of Harding and Hausman 2007, who use a Laplace approximation of the integral over the random coefficients in a mixed logit model without choice-specific random effects. Unlike them, we allow for endogeneous prices; our approach is also much simpler³.

Section 1 presents the model popularized by Berry–Levinsohn–Pakes (1995) and discusses some of the difficulties that practitioners have encountered when taking it to data. We give a detailed description of our algorithm in section 2; readers not interested in the derivation of our formulæ in fact can jump directly to our Monte Carlo simulations in section 7. The rest of the paper justifies our algorithm (sections 3 and 4); studies its properties (section 5); and discusses a variety of extensions (section 6).

1 The macro-BLP model

Our leading example is taken from empirical IO. Much work in this area is based on market share and price data. It has followed Berry et al (1995—hereafter BLP) in specifying a mixed multinomial logit model with product-level random effects that deals with the endogeneity of prices implied by these product-level random effects. BLP use a Generalized Method Moments (GMM) estimator that relies on the mean independence of the product-level random effects and a set of instruments.

To fix ideas, we define “the standard model” as follows⁴. Let J products be available on each of T markets. Each market contains an infinity of consumers who choose one of J products. Consumer i in market t is assumed to derive utility

$$\mathbf{X}_{jt}(\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_i) + \xi_{jt} + u_{ijt}$$

from choosing product j . There is also a good 0, the “outside good”, whose utility for consumer i is simply u_{i0t} . The random variables $\boldsymbol{\epsilon}$ represent individual variation in tastes for observed product characteristics, while the \mathbf{u} stand for idiosyncratic errors

³Harding and Hausman Ketz (2018) builds on a quadratic expansion in $\sigma_0 = 0$ to derive asymptotic distributions when the true σ_0 is on the boundary.

⁴While some of our exposition relies on it for simplicity, our methods apply to a more general model— see section 6.4.

observed by the individual, but unobserved by the econometrician. The vector $\boldsymbol{\epsilon}$ and \mathbf{u} are independent of each other, and of the covariates \mathbf{X} and product random effects $\boldsymbol{\xi}$. The vector $\mathbf{u}_{it} = (u_{i0t}, u_{i1t}, \dots, u_{iJt})$ is independently and identically distributed (iid) as standard type-I Extreme Value (EV); the product effects ξ_{jt} are unknown mean zero random variables conditional on a set of instruments, and the random variation in preferences $\boldsymbol{\epsilon}_i$ has a mean-zero distribution which is known up to its variance-covariance matrix $\boldsymbol{\Sigma}_0$. For instance, Berry et al. (1995) assume that the $\boldsymbol{\epsilon}_i$ are independent, identically distributed $N(0, \boldsymbol{\Sigma}_0)$ random vectors.

Some of the covariates in \mathbf{X}_{jt} may be correlated with the product-specific random effects. The usual example is a model of imperfect price competition where the prices firms set in market t depend on the value of the vector of unobservable product characteristics, $\boldsymbol{\xi}_t$.

The parameters to be estimated are the mean coefficients $\boldsymbol{\beta}_0$ and the variance-covariance matrix of the random coefficients $\boldsymbol{\Sigma}_0$. We collect them in $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$. The data available consists of the market shares (s_{1t}, \dots, s_{Jt}) and prices $(p_{1t}, \dots, p_{Jt})'$ of the J varieties of the good, of the covariates \mathbf{X}_t , and of additional instruments \mathbf{Z}_t , all for market t . Note that the market shares do not include information on the proportion S_{0t} of consumers who choose to buy good 0. Typically the analyst estimates this from other sources. Let us assume that this is done, so that we can deal with the augmented vector of market shares $(S_{0t}, S_{1t}, \dots, S_{Jt})$, with $S_{jt} = (1 - S_{0t})s_{jt}$ for $j \in \mathcal{J} = \{1, \dots, J\}$.

The market shares for market t are obtained by integration over the variation in preferences $\boldsymbol{\epsilon}$: for $j \in \mathcal{J}$,

$$S_{jt} = E_{\boldsymbol{\epsilon}} \left[\frac{\exp(\mathbf{X}_{jt}(\boldsymbol{\beta} + \boldsymbol{\epsilon}) + \xi_{jt})}{1 + \sum_{k=1}^J \exp(\mathbf{X}_{kt}(\boldsymbol{\beta} + \boldsymbol{\epsilon}) + \xi_{kt})} \right] \quad (2)$$

and $S_{0t} = 1 - \sum_{j=1}^J S_{jt}$.

Berry et al. (1995) assume that

$$E(\xi_{jt} | \mathbf{Z}_{jt}) = \mathbf{0}$$

for all $j \in \mathcal{J}$ and t . The instruments \mathbf{Z}_{jt} may for instance be the characteristics of competing products, or cost-side variables. The procedure is operationalized by showing that for given values of $\boldsymbol{\theta}$, the system (2) defines an invertible mapping⁵ in

⁵See Berry (1994).

\mathbb{R}^J . Call $\Xi(\mathbf{S}_t, \mathbf{X}_t, \boldsymbol{\theta})$ its inverse; a GMM estimator obtains by choosing functions \mathbf{Z}_{jt}^* of the instruments and minimizing a well-chosen quadratic norm of the sample analogue of:

$$E(\Xi(\mathbf{S}_t, \mathbf{X}_t, \boldsymbol{\theta})\mathbf{Z}_{jt}^*)$$

over $\boldsymbol{\theta}$.

These models have proved very popular; but their implementation has faced a number of problems. Much recent literature has focused on the sensitivity of the estimates to the instruments used in GMM estimation of the mixed multinomial logit model. Reynaert–Verboven (2014) showed that using linear combinations of the instruments can lead to unreliable estimates of the parameters of interest. They recommend using the optimal instruments given by the Amemiya formula (1975):

$$\mathbf{Z}_{jt}^* = E\left(\frac{\partial \Xi}{\partial \boldsymbol{\theta}}(\mathbf{S}_t, \mathbf{X}_t, \boldsymbol{\theta}_0) | \mathbf{Z}_{jt}\right).$$

Since the Amemiya formula relies on a consistent first-step estimate of the parameters, this is still problematic. Gandhi-Houde (2016) propose “differentiation IVs” to approximate the optimal instruments for the parameters $\boldsymbol{\Sigma}$ of the distribution of the random preferences $\boldsymbol{\epsilon}$. They also suggest a simple regression to detect weak instruments. An alternative is to use the Continuously Updating Estimator to build up the optimal instruments as minimization progresses. Armstrong (2016) points out that instruments based on the characteristics of competing products achieve identification through correlation with markups. But when there are a large number of products, many models of the cost-side of the market yield markups just do not have enough variation, relative to sampling error. This can give inconsistent or just uninformative estimates⁶.

Computation has also been a serious issue. The original BLP approach used a “nested fixed point” (NFP) approach: every time the objective function to be minimized was evaluated for the current parameter values, a contraction mapping/fixed pointed algorithm must be employed to compute the implied product effects $\boldsymbol{\xi}_t$ from the observed market shares \mathbf{S}_t and current value of $\boldsymbol{\theta}$. This was both very costly and

⁶Instruments that shift marginal cost directly (if available) do not need variation in the markup to shift prices, and therefore do not suffer from these issues. Variation in the number of products per market may also be used to restore identification, data permitting.

prone to numerical errors that propagate from the nested fixed point algorithm to the minimization algorithm. Dubé et al (2012) proposed a nonlinearly-constrained, nonlinear optimization problem to estimate θ . Their simulations suggest that this “MPEC” approach often outperforms the NFP method, sometimes by a large factor. Lee and Seo (2015) proposed an “approximate BLP” method that inverts a linearized approximation of the mapping from ξ_t to S_t . They argue that this can be even faster than MPEC.

Petrin and Train (2010) have proposed a maximum likelihood estimator that replaces endogenous regressors with a control function. This circumvents the need to compute the implied value of ξ for each value of θ , but still requires solving a nonlinear optimization problem to compute an estimate of θ_0 . Solving a nonlinear optimization problem for a potentially large set of parameters is time-consuming and typically requires starting values in the neighborhood of the optimal solution, closed-form gradients, and careful monitoring of optimization algorithm by the analyst because the objective function is not globally concave. The method we propose in this paper completely circumvents the need to solve a nonlinear optimization problem.

Our estimator relies on an approximate model that is exactly valid when there is no random variation in preferences, and becomes a coarser approximation as the amplitude of random variation grows. As such, our estimator is *not* a consistent estimator of the parameters of the BLP model. On the other hand, it has some very real advantages that may tip the scale in its favor. First, it requires a single linear 2SLS regression that can be computed in microseconds with off-the-shelf software. Second, our estimator needs to assume very little about the form of the distribution of the random variation in preferences ϵ (beyond its limited amplitude), justifying the “robust” in our title—where the quotes reflect our awareness that we are taking some liberties with the definition of robustness. Finally, because our estimating equation is linear, computing the “optimal” instruments for our estimator is also straightforward.

For those who find the “approximate correctness” of our estimator unsatisfying, it at least yields “nearly consistent” starting values for the classical nested-fixed point and MPEC nonlinear optimization procedures at a minimal cost. It also provides useful diagnoses about how well different parameters can be identified with a particular model and dataset; and a very simple way to select between models, as we discuss below.

2 2SLS Estimation in the Standard BLP Model

For the reader primarily interested in applying our method, this section provides a step-by-step guide to implementing the estimator in the standard macro-BLP model. This requires some notation. The dimensions of the vectors and matrices are as follows:

- for each $j \in \mathcal{J}$ and t , \mathbf{X}_{jt} is a row vector with n_X components
- $\boldsymbol{\beta}$ is a column vector with n_X components
- for each i , $\boldsymbol{\epsilon}_i$ is a row vector with n_e components; in the standard model, $n_e \leq n_X$.

We denote \mathcal{I} the set of pairs of indices (m, n) such that the variance-covariance element $\Sigma_{mn} = \text{cov}(\epsilon_{im}, \epsilon_{in})$ is *not* restricted to be zero⁷. For notational simplicity, we also assume that we use all conditional moment restrictions:

$$E(\xi_{jt} | \mathbf{Z}_{jt}) = 0,$$

for $j \in \mathcal{J}$ and $t = 1, \dots, T$.

Our procedure runs as follows:

Algorithm 1. *FRAC estimation of the standard BLP model*

1. on every market t , augment the market shares from (s_{1t}, \dots, s_{Jt}) to $(S_{0t}, S_{1t}, \dots, S_{Jt})$
2. for every product-market pair $(j \in \mathcal{J}, t)$:
 - (a) compute the market-share weighted covariate vector $\mathbf{e}_t = \sum_{k=1}^J S_{kt} \mathbf{X}_{kt}$;
 - (b) for every (m, n) in \mathcal{I} , compute the “artificial regressor”

$$K_{mn}^{jt} = \left(\frac{X_{jtm}}{2} - e_{tm} \right) X_{jtn}.$$

- (c) for every $j = 1, \dots, J$, define $y_{jt} = \log(S_{jt}/S_{0t})$

⁷E.g. if $n_e = n_X$ and $\boldsymbol{\Sigma}$ is assumed to be diagonal, $\mathcal{I} = \{(1, 1), \dots, (n_X, n_X)\}$.

3. run a two-stage least squares regression of \mathbf{y} on \mathbf{X} and \mathbf{K} , taking as instruments a flexible set of functions of the columns of \mathbf{Z}
4. (optional) run a three-stage least squares (3SLS) regression across the T markets stacking the J equations for each product with a weighting matrix equal to the inverse of the sample variance of the residuals from step 3.

Ideally, the “flexible set of functions of the columns of \mathbf{Z} ” in step 3 should be able to span the space of the optimal instruments $E(\mathbf{X}|\mathbf{Z})$ and $E(\mathbf{K}|\mathbf{Z})$ for our approximate model. Alternatively, these optimal instruments can be estimated by a nonparametric regressions of each the column of \mathbf{X} on the columns of \mathbf{Z} .

As is well-known, misspecification of one equation of the model can lead to inconsistency in 3SLS parameter estimates of all equations of the model. It is therefore unclear whether Step 4 is worth the additional effort. We intend to explore it in future work.

It is important to note here that \mathbf{e} is *not* a simple weighted average, as the weights do not sum to one, but only to $(1 - S_{0t})$. To illustrate, if $X_{jtm} \equiv 1$ is the constant, then e_{tm} is $(1 - S_{0t})$ and the artificial regressor that identifies the corresponding variance parameter is

$$K_{mm}^{jt} = S_{0t} - \frac{1}{2}.$$

More generally, if $X_{jtn} = \mathbf{1}(j \in \mathcal{J}_0)$ is a dummy that reflects whether variety j belongs to group $\mathcal{J}_0 \subset \mathcal{J}$, then it is easy to see that the corresponding variance parameter is the coefficient of the artificial regressor

$$K_{nn}^{jt} = \mathbf{1}(j \in K) \left(\frac{1}{2} - S_{\mathcal{J}_0t} \right)$$

where $S_{\mathcal{J}_0t}$ is the market share of group \mathcal{J}_0 on market t .

3 Second-order Expansions

The rest of the paper justifies algorithm 1 and discusses extensions. We first derive the small- σ expansions of the introduction.

We start from a specification of the utility of variety j for consumer i on market t as

$$\mathbf{X}_{jt}\beta + g(\mathbf{X}_{jt}, \boldsymbol{\epsilon}_i) + \xi_{jt} + u_{ijt} \quad (3)$$

for $j \in \mathcal{J}$; and $U_{i0t} = u_{i0t}$. Define the vectors $\mathbf{u}_{it} = (u_{i0t}, u_{i1t}, \dots, u_{iJt})$; $\mathbf{X}_t = (\mathbf{X}_{1t}, \dots, \mathbf{X}_{Jt})$; and $\boldsymbol{\xi}_t = (\xi_{1t}, \dots, \xi_{Jt})$. We assume that

1. the random terms $\boldsymbol{\epsilon}_i$ are i.i.d. across i with finite variance;
2. they are distributed independently of $(\mathbf{X}_t, \boldsymbol{\xi}_t)$;
3. $Eg(\mathbf{X}_{jt}, \boldsymbol{\epsilon}_i) = 0$ for all \mathbf{X}_{jt} ;
4. the random vectors \mathbf{u}_{it} are i.i.d. across i and t ; and they are distributed independently of $(\boldsymbol{\epsilon}_i, \mathbf{X}_t, \boldsymbol{\xi}_t)$.

These assumptions are all standard, except for the third one which is only a mild extension of the usual normalization $E\boldsymbol{\epsilon}_i = 0$. They allow for any type of codependence between the product effects $\boldsymbol{\xi}_t$ and the covariates \mathbf{X}_t . Note that the additive separability between β and $\boldsymbol{\epsilon}$ is not as strict as it seems. If for instance we start from a multiplicative model with utilities

$$\sum_{k=1}^{n_X} X_{jtk} \beta_k \zeta_{ki} + \xi_{jt} + u_{ijt}$$

we can always redefine $\epsilon_{ki} = \beta_k(\zeta_{ki} - 1)$ to recover (3).

Our crucial assumption, which we maintain throughout, is that the utilities are affine in β , and additive in the product effects $\boldsymbol{\xi}$ and in the idiosyncratic terms \mathbf{u} . On the other hand, we allow for any kind of distribution for $\boldsymbol{\epsilon}_i$ and \mathbf{u}_{it} . This encompasses most empirical specifications used, as well as many more. We will refer to three special cases for illustrative purposes:

1. The *standard model*, also known as the mixed multinomial logit model, has $g(\mathbf{X}, \boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\epsilon}$; and the vector \mathbf{u}_{it} is distributed as standard type-I EV iid.
2. The *standard binary model* (or mixed logit model) further imposes $J = 1$.
3. The *standard symmetric model* is a standard model with $\boldsymbol{\epsilon}$ distributed symmetrically around $\mathbf{0}$;

4. The *standard Gaussian model* is a standard model with ϵ jointly normal. It is probably the most commonly used in applications of the macro-BLP method.
5. Finally, the *standard Gaussian binary model* imposes both 2 and 4.

In order to do small- σ expansions, we need to introduce a scale parameter σ . We do this with Assumption 1, which fits the usual understanding of what a scale parameter is⁸ and also imposes that all moments of ϵ are finite-valued. The most common specification of the "macro-BLP" model has a Gaussian ϵ and of course obeys Assumption 1.

Assumption 1. *For some integer $L \geq 2$, all moments of order $1 \leq l \leq L + 1$ of the vector ϵ are finite; they are of order l in some non-negative scalar σ . The first moment is zero: $E\epsilon = \mathbf{0}$. We denote $\Sigma = E\epsilon\epsilon'$ the variance-covariance matrix of ϵ , and μ_l (for $l \geq 3$) its (uncentered) higher order moments.*

It will be convenient to write $\epsilon \equiv \sigma \mathbf{B}\mathbf{v}$ with \mathbf{v} a random vector of mean zero and variance identity, so that $\sigma \mathbf{B}$ is a square root of the variance-covariance matrix of ϵ : $\Sigma = \sigma^2 \mathbf{B}\mathbf{B}'$. We only use this decomposition for intermediate results. Note that \mathbf{B} is an $n_e \times n_v$ matrix, where \mathbf{v} is a row vector with n_v components. Our final expansions do not depend on how σ and \mathbf{B} are normalized, and we won't need to specify it.

We drop the index t from the notation in most of this section as we will only need to deal with one market at a time.

3.1 Second-order Expansions in the Standard Model

Much of the rest of the remainder of the paper focuses on the standard model, where the \mathbf{u} 's have iid Type I extreme value distributions. We will show in section 6.1 how to extend our results to more general distributions.

Recall that in the standard model, market shares are given by (2). If the scale parameter σ was zero, inverting (2) would simply give us

$$\xi_j = \log \frac{S_j}{S_0} - \mathbf{X}_j \boldsymbol{\beta} \text{ for } j \in \mathcal{J}. \quad (4)$$

⁸In principle it should be possible to use several scale parameters, say σ_1 for one part of the variance-covariance matrix and σ_2 for another one.

This is the starting point of the contraction algorithm described in Berry et al (1995).

Now let σ be positive. With $\boldsymbol{\epsilon} = \sigma \mathbf{B}\mathbf{v}$, a Taylor expansion of (4) at $\sigma = 0$ would give (assuming that the expansion is valid⁹)

$$\xi_j = \log \frac{S_j}{S_0} - \mathbf{X}_j \boldsymbol{\beta} + \sum_{l=1}^L a_{lj}(\mathbf{S}, \mathbf{X}, \boldsymbol{\beta}) \frac{\sigma^l}{l!} + O(\sigma^{L+1}). \quad (5)$$

In this equation, \mathbf{X} regroups the covariates of all products and \mathbf{S} is the vector of market shares. Market-share weighted sums will play a crucial role in what follows:

Definition 1. For any J -dimensional vector \mathbf{T} of J components, we define the scalar

$$e_{\mathbf{S}} \mathbf{T} = \sum_{k=1}^J S_k T_k.$$

By extension, if \mathbf{m} is a matrix with J columns $(\mathbf{m}_1, \dots, \mathbf{m}_J)$, we define the vector

$$e_{\mathbf{S}} \mathbf{m} = \sum_{k=1}^J S_k \mathbf{m}_k.$$

Finally, we denote $\hat{T}_j = T_j - e_{\mathbf{S}} \mathbf{T}$ and $\hat{\mathbf{m}}_j = \mathbf{m}_j - e_{\mathbf{S}} \mathbf{m}$.

Note that we are using the *observed* market shares of the J goods, so that these weighted sums are very easy to compute from the data. It is important to emphasize that the operator $e_{\mathbf{S}}$ is *not* an average, as the augmented market shares S_k for $k \in \mathcal{J}$ do not sum to one but to $(1 - S_0)$. Similarly, the \hat{T}_j terms are not residuals, and $e_{\mathbf{S}} \hat{\mathbf{T}} \neq 0$ in general.

Our first goal is to find explicit formulæ for the coefficients a_{lj} in (5). While this can be done at a high level of generality, let us start with a result that covers a large majority of applications.

In the standard model, $g(\mathbf{X}_j, \boldsymbol{\epsilon})$ is simply $\mathbf{X}_j \boldsymbol{\epsilon}$. Denote $\mathbf{x}_j = (\mathbf{X}_j \mathbf{B})'$, a vector of n_v components; and \mathbf{x} the matrix whose J columns are $(\mathbf{x}_1, \dots, \mathbf{x}_J)$. Then

$$g(\mathbf{X}_j, \boldsymbol{\epsilon}) = \sigma \mathbf{x}_j' \mathbf{v}.$$

We first derive the second-order expansion in σ in the standard model.

⁹We return to this point in section 5.1.

Theorem 1 (Intermediate expansion in the standard model). *In the standard model,*

- (i) *the a_{lj} coefficients only depend on \mathbf{S} and on \mathbf{x} ;*
- (ii) *the first-order coefficients are zero: $a_{1j} \equiv 0$ for all j ;*
- (iii) *the second-order coefficients are given by*

$$a_{2j} = 2\mathbf{x}_j \cdot e_{\mathbf{S}}\mathbf{x} - \|\mathbf{x}_j\|^2 = -\mathbf{x}_j \cdot \left(\mathbf{x}_j - 2 \sum_{k=1}^J S_k \mathbf{x}_k \right); \quad (6)$$

- (iv) *in the standard symmetric model, $a_{lj} = 0$ for all j and odd $l \leq L$. Therefore if $L \geq 3$,*

$$\xi_j = \log \frac{S_j}{S_0} - \mathbf{X}_j \boldsymbol{\beta} + \frac{a_{2j}}{2} \sigma^2 + O(\sigma^4). \quad (7)$$

Proof. See Appendix A. □

3.2 The Artificial Regressors in the Standard Model

When truncated of its remainder term, equation (7) becomes linear in the parameters $(\boldsymbol{\beta}, \sigma^2)$. The coefficients a_{2j} , however, are quadratic combinations of the vectors \mathbf{x}_j , which are themselves linear in the unknown coefficients of the matrix \mathbf{B} . Fortunately, the formula that gives a_{2j} can be transformed so that it becomes linear in the coefficients of the variance-covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{\epsilon}$.

To see this, note that since $\mathbf{x}_k = \mathbf{B}'\mathbf{X}'_k$,

$$\mathbf{x}'_k \mathbf{x}_l = \mathbf{X}_k \mathbf{B} \mathbf{B}' \mathbf{X}'_l.$$

But since $\boldsymbol{\Sigma} = \sigma^2 \mathbf{B} \mathbf{B}'$, we have

$$\sigma^2 \mathbf{x}'_k \mathbf{x}_l = \sum_{m,n=1}^{n_X} \Sigma_{mn} X_{km} X_{ln} = \text{Tr}(\boldsymbol{\Sigma} \mathbf{X}_l \mathbf{X}'_k)$$

where $\text{Tr}(\cdot)$ is the trace operator.

