Estimation of Random-Coefficient Demand Models: Two Empiricists' Perspective

ESTIMATION OF RANDOM-COEFFICIENT DEMAND MODELS: TWO EMPIRICISTS' PERSPECTIVE

Christopher R. Knittel and Konstantinos Metaxoglou*

Abstract—We document the numerical challenges we experienced estimating random-coefficient demand models as in Berry, Levinsohn, and Pakes (1995) using two well-known data sets and a thorough optimization design. The optimization algorithms often converge at points where the first- and second-order optimality conditions fail. There are also cases of convergence at local optima. On convergence, the variation in the values of the parameter estimates translates into variation in the models' economic predictions. Price elasticities and changes in consumer and producer welfare following hypothetical merger exercises vary at least by a factor of 2 and up to a factor of 5.

I. Introduction

ECONOMETRICS has become increasingly synonymous with the estimation of nonlinear models, where the objective function may not be globally concave or convex. This is very often the case in the so-called structural approach, which attempts to infer relationships between observable endogenous variables, and observable explanatory or unobservable variables based on optimizing behavior dictated by economic theory (Reiss & Wolak, 2007). Obtaining parameter estimates in these cases typically requires a nonlinear-search algorithm with a set of starting values and stopping rules for termination.

For the class of demand models introduced in the seminal work of Berry, Levinsohn, and Pakes (1995, henceforth, BLP), we find that convergence may occur in regions of the objective function where the first- and second-order optimality conditions fail. We also experience convergence at multiple local optima, as well as instances of convergence failure. Furthermore, parameter estimates and implied economic predictions, such as price elasticities, and changes in consumer welfare and firm profits due to hypothetical merger exercises, exhibit notable variation on convergence of the nonlinear search.

The difficulties surrounding proofs for the existence of a global optimum of a criterion function for nonlinear extremum estimators are widely known and well documented (McFadden & Newey, 1994). This is particularly true in the case of nonlinear GMM estimators like the one considered here. At the same time, the estimation problem that empirical economists like ourselves face often involves technical difficulties with material implications for the conclusions of their analyses, which may not be publicized to the extent that they should. Based on our reading of the literature, this has been largely the case for the class of demand models we consider in this paper. The purpose of our work is to show that a thorough optimization exercise and a clear documentation of its design and challenges are potentially as important to the conclusions of an empirical exercise as the identification approach. The estimation of typical BLP random-coefficient (RC) logit demand models based on two widely known data sets for automobiles and cereals gives us the opportunity to convey this message.

A typical BLP RC-logit demand model achieves more flexible substitution patterns compared to the simple or nested logit by allowing consumer heterogeneity in the valuation of product characteristics for differentiated products. At the same time, the model contains a product-specific demand shock in each market capturing all product characteristics that affect consumer choices but the econometrician cannot control for. Berry et al. (1995) introduced an estimation approach that also addresses endogeneity, as products with higher unmeasured quality probably sell at a higher price, using GMM and fixed-point iterations. These iterations allow the researcher to retrieve the product-specific demand shock in each market.

BLP RC-logit demand models, as well as their variants, are among the most popular state-of-the-art discrete-choice demand models and have provided answers to a variety of important questions in numerous empirical studies.[1] Measuring market power (Nevo, 2001), analyzing horizontal merger effects (Nevo, 2000a), evaluating international trade policies (Berry, Levinsohn, & Pakes 1999), welfare gains due to new products (Petrin, 2002), and construction of price indices that account for quality changes and product introductions (Nevo, 2003), to only name a few, are among the many important economic questions that have been addressed using BLP-type demand models. Table 1 lists articles using the same class of models, which have been published in prominent general-interest economic journals, as well as in the leading industrial organization journal.

The articles cited in the top panel of table 1 reflect our attempt to include only studies that used what we perceive as the main ingredients of the BLP approach in demand

Received for publication October 2, 2008. Revision accepted for publication November 28, 2012.

* Knittel: Sloan School of Management, MIT, and NBER; Metaxoglou: Bates White LLC.

An earlier version of this paper, "Estimation of Random Coefficient Models: Challenges, Difficulties and Warnings," was circulated as NBER working paper 14080, available at http://www.nber.org/papers/w14080. We received valuable comments from two anonymous referees and, especially, from the editor, Michael Greenstone. We have benefited from conversations with Steve Berry, Severin Borenstein, Matthew Gentzkow, Phil Haile, Aviv Nevo, Linh To, Hal White, Frank Wolak, Catherine Wolfram, and seminar participants at the University of Calgary, University of California at Berkeley, the University of California Energy Institute, and the 2008 NBER Winter IO meeting. We also thank Ken Judd for sharing some of his thoughts about our work. K.M. acknowledges financial support from Bates White, LLC. We are grateful to Bates White, LLC for making their computing resources available. The usual disclaimer applies.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00394.

[1] As of December 26, 2011, Berry et al. (1995) had 517 cites according to the Social Sciences Citation Index (SSCI). On the same day, Nevo's "Practitioner's Guide" (2000b), which popularized the BLP-type demand models with its accompanying Matlab code, had 96 cites using SSCI.

TABLE 1.—PAPERS USING BLP-TYPE DEMAND MODELS WITH AGGREGATE DATA

Journal	Paper
American Economic Review	Berry, Levinsohn, and Pakes (1999)
Econometrica	Berry et al. (1995); Goeree (2008); Nevo (2001)
Journal of Political Economy	Berry, Levinsohn, and Pakes (2004); Petrin (2002)
Review of Economics and Statistics	Nevo (2003); Rekkas (2007)
Review of Economic Studies	Berry, Linton, and Pakes (2004); Nakamura and Zerom (2010); Villas-Boas (2007)
Rand Journal of Economics	Ackerberg and Rysman (2005); Armantier and Richard (2008); Berry (1994); Bonnet and Dubois (2010); Chu (2010); Copeland, Dunn, and Hall (2011); Davis (2006); Iizuka (2007); Nevo (2000); Song (2007); Villas-Boas (2009)

Issues Discussed	Number of Papers
Optimization algorithm	6
Starting values	3
Fixed-point iteration settings	0
Market share evaluation draws	9
Gradient-based FOC diagnostics	0
Hessian-based SOC diagnostics	0
Multiple optima (Y/N)	2

We focus on leading general interest journals and the top industrial organization journal. We list only the papers that contain the main ingredients of the BLP approach for the estimation of random-coefficient logit demand models with aggregated data as discussed in section I. FOC and SOC refer, respectively, to first- and second-order conditions.

estimation for static models using aggregate data and a GMM framework to address endogeneity. From our point of view, these ingredients include heterogeneity in consumers' valuation of product characteristics, as well as the unobserved product- and market-specific demand shock and the associated fixed-point iterations to retrieve it.[2] There are also papers that have been published in the journals in table 1, such as Goolsbee and Petrin (2004) and Bayer, Ferreira, and McMillan (2007), which contain some of the main ingredients of the BLP approach for demand estimation in a static framework. We do not list these papers and, hence, we serve less than justice to a broad definition of BLP-type demand models. The bottom panel of table 1 indicates the number of papers discussing a subset of the issues we raise surrounding the computational challenges of the BLP-demand models throughout the paper; we refer to these entries in subsequent sections.

As we have already hinted in the last sentence of the previous paragraph, using the desirable features of the BLP RC-logit models requires the solution of a nontrivial optimization problem whose computational complexity has been publicized only recently. Starting with previous drafts of this paper, which were largely contemporaneous with the earlier drafts of Dube, Fox, and Su (2012), and more recently with the papers of Judd and Skrainka (2011) and Skrainka (2011), the optimization challenges underlying the estimation of this class of models are by now well documented. These challenges stem primarily from the combination of a nonlinear search to obtain parameters capturing heterogeneity in consumers' valuation for product characteristics and the fixed-point iterations to infer the product- and market-specific demand shock. This outer nonlinear search coupled with the fixed-point iterations has been by far the most popular approach in estimating BLP demand models to date.

[2] We include articles that assume both a continuous and a finite distribution for consumer heterogeneity.