Plugging this into (6) gives

$$\sigma^2 \frac{a_{2j}}{2} = \text{Tr} \left(\boldsymbol{\Sigma} \left(e_{\mathbf{S}} \mathbf{X} - \frac{\mathbf{X}_j}{2} \right) \mathbf{X}'_j \right).$$

Define the $n_X \times n_X$ matrices \mathbf{K}^j by

$$\mathbf{K}^j = \left(\frac{\mathbf{X}_j}{2} - e_{\mathbf{S}} \mathbf{X} \right) \mathbf{X}'_j$$

so that we can also write $\sigma^2 \frac{a_{2j}}{2} = -\text{Tr}[\boldsymbol{\Sigma} \mathbf{K}^j]$. The matrices \mathbf{K}^j can be constructed straightforwardly from the covariates \mathbf{X} and the market shares \mathbf{S} . We call their elements the ‘‘artificial regressors’’, for reasons that will soon become clear. Given that $\boldsymbol{\Sigma}$ is symmetric,

$$\text{Tr}[\boldsymbol{\Sigma} \mathbf{K}^j] = \sum_{m=1}^{n_X} \Sigma_{mm} K_{mm}^j + \sum_{m < n} \Sigma_{mn} (K_{mn}^j + K_{nm}^j).$$

Additional a priori restrictions can be accommodated very easily. It is for instance common to restrict $\boldsymbol{\Sigma}$ to be diagonal. Then only n_X artificial regressors enter in this sum; moreover,

$$K_{mm}^j = \left(\frac{X_{jm}}{2} - \sum_{k=1}^J S_k X_{km} \right) X_{jm}.$$

If $\boldsymbol{\Sigma}$ is not diagonal, then we need to also use the artificial regressors

$$K_{mn}^j = \left(\frac{X_{jm}}{2} - \sum_{k=1}^J S_k X_{km} \right) X_{jn}.$$

To summarize, we have:

Theorem 2 (Final expansion in the standard model). *In the standard model,*

$$\xi_j = \log \frac{S_j}{S_0} - \mathbf{X}_j \boldsymbol{\beta} - \sum_{m=1}^{n_X} \Sigma_{mm} K_{mm}^j - \sum_{m < n} \Sigma_{mn} (K_{mn}^j + K_{nm}^j) + O(\|\boldsymbol{\Sigma}\|^{k/2}), \quad (8)$$

where $k = 4$ if the model is symmetric, and $k \geq 3$ otherwise; and the artificial regressors are given by

$$\begin{aligned} K_{mm}^j &= \left(\frac{X_{jm}}{2} - \sum_{k=1}^J S_k X_{km} \right) X_{jm} \\ K_{mn}^j + K_{nm}^j &= X_{jm} X_{jn} - \left(\sum_{k=1}^J S_j X_{km} \right) X_{jn} - \left(\sum_{k=1}^J S_j X_{kn} \right) X_{jm}. \end{aligned}$$

4 2SLS Estimation

Equation (8) is linear in the parameters of interest $\theta = (\beta, \Sigma)$, up to the remainder term. This immediately suggests neglecting the remainder term and estimating the approximate model $\xi_j = \log \frac{S_j}{S_0} - \mathbf{X}_j \beta - \text{Tr}[\Sigma \mathbf{K}^j]$.

More precisely, assume we are given a sample of T markets, and instruments \mathbf{Z}_{jt} such that $E(\xi_{jt} | \mathbf{Z}_{jt})$ for all j and t . Then our proposed estimator $\hat{\theta}$ fits the approximate *linear* set of conditional moment restrictions:

$$E \left(\log \frac{S_{jt}}{S_{0t}} - (\mathbf{X}_{jt} \beta + \text{Tr}[\Sigma \mathbf{K}^{jt}]) | \mathbf{Z}_{jt} \right) = 0$$

which only differs from the original model by a term of order σ^3 (or σ^4 if the distribution of ϵ is symmetric). This can simply be done by choosing vector functions \mathbf{Z}_{jt}^* of the instruments and running two-stage least squares: for each $j = 1, \dots, J$, on the sample $t = 1, \dots, T$, we linearly regress $\log(S_{jt}/S_{0t})$ on \mathbf{X}_{jt} and the relevant¹⁰ variables \mathbf{K}^{jt} , with instruments \mathbf{Z}_{jt}^* .

5 Pros and Cons of the 2SLS Estimation Approach

The drawback of our method is obvious: since this is only an approximate model, the resulting estimator $\hat{\theta}$ will not converge to the true values as the number of markets T goes to infinity. We discuss this in much more detail in section 5.1. For now, let us note that this drawback is tempered by several considerations. First, the number of markets available in empirical IO is typically small, so that finite-sample performance of the estimator is most relevant, and we will examine that in Section 7. More importantly, our estimator has several useful features. Let us list six of them:

1. because the estimator is linear 2SLS, computing it is extremely fast and can be done in microseconds with any of-the-shelf software.
2. we do not have to assume any distributional form for the random variation in preferences \mathbf{v} . This is a notable advantage over other methods: while they

¹⁰E.g. only the n_X variables K_{mm}^{jt} if Σ is restricted to be diagonal, or even a subset if some coefficients are non random.

yield inconsistent estimates if the distribution of \mathbf{v} is misspecified, our estimator remains consistent for the parameters of the approximate model.

3. computing the optimal instruments does not require any first-step estimate because the estimating equation is linear. We can just use a flexible set of functions of the columns of \mathbf{Z} that span the space of the optimal instruments $E(\mathbf{X}|\mathbf{Z})$ and $E(\mathbf{K}|\mathbf{Z})$.
4. even if the econometrician decides to go for a different estimation method, our proposed 2SLS estimates obtained should provide a set of very good initial parameter values for a nonlinear optimization algorithm.
5. the confidence regions on the estimates will give useful diagnoses about the strength of identification of the parameters, both mean coefficients $\boldsymbol{\beta}$ and their random variation $\boldsymbol{\Sigma}$. This would be very hard to obtain otherwise, except by trying different specifications.
6. there has been much interest in systematic specification searches in recent years; see e.g. Horowitz–Nesheim 2018 for a Lasso-based selection approach in discrete choice models. With our method any number of variants can be tried in seconds, and model selection is drastically simplified

5.1 The Quality of the Approximation

Ideally, we would be able to bound the approximation error in the expansion of ξ_j , and use this bound to majorize the error in our estimator. While we have not gone that far, we can justify the local-to-zero validity of the expansion in the usual way. We are taking a mapping

$$\mathbf{S} = G(\boldsymbol{\xi}, \mathbf{X}, \sigma)$$

that is differentiable in both $\boldsymbol{\xi}$ and σ ; inverting it to $\boldsymbol{\xi} = \boldsymbol{\Xi}(\mathbf{S}, \mathbf{X}, \sigma)$; and taking an expansion to the right of $\sigma = 0$ for fixed market shares \mathbf{S} and covariates \mathbf{X} . The validity of the expansion for small σ and fixed (\mathbf{X}, \mathbf{S}) depends on the invertibility of the Jacobian $G_{\boldsymbol{\xi}}$.

First consider the standard model. It follows from Berry 1994 that $G_{\boldsymbol{\xi}}$ is invertible if no observed market share hits zero or one. Applying the Implicit Function Theorem

repeatedly shows that in fact the Taylor series of $\boldsymbol{\xi}$ converges over some interval $[0, \bar{\sigma}]$ if all moments of $\boldsymbol{\epsilon}$ are finite; and that the expansion is valid at order L if the moments of $\boldsymbol{\epsilon}$ are bounded to order $(L + 1)$. Characterizing this range of validity is trickier. Figure 1 in Appendix B plots the first four coefficients of the expansion in $(\sigma X_1)^2$ for the standard Gaussian binary model (that is, the Gaussian mixed logit) with one covariate X_1 as market shares vary between zero and one. While this simple example can only be illustrative, we find the figure encouraging as to the practical range of validity of the approximation.

5.2 “Robustness”

Our expansions only rely on the properties of the derivatives of the logistic cdf $L(t) = \frac{1}{1+\exp(-t)}$ and on the first two moments of $\boldsymbol{\epsilon}$. This has a distinct advantage over competing methods: the lower-order moments of $\boldsymbol{\epsilon}$ can be estimated by 2SLS, and nothing more needs to be known about its distribution.

Suppose for instance that the analyst does not want to assume that $\boldsymbol{\epsilon}$ has a symmetric distribution. Then the coefficients a_{1j} are still zero, and the coefficients a_{2j} are unchanged. In the absence of symmetry, the approximate model is only valid up to $O(\sigma^3)$; but running Algorithm 1 may still provide very useful estimators of the elements of $\boldsymbol{\Sigma}$.

6 Extensions

Our technique can easily be extended to different models as long as the utility remains additive in the product effects $\boldsymbol{\xi}$. Moreover, the calculations of these and higher-order terms can be automated with the help of a symbolic algebra system.

6.1 The Two-level Nested Logit

Campioni (2018) applies a nonparametric approach to the choice among a very large set of products. He shows that the mixed logit specification forces the price elasticity to become “too small” at high price levels. This raises the question of the appropriate choice of a distribution for the idiosyncratic terms u_{ijt} .

For the mixed logit ($J = 1$), it is very easy to compute the artificial regressors for any distribution of the idiosyncratic terms; we give the formulæ in Appendix B.3. When $J > 1$, the space of possible distributions increases dramatically. The computations also become more complicated. Finally, estimating the additional parameters of the distribution of \mathbf{u} requires (simple) nonlinear optimization.

For illustrative purposes, we give the estimating equation for the two-level nested logit model. Assume that there is a nest for good 0, and K nests N_1, \dots, N_K for the varieties of the good. For $k = 1, \dots, K$, we denote λ_k the corresponding distribution parameter—with the usual interpretation that $(1 - \lambda_k)$ proxies for the correlation between choices within nest k , and that the multinomial logit model obtains when all $\lambda_k = 1$.

We denote the market share of nest k by $S_{N_k} = \sum_{j \in N_k} S_j$. Take any variable $\mathbf{T} = (T_0, T_1, \dots, T_J)$. We define the within-nest- k share-weighted average as

$$\bar{T}_k = \sum_{j \in N_k} \frac{S_j}{S_{N_k}} T_j.$$

Note in particular that $e_{\mathbf{S}} \mathbf{T} = \sum_{k=1}^K S_{N_k} \bar{T}_k$.

Appendix C derives the equivalent of (6): for $j \in N_k$,

$$a_{2j} = \mathbf{x}_j \cdot \left(2e_{\mathbf{S}} \mathbf{x} - \frac{\mathbf{x}_j}{\lambda_k} + 2 \frac{1 - \lambda_k}{\lambda_k} \bar{\mathbf{x}}_k \right) - \frac{1 - \lambda_k}{\lambda_k} \|\bar{\mathbf{x}}_k\|^2.$$

Reintroducing the market index t , the corresponding artificial regressors are

$$K_{mn}^{jt} = \left(\frac{X_{jt,m}}{2} - \frac{1 - S_{0t} \lambda_k}{1 - S_{0t}} e_{tm} - 2 \frac{1 - \lambda_k}{\lambda_k} \bar{X}_{kt,m} \right) \frac{X_{jt,n}}{\lambda_k} + \frac{1 - \lambda_k}{\lambda_k} \bar{X}_{kt,m} \bar{X}_{kt,n}$$

where as in section 3, $e_{tm} = \sum_{j=1}^J S_{jt} X_{jtm}$.

If the λ_k parameters are known, then our procedure becomes:

Algorithm 2. FRAC estimation of the two-level nested logit BLP model

1. on every market t , augment the market shares from (s_{1t}, \dots, s_{Jt}) to $(S_{0t}, S_{1t}, \dots, S_{Jt})$
2. for every product-market pair $(j \in \mathcal{J}, t)$:

- (a) compute the market-share weighted covariate vector $\mathbf{e}_t = \sum_{l=1}^J S_{lt} \mathbf{X}_{lt}$ and the within-nest weighted average covariate vector

$$\bar{\mathbf{X}}_{k(j),t} = \sum_{l \in N_{k(j),t}} \frac{S_{lt}}{S_{N_{k(j),t}}} \mathbf{X}_{lt}$$

where $k(j)$ is the nest that variety j belongs to.

- (b) for every (m, n) in \mathcal{I} , compute the “artificial regressor”

$$K_{mn}^{jt} = \left(\frac{X_{jt,m}}{2} - \frac{1 - S_{0t} \lambda_{k(j)}}{1 - S_{0t}} e_{tm} - 2 \frac{1 - \lambda_{k(j)}}{\lambda_{k(j)}} \bar{X}_{k(j),t,m} \right) \frac{X_{jt,n}}{\lambda_{k(j)}} + \frac{1 - \lambda_{k(j)}}{\lambda_{k(j)}} \bar{X}_{k(j),t,m} \bar{X}_{k(j),t,n}$$

- (c) define

$$y_{jt} = \log \frac{S_{N_{k(j),t}}}{S_{0t}} + \lambda_{k(j)} \log \frac{S_{jt}}{S_{N_{k(j),t}}}$$

3. run a two-stage least squares regression of \mathbf{y} on \mathbf{X} and \mathbf{K} , taking as instruments a flexible set of functions of \mathbf{Z}
4. (optional) run a three-stage least squares (3SLS) regression across the T markets stacking the J equations for each product with a weighting matrix equal to the inverse of the sample variance of the residuals from step 4.

If the parameters $\boldsymbol{\lambda}$ are not known, then things are slightly more complicated: the formulæ cannot be made linear in $\boldsymbol{\lambda}$, and there are no corresponding artificial regressors. Estimation of $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ requires numerical minimization over the $\boldsymbol{\lambda}$.

More general distributions in the GEV family could also be accommodated. As the nested logit example illustrates, there is a cost to it: the approximate model becomes nonlinear in some parameters¹¹. Note however that if there is reason to believe that the true distribution is close to the multinomial logit (say $\boldsymbol{\lambda} \simeq \mathbf{1}$ in the example above), then one can take expansions in the same way we did for the random coefficients and use a 2SLS estimate again.

¹¹Technically, condition **C1** in the introduction still holds, but conditions **C2** and **C3** do not.

6.2 Higher-order terms

In Appendix B, we study in more detail the standard binary model. For this simpler case, calculations are easily done by hand for lower orders of approximation, or using symbolic software for higher orders.

More generally, return to the standard model and assume (as is often done in practice) that the ϵ_m are independent across the covariates $m = 1, \dots, n_X$. We denote as before $\Sigma_{mm} = E(\epsilon_m^2)$, and μ_{lm} the expected value of ϵ_m^l for $l \geq 3$. Tedious calculations¹² show that the second- to fourth-order terms of the expansion in σ are

$$\xi_j = \log \frac{S_j}{S_0} - \mathbf{X}_j \beta + \sum_{l=2}^4 A_{lj} + O(\sigma^5)$$

with

$$A_{2j} = \sum_m X_{jm} (e_{\mathbf{S}} \mathbf{X}_m - X_{jm}/2) \Sigma_{mm};$$

$$A_{3j} = \sum_m X_{jm} \left(X_{jm} \frac{e_{\mathbf{S}} \mathbf{X}_m}{2} + \frac{e_{\mathbf{S}}(\mathbf{X}_m^2)}{2} - \frac{X_{jm}^2}{6} - (e_{\mathbf{S}} \mathbf{X}_m)^2 \right) \mu_{3m};$$

and

$$\begin{aligned} A_{4j} = & \sum_m \mu_{4m} X_{jm} \left((e_{\mathbf{S}} \mathbf{X}_m)^3 - (e_{\mathbf{S}} \mathbf{X}_m)(e_{\mathbf{S}}(\mathbf{X}_m^2)) - X_{jm} \frac{(e_{\mathbf{S}} \mathbf{X}_m)^2}{2} - \frac{X_{jm}^3}{24} \right. \\ & \left. + \frac{e_{\mathbf{S}}(\mathbf{X}_m^3)}{6} + X_{jm} \frac{e_{\mathbf{S}}(\mathbf{X}_m^2)}{4} + X_{jm}^2 \frac{e_{\mathbf{S}} \mathbf{X}_m}{6} \right) \\ & + \frac{A_{2j}^2}{2} + \sum_m \Sigma_{mm} X_{jm} \left(e_{\mathbf{S}}(\mathbf{A}_2 \mathbf{X}_m) + (e_{\mathbf{S}} \mathbf{A}_2) \left(\frac{X_{jm}}{2} - 2(e_{\mathbf{S}} \mathbf{X}_m) \right) \right). \end{aligned}$$

First consider the third-order term A_{3j} . It is a linear function of the unknown skewnesses μ_{3m} ; in fact it can be rewritten as

$$- \sum_m T_m^j \mu_{3m}$$

where we introduced new artificial regressors

$$T_m^j \equiv X_{jm} \left(\frac{X_{jm}^2}{6} + (e_{\mathbf{S}} \mathbf{X}_m)^2 - X_{jm} \frac{e_{\mathbf{S}} \mathbf{X}_m}{2} - \frac{e_{\mathbf{S}}(\mathbf{X}_m^2)}{2} \right).$$

¹²Available from the authors.

Algorithm 1 can be adapted in the obvious way to take possible skewness of ϵ into account. Note that the procedure remains linear in the parameters (β, Σ, μ_3) , for which it generates approximate estimates by 2SLS.

The fourth-order term, on the other hand, contains terms that are linear in the μ_{4m} (the first two lines of the formula) as well as terms that are quadratic in Σ (the last line). The first group suggests introducing more artificial regressors

$$Q_m^j \equiv X_{jm} \left((e_S \mathbf{X}_m)(e_S(\mathbf{X}_m^2)) - (e_S \mathbf{X}_m)^3 + X_{jm} \frac{(e_S \mathbf{X}_m)^2}{2} + X_{jm}^3/24 - \frac{e_S(\mathbf{X}_m^3)}{6} - X_{jm} \frac{e_S(\mathbf{X}_m^2)}{4} - X_{jm}^2 \frac{e_S \mathbf{X}_m}{6} \right),$$

whose coefficients are the μ_{4m} . The second group yields

$$- \sum_{m,n=1} \Sigma_{mm} \Sigma_{nn} W_{mn}^j$$

where new artificial regressors \mathbf{W} are assigned products of the elements of Σ . Estimating the resulting regression requires nonlinear optimization (albeit a very simple one).

6.3 Bias correction

If the analyst is willing to make more distributional assumptions, she can resort to bootstrap or asymptotic approximations to correct for the bias of our 2SLS estimators.

6.3.1 Bootstrapping

Once we have approximate estimators $\hat{\beta}$ and $\hat{\Sigma}$, we can use them to solve the market shares equations for estimates of the product effects ξ and bootstrap them, *provided* that we are willing to impose a distribution for \mathbf{v} (beyond the normalization of its first two moments.)

We use Berry inversion to solve for $\hat{\xi}_t$ in the system

$$S_{jt} = E_{\mathbf{v}} \frac{\exp \left(\mathbf{X}_{jt} \left(\hat{\beta} + \hat{\Sigma}^{1/2} \mathbf{v} \right) + \hat{\xi}_{jt} \right)}{1 + \sum_{k=1}^J \exp \left(\mathbf{X}_{kt} \left(\hat{\beta} + \hat{\Sigma}^{1/2} \mathbf{v} \right) + \hat{\xi}_{kt} \right)}.$$

Even if the ξ_{jt} that generated the data are iid across varieties and markets, our approximation method necessarily induces heteroskedasticity at least as the remainder terms in the approximations depend on covariates. Therefore we use wild bootstrap to resample from the $\hat{\xi}$. For any resample ξ^* , we simulate the market shares from

$$S_{jt}^* = E_{\mathbf{v}} \frac{\exp\left(\mathbf{X}_{jt} \left(\hat{\beta} + \hat{\Sigma}^{1/2} \mathbf{v}\right) + \xi_{jt}^*\right)}{1 + \sum_{k=1}^J \exp\left(\mathbf{X}_{kt} \left(\hat{\beta} + \hat{\Sigma}^{1/2} \mathbf{v}\right) + \xi_{kt}^*\right)}$$

and we use our 2SLS method to get new estimates β^*, Σ^* . Finally, we compute bias-corrected estimates in the usual way, e.g.

$$\beta^C = 2\hat{\beta} - \frac{1}{B} \sum_{b=1}^B \beta_b^*.$$

More generally, the resampled estimates can be used to estimate the distribution of $\hat{\beta}$ and $\hat{\Sigma}$.

6.3.2 Asymptotic bias correction

Another way to use the third- and fourth-order terms is as a corrective term: that is, we run 2SLS on the second-order expansion and we use the formulæ for the higher-order terms to evaluate the bias due to the approximation.

Denote $\theta = (\Sigma, \beta)$, and θ_0 its true value. Let $\hat{\theta}_2$ be our 2SLS estimator based on a second-order expansion. That is, we estimate the approximate model $E(\xi_2 \mathbf{Z}) = 0$ with instruments \mathbf{Z} and weighting matrix \mathbf{W} , where

$$\xi_{2j} = \log \frac{S_j}{S_0} - \mathbf{X}_j \beta - \text{Tr} \Sigma \mathbf{K}^j. \quad (9)$$

As the number of markets T gets large, $\hat{\theta}_2$ converges to the solution θ_2 of $E f_2(\theta_2) = \mathbf{0}$, with

$$f_2(\theta) \equiv \frac{\partial \xi_2}{\partial \theta}(\theta, \mathbf{X}, \mathbf{S})' \mathbf{Z} \mathbf{W} \mathbf{Z}' \xi_2(\theta, \mathbf{X}, \mathbf{S}).$$

Alternatively, we could have estimated the model using inversion or MPEC, with an “exact” ξ_∞ . Let λ_0 denote additional parameters of the model (such as higher-order moments of the distribution of ϵ) that are identified using the exact ξ_∞ but not¹³ with our approximate ξ_2 .

¹³If the only free parameters of the distribution of ϵ are the elements of Σ , then λ_0 will be empty.

Since by assumption $E(\boldsymbol{\xi}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0, \mathbf{X}, \mathbf{S})\mathbf{Z}) = 0$, a fortiori $E f_\infty(\boldsymbol{\theta}_0; \boldsymbol{\lambda}_0) = \mathbf{0}$ with

$$f_\infty(\boldsymbol{\theta}; \boldsymbol{\lambda}_0) \equiv \frac{\partial \boldsymbol{\xi}_\infty}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda}_0, \mathbf{X}, \mathbf{S})' \mathbf{Z} \mathbf{W} \mathbf{Z}' \boldsymbol{\xi}_\infty(\boldsymbol{\theta}, \boldsymbol{\lambda}_0, \mathbf{X}, \mathbf{S}).$$

The dominant term in the asymptotic bias is given by expanding $E f_\infty(\boldsymbol{\theta}; \boldsymbol{\lambda}_0)$ around $\boldsymbol{\theta} = \boldsymbol{\theta}_2$, keeping $\boldsymbol{\lambda}_0$ fixed. It is

$$\boldsymbol{\theta}_2 - \boldsymbol{\theta}_0 \simeq \left(E \frac{\partial f_\infty}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2; \boldsymbol{\lambda}_0) \right)^{-1} E f_\infty(\boldsymbol{\theta}_2; \boldsymbol{\lambda}_0).$$

Denote \mathbf{X} the matrix with terms X_{jm} and \mathbf{K} the matrix whose row $j = 1, \dots, J$ contains the artificial regressors K_{mn}^j . We define $\mathbf{e}_2(\boldsymbol{\theta}; \boldsymbol{\lambda}_0) = \boldsymbol{\xi}_\infty(\boldsymbol{\theta}; \boldsymbol{\lambda}_0) - \boldsymbol{\xi}_2(\boldsymbol{\theta})$, the approximation error on $\boldsymbol{\xi}$. Under any assumption about the parameters in $\boldsymbol{\lambda}_0$, we can compute the higher-order terms $\boldsymbol{\xi}_3, \boldsymbol{\xi}_4, \dots$ to approximate \mathbf{e}_2 . If for instance we maintain the assumption that the model is symmetric, we can approximate $\mathbf{e}_2 \simeq \boldsymbol{\xi}_4 - \boldsymbol{\xi}_2$.

Let us suppose then that we have a reliable estimator $\hat{\mathbf{e}}_2(\boldsymbol{\theta}; \boldsymbol{\lambda}_0)$ of $\mathbf{e}_2(\boldsymbol{\theta}; \boldsymbol{\lambda}_0)$. Define \mathbf{V} by the Cholesky decomposition $\mathbf{Z} \mathbf{W} \mathbf{Z}' = \mathbf{V} \mathbf{V}'$, so that \mathbf{V} is a (J, J) matrix. We prove in Appendix D that asymptotic bias correction yields the following formula:

$$\boldsymbol{\theta}_0 \simeq \boldsymbol{\theta}_2 + \begin{pmatrix} E(\mathbf{X}' \mathbf{V} \mathbf{V}' \mathbf{X}) & E(\mathbf{X}' \mathbf{V} \mathbf{V}' \mathbf{K}) \\ E(\mathbf{K}' \mathbf{V} \mathbf{V}' \mathbf{X}) & E(\mathbf{K}' \mathbf{V} \mathbf{V}' \mathbf{K}) \end{pmatrix}^{-1} \begin{pmatrix} E(\mathbf{X}' \mathbf{V} \mathbf{V}' \hat{\mathbf{e}}_2) \\ E(\mathbf{K}' \mathbf{V} \mathbf{V}' \hat{\mathbf{e}}_2) - E\left(\frac{\partial \hat{\mathbf{e}}_2}{\partial \boldsymbol{\Sigma}} \mathbf{V} \mathbf{V}' \boldsymbol{\xi}_2\right) \end{pmatrix}.$$

To interpret this formula, note that if $\hat{\mathbf{e}}_2$ did not depend on $\boldsymbol{\Sigma}$ the corrective term on the right-hand-side would simply be the 2SLS estimate of the regression of $\hat{\mathbf{e}}_2$ on (\mathbf{X}, \mathbf{K}) with instruments \mathbf{V} . In fact if we are only interested in bias correction on $\boldsymbol{\beta}_2$, we can simply keep the corresponding part of the 2SLS estimate. The correction on $\boldsymbol{\Sigma}_2$ has an additional term as higher order terms in the expansion of $\boldsymbol{\xi}$ typically depend on $\boldsymbol{\Sigma}$. (Recall from Theorem 1.(i) that they do not depend on $\boldsymbol{\beta}$.)

6.4 Other Models with Random Coefficients

Let us return to our original structural equations (1).

6.4.1 Quasi-linear Random Coefficients Models

Consider the following class of models, whose defining characteristic is that the error term $\boldsymbol{\eta}$ and the mean coefficients $\boldsymbol{\beta}$ only enter via a linear combination $\boldsymbol{\eta} - \mathbf{f}_1(\mathbf{y})\boldsymbol{\beta}$:

$$\mathbf{G}(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\beta}, \sigma) \equiv \mathbf{G}^*(\mathbf{y}, E_{\mathbf{v}}\mathbf{A}^*(\mathbf{y}, \boldsymbol{\eta} - \mathbf{f}_1(\mathbf{y})\boldsymbol{\beta}, \sigma\mathbf{B}\mathbf{v})). \quad (10)$$

where \mathbf{v} is unobserved heterogeneity distributed independently of \mathbf{y} and $\boldsymbol{\eta}$ and normalized by $E\mathbf{v} = \mathbf{0}$ and $V\mathbf{v} = \mathbf{I}$; and both functions \mathbf{G}^* and \mathbf{A}^* are assumed to be known.

Note that the macro-BLP model takes this form, with $\mathbf{y} = (\mathbf{S}, \mathbf{X})$; $\mathbf{f}_1(\mathbf{y}) = -\mathbf{X}$; $\boldsymbol{\eta} = \boldsymbol{\xi}$; and

$$A_j^* = \Pr \left(j = \arg \max_{J=0,1,\dots,J} (\mathbf{X}_k\boldsymbol{\beta} + \xi_k + \sigma\mathbf{X}_k\mathbf{B}\mathbf{v}) \mid \mathbf{X}, \boldsymbol{\xi}, \mathbf{v} \right)$$

so that, denoting $\mathbf{a}_j \equiv \mathbf{X}_j$ and $b_j = \mathbf{X}_j\boldsymbol{\beta} + \xi_j$,

$$A^*(a, b, c) \equiv \frac{\exp(b_j + \mathbf{a}_j\mathbf{c})}{1 + \sum_{k=1}^J \exp(b_k + \mathbf{a}_k\mathbf{c})};$$

and $G_j^* \equiv S_j - E_{\mathbf{v}}A_j^*$.

We continue to assume that $E(\boldsymbol{\eta} \mid \mathbf{Z}) = \mathbf{0}$. The quasi-linear structure in (10) allows this class of models to be approximately estimated by 2SLS.

Theorem 3. *Consider a model of the class defined by (10) and assume that*

- \mathbf{G}^* is twice differentiable with respect to its second argument
- \mathbf{A}^* is twice differentiable with respect to its last two arguments
- the matrices $\mathbf{G}_2^*(\mathbf{y}, \mathbf{A}^*(\mathbf{y}, \boldsymbol{\eta} - \mathbf{f}_1(\mathbf{y})\boldsymbol{\beta}, \mathbf{0}))$ and $\mathbf{A}_2^*(\mathbf{y}, \boldsymbol{\eta} - \mathbf{f}_1(\mathbf{y})\boldsymbol{\beta}, \mathbf{0})$ are invertible for all $(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\beta})$.

Any such model satisfies the conditions **C1–C3** in the introduction. Moreover,

- $\mathbf{f}_1(\mathbf{y})$ appears directly in (10)
- the variables $\mathbf{f}_0(\mathbf{y})$ are defined by the system of equations

$$\mathbf{G}^*(\mathbf{y}, \mathbf{A}^*(\mathbf{y}, \mathbf{f}_0(\mathbf{y}), \mathbf{0})) = \mathbf{0}$$

- and the variables $\mathbf{f}_2(\mathbf{y})$ solve the linear system

$$\mathbf{A}_{33}^*(\mathbf{y}, \mathbf{f}_0(\mathbf{y}), \mathbf{0}) \mathbf{f}_2(\mathbf{y}) = -\mathbf{A}_2^*(\mathbf{y}, \mathbf{f}_0(\mathbf{y}), \mathbf{0}).$$

Proof: See Appendix E.

As explained in the introduction, these models can be estimated by regressing $f_0(y)$ on $f_1(y)$ and $f_2(y)$ with a set of flexible functions of \mathbf{Z} as instruments. Since the macro-BLP model belongs to this class, this confirms that conditions **C1–C3** hold in the BLP model; we had shown it implicitly in section 3 by deriving the expansions. Note also that we did not use *any* distributional assumption on the random coefficients and the idiosyncratic shocks—although of course the terms in the expansions do depend on these distributions. We give an illustration for a one-covariate mixed binary model without any distributional assumption in Appendix B.3.

6.4.2 Examples

It is easy to generate models in the quasi-linear class (10). Starting from any GLM model $g(y) = \mathbf{X}\boldsymbol{\beta} + \eta$, we can for instance transform the right-hand side by adding additive unobserved heterogeneity and another link function:

$$g(y) = E_\varepsilon h(\mathbf{X}\boldsymbol{\beta} + \eta, \sigma\varepsilon).$$

When the link functions g and h are both assumed to be known, all such models obey conditions **C1–C3** and can therefore be studied with our method. Note that in these models $f_1(y) \equiv -\mathbf{X}$ and $f_2(y) = -(h_1/h_{22})(f_0(y), 0)$ where

$$f_0(y) = h(\cdot, 0)^{-1}(g(y))$$

(assuming the inverse is well-defined.)

The nested logit of section 6.1 shows that our method remains useful beyond the class of quasi-linear models, at the cost of breaking condition **C2** and **C3** and requiring numerical optimization.

7 Simulations

This section presents the results of a Monte Carlo study of an aggregate discrete choice demand system with random coefficients. It compares the finite sample performance of our estimator of the parameters to estimators computed using the mathematical programming with equilibrium constraints (MPEC) approach recommended by Dubé, Fox and Su (2012) and the control function approach of Petrin and Train (2010). We also show results demonstrating some of the “robustness” of our estimation procedure to assumptions about the distribution of the random coefficients. Specifically, we find that even if the distribution of random coefficients is misspecified, our procedure still yields very good estimates of the means and variances of the random coefficients.

The basic set-up of our Monte Carlo study follows that in Dubé, Fox and Su (2012). It is a standard static aggregate discrete choice random coefficients demand system with $T = 50$ markets and $J = 25$ products in each market, and $K = 3$ observed product characteristics. Following Dubé, Fox, and Su (2012), let M_t denote the mass of consumers in market $t = 1, 2, \dots, T$. Each product is characterized by the vector $(\mathbf{X}'_{jt}, \xi_{jt}, p_{jt})'$, where \mathbf{X}_{jt} is a $K \times 1$ vector of observable attributes of product $j = 1, 2, \dots, J$ in market t , ξ_{jt} is the vertical product characteristic of product j in market t that is observed by producers and consumers, but unobserved by the econometrician, and p_{jt} is the price of product j in market t . Collect these variables for each product into the following market-specific variables: $\mathbf{X}_t = (\mathbf{X}'_{1t}, \dots, \mathbf{X}'_{Jt})'$, $\boldsymbol{\xi}_t = (\xi_{1t}, \xi_{2t}, \dots, \xi_{Jt})'$, and $\mathbf{p}_t = (p_{1t}, p_{2t}, \dots, p_{Jt})'$.

The conditional indirect utility of consumer i in market t from purchasing product j is

$$u_{ijt} = \beta_0 + \mathbf{X}'_{jt} \boldsymbol{\beta}_i^x - \beta_i^p p_{jt} + \xi_{jt} + \epsilon_{ijt}$$

The utility of the $j = 0$ good, the “outside” good, is equal to $u_{0jt} = \epsilon_{i0t}$. Each element of $\boldsymbol{\beta}_i^x = (\beta_{i1}^x, \dots, \beta_{iK}^x)'$ is assumed to be drawn independently from $N(\bar{\beta}_k^x, \sigma_k^2)$ distributions, and each β_i^p is assumed to be drawn independently from $N(\bar{\beta}_p, \sigma_p^2)$. We denote $\boldsymbol{\beta}_i = (\boldsymbol{\beta}_i^{x'}, \beta_i^p)'$.

We collect all parameters into

$$\boldsymbol{\theta} = (\beta_0, \bar{\beta}_1^x, \dots, \bar{\beta}_K^x, \bar{\beta}_p, \sigma_1^2, \dots, \sigma_K^2, \sigma_p^2)'$$

Our simulations have

$$(\beta_0, \bar{\beta}_1^x, \bar{\beta}_2^x, \bar{\beta}_3^x, \bar{\beta}_p) = (-1, 1.5, 1.5, 0.5, -1)$$

and varying variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_p^2)$. We also experiment with varying $\bar{\beta}_p$.

To compute the market shares for the J products, we assume that the ϵ_{ijt} are independently and identically distributed Type I extreme value random variables, so that the probability that consumer i with random preferences β_i purchases good j in market t is equal to

$$s_{ijt}(\mathbf{X}_t, \mathbf{p}_t, \boldsymbol{\xi}_t | \beta_i) = \frac{\exp(\beta^0 + \mathbf{X}'_{jt} \beta_i^x - \beta_i^p p_{jt} + \xi_{jt})}{1 + \sum_{k=1}^J \exp(\beta_0 + \mathbf{X}'_{kt} \beta_i^x - \beta_i^p p_{kt} + \xi_{kt})}$$

We compute the observed market share for all goods in market t by drawing $n_s = 1,000$ draws (ζ_{ikt}) from four $N(0, 1)$ random variables and constructing 1,000 draws from $\beta_i | \boldsymbol{\theta}$ as follows:

$$\beta_{ikt}^x = \bar{\beta}_k^x + \sigma_k \zeta_{ikt} \quad \text{and} \quad \beta_{it}^p = \bar{\beta}_p + \sigma_p \zeta_{ipt}.$$

We then use these draws to compute the observed market share of good j in market t as:

$$s_{jt}(\mathbf{X}_t, \mathbf{p}_t, \boldsymbol{\xi}_t | \boldsymbol{\theta}) = \frac{1}{n_s} \sum_{i=1}^{n_s} s_{ijt}(\mathbf{X}_t, \mathbf{p}_t, \boldsymbol{\xi}_t | \beta_i)$$

given the vectors \mathbf{X}_t , \mathbf{p}_t , and $\boldsymbol{\xi}_t$ for each market t .

Consistent with the experimental design in Dubé, Fox and Su (2012), we generate the values of \mathbf{X}_t , \mathbf{p}_t , $\boldsymbol{\xi}_t$ and a vector of 6 instruments \mathbf{Z}_{jt} as follows. First we draw \mathbf{X}_t for all markets $t = 1, 2, \dots, T$ from a multivariate normal distribution:

$$\begin{bmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 & 0.3 \\ -0.8 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix} \right)$$

The price of good j in market t is equal to

$$p_{jt} = |0.5 \xi_{jt} + e_{jt} + 1.1(x_{1j} + x_{2j} + x_{3j})|,$$

where $e_{jt} \sim N(0, 1)$, distributed independently across products and markets. The ξ_{jt} are $N(0, \sigma_\xi^2)$ random variables drawn independently across products and markets

for different values of σ_ξ^2 described below. The data generating process for the vector of instruments is:

$$z_{jtd} \sim U(0, 1) + 0.25(e_{jt} + 1.1(x_{1j} + x_{2j} + x_{3j}))$$

where $d = 1, \dots, 6$.

For a specified value of the parameter vector θ , following this process for $T = 50$ markets yields the dataset for one Monte Carlo draw.

7.1 MPEC Approach

The MPEC approach solves a nonlinear minimization problem subject to nonlinear equilibrium constraints. The first step of the estimation process constructs the following instrumental variables for all the products in all the markets. There are 42 instruments in total; they are constructed from product characteristics x_j and excluded instruments z_{jt} :

$$1, x_{kj}, x_{kj}^2, x_{kj}^3, x_{1j}x_{2j}x_{3j}, z_{jtd}, z_{jtd}^2, z_{jtd}^3, z_{jtd}x_{1j}, z_{jtd}x_{2j}, \prod_{d=1}^6 z_{jtd}$$

Let W denote this $(J \times T) \times 42$ matrix of instruments. In our case $J \times T = 1,250$ since $J = 25$ and $T = 50$.

The MPEC approach solves for θ by minimizing

$$\eta'W(W'W)^{-1}W'\eta$$

subject to the “equilibrium constraints”

$$s(\eta, \theta) = S$$

where S is the vector of observed market shares computed as described above given the values of x_t , p_t and ξ_t and η is a $(J \times T) \times 1$ vector defined by the following equation:

$$s_{jt}(\eta, \theta) = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{\exp(\theta_1 + x_{1j}\beta_{1i}^x + x_{2j}\beta_{2i}^x + x_{3j}\beta_{3i}^x + p_{jt}\beta_i^p + \eta_{jt})}{1 + \sum_{k=1}^J \exp(\theta_1 + x_{1k}\beta_{1i}^x + x_{2k}\beta_{2i}^x + x_{3k}\beta_{3i}^x + p_{kt}\beta_i^p + \eta_{kt})}$$

where each (β_i^x, β_i^p) is a random draw from the following normal distribution:

$$N \left(\begin{bmatrix} \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{bmatrix}, \begin{bmatrix} \theta_6 & 0 & 0 & 0 \\ 0 & \theta_7 & 0 & 0 \\ 0 & 0 & \theta_8 & 0 \\ 0 & 0 & 0 & \theta_9 \end{bmatrix} \right)$$

Note that θ_1 (like β_0) is not allowed to be random. For purposes of estimation we set $N_s = 1,000$. For each Monte Carlo simulation, we start the optimization with the following initial point: true values for θ , and a vector of zeros for the η vector.

7.2 A Control Function Approach

To implement the Petrin and Train (2010) control function approach, we first run a linear regression of the price p on all 42 instruments. We denote the residuals from this regression by $\hat{\epsilon}_{jt}$.

We then solve the following maximum likelihood problem:

$$\max_{\theta, \rho} \sum_{j=0}^J \sum_{t=1}^T S_{jt} \cdot \log(s'_{jt}(\theta, \rho))$$

where $j = 0$ refers to the outside product, S_{jt} is the observed market share, and s'_{jt} is defined by

$$s'_{jt}(\theta, \rho) = \frac{1}{NS} \sum_{i=1}^{NS} \frac{\exp(\beta^0 + x_{1j}\beta_{1i}^x + x_{2j}\beta_{2i}^x + x_{3j}\beta_{3i}^x + p_{jt}\beta_i^p + \rho\hat{\epsilon}_{jt})}{1 + \sum_{k=1}^J \exp(\beta^0 + x_{1k}\beta_{1i}^x + x_{2k}\beta_{2i}^x + x_{3k}\beta_{3i}^x + p_{kt}\beta_i^p + \rho\hat{\epsilon}_{kt})}$$

where β_0 and the (β_i^x, β_i^p) are generated as we did with MPEC.

7.3 Our 2SLS Approach

Our 2SLS approach resorts to a slight modification of the standard linear 2SLS estimator to account for the fact that the estimates of the σ_k^2 and σ_p^2 cannot be negative. First, we construct the instrumental variables as the MPEC approach. We then construct the artificial regressors K_1, K_2, K_3, K_p of Theorem 2 for each product in each

market by applying

$$\bar{X}_{it} = \sum_{k=1}^J x_{ik} S_{kt}$$

$$K_i^{jt} = x_{ij}(x_{ij}/2 - \bar{X}_{it})$$

for $i = 1, 2, 3, p$.

The next step performs an instrumental variable regression of $y_{jt} = \log(\frac{S_{jt}}{S_{0t}})$ on $1, x_1, x_2, x_3, x_4, K_1, K_2, K_3, K_p$ using all 42 instruments. If any coefficient for the last four variables is negative, we set that coefficient to 0 and rerun the regression without that variable. We iterate this process until all the coefficients are positive, or all four variables are excluded from the instrumental variables regression.

In addition to this standard 2SLS estimator, we compute a bias correction as explained in section 6.3.2. To evaluate it, we replace y_{jt} , the dependent variables for 2SLS estimates, with $y_{jt} - \xi_{2,jt} + \xi_{\text{inf},jt}$, where

- $\xi_{2,jt}$ is the residual from our initial 2SLS estimation procedure
- $\xi_{\text{inf},jt}$ is the value of ξ_{jt} that results from solving the equation $\mathbf{s}_t(\boldsymbol{\xi}_t, \hat{\theta}) = \mathbf{S}_t$, where $\hat{\theta}$ is the initial 2SLS estimate of θ .

We found that it worked as well as the bootstrap, at a lower computational cost. We also experimented with using the optimal instruments, obtained by a kernel regression of \mathbf{X} and of \mathbf{K} on the variables $x_1, x_2, x_3, z_1, \dots, z_6$.

7.4 Pseudo True Values for the 2SLS Approach

As explained earlier, the 2SLS estimator is not consistent for the true parameter values, as it estimates an approximate model. We constructed estimates of the pseudo true values to which our 2SLS estimators converge by simulating their probability limit. A first approach increases the number of markets and computes our 2SLS estimates for this large number of markets. The second approach computes estimates of the population values of the moments of our 2SLS estimator.

7.4.1 Increasing-number-of-markets Approach

For each simulation, we keep the size and distribution of product characteristics for each market fixed, but increase the number of markets. For each scenario, we calculate the pseudo true value (and its standard error) by 20 simulations of 100,000 markets. Note that across different simulations, we generate different product characteristics. Also, when calculating market shares, we use different random draws of β_i across different simulations, but the same random draws of β_i within a simulation. Estimates are calculated by the sample mean of the 20 simulations. Standard errors are calculated by the sample standard errors of the 20 simulations.

7.4.2 Moment-based Approach

We can also calculate the pseudo true values in a different way. We first run the first stage projection: $\hat{\Pi} = (W'W)^{-1}W'X$ for each simulation, where W is our matrix of instruments and X is our matrix of regressors. We then take the average across all the simulations to get our estimate of the population value of Π . Then in the second stage, we calculate $(W\Pi)'X$ and $(W\Pi)'Y$ for each simulation, and then take averages across all the simulations to get two matrices A and B . The final estimate is then $A^{-1}B$. In short, we have

$$\begin{aligned}\Pi &= E_{\text{all simulations}}[(W'W)^{-1}W'X] \\ A &= E_{\text{all simulations}}[(W\Pi)'X] \\ B &= E_{\text{all simulations}}[(W\Pi)'Y] \\ \text{Estimate} &= A^{-1}B\end{aligned}$$

With this method, we only have the estimates but cannot get the standard errors. We used 1000 simulations of 10,000 markets.

7.5 Monte Carlo Simulation Results

We used the SNOPT optimization package available from the Stanford Systems Optimization Laboratory to solve the nonlinear optimization problems for the MPEC estimator and the control function estimator. The software employs a sparse sequen-

tial quadratic programming (SQP) algorithm with limited-memory quasi-Newton approximations to the Hessian of the Lagrangian.

We run simulations for 9 scenarios obtained by setting three values for the variance of the product random effects: $\sigma_\xi^2 = \text{Var}(\xi) = 0.1, 0.5, 1$ and three values for the vector of variances of the coefficients $\beta_i = (\beta_0, \beta_{1i}^x, \beta_{2i}^x, \beta_{3i}^x, \beta_i^p)'$:

$$\text{Var}(\beta_i) = (0, 0.1, 0.1, 0.1, 0.05), (0, 0.2, 0.2, 0.2, 0.1), (0, 0.5, 0.5, 0.5, 0.2).$$

Note that the square roots of the elements of $\text{Var}(\beta_j)$ represent the relative values of the scale parameter σ of models 1, 2, and 5.

It is worth noting here that we explored other scenarii in which MPEC often failed to converge, even though we are starting it from the true values of the parameters. In particular, larger variances of ξ are problematic. It is also the reason why we reduced the highest value of σ_p^2 from 0.25 to 0.2.

All the other parameter specifications are as described above.

7.5.1 Distribution of the Estimates

We summarize the estimation results in Tables 1 to 9, where density plots are grouped by parameter for all scenarii. These plots suggest that if the researcher is interested in a precise estimate of the mean of the random coefficients, then using our 2SLS approach does not imply any significant bias or loss in efficiency relative to the MPEC approach. In contrast, the control function approach exhibits substantial bias in the estimate of the means of the random coefficients; and this bias increases with the variance of ξ_{jt} . This makes sense since the control function estimator in fact uses an invalid control function.

The MPEC approach appears to dominate the 2SLS approach for the variance of the random coefficients. The 2SLS estimators of the variances have a downward bias that increases with the variance of the random coefficients. However, larger values of the variance of ξ_{jt} do seem to improve the performance of the 2SLS estimator of the variance of the random coefficients.

The control function estimators of the variances of the random coefficients are often less biased than the 2SLS estimators. However, their distributions tend to have

a larger spread than those of the estimates from the MPEC estimation procedure, or of our 2SLS estimation procedure.

7.5.2 Starting Values

We are giving a big advantage to MPEC in our comparisons, since we allow the algorithm to start from the true values of the parameters. This is of course infeasible in practice. With this initial boost, MPEC converges 100% of the time, after 1,030 iterations on average; the minimization takes 110 seconds on average. Our 2SLS approach provides a more realistic alternative, in which we start MPEC from the results of our 2SLS regression. This appears to work very well: MPEC converges after an average 125 seconds and 1,280 iterations, again with a 100% success rate. The resulting estimates are very close to those obtained when starting from the true values: the difference is between 10^{-6} and 10^{-7} .

These results are very encouraging for the use of our approach as a method for finding very good starting values for the MPEC and nested-fixed point estimation procedures. Given that 2SLS takes no time at all, we would strongly recommend running it before a more sophisticated algorithm.

7.5.3 Price Elasticities

Based on the parameter estimates, we can estimate the own price elasticity of the demand for each product. The graphs in table 12 plot the distribution of the difference between the true price elasticity and the estimated price elasticity for the MPEC approach, our standard 2SLS approach, and our bias-corrected 2SLS approach. For space reasons, we only presents the results for five products: numbers 5, 10, 15, 20, 25. Our simulations have variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_p^2) = (0, 0.4, 0.4, 0.4, 0.2)$ with $\text{Var}(\xi) = 1$. Table 12 demonstrates that our procedure recovers nearly identical mean own-price elasticities for products as the MPEC approach, although the spread for our estimates is slightly larger than in the MPEC approach.

We also performed a set of simulations (with the same variances) to determine if changing the true value of the price coefficient $\bar{\beta}_p$ changes the performance of the estimators. The results in tables 13 and 14. reinforce our previous conclusions about our 2SLS approach. For a range of values of the mean value of the price coefficient,

our approach introduces minimal bias in the estimates of the means of the random coefficients. In contrast, the control function approach continues to show significant bias. The estimates of the variances of the random coefficients for our 2SLS estimate continue to be downwards biased in general, but the bias is smaller for larger price coefficients.

7.5.4 Pseudo-true Values

Tables 10 and 11 demonstrate that for most scenarios and coefficients, the pseudo true values implied by our 2SLS procedure are not substantially different from the true values. Based on these results, it is difficult to argue that a researcher would draw conclusions from 2SLS estimates that differ in an economically or even statistically meaningful way from those obtained with MPEC estimates.

7.5.5 Variable Selection Tests

Researchers in empirical IO have little guidance on the list of characteristics \mathbf{X} they should include, or how to specify the matrix Σ . Experimenting with different specifications is costly with the usual estimators. Our 2SLS approach, on the other hand, makes variable selection very easy. We can decide whether a characteristic simply by testing whether the corresponding covariate can be dropped from the estimating equation; and to decide whether we should allow for a random coefficient, we only need to test whether the associated artificial regressor can be dropped from the equation. We experimented with this approach to detecting random coefficients by setting $\bar{\beta}_1^x = \sigma_1^x = 0$ in the data generating process and applying standard tests that the covariate x_1 and/or the artificial regressor K_1 has a zero coefficient in the 2SLS regression. We also performed this test using our bias-corrected 2SLS estimates. Tables 17 to 20 give the probability that the null hypothesis is not rejected, where the null hypothesis is

- $\bar{\beta}_1^x = 0$ (Tables 15 and 16)
- $\sigma_1^x = 0$ (Tables 17 and 18)
- $\bar{\beta}_1^x = \sigma_1^x = 0$ (Tables 19 and 20).