To highlight the difficulties in the solution of the optimization problem underlying the estimation of the BLP RC-logit models, we use two widely known data sets, three classes of optimization algorithms with fifty different sets of starting values, and two different implementations of the fixed-point iterations. The first data set, which is for automobiles, is the one used in Berry et al. (1995). The second data set contains the pseudo-real data for cereals used in Nevo (2000b). We use eleven optimization algorithms that may be divided into three categories: derivative based, deterministic direct search, and stochastic direct search. The two implementations of the fixed-point iterations differ in the tolerance that we use to declare convergence—we use a loose and a tight tolerance. Following the recommendations of McCullough and Vinod (2003), we discuss a number of diagnostics regarding first- and second-order optimality conditions at the terminal points of the nonlinear searches.

Our findings point to substantial variation in the value of the objective function of the underlying optimization problem both within and across algorithms for those combinations of starting values and fixed-point tolerances that converge. The variation is present in both data sets and slightly more pronounced in the case of loose tolerances. Although the derivative-based algorithms exhibit a superior performance relative to their direct-search counterparts in the case of cereals, this does not seem to be the case for automobiles. Two of the publicly available gradient-based optimization algorithms give rise to the smallest objective function value in both data sets. In addition, a nonnegligible number of combinations of starting values, optimization algorithms, and fixed-point tolerances fail to converge; this is particularly true in the case of cereals with an almost equal split between loose and tight tolerances.

The variation in objective function values leads to substantial variation in parameter estimates even when we limit our attention to the set of parameters that give rise to the

smallest objective function value for each optimization algorithm under tight tolerance for the fixed-point iterations. Using the gradient norm, the Hessian eigenvalues, and a scale-invariant weighted-gradient stopping criterion, we identify several local optima in the case of automobiles. Using the same criteria, we identify a single local optimum and a saddle point in the case of cereals.

The variation in parameter estimates translates into variation in economic variables of interest for both data sets. We use own-price elasticities to document the variation in economic predictions at the product level. We use aggregate elasticity, which we calculate simulating a price increase of all products, as well as the change in consumer welfare and firm profits following a hypothetical merger in the two industries to document variation in economic predictions at the market level. All of these economic variables exhibit substantial variation even when we focus on those combinations of optimization algorithms and starting values that converged under tight tolerance for the fixed-point iterations and remove any outliers.

In the case of automobiles, the range of the own-price elasticity for the product with the highest market share is -3.48 to -0.93 with a mean of -2.77 and standard deviation of 0.53 when we exclude a handful of observations exceeding 0. When we limit our attention to those combinations of algorithms, starting values, and fixed-point iterations under tight tolerance that gave rise to the local optima, the own-price elasticities for the same product are between -3.21 and -1.29, excluding a single positive value associated with one of those optima. In addition, the average aggregate elasticity across twenty markets is between -1.74 and -0.41. The mean is -1.20, and the standard deviation is 0.24. If we focus on the aggregate elasticity values associated with the local optima, we see the values between -1.39 and -0.41.

The average change in profits due to the hypothetical merger for automobiles exhibits a range between \$485 million and \$2,548 million with a mean of \$787 million and a standard deviation of \$316 million. In the case of local optima, the average change in profits is as low as \$568 million and as high as \$1,697 million, excluding observations associated with local optima that we treat as outliers. The average change in consumer welfare for the same exercise is between $-$\$4,328 million and $-$\$1,165 million. The mean is $-$\$2,183, and the standard deviation is \$531 million. Limiting our attention only to those values associated with the local optima and excluding ones that we treat as outliers, the range is between $-$\$3,550 million and $-$\$1,926 million. The values of the economic variables discussed here that are associated with the smallest local optimum appear to be outliers in the corresponding distributions that emerge from combinations of optimization algorithms, starting values, and fixed-point iterations.

We also find that the variation in economic variables of interest need not coincide with large differences in the objective function value. For example, we uncover two local minima in the automobile data with highly comparable objective function values, but the change in profits and consumer welfare due to the hypothetical merger exercise at these two minima are notably different by a factor of 2.5 and almost 2, respectively. This brings up the possibility of a horse race between two local minima, where the lowest objective function value may oscillate between the two depending on the sample.

In the case of cereals, the variation in the economic variables of interest is smaller compared to automobiles. The own-price elasticity of the product with the highest market share is between -2.47 and -1.34 among those combinations of optimization algorithms and starting values that converged, with a mean of -1.98 and a standard deviation of 0.14. Similarly, the average aggregate elasticity across the 94 markets is between -1.78 and -0.84. The mean is -1.34, and the standard deviation is 0.11. The average change in profits across markets due to the hypothetical merger is between \$104.3 million and \$229.7 million, with a mean of \$170.8 million and a standard deviation of \$15.7 million. The average change in consumer welfare across markets for the same exercise also exhibits substantial variation: $-$\$975 million to $-$\$469 million, with a mean of $-$\$671 million and a standard deviation of about \$60 million.

The remainder of the paper is organized as follows. Section II provides an overview of the BLP RC-logit demand model. Section III offers an outline of the methodology for the merger simulation and the calculation of the implied change in consumer welfare, along with a discussion of some recent developments in issues surrounding this exercise. The details of our optimization design are discussed in Section IV. Section V is an overview of the data and the demand-model specifications we employ. Section VI describes the optimization results documenting the variation in objective function values due to combinations of optimization algorithms, starting values, and tolerances for the fixed-point iterations. In Section VII we illustrate the implication of such variation for economic variables of interest. We offer conclusions and some recommendations to practitioners in the final section. The online appendix provides a discussion of the parameter estimates for the demand models and results for some additional optimization exercises we considered.

II. The Demand Model

In this section, we describe the standard BLP-type RC-logit demand model with aggregate data. Following standard notation in the literature, we assume that a consumer i derives utility from a product j in market t that may be written as

$$u_{ijt} = x_j\beta_i - \alpha_i p_{jt} + \xi_{jt} + \varepsilon_{ijt} = V_{ijt} + \varepsilon_{ijt}, \tag{1}$$

where p_{jt} is the price of product j in market t and x_j is a (row) vector of nonprice product characteristics. The vector ξ_{jt} captures the product-specific demand shock in each market. Each individual is assumed to choose one of the $1, \ldots, J$ products available in the market or not to purchase at all.

The no-purchase option is usually termed the outside good, and its associated utility is $u_{i0t} = \varepsilon_{i0t}$. The logit error term ε_{ijt} is the first source of consumer heterogeneity in the utility function. The second source of consumer heterogeneity, the random coefficients α_i and β_i, may be written as

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \Pi D_i + \Sigma v_i,\ D_i \backsim P_D(D), v_i \backsim P_v(v). \tag{2}$$

The decomposition in equation (2) leads to terms that are common across consumers, such as α and β, as well as to terms D_i and v_i, which are vectors of observed and unobserved consumer characteristics that affect purchasing decisions and follow the distributions P_D and P_v. Although the random coefficients are desirable because they generate more realistic substitution patterns, their presence has direct implications for the computational complexity of the model as we discuss below. After combining equations (1) and (2), we obtain

$$\begin{aligned} u_{ijt} &= \delta_{jt}(x_j, p_{jt}, \xi_{jt}; \theta_1) + \mu_{ijt}(x_j, p_{jt}, D_i, v_i, \theta_2) + \varepsilon_{ijt}, \\ \delta_{jt} &= x_j\beta - \alpha p_{jt} + \xi_{jt},\ \mu_{ijt} = [p_{jt}, x_j]'(\Pi D_i + \Sigma v_i). \end{aligned} \tag{3}$$

We use $[p_{jt}, x_j]$ to denote a column vector of appropriate dimension. The mean utility associated with the consumption of good j that is common across consumers in market t is captured by δ_{jt}. Deviations from this mean utility are reflected in μ_{ijt} and ε_{ijt}. The vectors θ_1 and θ_2 differ in that the former contains α and β, while the latter contains the elements of matrices Π and Σ. Under independence of consumer idiosyncrasies for characteristics, the market share of product j is given by

$$\begin{aligned} s_{jt}(x, p_{\cdot t}, \delta_{\cdot t}; \theta_2) &= \int_{A_{jt}} dP(D, v, \varepsilon) \\ &= \int_{A_{jt}} dP_\varepsilon(\varepsilon) dP_v(v) dP_D(D), \\ A_{jt}(x, p_{\cdot t}, \delta_{\cdot t}; \theta_2) &= \{(D_i, v_i, \varepsilon_{it}) | u_{ijt} \geq u_{ilt}\}, \forall\, l = 1, \ldots, J. \end{aligned} \tag{4}$$