The row labelled “2SLS with heteroskedasticity-robust standard error” is our 2SLS estimate, using a standard heteroskedasticity-robust covariance matrix to compute standard errors. The row labelled “GLS estimator and standard errors” uses Cragg’s (1983) generalized least squares estimator and his recommended standard error estimates. The row labelled “2SLS with clustered standard errors” uses our 2SLS estimates with standard errors clustered at market level.

Since the null hypothesis is true, each row in Tables 15-20 would ideally contain 0.99, 0.95, and 0.90. Clearly, our test rejects the null too often. In this particular application, this is probably better than the alternative: better to include more variables and lose some efficiency than to incur bias by leaving them out. The size distortion is smaller for tests on the means (Tables 15 and 16); it is also smaller when we use bias-corrected estimates. The clustered standard error estimates appear to have the largest size distortions. On the whole, we take this to suggest that demonstrate that our estimator can be used to good effect in order to decide which coefficients should be modelled as random.

7.6 Lognormal Distribution for β

As explained in section 5.2, our estimating equation is the same whether the distribution of the random coefficients is normal or not. To illustrate this, we modify the data-generating process so that the consumer preference parameters β_i have a lognormal distribution:

$$\begin{aligned}\beta_i &= \bar{\beta}_i \epsilon_i \\ \bar{\beta}_i &= (1, 1.5, 1.5, 0.5, 1) \\ \ln(\epsilon_i) &\sim N(-0.5\sigma^2, \sigma^2).\end{aligned}$$

We study several cases, with $\sigma = 0.3, 0.4, 0.5$ and $\xi_{jt} \sim N(0, 0.1)$. The rest of the specification is as before. Lognormality induces significant skewness and kurtosis into the distribution of the random coefficients. The standard 2SLS approach gives us estimates of the first and second moments. We can also introduce the additional artificial regressors T of section 6.2, either to control for skewness or to estimate it. We experimented with both possibilities. Each plot in Tables 21, 22, and 23 shows the distributions when we use only X and K (“only include 2nd moment”) and when

we add T (“include third moment”). These three tables report the distributions of the estimates of the first, second, and third moments of $\beta_1, \beta_2, \beta_3$, and β_p . Table 24 provides the corresponding summary statistics.

For a variety of values for the parameter σ of our lognormal distribution, the 2SLS estimates are just as good as they were in the normal setup. The additional information in the third moment of the random coefficients does not appreciably increase the precision in our estimates of the means and variances of the random coefficients. In fact, for some of the coefficients, including the third-order artificial regressors T leads to significantly less efficient estimates. This is likely due to the fact that our procedure has a difficult time estimating the third moment of the random coefficients, as Table 23 shows.

7.6.1 Bias Correction and Kernel

Our paper suggested two potential improvements to the standard 2SLS regression: bias correction, and using a kernel regression to estimate the optimal instruments $E(\mathbf{X}|\mathbf{Z})$ and $E(\mathbf{K}|\mathbf{Z})$. We compare both methods, when coefficients are normal with variances $(0, 0.5, 0.5, 0.5, 0.2)$ and when they are lognormally distributed with $\sigma = 0.4$ for $\ln(\epsilon_i) \sim N(-0.5\sigma^2, \sigma^2)$. In both cases we took $Var(\xi) = 0.1$.

Tables 25 and 26 plot the distributions of the estimators. They suggest that our bias-correction does to reduce the bias, both for the means and the variances. This holds whether the random coefficients are normally or log-normally distributed. Using kernel regressions to approximate the optimal instruments appears to slightly reduce both the bias and the variance of some of the estimates.

Concluding Comments

Our FRAC estimation procedure applies directly to the random coefficients demand models commonly used in empirical IO. For the most part, our Monte Carlo results confirm the findings from the expansions. The 2SLS approach yields reliable estimates of the parameters of the model and of economically meaningful quantities such as price elasticities; and it does so at a very minimal cost. It is “robust” to variations on the distribution of the random coefficients. In addition, it provides straightforward tests

that help in for variable selection, especially as a guide to determine which coefficients in the demand system should be modeled as random.

Some of our simulation results are unexpected and point to directions for future research. We hope to report more general analytical results that illuminate these findings.

References

Amemiya, T. (1975), “The nonlinear limited-information maximum-likelihood estimator and the modified nonlinear two-stage least-squares estimator,” *Journal of Econometrics*, 3, 375–386.

Armstrong, T. (2016), “Large Market Asymptotics for Differentiated Product Demand Estimators With Economic Models of Supply”, *Econometrica*, 84, 1961-1980.

Berry, S. (1994), “Estimating Discrete Choice Models of Product Differentiation”, *Rand Journal of Economics*, 23, 242-262.

Berry, S., Levinsohn, J., and A. Pakes (1995), “Automobile Prices in Market Equilibrium”, *Econometrica*, 60, 889-917.

Campioni, G. (2018), “Nonparametric Demand Estimation in Differentiated Products Markets”, mimeo Yale.

Chesher, A. (1991), “The Effect of Measurement Error”, *Biometrika*, 78, 451–462.

Chesher, A. and J. Santos-Silva (2002), “Taste Variation in Discrete Choice Models”, *Review of Economic Studies*, 69, 147–168.

Cragg, J.G. (1983), “More Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form”, *Econometrica*, 51, 3, 751–763.

Dubé, J.-P., Fox, J., and C.-L. Su (2012), “Improving the Numerical Performance of BLP Static and Dynamic Discrete Choice Random Coefficients Demand Estimation”, *Econometrica*, 80, 2231-2267.

Gandhi, A. and J.-F. Houde (2016), “Measuring Substitution Patterns in Differentiated Products Industries”, mimeo.

Harding, M. and J. Hausman (2007), “Using a Laplace Approximation to Estimate the Random Coefficients Logit Model by Nonlinear Least Squares”, *International Economic Review*, 48, 1311–1328.

Horowitz, J. and L. Nesheim (2018), “Using Penalized Likelihood to Select Parameters in a Random Coefficients Multinomial Logit Model”, mimeo UCL.

Kadane, J. (1971), “Comparison of k -class Estimators when the Variance is Small”, *Econometrica*, 39, 723–737.

Ketz, P. (2018), “On Asymptotic Size Distortions in the Random Coefficients Logit Model”, mimeo Paris School of Economics.

Lee, J. and K. Seo (2015), “A computationally fast estimator for random coefficients logit demand models using aggregate data”, *Rand Journal of Economics*, 46, 86-102.

Reynaert, M. and F. Verboven (2014), “Improving the Performance of Random Coefficients Demand Models: The Role of Optimal Instruments”, *Journal of Econometrics*, 179, 83-98.

Su, C.-L. and K. Judd (2012), “Constrained Optimization Approaches to Estimation of Structural Models”, *Econometrica*, 80, 2213-2230.

A Proof of Theorem 1

Proof. We start from (2); we drop the market index t and the bold letters. Since we now denote $X_\epsilon = \sigma x \cdot v$, we can rewrite (2) in the standard model as

$$S_j = E_v \frac{\exp(X_j \beta + \sigma x_j \cdot v + \xi_j)}{1 + \sum_{k=1}^J S_k \exp(X_k \beta + \sigma x_k \cdot v + \xi_k)}.$$

Given that

$$\xi_j = \log \frac{S_j}{S_0} - \mathbf{X}_j \boldsymbol{\beta} + a_{1j} \sigma + a_{2j} \frac{\sigma^2}{2} + O(\sigma^3),$$

we get

$$S_j = E_v \frac{S_j \exp\left(\sigma(x_j \cdot v + a_{1j}) + a_{2j} \frac{\sigma^2}{2} + O(\sigma^3)\right)}{S_0 + \sum_{k=1}^J S_k \exp\left(\sigma(x_k \cdot v + a_{1k}) + a_{2k} \frac{\sigma^2}{2} + O(\sigma^3)\right)}$$

Eliminating S_j gives

$$1 = E_v \frac{\exp\left(\sigma(\mathbf{x}'_j \mathbf{v} + a_{1j}) + \frac{\sigma^2}{2} a_{2j} + O(\sigma^3)\right)}{S_0 + \sum_{k=1}^J S_k \exp\left(\sigma(\mathbf{x}'_k \mathbf{v} + a_{1k}) + \frac{\sigma^2}{2} a_{2k} + O(\sigma^3)\right)}. \quad (11)$$

In this form, Theorem 1.(i) is obvious since only the vectors \mathbf{x}_k and market shares S_k enter the system of equations.

Now use the notation $e_S Z \equiv \sum_{k=1}^J S_k Z_k$ and $\hat{Z}_j = Z_j - e_S Z$ to rewrite (11) as

$$0 = E_v \left(\frac{\hat{V}_j}{1 + e_S V} \right) \quad (12)$$

where $V_j \equiv f_j + \alpha_j + f_j \alpha_j$, with

$$f_j \equiv \exp(\sigma x_j \cdot v) - 1 = \sigma x_j \cdot v + \frac{\sigma^2}{2} (x_j \cdot v)^2 + O_P(\sigma^3)$$

and

$$\alpha_j = \exp(a_{1j} \sigma + a_{2j} \sigma^2 / 2 + O(\sigma^3)) - 1 = a_{1j} \sigma + (a_{2j} + a_{1j}^2) \frac{\sigma^2}{2} + O(\sigma^3).$$

We note that f_j is $O_P(\sigma)$ and α_j is $O(\sigma)$, so that V_j is also $O_P(\sigma)$.

Now expanding (12) gives

$$O(\sigma^3) = E_v \hat{V}_j - E_v \left(\hat{V}_j (e_S V) \right) \quad (13)$$

$$= E_v \hat{f}_j + \hat{\alpha}_j \quad (14)$$

$$+ \widehat{\alpha_j E_v f_j} - (e_S \alpha) E_v \hat{f}_j - \hat{\alpha}_j E_v (e_S f) - \hat{\alpha}_j (e_S \alpha) - E_v \hat{f}_j (e_S f). \quad (15)$$

Only the terms on line (14) can be of order 1 in σ . But using $E_v v = 0$ and $E_v(x_j \cdot v)(x_k \cdot v) = x_j \cdot x_k$ gives us $E_v f_j = \frac{\sigma^2}{2} \|x_j\|^2 + O(\sigma^3)$. Therefore the only term of order 1 is in $\hat{\alpha}_j = \hat{a}_{1j}\sigma + O(\sigma^2)$, and we must have $\hat{a}_{1j} = 0$. We note that the “hat” operator is linear and invertible:

Lemma 1. *If $\hat{Z}_j = \hat{W}_j$ for all j and $S_0 < 1$, then $Z_j = W_j$.*

Proof. $Z_j - e_S Z = W_j - e_S W$ implies $Z_j = W_j + \lambda$, where $\lambda = e_S Z - e_S W$. But then $e_S Z = e_S W + e_S \lambda = (1 - S_0)\lambda$, so that $\lambda = (1 - S_0)\lambda = 0$. \square

Applying the lemma gives $a_{1j} = 0$. As a consequence, $\alpha_j = a_{2j}\sigma^2/2 + O(\sigma^3)$; and all terms on line (15) except the last one are of order at least 3 in σ . Since

$$E_v \hat{f}_j + \hat{\alpha}_j = \frac{\sigma^2}{2} \widehat{\|x_j\|^2} + \frac{\sigma^2}{2} \hat{a}_{2j} + O(\sigma^3)$$

and

$$E_v \hat{f}_j(e_S f) = \sigma^2 E_v(\hat{x}_j \cdot v)((e_S x) \cdot v) + O(\sigma^3) = \sigma^2(\hat{x}_j \cdot (e_S x)) + O(\sigma^3)$$

applying the lemma again gives us $(\|x_j\|^2 + a_{2j})/2 - x_j \cdot (e_S x) = 0$.

Finally, if the distribution of \mathbf{v} is symmetric around $\mathbf{0}$ changing σ to $-\sigma$ in (2) must leave all market shares unchanged; therefore all expansions can only contain even-degree terms in σ . \square

B Detailed Examination of the Mixed Logit

The standard binary model is simply a mixed logit. Applying Theorem 1 with $J = 1$ and using $S_0 + S_1 = 1$, we obtain

$$a_{21} = (2S_1 - 1)\|\mathbf{x}_1\|^2$$

and $\mathbf{K}^1 = (1/2 - S_1)\mathbf{X}_1\mathbf{X}_1'$. Therefore

$$\xi_1 = \log \frac{S_1}{S_0} - \mathbf{X}_1\boldsymbol{\beta} - \left(\frac{1}{2} - S_1\right) \text{Tr} \boldsymbol{\Sigma} \mathbf{X}_1\mathbf{X}_1' + O(\sigma^k)$$

where $k = 3$ in general, and $k = 4$ if the distribution of $\boldsymbol{\epsilon}$ is symmetric around zero.

The presence of the term $(1/2 - S_1)$ in this formula is a consequence of the symmetry of the distribution of \mathbf{v} around $\mathbf{0}$ and of the logistic distribution around 0. Taken together, this implies that market shares around one half vary very little with σ . The random variation in tastes can only be identified from nonlinearities in the market shares; but since the cdf of the logistic has an inflexion point when its value is one half, market shares are essentially linear around that point. It is easy to check that this is specific to the one-product case; when $J > 1$, the mixed multinomial logit does not face any such difficulty.

Let us focus for simplicity on the case when random variation in preferences is uncorrelated across covariates: Σ is the $n_X \times n_X$ diagonal matrix with elements Σ_{mm} . Then given instruments such that $E(\xi_1 | \mathbf{Z}) = 0$, the approximate model is

$$E \left(\log \frac{S_1}{S_0} - \mathbf{X}_1 \boldsymbol{\beta} - \left(\frac{1}{2} - S_1 \right) \sum_{m=1}^{n_X} \Sigma_{mm} X_{1m}^2 \mid \mathbf{Z} \right) = 0. \quad (16)$$

B.1 Identification

The form of the estimating equation holds interesting insights about identification. First note that the optimal instruments are

$$\mathbf{f}(\mathbf{Z}) = E(\mathbf{X}_1 | \mathbf{Z}), E \left(\left(\frac{1}{2} - S_1 \right) \mathbf{X}_1^2 | \mathbf{Z} \right)$$

where \mathbf{X}_1^2 is the vector with components X_{1m}^2 . The asymptotic variance-covariance matrix of our estimator $\hat{\boldsymbol{\theta}}$ is given by the usual formula:

$$T V_{\text{as}} \hat{\boldsymbol{\theta}} \simeq \mathbf{J}^{-1} V(\xi_1 \mathbf{f}(\mathbf{Z})) \mathbf{J}^{-1},$$

where

$$\mathbf{J} = E \left(\left(\mathbf{X}_1, \left(\frac{1}{2} - S_1 \right) \mathbf{X}_1^2 \right) \mathbf{f}(\mathbf{Z}) \right).$$

The identifying power of the (approximate) model relies on the full-rank of the matrix \mathbf{J} . Suppose for instance that after projecting (via nonparametric regression) the regressors on the instruments, the residual variation in the artificial regressor $(1/2 - S_1) X_{1m}^2$ is very well explained in a linear regression on the other covariates. Then the estimate of Σ_{mm} will be very imprecise, and random taste variation on the characteristic X_{1m} is probably best left out of the model. Of course, this can be diagnosed immediately by looking at the precision of the 2SLS estimates.

B.2 Higher-order terms

It is easy to program a symbolic algebra system to compute higher-order terms a_{lj} for $l > 2$. We show here how to compute the fourth-order term in the mixed logit model. This will also illustrate the “robustness” of our method to distributional assumptions.

Assume that ϵ has a distribution that is symmetric around zero, and that its components are independent of each other with variances Σ_{mm} and fourth-order moments k_m . As before, we assume that Σ_{mm} is of order σ^2 and k_m is of order σ^4 . We also assume that we can take expansions to order $L \geq 5$.

Since the distribution is symmetric, we already know that

$$\xi_1 = \log \frac{S_1}{S_0} - \mathbf{X}_1 \boldsymbol{\beta} + \frac{a_{21}}{2} \sigma^2 + \frac{a_{41}}{24} \sigma^4 + O(\sigma^6).$$

Define $L(t) = 1/(1 + \exp(-t))$ the cdf of the logistic distribution. Note that $L' = L(1 - L)$, and that higher-order derivatives follow easily:

$$\begin{aligned} L'' &= L(1 - L)(1 - 2L) \\ L^{(3)} &= L(1 - L)(1 - 6L + 6L^2) \\ L^{(4)} &= L(1 - L)(1 - 2L)(1 - 12L + 12L^2). \end{aligned}$$

Since the market share of good 1 is

$$S_1 = E_{\epsilon} L(\mathbf{X}_1(\boldsymbol{\beta} + \epsilon) + \xi_1)$$

we obtain, much as in Appendix A,

$$S_1 = E_{\epsilon} L \left(\log \frac{S_1}{S_0} + \mathbf{X}_1 \epsilon + \alpha_2 \sigma^2 + \alpha_4 \sigma^4 + O(\sigma^6) \right)$$

where we defined $\alpha_l = a_{l1}/l!$ for $l = 2, 4$.

Let a 0 subscript indicate that we take the value and derivatives of $L(t)$ at $t = \log(S_1/S_0)$. Defining $u(\epsilon) = \mathbf{X}_1 \epsilon + \alpha_2 \sigma^2 + \alpha_4 \sigma^4 + O(\sigma^6)$ and expanding gives

$$L \left(\log \frac{S_1}{S_0} + u \right) = L_0 + L_0' u + \frac{L_0''}{2} u^2 + \frac{L_0^{(3)}}{6} u^3 + \frac{L_0^{(4)}}{24} u^4 + O(u^5).$$

Incorporating $L_0 = S_1$, $L'_0 = S_1(1 - S_1)$, up to $L_0^{(4)}$ gives

$$S_1 = E_\epsilon \left(S_1 + S_1(1 - S_1)u(\epsilon) + S_1(1 - S_1)(1 - 2S_1)\frac{u(\epsilon)^2}{2} \right. \\ \left. + S_1(1 - S_1)(1 - 6S_1 + 6S_1^2)\frac{u(\epsilon)^3}{6} + S_1(1 - S_1)(1 - 2S_1)(1 - 12S_1 + 12S_1^2)\frac{u(\epsilon)^4}{24} + O(u(\epsilon)^5) \right);$$

dividing by $S_1(1 - S_1)$ yields

$$E_\epsilon u + (1 - 2S_1)E_\epsilon u^2/2 + (1 - 6S_1 + 6S_1^2)E_\epsilon u^3/6 + (1 - 2S_1)(1 - 12S_1 + 12S_1^2)E_\epsilon u^4/24 = E_\epsilon O(u^5). \quad (17)$$

Finally, up to order 6 in σ :

$$E_\epsilon u = \alpha_2 \sigma^2 + \alpha_4 \sigma^4 \\ E_\epsilon u^2 = \sigma^2 E_\epsilon (\mathbf{X}_1 \epsilon)^2 + \alpha_2^2 \sigma^4 = \sum_{m=1}^{n_X} \Sigma_{mm} x_{1m}^2 + \alpha_2^2 \sigma^4 \\ E_\epsilon u^3 = 3\alpha_2 \sigma^4 E_\epsilon (\mathbf{X}_1 \epsilon)^2 = 3\alpha_2 \sum_{m=1}^{n_X} \Sigma_{mm} x_{1m}^2 \\ E_\epsilon u^4 = \sigma^4 E_\epsilon (\mathbf{X}_1 \epsilon)^4 = \sum_{m=1}^{n_X} k_m x_{1m}^4.$$

Regrouping terms in σ^2 in (17) confirms that

$$\alpha_2 \sigma^2 = (S_1 - 1/2) \sum_{m=1}^{n_X} \Sigma_{mm} x_{1m}^2,$$

which we knew from Theorem 1. The terms in σ^4 give us

$$\alpha_4 \sigma^4 = \alpha_2^2 \sigma^4 (S_1 - 1/2) - \alpha_2 \sigma^2 (1 - 6S_1 + 6S_1^2) \sum_{m=1}^{n_X} \Sigma_{mm} x_{1m}^2 / 2 \\ - (1 - 2S_1)(1 - 12S_1 + 12S_1^2) \sum_{m=1}^{n_X} k_m x_{1m}^4 / 24.$$

This simplifies to

$$\alpha_4 \sigma^4 = \left(\frac{1}{2} - S_1 \right) \times \\ \left(\left(\frac{1}{4} - 2S_1(1 - S_1) \right) \left(\sum_{m=1}^{n_X} \Sigma_{mm} x_{1m}^2 \right)^2 - \left(\frac{1}{12} - S_1(1 - S_1) \right) \sum_{m=1}^{n_X} k_m x_{1m}^4 \right).$$

This formula may not seem especially enlightening, but it shows several important points. First, terms of higher orders can be computed without much difficulty. Second, each additional term adds information on lower-order moments (here σ_m^2), as well as on the moments of higher order (here k_m). The model remains linear in the highest order moments; here for k_m we have new artificial regressors

$$\left(\frac{1}{2} - S_1\right) \left(\frac{1}{12} - S_1(1 - S_1)\right) x_{1m}^4.$$

On the other hand, the higher-order expansions introduce nonlinear functions of the lower-order moments, here represented by

$$\left(\frac{1}{2} - S_1\right) \left(\frac{1}{4} - 2S_1(1 - S_1)\right) \left(\sum_{m=1}^{n_X} \Sigma_{mm} x_{1m}^2\right)^2,$$

and the model is not linear in these parameters any more. This could be dealt with in several ways: by nonlinear optimization (of a very simple kind), or by iterative methods. In any case, we will see in our simulations that stopping with the second-order expansion often gives results that are already very reliable.

Finally, while the estimator based on the second-order expansion is “robust” to any (well-behaved) distribution, the estimator based on this fourth-order expansion also assumes symmetry: a skewed distribution would generate terms in σ^3 . Making more assumptions changes the form of the artificial regressors. To illustrate this, consider a mixed logit with one covariate only ($n_X = 1$). The expansion to order 8 can be written

$$\xi_1 = \log \frac{S_1}{S_0} - \beta X_1 + \sum_{k=1}^4 t_k(S_1) (\Sigma_{11} X_1^2)^k + O(\sigma^{10}).$$

Assume that ϵ has normal kurtosis. Then $k_1 = 3\Sigma_{11}^2$ and we find the simpler formula

$$t_2 = \alpha_4 = \left(\frac{1}{2} - S_1\right) S_1(1 - S_1).$$

Specializing further, Figure 1 plots the terms $t_k(S)$ for $k = 1, 2, 3, 4$ as the market share goes from zero to one for the particular case of a Gaussian ϵ . The visual impression is clear: the coefficients damp quickly. Beyond the first term (which corresponds to our 2SLS method), the coefficients are always smaller than 0.05 in

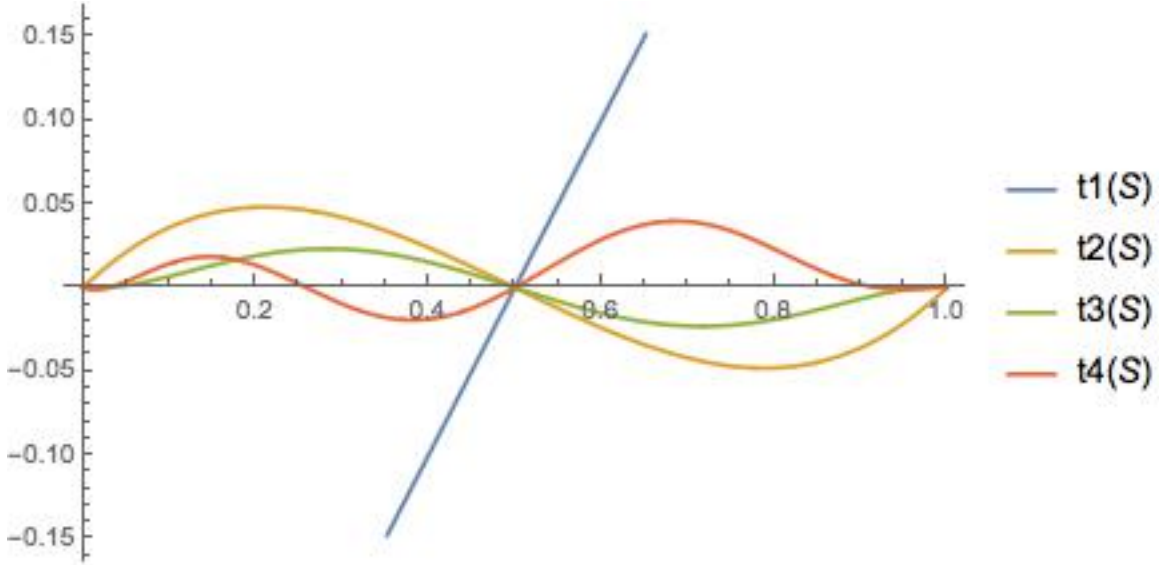


Figure 1: Coefficients $t_{1,2,3,4}(S)$

absolute value. Of course, the approximation error also depends on the values taken by the covariates.

To be more precise, if the components of ϵ are independently distributed and have third moments (s_1, \dots, s_{n_X}) , then it is easy to see that an additional term

$$\left(S_1(1 - S_1) - \frac{1}{6}\right) \sum_{m=1}^{n_X} s_m X_{1m}^3$$

enters the expansion. To test for skewness on covariate m , one could simply test for the significance of the regressor $(S_1(1 - S_1) - \frac{1}{6}) \sum_{m=1}^{n_X} s_m X_{1m}^3$.

B.3 Beyond Logit and Gaussian

The properties of the logistic function may seem to have been more central to our calculations; but in fact they are quite ancillary. Suppose that $u_{i1t} - u_{i0t}$ has some distribution with cdf Q instead of L . While the derivatives of Q may not obey the nice polynomial formulæ we used for L , it is still true that if Q is invertible and smooth then we can define functions F_k by

$$Q^{(k)}(t) = F_k(Q(t)).$$

This is all we need to carry out the expansions. One can show for instance that the factor $(S_1 - \frac{1}{2})$ that appears in (16) just needs to be replaced with

$$-\frac{F_2(S_1)}{2F_1(S_1)}.$$

Take for instance a mixed binary model with such a general distribution for $u_1 - u_0$, and a distribution of the random coefficient on the single covariate X_1 that has successive moments $0, \Sigma, \mu_3, \mu_4$. Then it is easy to derive the following fourth-order expansion, which could perhaps serve as the basis for a semiparametric estimator:

$$\begin{aligned} \xi_2 = & \log \frac{S_1}{S_0} - \beta X_1 - \frac{F_2(S_1)}{F_1(S_1)} X_1^2 \Sigma \\ & - \frac{F_3(S_1)}{F_1(S_1)} X_1^3 \mu_3 \\ & + \frac{F_2(S_1)}{F_1(S_1)} \left(3 \frac{F_3(S_1)}{F_1(S_1)} - \left(\frac{F_2(S_1)}{F_1(S_1)} \right)^2 \right) X_1^4 \Sigma^2 - \frac{F_4(S_1)}{F_1(S_1)} \mu_4 X_1^4 + O(\sigma^5). \end{aligned}$$

This can be extended in the obvious way to make v heteroskedastic (just replace Σ with $E(\epsilon^2|X_1)$ and μ_m with $E(\epsilon^m|X_1)$ in the above formula.)

C The Two-level Nested Logit

In the unmixed model ($\sigma = 0$) the mean utility of alternative j is $U_j = I_k + \lambda_k \log S_{j|N_k}$ if $j \in N_k$, with $I_k \equiv \log(S_{N_k}/S_0)$ and $S_{j|N_k} \equiv S_j/S_{N_k}$. This gives

$$\xi_j^0 = -\mathbf{X}_j \boldsymbol{\beta} + \log(S_{N_k}/S_0) + \lambda_k \log S_{j|N_k}.$$

We write (imposing $a_{1j} = 0$ from the start as this is a general property of models with $E\mathbf{v} = \mathbf{0}$)

$$U_j(\mathbf{v}) = \log(S_{N_k}/S_0) + \lambda_k \log S_{j|N_k} + \sigma \mathbf{x}_j \cdot \mathbf{v} + \frac{\sigma^2}{2} a_{2j}$$

and

$$\exp(I_k(\mathbf{v})/\lambda_k) = \sum_{j \in N_k} \exp(U_j(\mathbf{v})/\lambda_k) = (S_{N_k}/S_0)^{1/\lambda_k} \bar{f}_k(\mathbf{v})$$

where we denote $\bar{X}_k = \sum_{j \in N_k} S_{j|N_k} X_j$ and

$$f_j(\mathbf{v}) = \exp\left(\frac{\sigma}{\lambda_k} \left(\mathbf{x}_j \cdot \mathbf{v} + \sigma \frac{a_{2j}}{2}\right)\right) \simeq 1 + \frac{\sigma}{\lambda_k} (\mathbf{x}_j \cdot \mathbf{v}) + \frac{\sigma^2}{2\lambda_k^2} (\lambda_k a_{2j} + (\mathbf{x}_j \cdot \mathbf{v})^2)$$

so that

$$\bar{f}_k(\mathbf{v}) \simeq 1 + \frac{\sigma}{\lambda_k} \bar{\mathbf{x}}_k \cdot \mathbf{v} + \frac{\sigma^2}{2\lambda_k^2} (\lambda_k \bar{a}_{2k} + (\bar{\mathbf{x}} \cdot \mathbf{v})^2_k).$$

Now using

$$S_j = E_{\mathbf{v}} \exp((U_j(\mathbf{v}) - I_k(\mathbf{v}))/\lambda_k) \frac{\exp(I_k(\mathbf{v}))}{1 + \sum_{l=1}^K \exp(I_l(\mathbf{v}))}$$

we get

$$1 = E_{\mathbf{v}} \left(\frac{f_j(\mathbf{v})}{\bar{f}_k(\mathbf{v})} \frac{(\bar{f}_k(\mathbf{v}))^{\lambda_k}}{S_0 + \sum_{l=1}^K S_{N_l} (\bar{f}_l(\mathbf{v}))^{\lambda_l}} \right).$$

We note that

$$\frac{1 + a\sigma + b\sigma^2}{1 + c\sigma + d\sigma^2} = 1 + (a - c)\sigma + (b - d - c(a - c))\sigma^2 + O(\sigma^3). \quad (18)$$

Denote $\hat{A}_{j|k} = A_j - \bar{A}_k$. Applying (18) gives

$$\frac{f_j(\mathbf{v})}{\bar{f}_k(\mathbf{v})} \simeq 1 + \frac{\sigma}{\lambda_k} C_j(\mathbf{v}) + \frac{\sigma^2}{2\lambda_k^2} D_j(\mathbf{v}).$$

with

$$C_j(\mathbf{v}) = \hat{\mathbf{x}}_{j|k} \cdot \mathbf{v}$$

and

$$D_j(\mathbf{v}) = \lambda_k \widehat{a_{2j|k}} + (\widehat{\mathbf{x} \cdot \mathbf{v}})_{j|k}^2 - 2(\bar{\mathbf{x}}_k \cdot \mathbf{v})(\hat{\mathbf{x}}_{j|k} \cdot \mathbf{v}).$$

Moreover,

$$(\bar{f}_l(\mathbf{v}))^{\lambda_l} \simeq 1 + \sigma \bar{\mathbf{x}}_l \cdot \mathbf{v} + \frac{\sigma^2}{2} \left(\frac{\lambda_l - 1}{\lambda_l} (\bar{\mathbf{x}}_l \cdot \mathbf{v})^2 + \bar{a}_{2l} + \frac{(\bar{\mathbf{x}} \cdot \mathbf{v})_{l}^2}{\lambda_l} \right)$$

and

$$\frac{(\bar{f}_k(\mathbf{v}))^{\lambda_k}}{S_0 + \sum_{l=1}^K S_{N_l} (\bar{f}_l(\mathbf{v}))^{\lambda_l}} \simeq \frac{1 + \sigma \bar{\mathbf{x}}_k \cdot \mathbf{v} + \frac{\sigma^2}{2} \left(\bar{a}_{2k} + \frac{\lambda_k - 1}{\lambda_k} (\bar{\mathbf{x}}_k \cdot \mathbf{v})^2 + \frac{(\bar{\mathbf{x}} \cdot \mathbf{v})_{k}^2}{\lambda_k} \right)}{1 + \sigma e_{\mathbf{S}} \mathbf{x} \cdot \mathbf{v} + \frac{\sigma^2}{2} \left(e_{\mathbf{S}} a_2 + \sum_{l=1}^K S_{N_l} \left(\frac{\lambda_l - 1}{\lambda_l} (\bar{\mathbf{x}}_l \cdot \mathbf{v})^2 + \frac{(\bar{\mathbf{x}} \cdot \mathbf{v})_{l}^2}{\lambda_l} \right) \right)}$$

where as usual $e_{\mathbf{S}} \mathbf{T} = \sum_{j=1}^J S_j T_j = \sum_{k=1}^K S_{N_k} \bar{T}_k$.

Then, using (18) again,

$$\frac{(\bar{f}_k(\mathbf{v}))^{\lambda_k}}{S_0 + \sum_{l=1}^K S_{N_l} (\bar{f}_l(\mathbf{v}))^{\lambda_l}} \simeq 1 + \sigma E_k(\mathbf{v}) + \frac{\sigma^2}{2} F_k(\mathbf{v})$$

with

$$E_k(\mathbf{v}) = (\bar{\mathbf{x}}_k - e_{\mathcal{S}} \mathbf{x}) \cdot \mathbf{v}$$

and

$$\begin{aligned} F_k(\mathbf{v}) &= \bar{a}_{2k} - e_{\mathcal{S}} a_2 \\ &+ \frac{\lambda_k - 1}{\lambda_k} (\bar{\mathbf{x}}_k \cdot \mathbf{v})^2 - \sum_{l=1}^K S_{N_l} \frac{\lambda_l - 1}{\lambda_l} (\bar{\mathbf{x}}_l \cdot \mathbf{v})^2 \\ &+ \frac{(\overline{\mathbf{x} \cdot \mathbf{v}})_k^2}{\lambda_k} - \sum_{l=1}^K S_{N_l} \frac{(\overline{\mathbf{x} \cdot \mathbf{v}})_l^2}{\lambda_l} \\ &- 2(e_{\mathcal{S}} \mathbf{x} \cdot \mathbf{v})((\bar{\mathbf{x}}_k - e_{\mathcal{S}} \mathbf{x}) \cdot \mathbf{v}). \end{aligned}$$

This allows us to write

$$\begin{aligned} 1 &\simeq E_{\mathbf{v}} \left(1 + \frac{\sigma}{\lambda_k} C_j + \frac{\sigma^2}{2\lambda_k^2} D_j \right) \left(1 + \sigma E_k + \frac{\sigma^2}{2} F_k \right) \\ &\simeq E_{\mathbf{v}} \left(1 + \sigma \left(\frac{C_j}{\lambda_k} + E_k \right) + \frac{\sigma^2}{2\lambda_k^2} (D_j + \lambda_k^2 F_k + 2\lambda_k C_j E_k) \right). \end{aligned}$$

We have $E_{\mathbf{v}} C_j = E_{\mathbf{v}} E_k = 0$; also,

$$\begin{aligned} ED_j &= \lambda_k \hat{a}_{2j|k} + \|\mathbf{x}_j\|^2 - \|\overline{\mathbf{x}}\|_k^2 - 2\bar{\mathbf{x}}_k \cdot \hat{\mathbf{x}}_{j|k} \\ EF_k &= \bar{a}_{2k} - e_{\mathcal{S}} a_2 \\ &+ \frac{\lambda_k - 1}{\lambda_k} \|\bar{\mathbf{x}}_k\|^2 - \sum_{l=1}^K S_{N_l} \frac{\lambda_l - 1}{\lambda_l} \|\bar{\mathbf{x}}_l\|^2 \\ &+ \frac{\|\overline{\mathbf{x}}\|_k^2}{\lambda_k} - \sum_{l=1}^K S_{N_l} \frac{\|\overline{\mathbf{x}}\|_l^2}{\lambda_l} \\ &- 2(e_{\mathcal{S}} \mathbf{x}) \cdot (\bar{\mathbf{x}}_k - e_{\mathcal{S}} \mathbf{x}) \\ E(C_j E_k) &= \hat{\mathbf{x}}_{j|k} \cdot (\bar{\mathbf{x}}_k - e_{\mathcal{S}} \mathbf{x}). \end{aligned}$$

Writing $E(D_j + \lambda_k^2 F_k + 2\lambda_k C_j E_k) = 0$ gives us an equation of the form

$$\lambda_k (a_{2j} - \bar{a}_{2k}) + \lambda_k^2 (\bar{a}_{2k} - e_{\mathcal{S}} a_2) = \lambda_k^2 M + \nu_k + \mu_j$$

where

$$\begin{aligned}
M &= \sum_{l=1}^K S_{N_l} \frac{\lambda_l - 1}{\lambda_l} \|\bar{\mathbf{x}}_l\|^2 + \sum_{l=1}^K S_{N_l} \frac{\|\bar{\mathbf{x}}\|_l^2}{\lambda_l} - 2\|e_{\mathcal{S}}\mathbf{x}\|^2 \\
\nu_k &= \|\bar{\mathbf{x}}\|_k^2 - 2\|\bar{\mathbf{x}}_k\|^2 - \lambda_k(\lambda_k - 1)\|\bar{\mathbf{x}}_k\|^2 - \lambda_k\|\bar{\mathbf{x}}\|_k^2 + 2\lambda_k^2 e_{\mathcal{S}}\mathbf{x} \cdot \bar{\mathbf{x}}_k + 2\lambda_k\|\bar{\mathbf{x}}_k\|^2 - 2\lambda_k\bar{\mathbf{x}}_k \cdot e_{\mathcal{S}}\mathbf{x} \\
&= (1 - \lambda_k) (\|\bar{\mathbf{x}}\|_k^2 - (2 - \lambda_k)\|\bar{\mathbf{x}}_k\|^2 - 2\lambda_k\bar{\mathbf{x}}_k \cdot e_{\mathcal{S}}\mathbf{x}) \tag{19}
\end{aligned}$$

$$\begin{aligned}
\mu_j &= -\|\mathbf{x}_j\|^2 + 2\mathbf{x}_j \cdot \bar{\mathbf{x}}_k - 2\lambda_k\mathbf{x}_j \cdot (\bar{\mathbf{x}}_k - e_{\mathcal{S}}\mathbf{x}) \\
&= \mathbf{x}_j \cdot (2\lambda_k e_{\mathcal{S}}\mathbf{x} - \mathbf{x}_j + 2(1 - \lambda_k)\bar{\mathbf{x}}_k). \tag{20}
\end{aligned}$$

It is easy to aggregate from $a_{2j} = (1 - \lambda_k)\bar{a}_{2k} + \lambda_k e_{\mathcal{S}}a_2 + \lambda_k M + (\nu_k + \mu_j)/\lambda_k$ to

$$\bar{a}_{2k} = e_{\mathcal{S}}a_2 + M + \frac{\nu_k + \bar{\mu}_k}{\lambda_k^2}$$

and then to

$$S_0 e_{\mathcal{S}}a_2 = (1 - S_0)M + \sum_{k=1}^K S_{N_k} \frac{\nu_k + \bar{\mu}_k}{\lambda_k^2},$$

which gives

$$\begin{aligned}
a_{2j} &= e_{\mathcal{S}}a_2 + M + (1 - \lambda_k) \frac{\nu_k + \bar{\mu}_k}{\lambda_k^2} + \frac{\nu_k + \mu_j}{\lambda_k} \\
&= \frac{M}{S_0} + \frac{1}{S_0} \sum_{l=1}^K S_{N_l} \frac{\nu_l + \bar{\mu}_l}{\lambda_l^2} + (1 - \lambda_k) \frac{\nu_k + \bar{\mu}_k}{\lambda_k^2} + \frac{\nu_k + \mu_j}{\lambda_k} \\
&= \frac{M}{S_0} + \frac{1}{S_0} \sum_{l=1}^K S_{N_l} \frac{\nu_l + \bar{\mu}_l}{\lambda_l^2} + \frac{\nu_k + (1 - \lambda_k)\bar{\mu}_k}{\lambda_k^2} + \frac{\mu_j}{\lambda_k}.
\end{aligned}$$

Finally, using equations (19) and (20) we aggregate

$$\bar{\mu}_k = 2\lambda_k\bar{\mathbf{x}}_k \cdot e_{\mathcal{S}}\mathbf{x} + 2(1 - \lambda_k)\|\bar{\mathbf{x}}_k\|^2 - \|\bar{\mathbf{x}}\|_k^2,$$

which gives

$$\nu_k + \bar{\mu}_k = 2\lambda_k^2\bar{\mathbf{x}}_k \cdot e_{\mathcal{S}}\mathbf{x} + \lambda_k(1 - \lambda_k)\|\bar{\mathbf{x}}_k\|^2 - \lambda_k\|\bar{\mathbf{x}}\|_k^2$$

and

$$\nu_k + (1 - \lambda_k)\bar{\mu}_k = -\lambda_k(1 - \lambda_k)\|\bar{\mathbf{x}}_k\|^2.$$

Putting everything together, we get

$$\begin{aligned}
a_{2j} &= \frac{M}{S_0} + \frac{1}{S_0} \sum_{l=1}^K S_{N_l} \frac{\nu_l + \bar{\mu}_l}{\lambda_l^2} + \frac{\nu_k + (1 - \lambda_k)\bar{\mu}_k}{\lambda_k^2} + \frac{\mu_j}{\lambda_k} \\
&= \frac{1}{S_0} \left(\sum_{l=1}^K S_{N_l} \frac{\lambda_l - 1}{\lambda_l} \|\bar{\mathbf{x}}_l\|^2 + \sum_{l=1}^K S_{N_l} \frac{\|\mathbf{x}\|_l^2}{\lambda_l} - 2\|e_{\mathbf{S}}\mathbf{x}\|^2 \right) \\
&\quad + \frac{2}{S_0} \|e_{\mathbf{S}}\mathbf{x}\|^2 + \frac{1}{S_0} \sum_{l=1}^K S_{N_l} \frac{-\|\mathbf{x}\|_l^2 + (1 - \lambda_l)\|\bar{\mathbf{x}}_l\|^2}{\lambda_l} \\
&= \mathbf{x}_j \cdot \left(2e_{\mathbf{S}}\mathbf{x} - \frac{\mathbf{x}_j}{\lambda_k} + 2\frac{1 - \lambda_k}{\lambda_k} \bar{\mathbf{x}}_k \right) - \frac{1 - \lambda_k}{\lambda_k} \|\bar{\mathbf{x}}_k\|^2.
\end{aligned}$$

D Bias Correction Formula

Remember from section 6.3.2 that

$$\boldsymbol{\theta}_0 \simeq \boldsymbol{\theta}_2 - \left(E \frac{\partial f_\infty}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2; \boldsymbol{\lambda}_0) \right)^{-1} f_\infty(\boldsymbol{\theta}_2; \boldsymbol{\lambda}_0). \quad (21)$$

The term in the inverse is easily proxied:

$$E \frac{\partial f_\infty}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2; \boldsymbol{\lambda}_0) \simeq E \frac{\partial f_2}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2) = E \left(\frac{\partial \boldsymbol{\xi}_2}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2) \right)' \mathbf{V} \mathbf{V}' \frac{\partial \boldsymbol{\xi}_2}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2),$$

since $\boldsymbol{\xi}_2$ is linear in $\boldsymbol{\theta}$. Note that this is $E \boldsymbol{\mathcal{X}}' \boldsymbol{\mathcal{X}}$, where

$$\boldsymbol{\mathcal{X}} \equiv \mathbf{V}' \frac{\partial \boldsymbol{\xi}_2}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2) = -\mathbf{V}'(\mathbf{X}, \mathbf{K})$$

and row $j = 1, \dots, J$ of (\mathbf{X}, \mathbf{K}) lists the covariates and artificial regressors for this product. It follows that

$$E \frac{\partial f_\infty}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2; \boldsymbol{\lambda}_0) \simeq \begin{pmatrix} E(\mathbf{X}' \mathbf{V} \mathbf{V}' \mathbf{X}) & E(\mathbf{X}' \mathbf{V} \mathbf{V}' \mathbf{K}) \\ E(\mathbf{K}' \mathbf{V} \mathbf{V}' \mathbf{X}) & E(\mathbf{K}' \mathbf{V} \mathbf{V}' \mathbf{K}) \end{pmatrix}.$$

To the second-order in \mathbf{e}_2 , $E f_\infty(\boldsymbol{\theta}_2; \boldsymbol{\lambda}_0)$ equals

$$E \left(\frac{\partial \boldsymbol{\xi}_2}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2)' \mathbf{V} \mathbf{V}' \hat{\mathbf{e}}_2(\boldsymbol{\theta}_2; \boldsymbol{\lambda}_0) \right) + E \left(\frac{\partial \hat{\mathbf{e}}_2}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2; \boldsymbol{\lambda}_0)' \mathbf{V} \mathbf{V}' \boldsymbol{\xi}_2(\boldsymbol{\theta}_2) \right). \quad (22)$$

The first term in (22) is simply $E(\mathbf{X}'\mathbf{V}'\hat{\boldsymbol{\epsilon}}_2)$. Going back to (21), we get

$$\boldsymbol{\theta}_0 \simeq \boldsymbol{\theta}_2 - \begin{pmatrix} E(\mathbf{X}'\mathbf{V}\mathbf{V}'\mathbf{X}) & E(\mathbf{X}'\mathbf{V}\mathbf{V}'\mathbf{K}) \\ E(\mathbf{K}'\mathbf{V}\mathbf{V}'\mathbf{X}) & E(\mathbf{K}'\mathbf{V}\mathbf{V}'\mathbf{K}) \end{pmatrix}^{-1} \left(E(\mathbf{X}'\mathbf{V}'\hat{\boldsymbol{\epsilon}}_2) + E\left(\frac{\partial \hat{\boldsymbol{\epsilon}}_2'}{\partial \boldsymbol{\theta}} \mathbf{V}\mathbf{V}'\boldsymbol{\xi}_2\right) \right).$$

Finally, using Theorem 1(i), we know that $\frac{\partial \hat{\boldsymbol{\epsilon}}_2}{\partial \boldsymbol{\beta}} = \mathbf{0}$. Therefore

$$E(\mathbf{X}'\mathbf{V}'\hat{\boldsymbol{\epsilon}}_2) + E\left(\frac{\partial \hat{\boldsymbol{\epsilon}}_2'}{\partial \boldsymbol{\theta}} \mathbf{V}\mathbf{V}'\boldsymbol{\xi}_2\right) = \begin{pmatrix} -E(\mathbf{X}'\mathbf{V}\mathbf{V}'\hat{\boldsymbol{\epsilon}}_2) \\ -E(\mathbf{K}'\mathbf{V}\mathbf{V}'\hat{\boldsymbol{\epsilon}}_2) + E\left(\frac{\partial \hat{\boldsymbol{\epsilon}}_2'}{\partial \boldsymbol{\Sigma}} \mathbf{V}\mathbf{V}'\boldsymbol{\xi}_2\right) \end{pmatrix}. \quad \square$$

E Proof of Theorem 3

We drop the bold letters in this proof to alleviate the notation, and without loss of generality we normalize $B = 1$.