In the share equation (4), x includes the characteristics of the products, while $p_{\cdot t} = (p_{1t}, \ldots, p_{Jt})'$ and $\delta_{\cdot t} = (\delta_{1t}, \ldots, \delta_{Jt})'$. In the same equation, the error term ε can be integrated out analytically, giving rise to the well-known logit probabilities. Given distributional assumptions for v and D, the integral associated with market shares is commonly evaluated using Monte Carlo integration assuming a number ns of individuals:

$$\begin{aligned} s_{jt}(x, p_{\cdot t}, \delta_{\cdot t}; \theta_2) &= \frac{1}{ns} \sum_{i=1}^{ns} s_{ijt} \\ &= \frac{1}{ns} \sum_{i=1}^{ns} \frac{\exp(\delta_{jt} + \mu_{ijt})}{\sum_{j=0}^{J} \exp(\delta_{jt} + \mu_{ijt})}. \end{aligned} \tag{5}$$

The shock ξ_{jt} that was Berry (1994) introduced plays the role of the structural error in a demand system. In its absence, the market shares given by equation (5) are smooth and continuous deterministic functions of the product characteristics and price. The presence of ξ implies likely endogeneity of prices because both consumers and firms observe ξ, and therefore its value enters into the firms' pricing decisions. The standard approach to address endogeneity is nonlinear GMM with the moments implemented as

$$E[\xi_{jt} \cdot f(x_j, z_{jt})] = 0, \tag{6}$$

where z_{jt} is an appropriate vector of excluded instruments and $f(\cdot)$ is a known vector-valued function. Given a vector of mean utilities δ, a sample analog of the moment condition can be constructed and the researcher may proceed with estimation. The vector of mean utilities δ is retrieved by equating the observed market shares from the data with those implied by the model for a given vector of parameters θ_2:

$$s_{\cdot t}^{obs} = s_{\cdot t}^{pred}(x, p_{\cdot t}, \delta_{\cdot t}; \theta_2). \tag{7}$$

As opposed to the simple logit and nested logit, where analytical solutions for δ are available for the system of equations in equation (7), the RC-logit requires a numerical solution of a highly nonlinear system of equations whose dimension equals the number of products in the market under consideration. The econometrician may retrieve δ using the following fixed-point iterations,

$$\delta_{\cdot t}^{(k+1)} = \delta_{\cdot t}^{(k)} + \ln s_{\cdot t}^{obs} - \ln s_{\cdot t}^{pred}(x, p_t, \delta_{\cdot t}^{(k)}, \theta_2), \tag{8}$$

where $\delta_{\cdot t}^{(k)}$ denotes the kth iterate. For a given value of θ_2, the fixed-point iterations in equation (8) can be initiated with the logit solution $\delta_{\cdot t}^{(0)} = \ln(s_{\cdot t}) - \ln(s_{0t})$, where s_{0t} is the share of the outside good. The iterations continue until some norm of the difference between two consecutive iterates is smaller than some prespecified tolerance. Once δ is retrieved, ξ can be inferred from

$$\xi_{jt} = \delta_{jt} - x_j\beta - ap_{jt}. \tag{9}$$

The elements of θ_1, namely, α and β, in equation (9) are retrieved using linear instrumental variables (IVs). Having defined $\theta = (\theta_1, \theta_2)$, and with the aggregate demand shock playing the role of a structural error term that is a function of θ, the econometrician faces a nonlinear GMM problem with objective function given by

$$Q_T(\theta) = \{T^{-1}\xi(\theta)'Z\} W_T \{T^{-1}Z'\xi(\theta)\} \tag{10}$$

for an appropriate weighting matrix W_T assuming a sample size T. Inference is performed using results in Berry et al. (1995), with the asymptotics working as $J \to \infty$; Berry et al. (2004) offer additional details. This methodology allows the econometrician to perform a nonlinear search in the parameter space only for θ_2 by concentrating out θ_1. For a given value of θ_2, we infer δ using equations (7) and (8). With δ in hand, we obtain θ_1 using linear IVs. Having δ and θ_1 available, the

researcher constructs the econometric error that appears in equation (10). Draws from P_v and P_D required in equation (5) are made once and are kept constant through the estimation exercise.

The algorithm just described, which consists of an outer loop that minimizes the objective function with respect to θ_2 and an inner loop that uses fixed-point iterations to infer δ, may be termed a nested fixed point (NFP) algorithm in the language of Rust (1987). We follow this nomenclature in the remainder of our discussion. The publication of computer code by Nevo (2000b) undoubtedly contributed to the popularity of the NFP algorithm for the estimation of BLP-type demand models for the last ten years or so. Recently studies have identified issues regarding computational aspects of the methodology outlined here.

Dube et al. (2012) show that the temptation to implement loose stopping criteria for the fixed-point iterations to speed up the estimation process may cause two types of problems. First, the approximation error of the inner fixed-point iterations propagates into the outer GMM objective function and its derivatives that may cause an optimization routine to fail to converge. Second, to induce the convergence of an optimization routine, the researcher may then loosen the outer-loop stopping criterion. Consequently, even when an optimization run converges, it may falsely stop at a point that is not a local minimum. The authors offer an alternative formulation of the GMM problem as a mathematical program with equilibrium constraints (MPEC) building on the results in Su and Judd (2012), who show that the MPEC and NFP algorithms produce estimators with similar statistical properties.

III. Merger Simulation and Consumer Welfare

In many empirical exercises, demand estimation serves as an immediate input to study the effects of changes in the structure of an industry, such as price increases implied by a merger. Measures of changes in consumer welfare implied by the new market structure, such as compensating variation, are immediately available following the demand estimation exercise, as the following discussion shows.

More specifically, with demand estimates available, constructing a matrix of price derivatives emerging from the first-order conditions implied by profit maximization is straightforward. Combined with information on the ownership structure of the market and a model of competition, inferring marginal cost is possible. For example, under static Bertrand and constant marginal costs, the first-order conditions associated with the firms' profit-maximization problem imply

$$p - mc = \Omega\,(p)^{-1}\, s\,(p)\,, \tag{11}$$

where p is the price vector, $s(\cdot)$ is the vector of market shares, and mc denotes the corresponding marginal costs. The dimension of these vectors is equal to the number of products available in the market, say, J. The Ω matrix is the Hadamard product of the (transpose) of the matrix of the share derivatives with respect to prices and an ownership structure matrix. The ownership structure matrix is of dimension $J \times J$ with its (i,j) element equal to 1 if products i and j are produced by the same firm and 0 otherwise. Because prices are observed and demand estimation allows us to retrieve the elements of Ω, estimates of marginal costs, $\widehat{mc}$, are directly obtained using equation (11).

A simple change of 0s and 1s in the ownership structure matrix, along with a series of additional assumptions, allows the simulation of a change in the industry's structure, as the one implied by horizontal mergers among competitors (see Nevo, 2001). Simply put, a merger simulation implies the same Bertrand game with a smaller number of firms. The vector of postmerger prices p^* is the solution to the following system of nonlinear equations:

$$p^* - \widehat{mc} = \hat{\Omega}^{post}\left(p^*\right)^{-1}\hat{s}\left(p^*\right). \tag{12}$$

The elements of $\hat{\Omega}^{post}$ reflect changes in the ownership structure implied by the hypothetical merger. Solving for the postmerger prices is equivalent to solving a system of nonlinear equations of dimension equal to the number of products in the market under consideration. For example, using the cereal data set, we have 94 markets with 24 products in each market. As a result, solving equation (12) requires the solution of 94 systems of nonlinear equations of dimension 24. An approximate solution for the postmerger prices, which avoids the need to solve the systems of nonlinear equations and is discussed in Nevo (1997), is given by

$$p^{approx} = \widehat{mc} + \hat{\Omega}^{post}\,(p^{pre})^{-1}\,\hat{s}(p^{pre}), \tag{13}$$

where $\hat{s}(p^{pre})$ is the premerger vector of market shares and the elements of $\hat{\Omega}$ associated with share price derivatives are evaluated at the premerger prices. Thus, we avoid dealing with issues related to the potential numerical instabilities of Newton routines used in the solution of nonlinear first-order conditions, as well as with issues related to the existence and the uniqueness of equilibrium.

To the best of our knowledge, in the case of Bertrand competition with multiproduct firms facing RC-logit demand functions of the type discussed here, there is no result showing the (a) existence of an equilibrium in pure strategies and (b) whether the equilibria are unique solutions to the system of equations implied by the first-order conditions (FOCs) of the underlying game. In the papers we are aware of, both existence and uniqueness have been assumed (for example, note 12 in Berry et al., 1995).[3]

[3] Allon, Federgruen, and Pierson (2011) provide a sufficient condition under which a Bertrand equilibrium exists and the set of Bertrand equilibria coincides with the solutions of FOCs in the case of single-product firms facing RC-logit demand functions. This condition precludes a very high degree of market concentration: no firm captures more than 50% of the potential market in any of the consumer segments that it serves. A somewhat stronger version of the same condition, firm shares below 30%, establishes uniqueness. Allon et al. also provide a sufficient condition for a (unique) equilibrium for markets with an arbitrary degree of concentration in the presence of an exogenous price limit. However, in this case, the equilibrium

TABLE 2.—OPTIMIZATION ALGORITHMS

Class	Description	Source	Acronym
Derivative-based	Quasi-Newton 1	MathWorks	DER1-QN1
	Quasi-Newton 2	Publicly available	DER2-QN2
	Conjugate gradient	Publicly available	DER3-CGR
	SOLVOPT	Publicly available	DER4-SOL
	KNITRO	Ziena Optimization	DER5-KNI
Deterministic direct search	Simplex	MathWorks	DIR1-SIM
	Mesh adaptive direct search	MathWorks	DIR2-MAD
	Generalized pattern search	MathWorks	DIR3-GPS
Stochastic direct search	Simulated annealing	Publicly Available	STO1-SIA
	Genetic algorithm GADS	MathWorks	STO2-GAL
	Simulated annealing GADS	MathWorks	STO3-SIG

With the postmerger prices in hand, we can estimate expected consumer welfare changes due to the mergers under consideration. One such measure of change in consumer welfare is the compensating variation. Assuming away nonlinear income effects, as is the case in the demand models we consider here, following McFadden (1981) and Small and Rosen (1985), the compensating variation for individual i is given by

$$CV_i = \frac{\ln\left[\sum_{j=0}^{j=J} \exp\left(V_{ij}^{post}\right)\right] - \ln\left[\sum_{j=0}^{j=J} \exp\left(V_{ij}^{pre}\right)\right]}{\alpha_i}, \tag{14}$$

where V_{ij}^{pre} and V_{ij}^{post} are defined in equation (1) using the pre- and postmerger prices. Integrating over the density of consumers and multiplying by their total number (market size) yields the average compensating variation in the population.

IV. Optimization Design

We estimated the automobile and cereal demand models adapting the code used by Nevo (2000b), written in the Matlab matrix language developed by Mathworks.[4] The main body of Nevo's code had to be altered to accommodate the setup of an exercise that involved eleven optimization algorithms using fifty sets of starting values and various stopping rules described below.

The starting values for the mean utility vector δ are the fitted values of a simple logit after adding draws from a zero-mean normal distribution with a standard deviation equal to the standard error of the logit regression; therefore, the variation in the starting values represents regression error plausibly obtained across researchers. The starting values for the vector θ_2 entering the nonlinear part μ_{ijt} of the utility function in equation (3) are draws from a standard normal distribution; this represents the fact that little is known about the magnitude of θ_2 a priori.

Table 2 lists the eleven optimization algorithms we used to estimate the cereal and automobile demand models. The table also contains an acronym for each of the algorithms that we will use for the remainder of our discussion when we refer to them. Five of the algorithms are derivative based. The remaining six are either deterministic or stochastic direct search. The derivative-based algorithms use some information about the steepness and the curvature of the objective function without necessarily keeping track of information associated with the Hessian matrix while searching for a minimum of the objective function. The direct-search algorithms are based on function evaluations and are divided into deterministic and stochastic depending on whether they include a random component in their searches of the optimum of the objective function.

All of the algorithms are coded in Matlab. Seven of the algorithms are commercially available as part of the Matlab Optimization and Genetic Algorithm and Direct Search (GADS) toolboxes. The codes for the remaining four of the algorithms are publicly available from their authors. Two of the derivative-based algorithms (DER1-QN1 and DER2-QN2) are quasi-Newton, the third (DER3-CGR) is a conjugate gradient, and the fourth (DER4-SOL) is an implementation of Shor's r-algorithm. The last of the derivative-based algorithms (DER5-KNI) implements interior point and active set methods for solving continuous, nonlinear optimization problems.

The Matlab routines for the two quasi-Newton algorithms, DER1-QN1 and DER2-QN2, are available in the Matlab optimization toolbox and on the website maintained by Hans Bruun Nielsen, respectively. The routine for the conjugate-gradient algorithm, DER3-CGR, is also posted on Nielsen's website. Alexei Kuntsevich and Franz Kappel provide the routines for DER4-SOL.[5] The KNITRO routines for DER5-KNI are available in the Matlab optimization toolbox as add-ons. For the purpose of estimation, the derivative-based algorithms were implemented using analytical gradients and numerical Hessians (when necessary).

The routines for the three deterministic direct-search algorithms are available in the Matlab optimization and GADS

may not necessarily reside in the interior of the feasible price region and, hence, not be characterized by the FOCs.

[4] Nevo's original code is available at http://faculty.wcas.northwestern.edu/~ane686. Our adaptation of Nevo's original code is available at http://web.mit.edu/knittel/www.

[5] Burke, Lewis, and Overton (2008) provide a compact self-contained discussion of Shor's r-algorithm. See Kappel and Kuntsevich (2000) for additional details. Furlong (2012) reported a better performance for DER4-SOL relative to DER1-QN1 and DIR1-SIM in estimating a BLP-type demand model for hybrid vehicles.

toolboxes, respectively. They include an application of the Nelder-Mead simplex, DIR1-SIM, the Mesh adaptive direct search (MADS), DIR2-MAD, and the generalized pattern search (GPS), DIR3-GPS. (We refer readers to Lagarias, Reeds, & Wright, 1998, for the mechanics of the Nelder-Mead simplex.) Torczon (1997) provides a detailed description of the GPS. Material related to MADS, a generalization of the GPS algorithm, is available in Audet and Dennis (2006).

The stochastic direct search routines implement one genetic and two simulated-annealing algorithms. The code for the first simulated-annealing algorithm, STO1-SIA, is our translation of the code originally developed for the GAUSS matrix language by E. G. Tsionas. The routines for the second simulated-annealing algorithm STO3-SIG, as well as for the genetic algorithm STO2-GAL, are available in the Matlab GADS toolbox. We refer to Dorsey and Mayer (1995) and Goffe, Ferrier, and Rogers (1994) for a compact discussion of the genetic and simulated annealing algorithms in the context of econometrics, respectively.

We experimented with a number of stopping rules for the various optimization algorithms we employed. For the majority of the algorithms, convergence is dictated by the change in the objective function and the parameter vector (in some norm) between two consecutive iterations of an algorithm on the basis of a specified tolerance; the gradient norm is another metric. A maximum number of iterations or function evaluations are also employed as stopping rules. We used a tolerance of 1E–03 for changes in both the parameter vector and the objective function. We limited the number of function evaluations to 4,000. Imposing an upper bound on the number of function evaluations was largely dictated by the use of the direct-search algorithms, which tend to be more time-consuming relative to the gradient-based algorithms. If an algorithm exceeded the maximum number of function evaluations, it was terminated.[6]

A stopping rule is also required for the NFP iterations (see equation [8]), which introduce an additional layer of computational burden given their linear rate of convergence. More specifically, the NFP rate of convergence is measured by the Lipschitz constant, which is the norm of a matrix involving the own- and cross-demand elasticities with respect to the demand shock ξ (Dube et al., 2012). We present results using a loose and a tight tolerance for the fixed-point iterations. The loose tolerance reflects the approach in Nevo (2000b). More specifically, the tolerance is initially set to 1E–06. After that, the tolerance level becomes less stringent by a factor of ten every 50 iterations if $||\theta_2^{k+1} - \theta_2^k|| \geq 0.01$, where k and $k+1$ are used to denote two successive iterates of θ_2. If $||\theta_2^{k+1} - \theta_2^k|| < 0.01$, then the tolerance level is set to 1E–09 and no adjustment takes place. As a result, a loose (tight) tolerance is implemented when the parameter estimates are far from (close to) the solution. We also present results fixing the tolerance associated with the automobile and cereal data to 1E–16 and 1E–14, respectively, imposing an upper bound of 2,500 NFP iterations.[7]

We simulated the market shares using fifty and twenty individual draws from a standard normal distribution for the automobile and cereal data, respectively. For automobiles, the number of draws is representative of what we perceive to be a standard approach among practitioners.[8] The number of draws in the case of cereals reflects the setting in Nevo's code. The standard-normal draws are made once in the beginning and are held fixed during estimation such that the limit theorems of Pakes and Pollard (1989) hold. Following Nevo (2000b), we use Monte Carlo integration to approximate the integrals associated with the market share calculations.