Remember that $G(y, F(y, \beta, \sigma), \beta, \sigma) = 0$, so that $G(y, F(y, \beta, 0), \beta, 0) = 0$. Given (10), this gives $G^*(y, A^*(y, F(y, \beta, 0) - f_1(y)\beta, 0) - f_1(y)\beta, 0) = 0$ for all β . This can only hold if $F(y, \beta, 0) - f_1(y)\beta$ does not depend on β , which implies condition **C2**. Denoting $f_0(y) = F(y, \beta, 0) - f_1(y)\beta$, we obtain

$$G^*(y, A^*(y, f_0(y), 0)) = 0.$$

Now writing $G^*(y, E_v A^*(y, F(y, \beta, \sigma) - f_1(y)\beta, \sigma v)) = 0$ as an identity in σ and taking derivatives with respect to σ , we get

$$\begin{aligned} G_2^* E_v (A_2^* F_\sigma + A_3^* v) &= 0 \\ G_{22}^* [E_v (A_2^* F_\sigma + A_3^* v)] [E_v (A_2^* F_\sigma + A_3^* v)] \\ + G_2^* E_v (A_2^* F_{\sigma\sigma} + A_{22}^* [F_\sigma, F_\sigma] + 2A_{23}^* [F_\sigma, v] + A_{33}^* [v, v]) &= 0. \end{aligned}$$

Fortunately, this simplifies greatly at $\sigma = 0$. The first equation gives

$$G_2^* E_v (A_2^* F_\sigma(y, \beta, 0) + A_3^* v) = 0,$$

where the derivatives A_2^* and A_3^* do not depend on v since $\sigma = 0$. It follows that $G_2^* A_2^* F_\sigma(y, \beta_0, 0) = 0$ since $E v = 0$. Given our invertibility assumption, condition **C1** also holds. Using the second equation at $\sigma = 0$, and given that $F_\sigma(y, \beta_0, 0) = 0$, we get

$$G_2^* E_v (A_{22}^* [F_\sigma, F_\sigma] + 2A_{23}^* [F_\sigma, v]) = 0$$

so that

$$G_2^* (A_2^* F_{\sigma\sigma} + A_{33}^*) = 0.$$

Given that G_2^* is invertible, this gives (reintroducing the arguments)

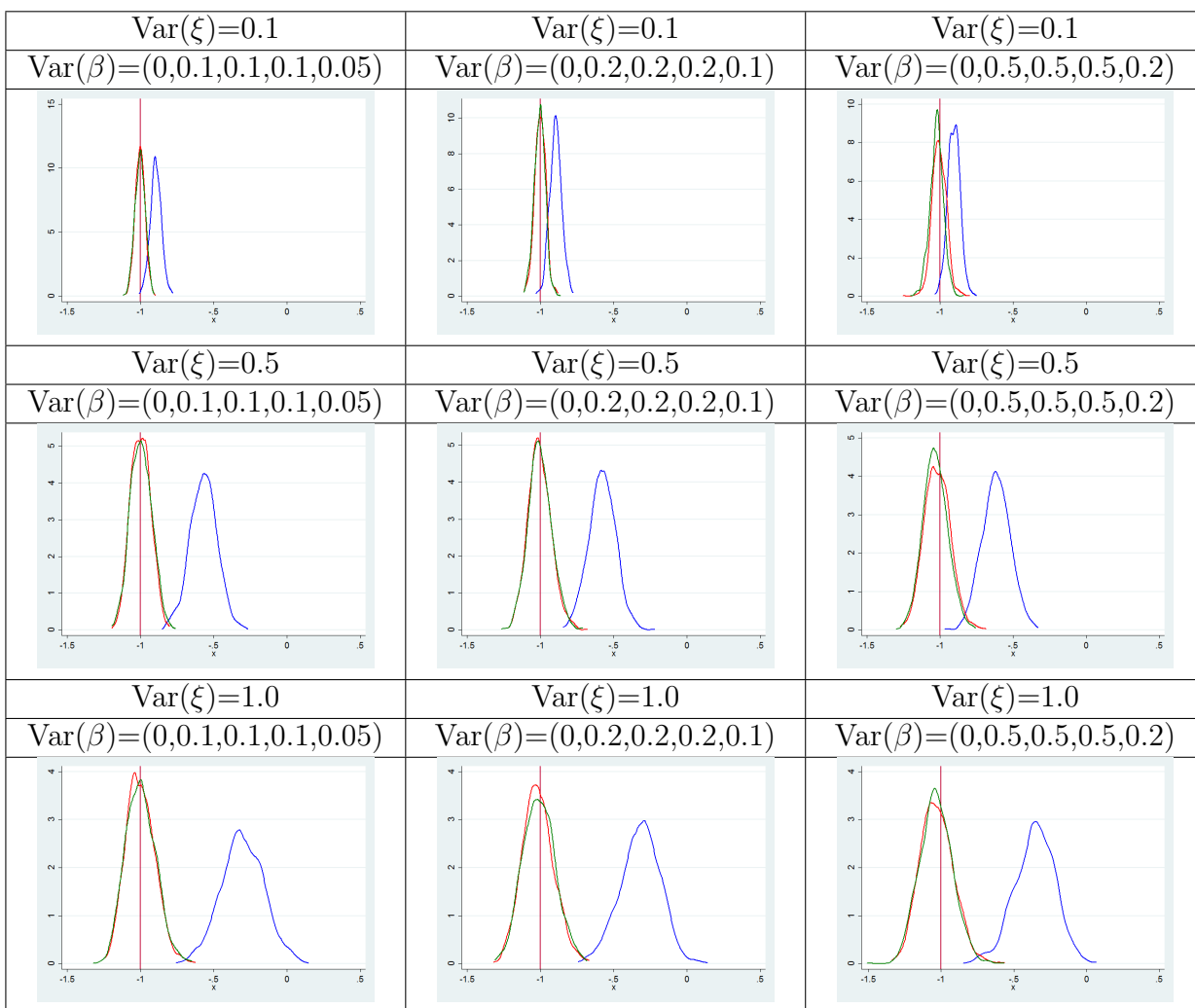
$$A_2^*(y, f_0(y), 0)F_{\sigma\sigma}(y, \beta, 0) + A_{33}^*(y, f_0(y), 0) = 0.$$

Therefore $F_{\sigma\sigma}(y, \beta, 0)$ is independent of β and condition **C3** holds. Noting that $f_2(y) = -F_{\sigma\sigma}(y, \beta, 0)$ completes the proof.

List of Tables

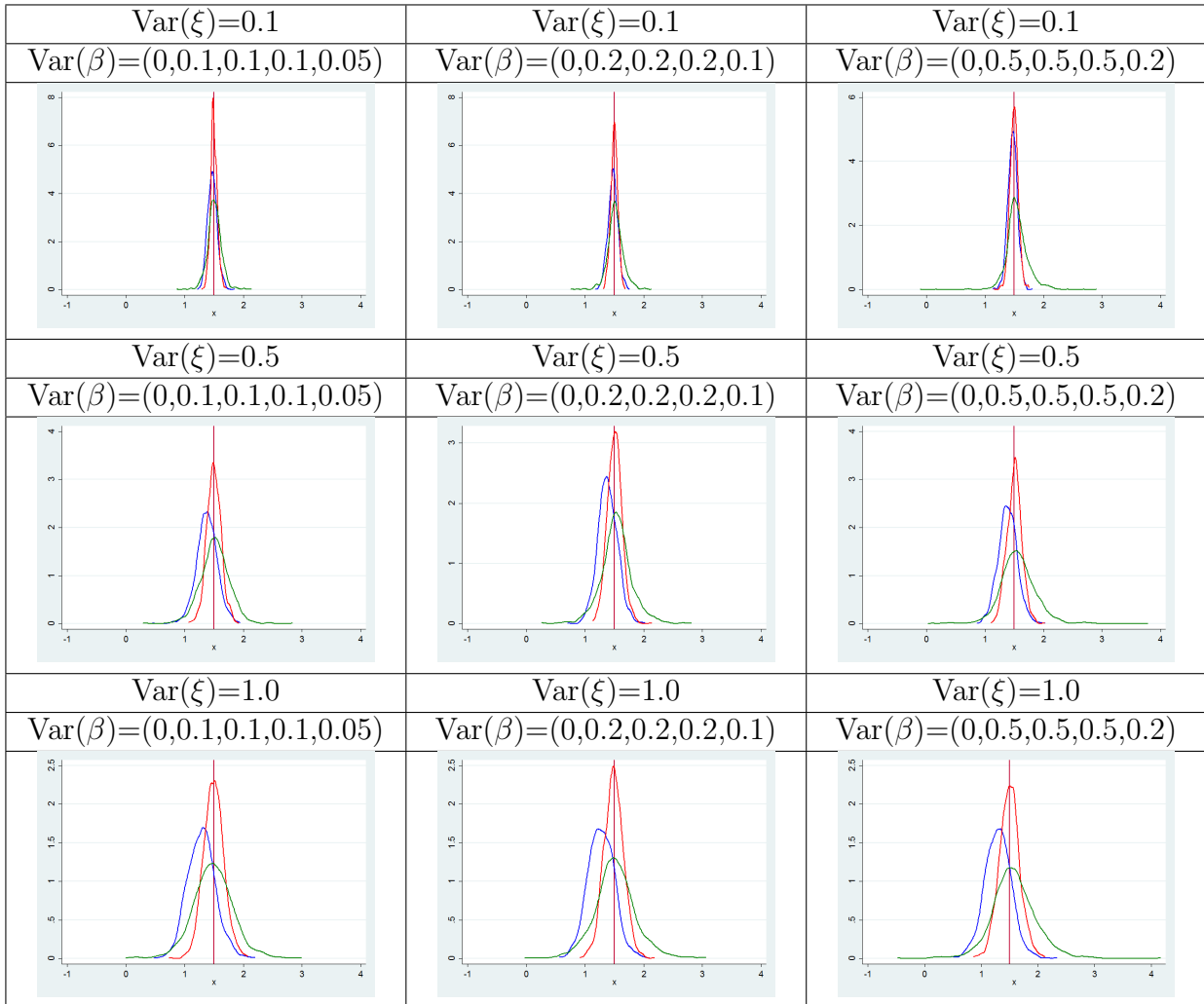
1	Distribution of the Estimates of β_0	55
2	Distribution of the Estimates of $\bar{\beta}_1^x$	56
3	Distribution of the Estimates of $\text{Var}(\beta_1^x)$	57
4	Distribution of the Estimates of $\bar{\beta}_2^x$	58
5	Distribution of the Estimates of $\text{Var}(\beta_2^x)$	59
6	Distribution of the Estimates of $\bar{\beta}_3^x$	60
7	Distribution of the Estimates of $\text{Var}(\beta_3^x)$	61
8	Distribution of the Estimates of $\bar{\beta}^p$	62
9	Distribution of the Estimates of $\text{Var}(\beta^p)$	63
10	Pseudo True Value: Increasing-number-of-markets Approach	64
11	Pseudo True Value: Moment-based Approach	65
12	Distribution of the Difference between True and Estimated Elasticity	66
13	Distribution of the Estimates of the Means — Different $\bar{\beta}^p$	67
14	Distribution of the Estimates of the Variances — Different $\bar{\beta}^p$	68
15	Testing for Zero Means — Standard 2SLS	69
16	Testing for Zero Means — Bias Corrected 2SLS	70
17	Testing for Zero Variances — Standard 2SLS	71
18	Testing for Zero Variances — Bias Corrected 2SLS	72
19	Joint Test of Zero Means and Variances — Standard 2SLS	73
20	Joint Test of Zero Means and Variances — Bias-Corrected 2SLS	74
21	Distribution of the Estimates of the Means (Lognormal Case)	75
22	Distribution of the Estimates of the Variances (Lognormal Case)	76
23	Distribution of the Estimates of the Third-order Moments (Lognormal Case)	77
24	Summary Statistics for the Lognormal Case	78
25	Distribution of Three Estimates of the Means — Normal and Lognormal	79
26	Distribution of Three Estimates of the Variances — Normal and Lognormal	80

Table 1: Distribution of the Estimates of β_0



Control Function 
MPEC 
2SLS 

Table 2: Distribution of the Estimates of $\bar{\beta}_1^x$






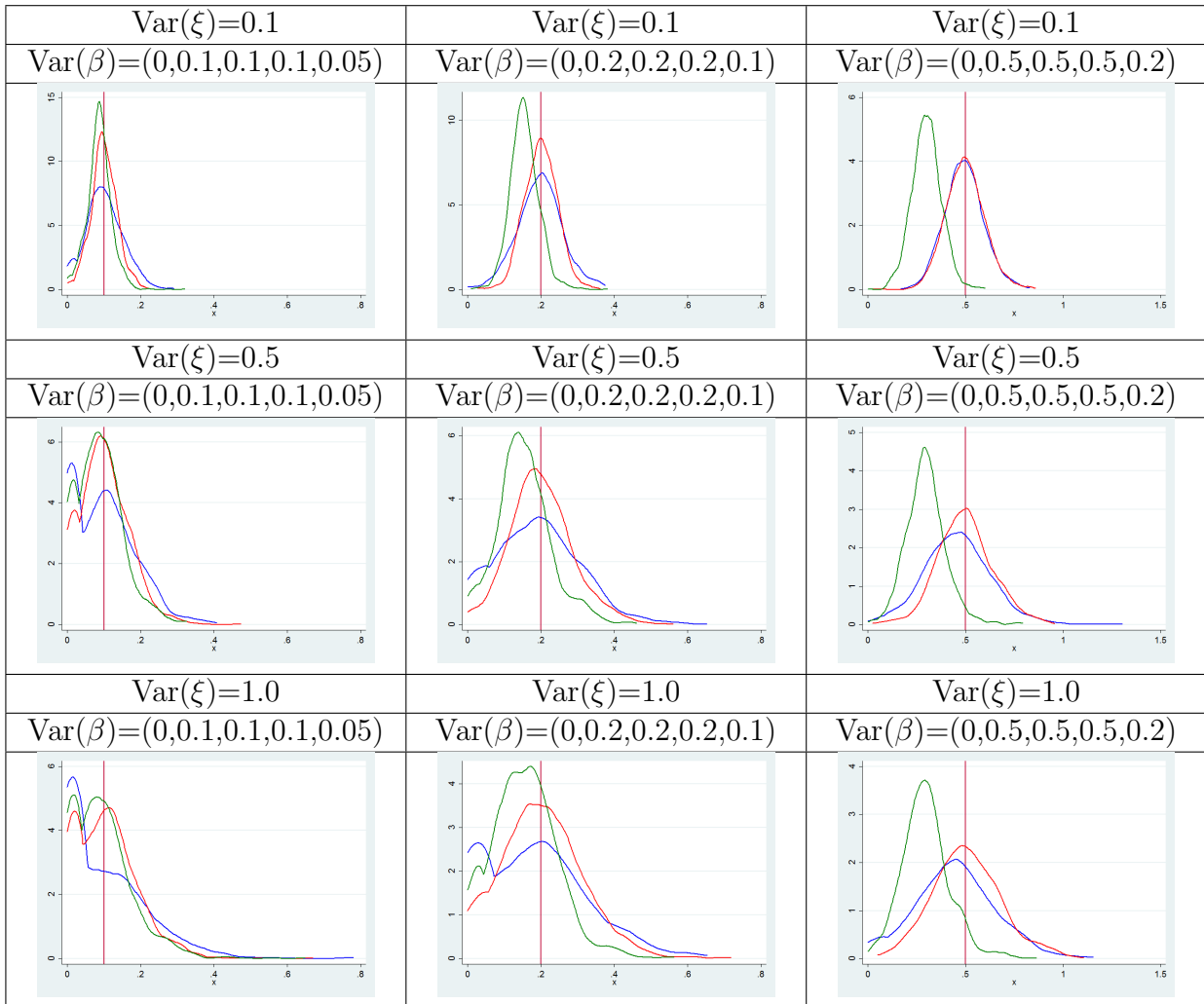
Control Function 
MPEC 
2SLS 

Table 3: Distribution of the Estimates of $\text{Var}(\beta_1^x)$






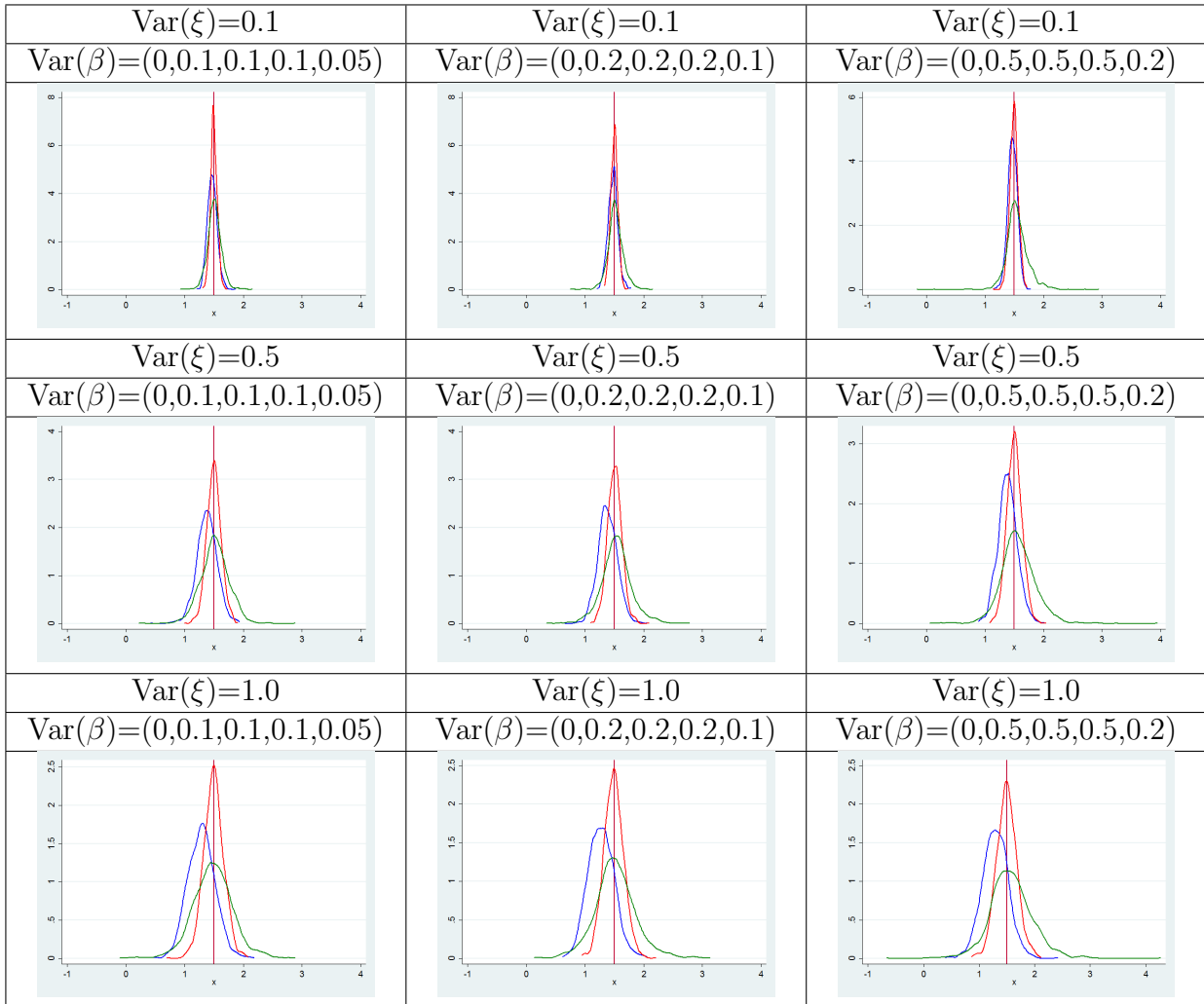
Control Function 
MPEC 
2SLS 

Table 4: Distribution of the Estimates of $\bar{\beta}_2^x$




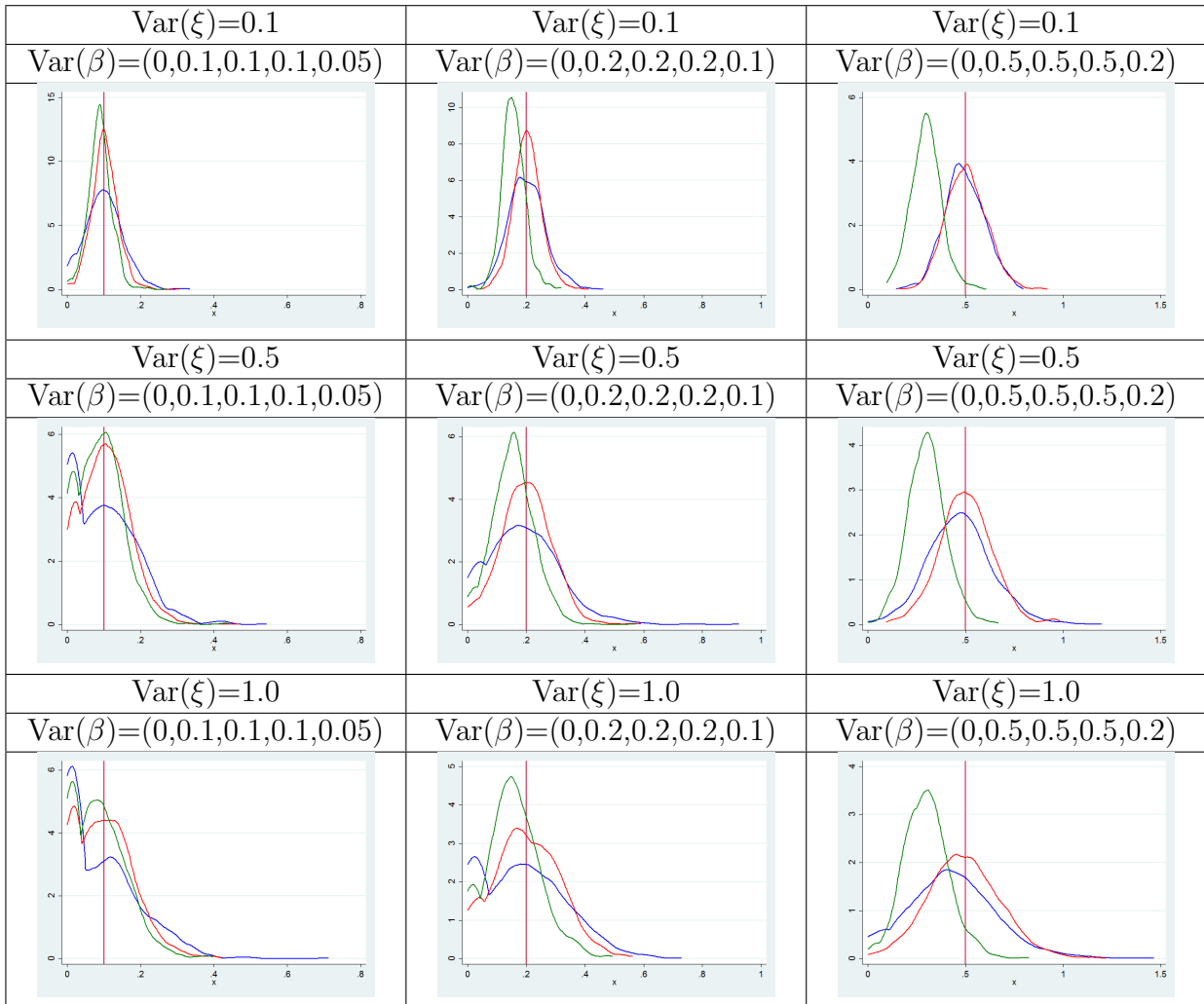
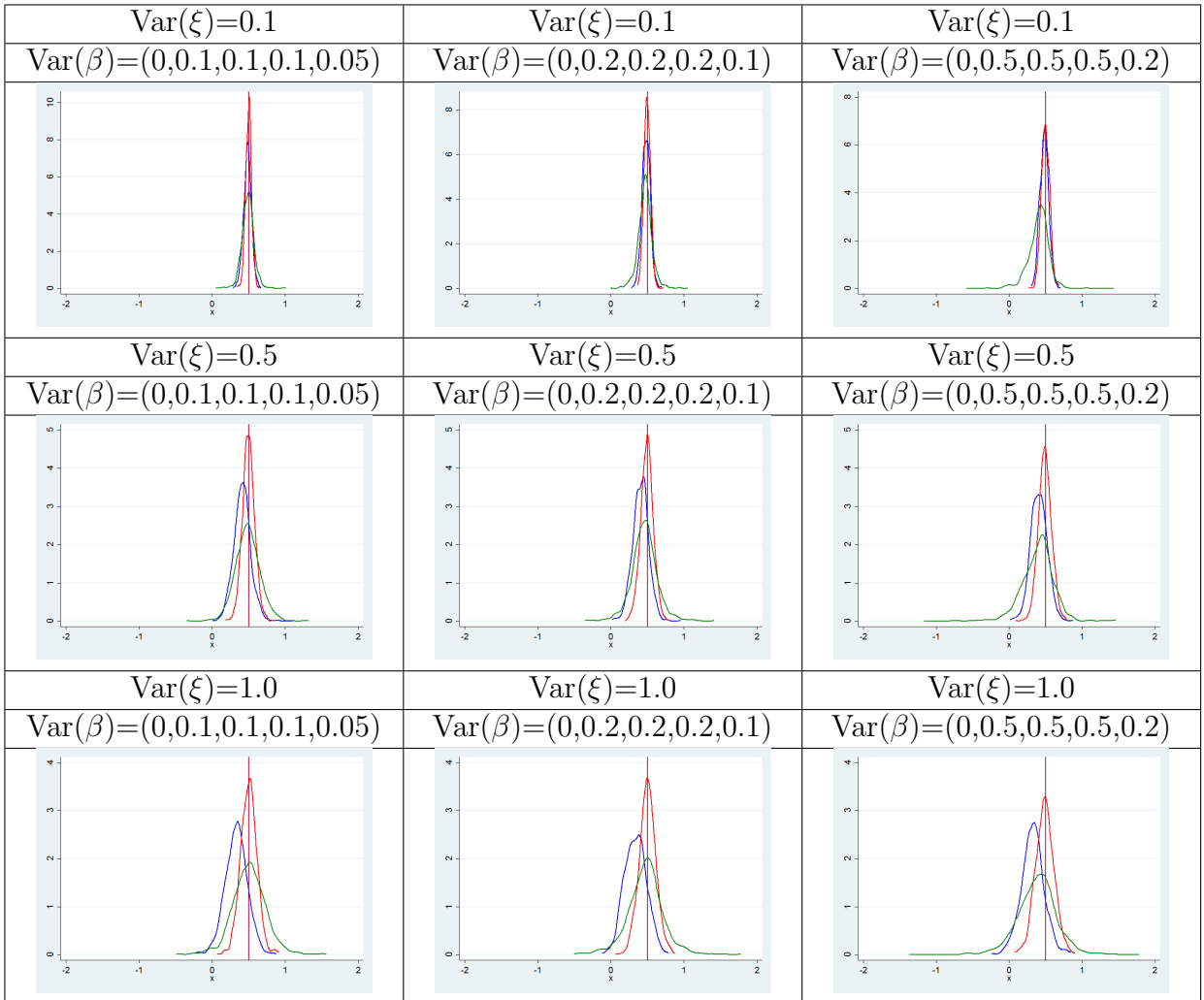
Control Function 
MPEC 
2SLS 