Concluding this section, we mention that various aspects of the optimization design discussed in this section are also discussed in various papers listed in table 1. Six papers provide information regarding the optimization algorithm used in estimating the demand models.[9] Three explicitly mention the use of multiple starting values for their optimization algorithms, and nine discuss the number of draws for the simulation of the market share integrals. None of the papers provide any details regarding the settings of the fixed-point iterations.

V. Data and Specifications

We use two data sets for implementing the BLP GMM algorithm: the automobile data set used by Berry et al. (1995, 1999) and the cereal pseudo-real data set from Nevo (2000b). Much of our motivation for the use of these data was due to the fact that they were publicly available at the time of our first draft.[10] Readers should also keep in mind that none of the exercises undertaken throughout the paper should be viewed as replication or validation exercises.

The automobile data consist of 2,217 observations for all models marketed between 1971 and 1990 in the United States Each model-year combination is treated as a separate product, and each year between 1971 and 1990 is treated is a separate market. The number of products, in each market lies between 72 (market 4, 1974) and 150 (market 18, 1988). The 2,217 model/years represent 997 distinct models. We use

[6] The online appendix provides results using a tolerance of 1E-06 for changes in the parameter vector and the objective function value for the derivative-based algorithms for both cereals and automobiles.

[7] Dube et al. (2012) recommend a best-practice tolerance of 1E-14 for the fixed-point iterations. The script `invertshares.m` in the most recent version of their Matlab code indicates an upper bound of 2,500 NFP iterations.

[8] For example, Jiang, Manchandab, and Rossi (2009) argue that the commonly used values in the literature are between 20 and 50 in their note 1.

[9] For two of the papers, we had to refer to a journal online appendix and the NBER working paper version to retrieve such information.

[10] To the best of our knowledge, the code for BLP (1995) is not publicly available. However, GAUSS code for BLP (1999) was publicly available at James Levinsohn's website at the University of Michigan at the time of the first draft of this paper, circa fall 2005. We used the GAUSS code to extract the data used in this paper. Table 1 in BLP (1995) and table 2 in BLP (1999) contain descriptive statistics indicating that the authors used the same data for both papers.

four observable product characteristics other than price when we model the demand for automobiles: (a) the ratio of the vehicle's horsepower to its weight (HP/WT); (b) the vehicle space, which is measured as length times width (space); (c) a dummy indicating whether air-conditioning is standard; and (d) tens of miles per gallon of gasoline (MPG).[11]

Our specification of the automobile demand model is not identical to the specification in either BLP 1995 or BLP 1999. In addition, both BLP 1995 and BLP 1999 estimate demand jointly with supply, but we do not.[12] In our specification, a constant term, the price, and the four product characteristics of the previous paragraph, enter the utility function. All of these variables with the exception of space are also assigned random coefficients that correspond to the standard deviation of normal draws. Based on our specification, the outer loop of the BLP GMM algorithm for automobiles involves a parameter vector θ_2 of dimension 5. The six coefficients corresponding to the elements of the vector θ_1 are concentrated out and are retrieved using linear IVs.

In the case of cereals, the data consist of 2,256 observations for 24 products (brands) in 47 cities over two quarters. The 24 brands are present in each of the 94 markets. Our specification of the demand equation is identical to Nevo's. More precisely, the specification includes cereal brand dummies, which subsume product characteristics other than prices, as well as unobservable consumer characteristics interacted with a constant term, price, sugar content (sugar), and a mushy dummy indicating whether the cereal gets soggy in milk (mushy).

The specification also includes interactions of product characteristics with consumer demographics drawn from the Current Population Survey. For example, the price is interacted with the individual's log of income (income), the log of income squared (income sq), and a child dummy indicating whether the individual is less than 16 years old (child). The constant, sugar, and mushy are all interacted with income and age. The mean taste parameters associated with the constant, price, sugar, and mushy are retrieved using a minimum-distance procedure given the presence of the brand dummies.[13] The interaction of the product characteristics with the consumer unobservables and demographics gives rise to thirteen terms in total. Therefore, the outer loop of the BLP GMM algorithm involves a parameter vector θ_2 of dimension 13. The 24 coefficients associated with the brand dummies, corresponding to the elements of the vector θ_1, are concentrated out and are retrieved using linear IVs.

Our identification strategy regarding automobiles is similar to that in BLP 1995. Our instruments consist of the five nonprice automobile characteristics, their sums across other automobiles produced by the same firm, and their sums across automobiles produced by the rival firms. In the case of cereals, we use the 44 instruments readily available in Nevo's data set.

In terms of inference, the econometrician faces in principle three sources of error in the BLP demand models: sampling error in estimating market shares, simulation error in approximating the shares predicted by the model, and the underlying model error. There is no doubt that addressing all three sources of error is necessary to obtain the correct standard errors (see BLP, 1995, and Berry et al., 2004, for a detailed discussion). We report standard errors that account only for heteroskedasticity in the underlying model error for both the cereal and the automobile data. In section A.1.1 of the online appendix, we use the standard errors for a subset of the automobile parameter estimates to compare their variation due to the underlying model error with their variation due to the optimization design.

VI. Optimization Results

A. *Objective Function Values*

Figure 1 contains box-and-whisker (BaW) plots of the objective function values by optimization algorithm and NFP tolerance. The top panel refers to automobiles and the bottom panel refers to cereals.[14] The naming convention for the algorithms on the vertical axis follows table 2. Each point in a BaW plot is a combination of a starting value and an NFP tolerance for which the optimization algorithm under consideration converged on the basis of some stopping criteria. Depending on the algorithm's implementation, convergence does not necessarily imply that the first- and second-order conditions for a local minimum are met (see our discussion below). The prominent vertical lines on the left part of the figures indicate the smallest objective function values across all such combinations.

The various algorithms implemented here use a number of similar but not identical stopping criteria. These criteria include, for example, the change in the parameter vector or the associated gradient, the change in the objective function value, the maximum number of iterations, or function evaluations. Based on such stopping criteria, the algorithms generate exit codes to indicate the conditions under which they stopped. The algorithms in the Matlab optimization and GADS toolboxes produce a 0 exit code to indicate that the maximum number of iterations or function evaluations was reached and a negative exit code to indicate their failure to converge. We use such exit codes to declare convergence. In the case of the deterministic and stochastic direct-search algorithms in these two toolboxes, we used exit codes

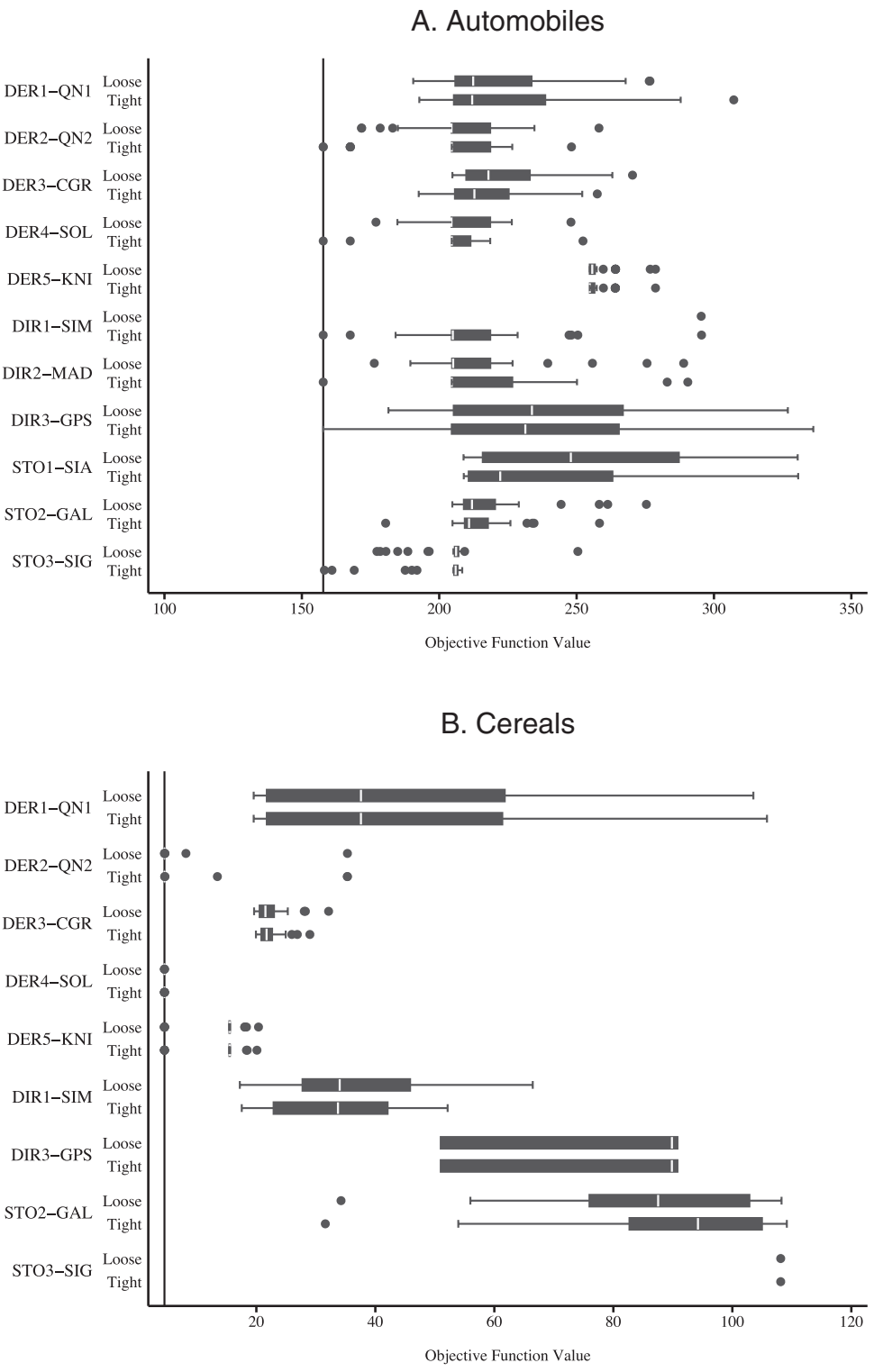
[11] For additional details regarding the data, see section 7.1 in BLP (1995).

[12] The demand and supply specifications in the two papers are highly similar but not identical. See table IV on page 876 in BLP (1995) and table 5 on page 416 in BLP (1999) for comparisons.

[13] See section 4 in Nevo (2000b) for additional information regarding the data and variables used in the demand specification. The results are available in table I. See section 3.5 for additional details about the minimum-distance procedure used to retrieve the mean taste parameters in the same paper.

[14] For each of the BaW plots, the boxes cover the interquartile range, from the lower quartile to the upper quartile, and contain a vertical white line indicating the median. The whiskers, denoted by horizontal lines, intend to cover most or all the range of the data. The left whisker extends to a value that is the lower quartile minus 1.5 times the interquartile range, or the minimum should this be smaller. The right whisker extends to a value that is the upper quartile plus 1.5 times the interquartile range, or at the maximum, if this is smaller. Data points outside the whiskers are represented with dots.

FIGURE 1.—BOX-AND-WHISKER PLOTS OF THE OBJECTIVE FUNCTION VALUE



The naming convention of the optimization algorithms on the vertical axis of the box-and-whisker plots (BaW) in this figure follows table 2. "Loose" and "tight" refer to the alternative nested fixed-point (NFP) tolerances described in section IV. The vertical line indicates the smallest objective function value for the combinations of optimization algorithms, NFP tolerances, and starting values that converged. We implemented DER5-KNI using a lower bound of 0 on the parameter vector over which the algorithm performed the nonlinear search.

exceeding 0 to declare convergence.[15] Once again, convergence is not necessarily synonymous to termination at a point where the first- and second-order conditions of a local minimum are met.

In the case of automobiles, 988 out of 1,100 combinations converged, with more than half of them (513) associated with tight NFP tolerance. For cereals, 883 combinations of starting values, optimization algorithms, and NFP tolerances led to convergence. Almost half of them (440) were associated with tight NFP tolerance. All 100 combinations for the STO1-SIA algorithm did not converge for the cereal data set, while 50 of them failed to converge in the case of the automobiles. In all 50 cases for automobiles, STO1-SIA reached the maximum number of function evaluations: 22 times under loose NFP tolerance and 28 times under tight NFP tolerance. Furthermore, DIR1-SIM did not converge 69 times in the case of cereals—33 times under loose NFP tolerance and 36 times under tight NFP tolerance. DIR1-SIM did not converge 50 times in the case of automobiles—49 times under loose NFP tolerance and 1 time under tight NFP tolerance. In all these instances, DIR1-SIM reached the maximum number of function evaluations. In addition, DIR2-MAD did not converge in 28 cases for the cereal data set. The algorithm reached the maximum number of function evaluations in all 28 cases with an equal split between loose and tight NFP tolerances.[16]

The BaW plots in figure 1 account for extreme objective function values due to rather meaningless stopping points of various combinations of optimization algorithm and starting values. They also account for combinations of algorithms and starting values that gave rise to extraneous observations for market shares (for example, NaNs). For example, in the case of automobiles, the maximum across the 988 observations lies above 62,000, the median is 213.37, and the 90th percentile is 336.17. The minimum value is 157.79. Hence, we constructed that BaW plots excluding values above 340, which is very close to the 90th percentile. For cereals, the maximum across the 883 observations exceeds 89,000, the 75th percentile is 882.81, and the median is just 101.09. The minimum value is 4.56, which is consistent with the value reported by Dube et al. (2012). As a consequence, we constructed the cereal BaW plots using values that lie below 110, a value very close to the median under tight tolerance (105.53). By truncating the distributions, we are obviously understating the variation in the values of the objective function.

In the case of automobiles, we see substantial variation in the objective function values both across and within optimization algorithms. Overall, the derivative-based algorithms do not exhibit a superior performance relative to the their deterministic or stochastic direct-search counterparts in terms of reaching regions of the parameter space, with low objective function values once we focus on values not exceeding 340. There are 890 observations with values less than 340, with 475 of them associated with the tight NFP tolerance. Four combinations of optimization algorithms and starting values under tight NFP tolerance led to the smallest objective function value of 157.79, with two others (DIR3-GPS and STO3-SIG) reaching the same parameter space, achieving an objective function value below 158.21. Two of the algorithms are derivative based: DER2-QN2 and DER4-SOL. The other two algorithms are deterministic direct search: DIR1-SIM and DIR2-MAD. None of the 100 values for DER5-KNI falls below 254.76.

The pattern of substantial variation in the objective function value across and within optimization algorithms is also present in the cereal data. In contrast with the automobile data, the derivative-based algorithms exhibit superior performance, in terms of reaching regions of the objective function with low values, relative to their deterministic or stochastic direct-search counterparts. Among the 464 values not exceeding 110, 238 of them are related to combinations with loose NFP tolerance. Only six (two) combinations of starting values, and NFP tolerance for DIR3-GPS (STO3-SIG) lead to objective function values that do not exceed 110. Six combinations of optimization algorithms, starting values, and NFP tolerances achieve the smallest objective function value of 4.56. All six combinations are associated with the derivative-based algorithms DER2-QN2, DER4-SOL, and DER5-KNI. Three of the six combinations use tight NFP tolerance. Recall that DER2-QN2 and DER4-SOL also led to the smallest objective function value for the automobile data. Interestingly, all 100 combinations of starting values and NFP tolerances for DER4-SOL give rise to the value of 4.56.

For DER1-QN1, the objective function values are between 19.55 and 105.82. The last of the derivative-based algorithms, DER3-CGR, exhibits values between 19.62 and 32.15. The deterministic direct-search algorithms achieve objective function values between 17.22 and 90.78. The values for DIR1-SIM range from 17.22 to 66.46, while those for DIR3-GPS are between 50.99 and 90.78. Although DIR1-SIM was able to identify the point with the smallest objective function value for the automobile demand model, this is not the case for the cereal demand model. The objective function values for the stochastic-search algorithms also exhibit substantial variation—between 31.59 and 109.16. This range, which is similar to that of DER1-QN1, is largely attributable to 56 values associated with STO2-GAL. The two values associated with STO3-SIG equal 108.13.