Table 5: Distribution of the Estimates of $\text{Var}(\beta_2^x)$



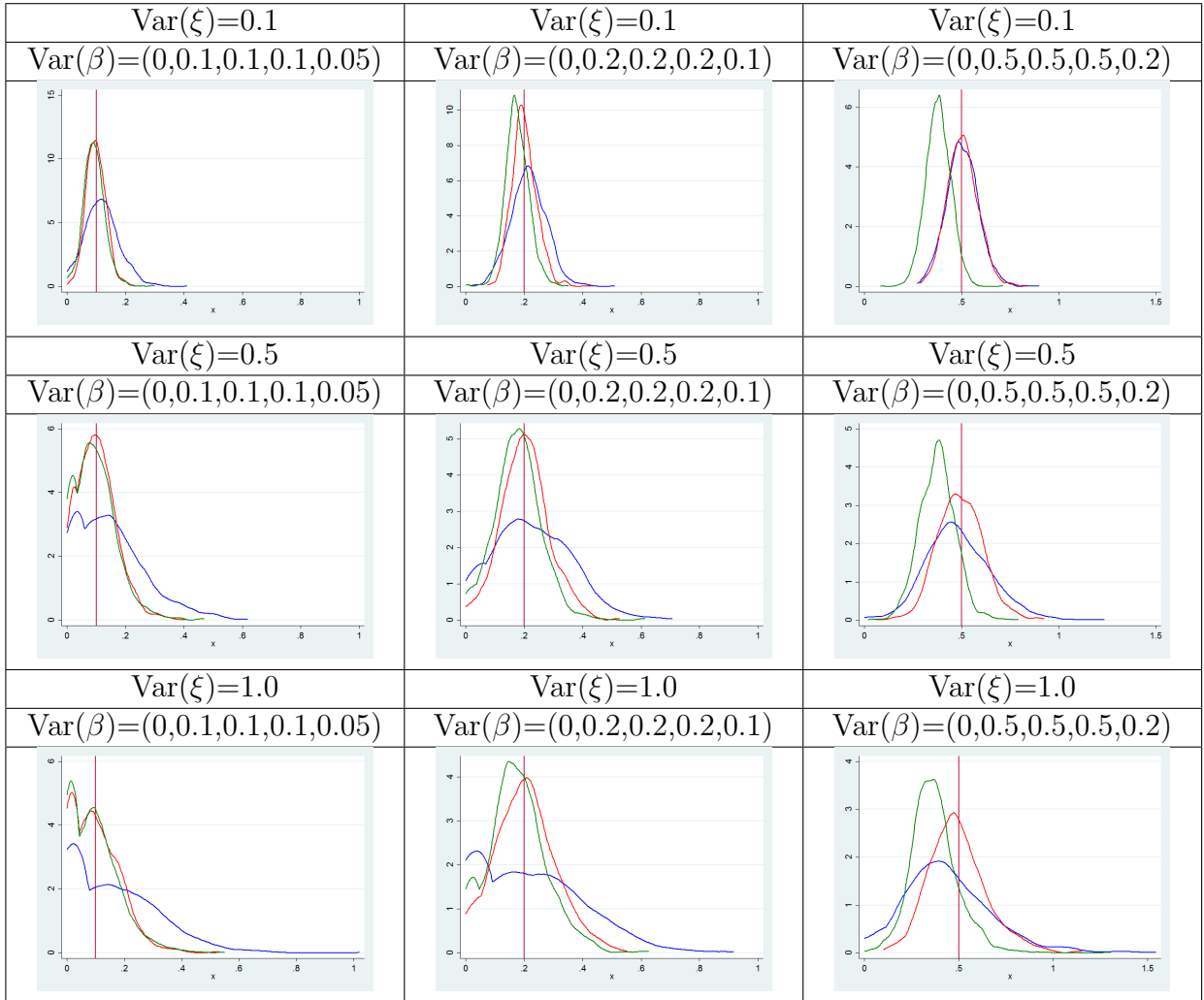
Control Function
MPEC
2SLS

Table 6: Distribution of the Estimates of $\bar{\beta}_3^x$



Control Function 
MPEC 
2SLS 

Table 7: Distribution of the Estimates of $\text{Var}(\beta_3^x)$






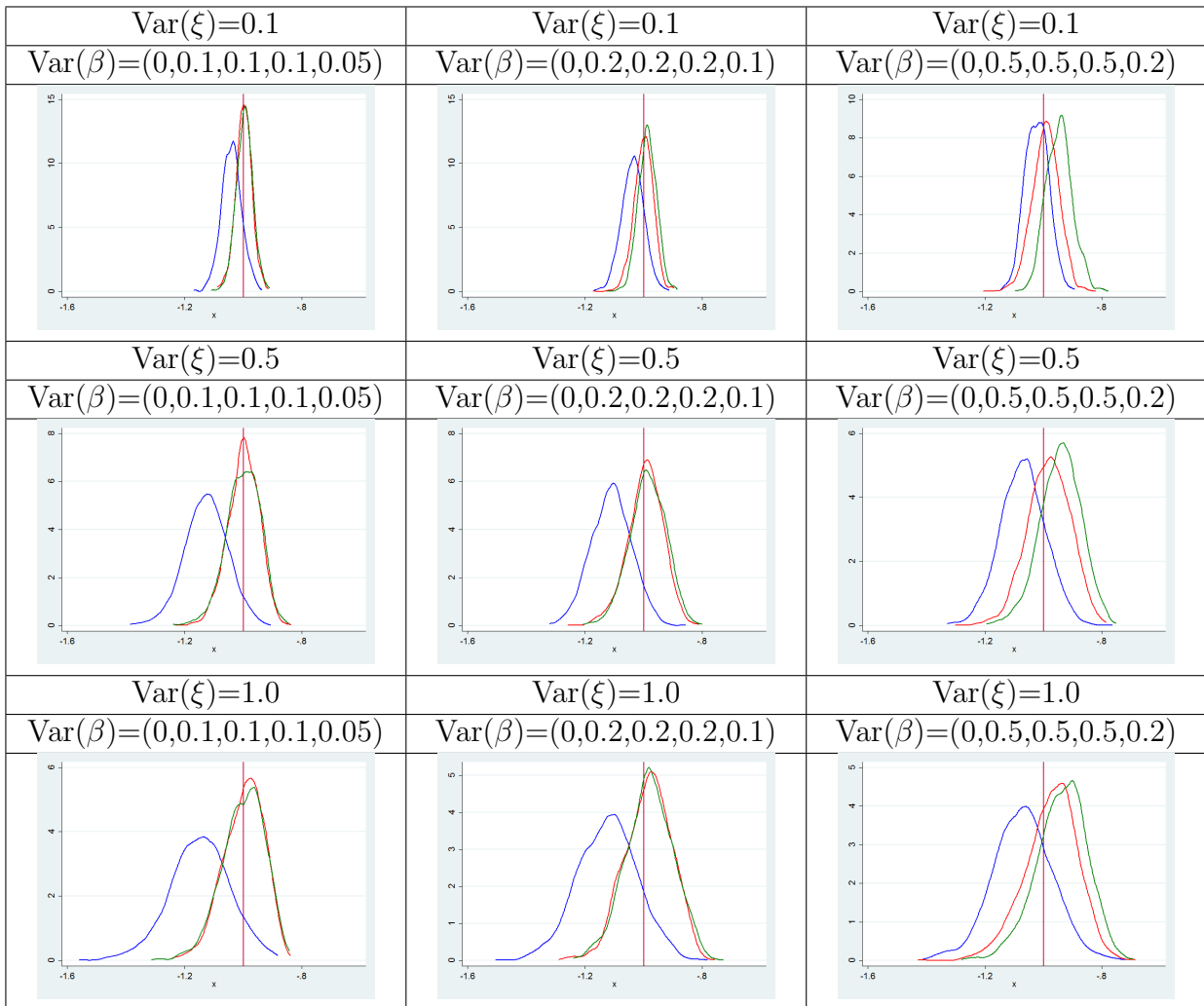
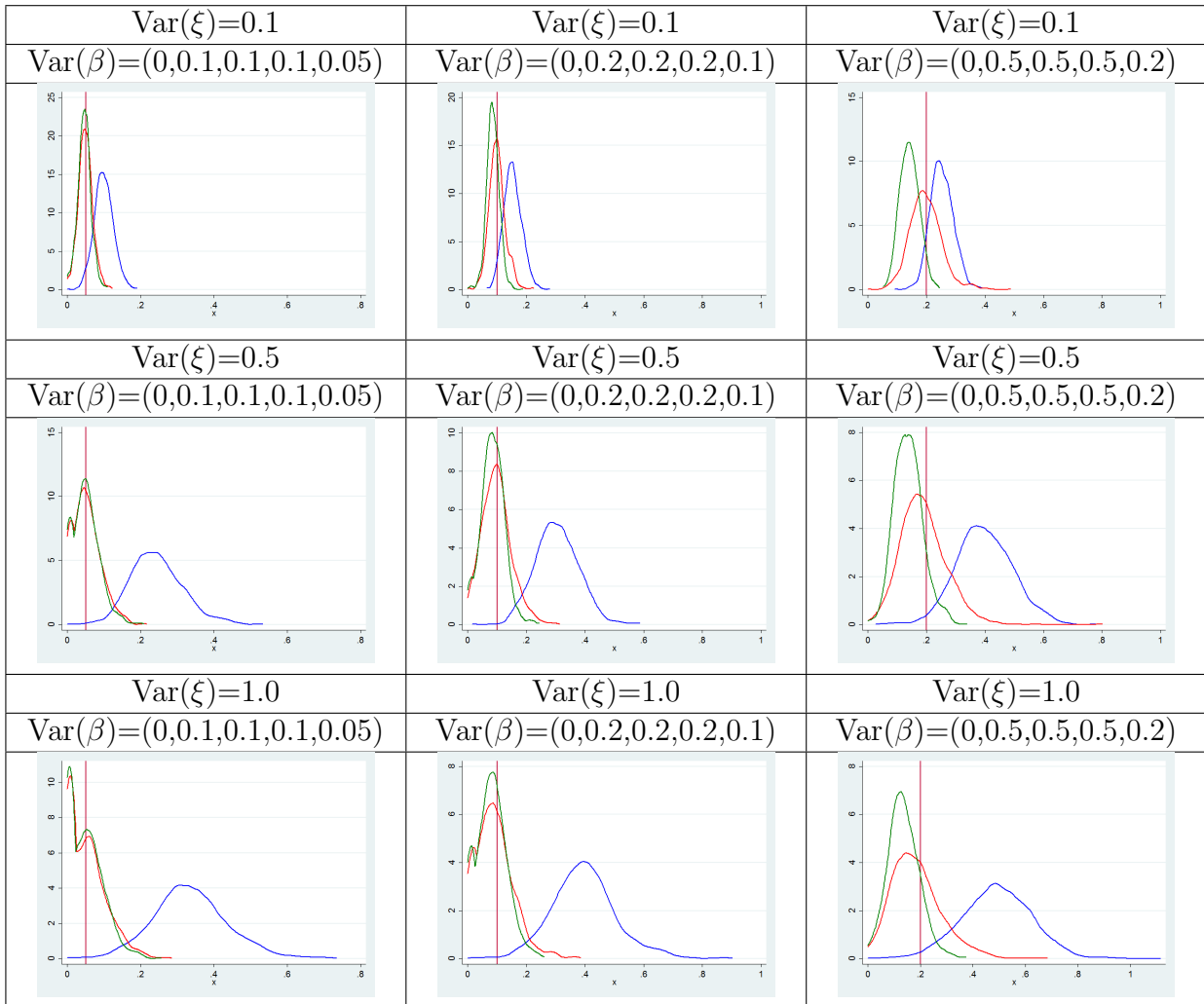
Control Function 
MPEC 
2SLS 

Table 8: Distribution of the Estimates of $\bar{\beta}^p$



Control Function 
MPEC 
2SLS 

Table 9: Distribution of the Estimates of $\text{Var}(\beta^p)$



Control Function
MPEC
2SLS

Table 10: Pseudo True Value: Increasing-number-of-markets Approach

Parameter	Scenarios								
	(0, 0.1, 0.1, 0.1, 0.05)			(0, 0.2, 0.2, 0.2, 0.1)			(0, 0.5, 0.5, 0.5, 0.2)		
True Var(β) :	0.1	0.5	1	0.1	0.5	1	0.1	0.5	1
$\beta_0 = -1$	-1.00 (.0043)	-1.00 (.0050)	-1.00 (.0058)	-1.00 (.011)	-1.00 (.012)	-1.00 (.013)	-1.02 (.032)	-1.03 (.035)	-1.03 (.038)
$\bar{\beta}_1^x = 1.5$	1.51 (.022)	1.51 (.023)	1.51 (.024)	1.53 (.050)	1.53 (.050)	1.53 (.050)	1.56 (.13)	1.57 (.13)	1.57 (.13)
$\bar{\beta}_2^x = 1.5$	1.51 (.023)	1.51 (.024)	1.51 (.025)	1.52 (.048)	1.52 (.049)	1.52 (.049)	1.55 (.12)	1.56 (.12)	1.56 (.12)
$\bar{\beta}_3^x = 0.5$	0.487 (.022)	0.487 (.022)	0.487 (.022)	0.465 (.048)	0.465 (.047)	0.464 (.047)	0.403 (.12)	0.400 (.12)	0.398 (.11)
$\bar{\beta}^p = -1$	-0.999 (.0086)	-0.999 (.0088)	-0.999 (.0090)	-0.990 (.0184)	-0.990 (.0186)	-0.990 (.0188)	-0.954 (.043)	-0.955 (.044)	-0.956 (.045)
Var(β_1^x)	0.0857 (.011)	0.0856 (.011)	0.0856 (.011)	0.152 (.028)	0.152 (.027)	0.152 (.027)	0.288 (.078)	0.290 (.076)	0.291 (.075)
Var(β_2^x)	0.0863 (.0086)	0.0865 (.0086)	0.0866 (.0087)	0.152 (.0205)	0.152 (.020)	0.153 (.020)	0.284 (.059)	0.286 (.057)	0.288 (.056)
Var(β_3^x)	0.0952 (.0097)	0.0949 (.010)	0.0946 (.010)	0.182 (.024)	0.181 (.023)	0.181 (.023)	0.400 (.063)	0.399 (.063)	0.397 (.062)
Var(β^p)	0.0480 (.0056)	0.0479 (.0057)	0.0478 (.0059)	0.0888 (.013)	0.088 (.013)	0.088 (.014)	0.148 (.031)	0.147 (.032)	0.147 (.033)

Table 11: Pseudo True Value: Moment-based Approach

Parameter	Scenarios								
True $\text{Var}(\beta)$:	(0, 0.1, 0.1, 0.1, 0.05)			(0, 0.2, 0.2, 0.2, 0.1)			(0, 0.5, 0.5, 0.5, 0.2)		
True $\text{Var}(\xi)$:	0.1	0.5	1	0.1	0.5	1	0.1	0.5	1
$\beta_0 = -1$	-1.01	-1.01	-1.01	-1.04	-1.04	-1.04	-1.11	-1.11	-1.12
$\beta_1^x = 1.5$	1.49	1.49	1.49	1.48	1.48	1.48	1.43	1.43	1.43
$\beta_2^x = 1.5$	1.49	1.49	1.49	1.48	1.48	1.48	1.43	1.43	1.43
$\beta_3^x = 0.5$	0.496	0.496	0.496	0.486	0.486	0.486	0.455	0.455	0.455
$\beta^p = -1$	-0.989	-0.988	-0.988	-0.958	-0.957	-0.955	-0.873	-0.869	-0.864
$\text{Var}(\beta_1^x)$	0.0854	0.0855	0.0855	0.149	0.149	0.149	0.275	0.275	0.276
$\text{Var}(\beta_2^x)$	0.0855	0.0855	0.0856	0.149	0.149	0.149	0.273	0.274	0.274
$\text{Var}(\beta_3^x)$	0.0938	0.0938	0.0937	0.176	0.175	0.175	0.369	0.368	0.366
$\text{Var}(\beta^p)$	0.0421	0.0421	0.0419	0.0685	0.0681	0.0676	0.0920	0.0906	0.0888

Table 12: Distribution of the Difference between True and Estimated Elasticity

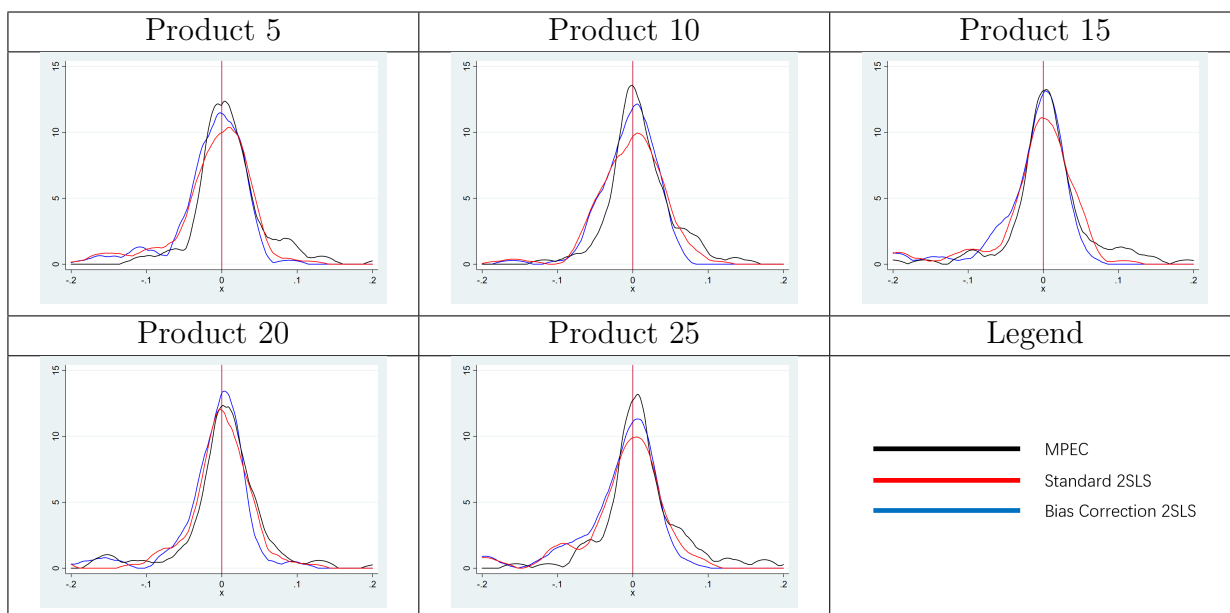
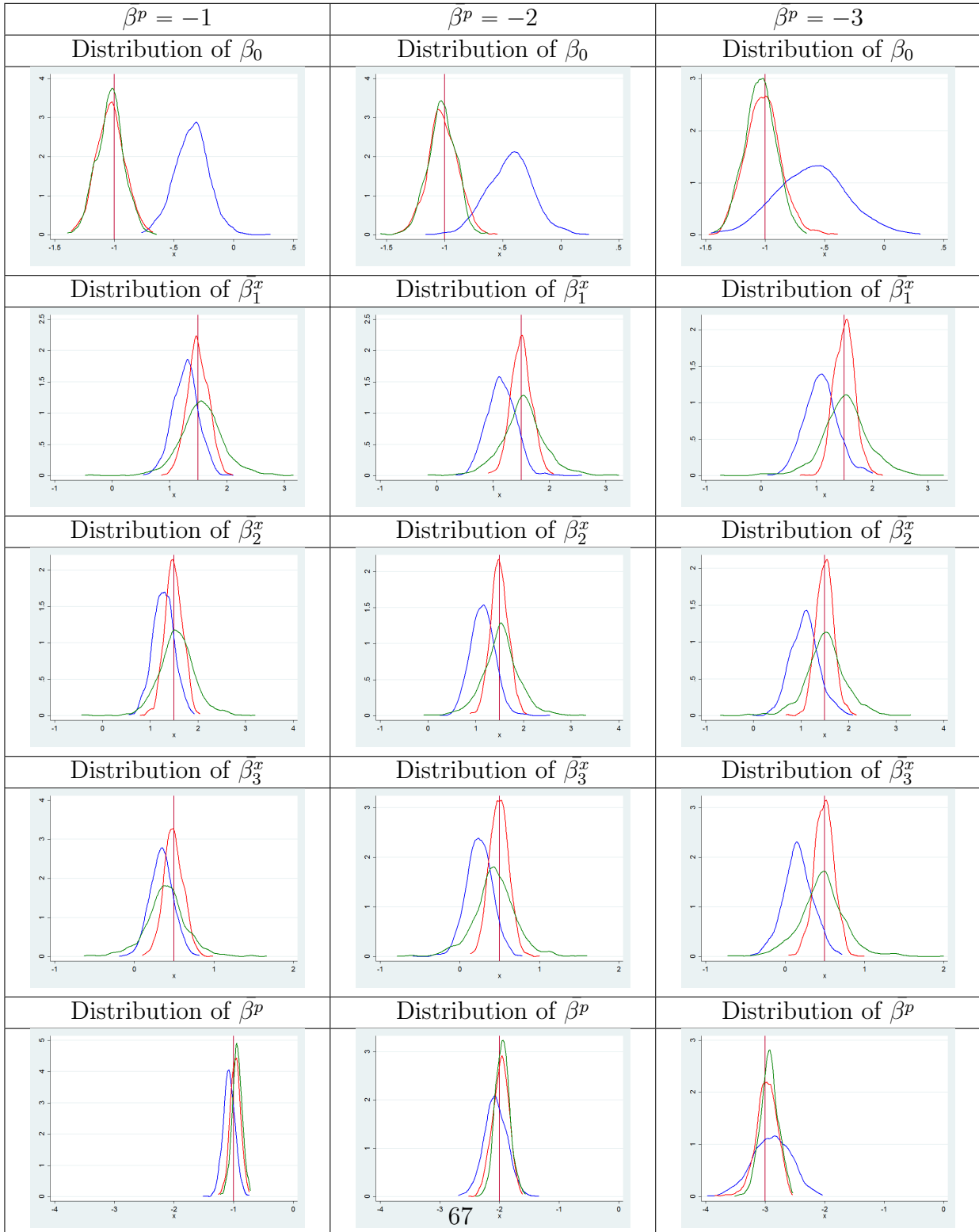
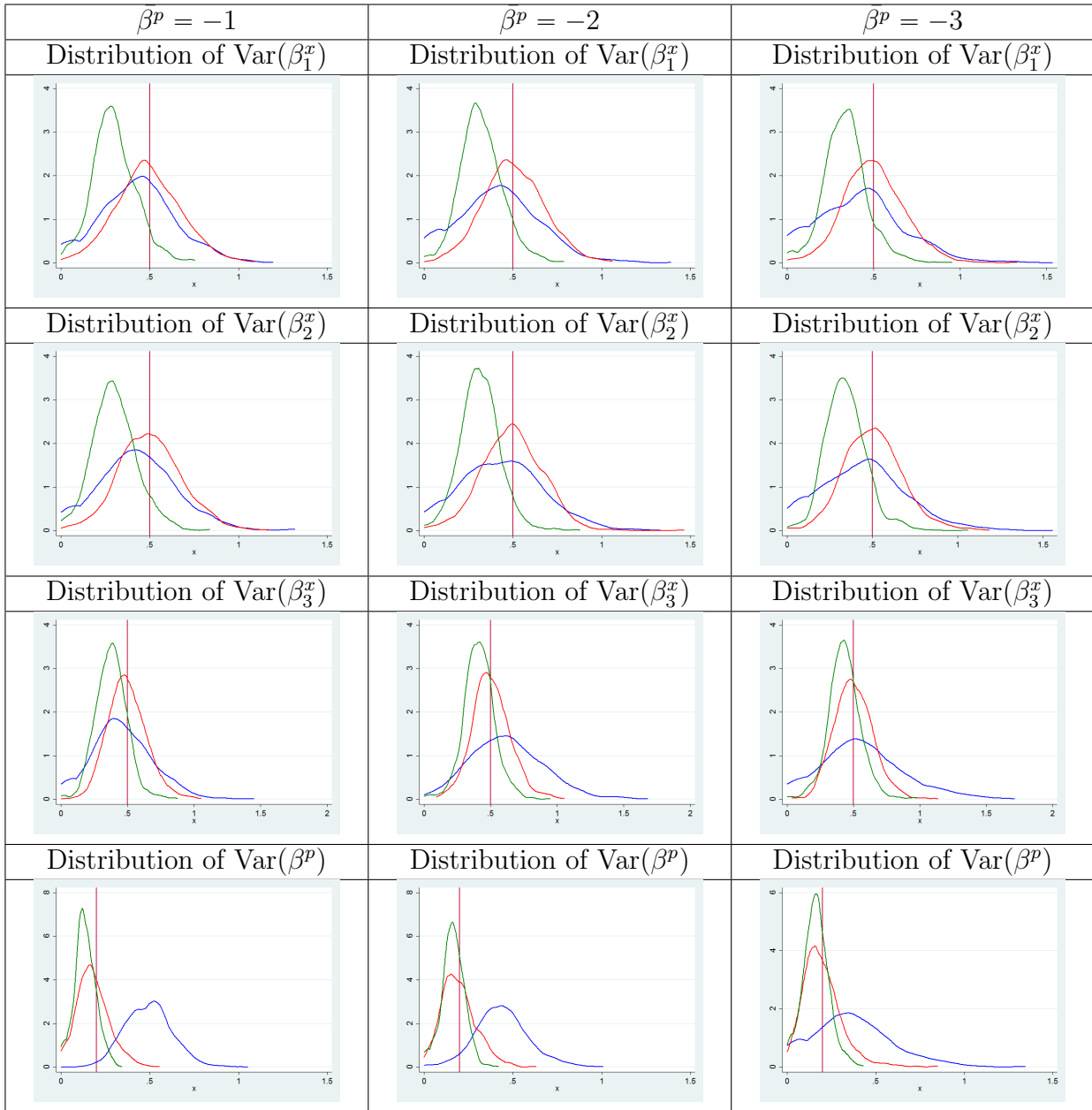