Overall, the box-and-whisker plots for the objective function values in figure 1 show variation both within algorithm and across algorithms. This variation, although ameliorated to some degree, continues to be present even with tight NFP tolerance. It is difficult to judge the relative performance

[15] Refer to our publicly available code for a detailed treatment of exit codes of the various algorithms in order to declare convergence.

[16] In the case of automobiles, DER5-KNI generated the exit codes −100 (−101) 41 (18) times. According to appendix A in Waltz and Plantenga (2009), exit codes −100 to −199 indicate that "a feasible approximate solution was found." More specifically, the exit code −100 implies that "no more progress can be made, but the stopping tests are closed to be satisfied (within a factor of 100) and so the current approximate solution is believed to be optimal." The exit code −101 implies that "it is possible the approximate feasible solution is optimal, but perhaps the stopping tests cannot be satisfied because of degeneracy, ill-conditioning or bad scaling."

of classes of algorithms since converging to points in the objective function that have low values, but are not the consistent root, does not necessarily yield results that are closer to the truth. Given the variation that continues to exist even under tight NFP tolerances, we would recommend practitioners to experiment not only with multiple starting values, but also with more than one class of algorithms when estimating BLP-type demand models and report their experiences.

B. Gradients and Hessians

In this section, we investigate whether the objective function values reported in tables A2 and A4 in the online appendix correspond to local minima as opposed to other critical points, such as saddles, by examining the gradient (g) and the Hessian (H) of the objective function. We focus on results associated with tight NFP tolerance given the findings in Dube et al. (2012). As in the previous section, our discussion excludes STO1-SIA for automobiles and DIR2-MAD plus STO1-SIA for cereals because these algorithms stopped by exceeding the maximum number of function evaluations with tight NFP tolerance. When discussing our results, readers should keep in mind that finding the global optimum of a function, or even proving that a given local optimum is a global optimum, is a difficult problem.[17] Furthermore, the discussion of gradient and Hessian diagnostics almost never appears in empirical work involving BLP-type demand models (for a notable recent exception, see Goldberg & Hellerstein, 2013). As table 1 indicates, none of the papers listed provides any diagnostics regarding the gradient or the Hessian of the objective function when they discuss estimation results.

We measure the length of the analytical gradient using its inf-norm $\|g\|_\infty$. The inf-norm is equivalent to finding the maximum of the absolute values of the gradient elements. We refer to $\|g\|_\infty$ as the gradient norm for the remainder of our discussion. While Nevo's code provides analytical expressions for the gradient, it does not provide analytical expressions for the Hessian of the objective function. Following what we perceive to be common practice, we evaluated H using the DER1-QN1 algorithm that offers numerical approximations to the Hessian as a by-product. In addition, we constructed a scale-invariant weighted-gradient stopping criterion, $g'H^{-1}g$.[18] Using the Matlab built-in eigenvalue function (`eig`), we calculated the eigenvalues of H to determine whether it is positive definite.

We also examined the condition number of the Hessian at various optima because an ill-conditioned Hessian puts the accuracy of the reported results into question. More specifically, the Hessian condition number is given by $\kappa(H) = \lambda_{\max}(H)/\lambda_{\min}(H)$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and the smallest Hessian eigenvalues, respectively. McCullough and Vinod (2003) recommend that solutions for which the Hessian condition number exceeds $1/\sqrt{\epsilon}$, where ϵ is the machine precision, should not be accepted uncritically. For Matlab, $\epsilon = 2.2\text{E}{-16}$, such that $1/\sqrt{\epsilon} = 6.7\text{E}{+07}$.[19] However, we caution that the Hessian eigenvalues often seem to be particularly sensitive to small changes in the parameter estimates in terms of their magnitudes and signs.

For automobiles, the implied gradient norm for the estimates in table A2 of the online appendix is between 0.08 for DIR1-SIM and 440.39 for STO2-GAL (see the "automobiles" section of table 3). The weighted-gradient criterion does not exceed 0.001 for the four algorithms that stopped at 157.79: DER2-QN2, DER4-SOL, DIR1-SIM, and DIR2-MAD. The Hessians for all four algorithms are positive definite, and their condition numbers, which do not exceed 7.35E+03, are of the same order of magnitude. The Hessian condition number does not seem to raise concerns regarding the numerical precision of the solution, at least based on the metric proposed by McCullough and Vinod. Therefore, the point at 157.79 may be treated as a local minimum. We may call the minimum at 157.79 the global minimum although we have not proved or formally tested that we have indeed found a global minimum. To the best of our knowledge, although tests for the null of whether a global optimum of a criterion function has been identified are available, they are not widely used by practitioners. This statement is particularly true for practitioners undertaking empirical exercises involving BLP-type demand models.[20]

None of the other terminal points, which correspond to objective function values between 180.51 (STO2-GAL) and 254.83 (DER5-KNI), can safely be treated as local minima.

[17] There are three cases in which the problem is somewhat easier: when the objective function has one critical point in its domain, when the function is globally concave or convex in its domain (see Simon & Blume, 1994, p. 55), and when the domain of the objective function is a compact subset, say C of $\mathbb{R}^n$, assuming an nth-dimensional parameter space. Based on Weierstrass's theorem, every continuous function whose domain is a compact subset C achieves its global maximum and its global minimum on C (see Simon & Blume, p. 823).

[18] Section 7.3 in McFadden and Newey (1994) provides an illustrative numerical Hessian approximation. See also section 9.2.5 in Davidson (2000). The weighted-gradient stopping criterion is largely inspired from ML estimation. See the discussion about stopping criteria in section 6.3 in Quandt (1983) and in section 16.5 in Ruud (2000). Figure 16.5 in Ruud provides an intuitive explanation for the use of the criterion. Dube et al. (2012) derive analytically the elements of the Hessian for the GMM estimation approach employed here in an online appendix.

[19] For a matrix A, the rough rule of thumb is that as $\kappa(A)$ increases by a factor of 10, you lose one significant digit in the solution of the linear system $Ax = b$ (see Judd, 1998, p. 68). Judd argues that a condition number is small if its base 10 logarithm is about 2 or 3 for a computer that carries about sixteen significant decimal digits. The implications of ill conditioning for a wide class of optimization algorithms can be seen using the iterations in Newton's method given by $x_{k+1} = x_k - H^{-1}(x_k)g(x_k)$, where H is the Hessian and g is the gradient. In the linear system $Ax = b$, $\kappa(A)$ indicates the maximum effect of a perturbation in b or A in the solution. It can be shown that $||\delta x||/||x|| = \kappa(A) \times ||\delta b||/||b||$ and $||\delta x||/||x + \delta x|| = \kappa(A) \times ||\delta A||/||A||$.

[20] According to Andrews (1997), the best-specified definitions of GMM estimators in the literature require a value $\widehat{\theta}$ of the estimator that is close to minimizing the criterion function. For example, Pakes and Pollard (1989) require that $\widehat{\theta}$ yield a value of the criterion function that is within $o_p(1)$ of the minimum for consistency and within $o_p(n^{-1/2})$ of the minimum for asymptotic normality of $\widehat{\theta}$. Veall (1990) provides a diagnostic test of the null that a maximum already found is global using results from extreme-value asymptotic theory due to de Haan (1981). Andrews (1997) identifies a serious problem with the power of Veall's test and proposes a stopping rule (SR) procedure for the computation of GMM estimators.