Table 13: Distribution of the Estimates of the Means — Different $\bar{\beta}^p$



Control Function —————
MPEC —————
2SLS —————

Table 14: Distribution of the Estimates of the Variances — Different $\bar{\beta}^p$



Control Function ————
MPEC ————
2SLS ————

Table 15: Testing for Zero Means — Standard 2SLS

Significance level	1%	5%	10%
2SLS with heteroskedasticity-robust standard error	0.904	0.793	0.711
GLS estimator and standard errors	0.889	0.765	0.678
2SLS with clustered standard error	0.904	0.780	0.702

Table 16: Testing for Zero Means — Bias Corrected 2SLS

Significance level	1%	5%	10%
2SLS with heteroskedasticity-robust standard error	0.915	0.819	0.725
GLS estimator and standard errors	0.882	0.767	0.669
2SLS with clustered standard error	0.906	0.809	0.731

Table 17: Testing for Zero Variances — Standard 2SLS

Significance level	1%	5%	10%
2SLS with heteroskedasticity-robust standard error	0.746	0.625	0.556
GLS estimator and standard errors	0.738	0.618	0.542
2SLS with clustered standard error	0.740	0.617	0.547

Table 18: Testing for Zero Variances — Bias Corrected 2SLS

Significance level	1%	5%	10%
2SLS with heteroskedasticity-robust standard error	0.792	0.688	0.626
GLS estimator and standard errors	0.773	0.680	0.613
2SLS with clustered standard error	0.775	0.679	0.620

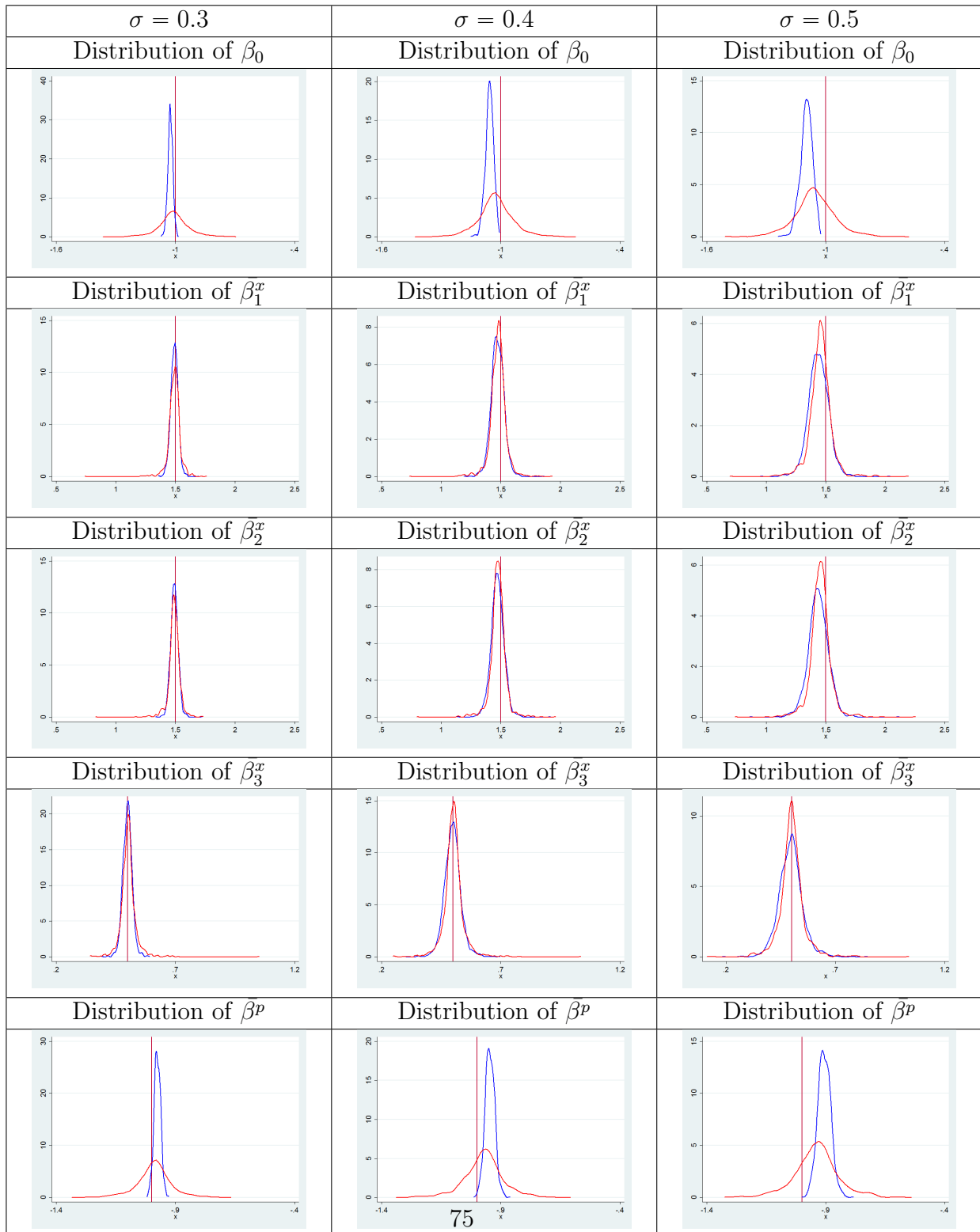
Table 19: Joint Test of Zero Means and Variances — Standard 2SLS

Significance level	1%	5%	10%
2SLS with heteroskedasticity-robust standard error	0.731	0.578	0.507
GLS estimator and standard errors	0.699	0.545	0.483
2SLS with clustered standard error	0.702	0.565	0.498

Table 20: Joint Test of Zero Means and Variances — Bias-Corrected 2SLS

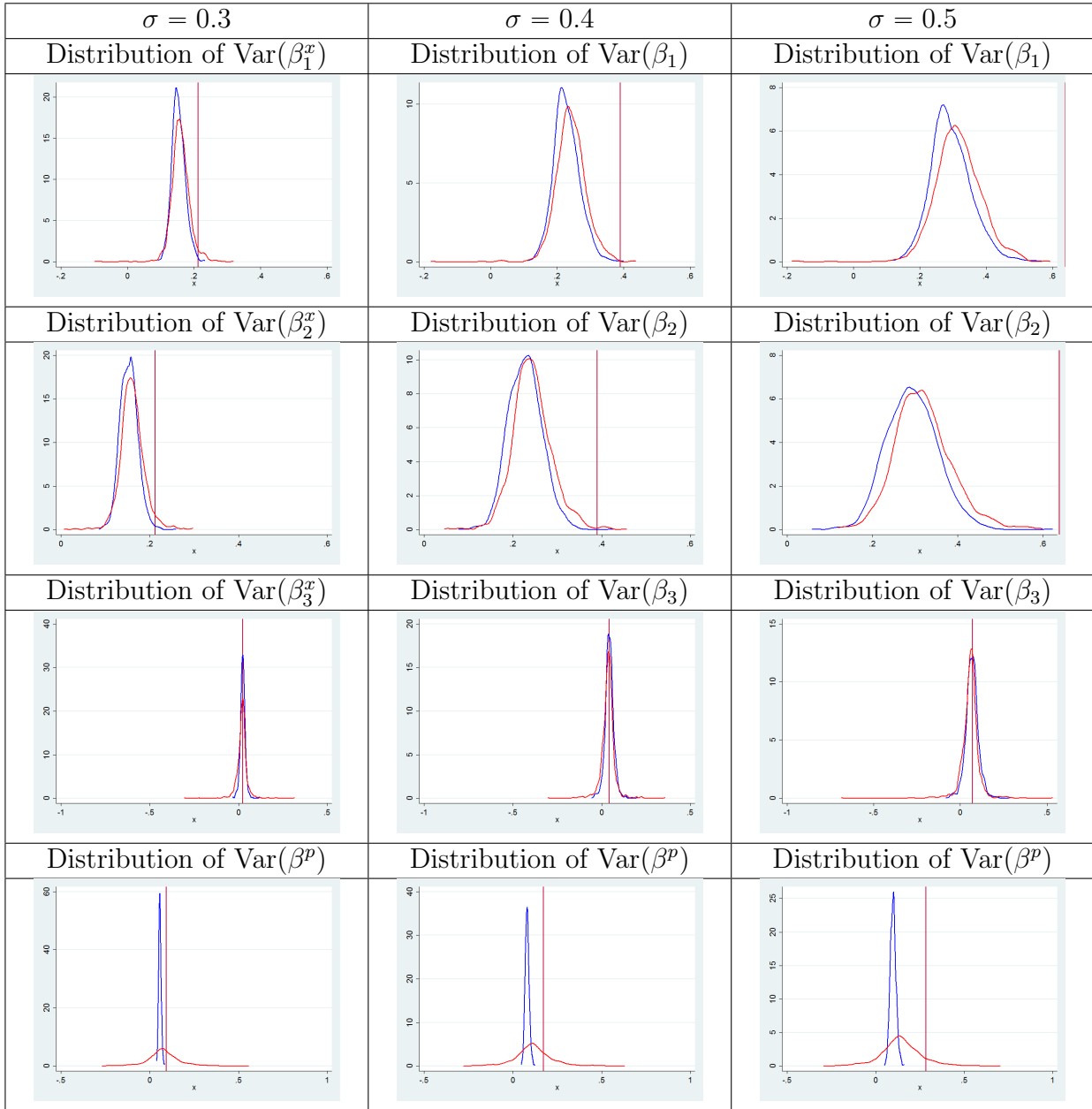
Significant level	1%	5%	10%
2SLS with heteroskedasticity-robust standard error	0.756	0.642	0.568
GLS estimator and standard errors	0.738	0.631	0.548
2SLS with clustered standard error	0.749	0.617	0.546

Table 21: Distribution of the Estimates of the Means (Lognormal Case)



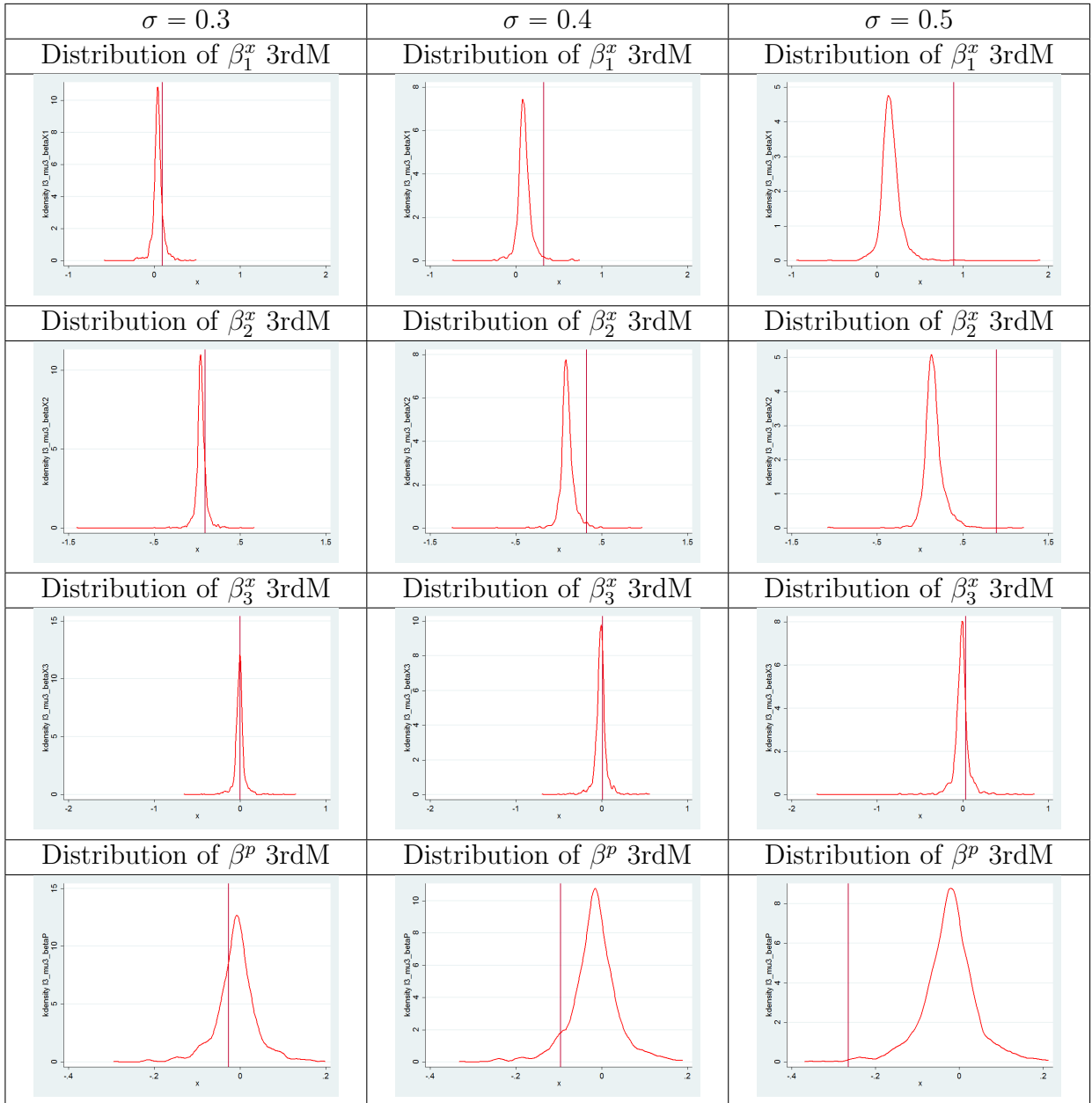
Only include 2nd moment ———— (blue line)
 Include 3rd moment ———— (red line)

Table 22: Distribution of the Estimates of the Variances (Lognormal Case)



Only include 2nd moment —
 Include 3rd moment —

Table 23: Distribution of the Estimates of the Third-order Moments (Lognormal Case)



Only include 2nd moment —————
 Include 3rd moment —————

Table 24: Summary Statistics for the Lognormal Case

Parameter σ Moments included:	Scenarios					
	$\sigma = 0.3$		$\sigma = 0.4$		$\sigma = 0.5$	
	2	3	2	3	2	3
$\beta_0 = -1$	-1.03 (0.013)	-1.01 (0.080)	-1.06 (0.021)	-1.03 (0.092)	-1.10 (0.031)	-1.06 (0.107)
$\bar{\beta}_1^x = 1.5$	1.49 (0.032)	1.49 (0.057)	1.47 (0.056)	1.48 (0.072)	1.44 (0.086)	1.45 (0.096)
$\bar{\beta}_2^x = 1.5$	1.49 (0.033)	1.49 (0.054)	1.47 (0.057)	1.47 (0.070)	1.44 (0.088)	1.45 (0.093)
$\bar{\beta}_3^x = 0.5$	0.499 (0.020)	0.502 (0.039)	0.497 (0.036)	0.502 (0.047)	0.496 (0.056)	0.501 (0.060)
$\bar{\beta}^p = -1$	-0.976 (0.014)	-0.991 (0.076)	-0.946 (0.020)	-0.972 (0.084)	-0.906 (0.027)	-0.940 (0.095)
$\text{Var}(\beta_1^x) = 0.212/0.390/0.639$	0.153 (0.020)	0.159 (0.030)	0.229 (0.039)	0.241 (0.048)	0.295 (0.062)	0.316 (0.072)
$\text{Var}(\beta_2^x) = 0.212/0.390/0.639$	0.153 (0.020)	0.160 (0.028)	0.228 (0.038)	0.244 (0.045)	0.294 (0.060)	0.319 (0.067)
$\text{Var}(\beta_3^x) = 0.04/0.043/0.071$	0.025 (0.014)	0.030 (0.033)	0.045 (0.025)	0.036 (0.041)	0.068 (0.038)	0.056 (0.059)
$\text{Var}(\beta^p) = 0.094/0.174/0.284$	0.0579 (0.007)	0.077 (0.095)	0.081 (0.011)	0.115 (0.11)	0.099 (0.016)	0.145 (0.12)
$3rdM(\beta_1^x) = 0.093/0.322/0.894$		0.044 (0.067)		0.096 (0.091)		0.160 (0.135)
$3rdM(\beta_2^x) = 0.093/0.322/0.894$		0.043 (0.082)		0.097 (0.101)		0.161 (0.130)
$3rdM(\beta_3^x) = 0.003/0.012/0.033$		-0.0073 (0.070)		-0.012 (0.084)		-0.015 (0.120)
$3rdM(\beta^p) = -0.027/ - 0.096/ - 0.265$		-0.0095 (0.050)		-0.018 (0.058)		-0.024 (0.066)

Table 25: Distribution of Three Estimates of the Means — Normal and Lognormal

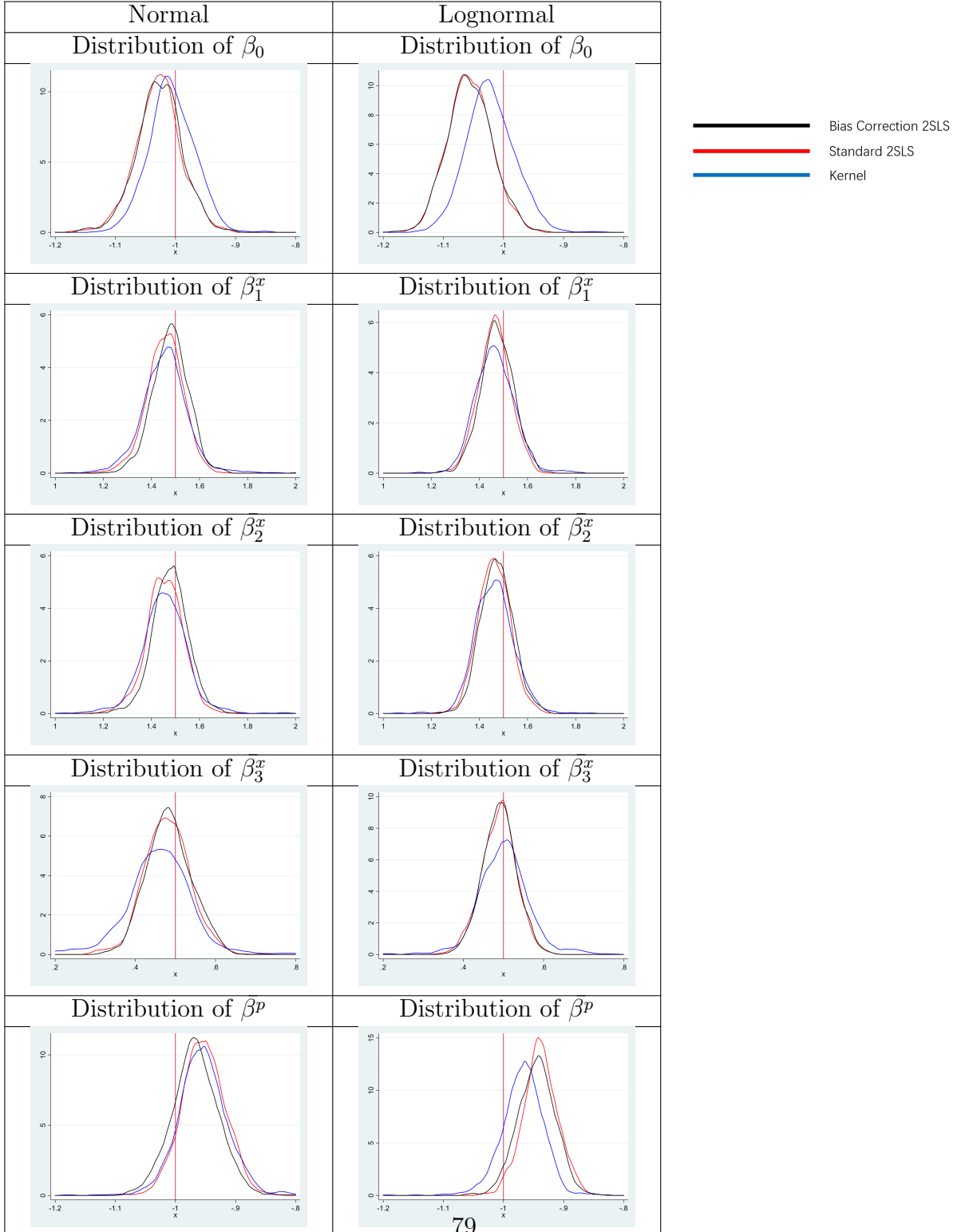


Table 26: Distribution of Three Estimates of the Variances — Normal and Lognormal