TABLE 3.—GRADIENT AND HESSIAN DIAGNOSTICS USING TIGHT NFP TOLERANCE

Algorithm	Objective Function Value	$\|\|g\|\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
Automobiles				
DER1-QN1	192.751	7.661	18.700	1.52E+04*
DER2-QN2	157.786	0.214	0.000	7.35E+03*
DER3-CGR	192.541	7.337	13.987	9.51E+03*
DER4-SOL	157.786	0.109	0.000	3.70E+03*
DER5-KNI	254.825	24.006	−27.550	−8.00E + 02
DIR1-SIM	157.786	0.082	0.000	3.77E+03*
DIR2-MAD	157.787	0.596	0.001	6.31E+03*
DIR3-GPS	157.797	16.281	0.024	3.83E+03*
STO1-SIA	163.629	138.377	14.306	8.02E+03*
STO2-GAL	180.505	440.386	42.376	4.02E+03*
STO3-SIG	158.206	83.909	0.691	3.81E+03*
Cereals				
DER1-QN1	19.554	1.088	26.747	3.97E+07*
DER2-QN2	4.562	0.230	0.000	5.80E+08*
DER3-CGR	19.929	13.231	39.737	2.88E+07*
DER4-SOL	4.562	0.009	0.000	5.76E+08*
DER5-KNI	4.562	0.041	0.000	5.82E+08*
DIR1-SIM	17.549	7.787	15.865	1.00E+08*
DIR2-MAD	16.055	17.302	15.513	1.51E+08*
DIR3-GPS	50.993	472.767	34.194	8.72E+06*
STO1-SIA	131.320	493.765	365.304	2.53E+07*
STO2-GAL	31.594	86.113	25.822	−2.25E + 06
STO3-SIG	108.130	412.090	−2.850	−9.61E + 01

The naming convention for the optimization algorithms in the "Algorithm" column follows table 2. We use g and H to denote the gradient and the Hessian of the objective function value evaluated at the parameter estimates in table A2 for automobiles and in table A4 for cereals in the online appendix. κ(H) denotes the Hessian condition number. * indicates a positive definite Hessian.

Their weighted-gradient criterion exceeds 13, and in the case of DER5-KNI, the Hessian is not positive definite.

The global minimum at 157.79 associated with the 49th set of starting values for DIR1-SIM implies $\lambda_{max} = 2.30E{+}04$ and $\lambda_{min} = 6.11E{+}00$, such that $\kappa(H) = 3.77E{+}03$. The eigenvector corresponding to the smallest eigenvalue has an extremal element of −0.7747 in the direction of the standard deviation term of HP/WT. The eigenvector corresponding to the largest eigenvalue has an extremal element of −0.9992 in the direction of the standard deviation term for price. The Hessian condition number does not seem to raise concerns regarding the numerical precision of the solution, at least, based on the metric proposed by McCullough and Vinod.

For cereals, the implied gradient norm for the estimates in table A4 of the online appendix is between 0.009 (DER4-SOL) and 472.77 (DIR3-GPS) (see the "cereals" section of table 3). The Hessian is positive definite for all the algorithms that terminated successfully with the exception of STO2-GAL and STO3-SIG. The weighted-gradient criterion is almost identical to 0 for DER2-QN2, DER4-SOL, and DER5-KNI. Therefore, based on the first- and second-order optimality conditions, the objective function value of 4.56 implied by the parameter estimates in table A4 of the online appendix corresponds to a local minimum. In the spirit of our earlier discussion for automobiles, we may call the local minimum at 4.56 a global minimum, although we have not proved formally that we have found a global minimum.

The global minimum at 4.56 corresponding to the 42nd set of starting values for DER2-QN2 implies $\lambda_{max} = 1.65E{+}04$ and $\lambda_{min} = 2.84E{-}05$, which in turn give $\kappa(H) = 5.80E{+}08$. The eigenvector corresponding to the smallest eigenvalue has an extremal element of 0.9986 in the direction of the interaction of price with the log of income. The eigenvector corresponding to the largest eigenvalue has an extremal element of 0.9861 in the direction of the standard deviation term for sugar. This Hessian condition number exceeds the threshold in McCullough and Vinod by almost an order of magnitude.

To sum up, we provided diagnostics for the analytical gradients and the numerical Hessian of the objective function for the estimates in tables A2 and A4 of the online appendix that gave rise to the smallest objective function value using tight NFP tolerance across all starting values. For automobiles, the point at which the objective function value equals 157.79 satisfies the criteria of a local minimum. Six of the ten algorithms that converged stopped at or near this minimum. The other four algorithms converged at points in the parameter space that cannot be treated as local minima. In the case of cereals, the point at which the objective function value is 4.56 satisfies the criteria of a local minimum. Only four of the nine algorithms that converged achieved this minimum. The Hessian condition number may raise some concerns about the numerical precision of the solution based on the metric suggested by McCullough and Vinod (2003).

C. *Additional Local Optima*

In this section, we examine whether any terminal points of the nonlinear searches for the various algorithms beyond the ones reported in tables A1 through A4 qualify as local optima. We do so by using the first- and second-order diagnostics that we discussed in the previous section. We first discuss our findings with tight NFP tolerance. We subsequently provide a discussion of our findings with loose NFP tolerance. Before moving to any of these details, we offer a brief motivation for our discussion of the local optima.

Overall, it can be difficult to show that the criterion function of an extremum estimator attains a unique (global) minimum at the true parameter vector. For example, in nonlinear GMM, conditions for identification are like conditions for unique solutions of nonlinear equations that are known to be difficult (see section 2.2.3 in McFadden & Newey, 1994). In addition, it is often challenging to find the unique global minimum of a criterion function setting aside the case of a globally convex criterion function, where there can be at most one local minimum that is also the global minimum. As a result, two consistency theorems for extremum estimators are available—one for a global optimum and one for a local optimum.[21]

When there is more than one local optimum, the consistency theorem for a local optimum states that one of the local optima is consistent, but provides no guidance to which one is consistent.[22] Therefore, although any of the plausible, from an economic-theoretic viewpoint, local optima can be a consistent root, the studies in table 1 seem not to have examined the possibility of such local optima with two exceptions.[23] At least, they do not discuss such a possibility explicitly. As we mentioned in the previous section, we are not aware of a study formally testing that has indeed found a global minimum following, say, the procedure in Andrews (1997), either.

In the case of cereals, we did not identify any local minima that were different from the one at 4.56 with tight NFP tolerance. For automobiles, 59 terminal points of the nonlinear searches indicate an objective function value of 204.56 when rounded to the second decimal point. Six algorithms reached this parameter space: DER1-QN1 (3 times), DER2-QN2 (9), DER4-SOL (32), DIR1-SIM (11), DIR2-MAD (3), and DIR3-GPS (1). The associated parameter estimates are qualitatively very similar across these 59 points. The maximum gradient norm across the 59 terminal points is 0.1, the maximum weighted gradient criterion is 0.001, and the implied Hessians are all positive definite. For the same set of local minima, the range for $\kappa(H)$ is 2.61E+03 to 2.66E+03. We also identified additional local optima with objective function values of 167.63, 211.36, 218.56, 226.59, and 248.14. Across all of these six local optima, the maximum of the implied gradient norms does not exceed 0.001, the weighted-gradient criterion is below .0001, and the implied Hessians are all positive definite, with, $\kappa(H)$, between 2.08E+03 and 2.51E+04.

Our results using a loose NFP tolerance underscore the findings in Dube et al. (2012). In the case of automobiles with loose NFP tolerance, a single set of starting values for DER2-QN2 gave rise to an objective function value of 204.54 when rounded to the second decimal point. Although the gradient norm and the weighted-gradient criterion are 0.08 and 0.02, the Hessian exhibits both positive and negative eigenvalues, which is consistent with a saddle point. As a note of caution, both the size and the sign of the Hessian eigenvalues seemed to be particularly sensitive to parameter estimates generating objective function values very close—within decimal points—to 204.54.

In the case of cereals with loose NFP tolerance, the DER5-KNI algorithm hovered around an objective function value of about 15.5. More precisely, sixteen sets of starting values gave rise to function values between 15.46 and 15.60 when rounded to the second decimal point. The coefficient estimates implied by these sixteen sets of starting values are largely identical with the exception of the two coefficients associated with the interaction of price with log income and log income squared. The range of these two coefficients is -1.42 to 2.00 and 0.09 to 0.27, respectively.

In all sixteen instances, although the gradient norm (weighted-gradient criterion) is between 0.0393 (-0.0002) and 0.0968 (0.0007), the Hessians exhibit both positive and negative eigenvalues. These diagnostics are consistent with a saddle point.[24] The range of the condition number $\kappa(H)$ is between -1.50E+03 and -2.75E+02 across the sixteen points. Hence, the numerical accuracy of the solutions does not seem to be a concern.[25]

VII. Implications for Economic Variables of Interest

A. *Preliminaries*

In this section, we examine the implications of the variation in the terminal points of the nonlinear searches for economic variables of interest routinely studied in the literature, such as elasticities, consumer welfare, and firm profits. Following our extensive optimization exercise, the number of observations for the analysis of the variation in such variables, especially at the product level, is immense. Every combination of parameter starting values, optimization algorithm, and NFP tolerance yields an elasticity matrix for each market; of

[21] See theorems 4.1.1 and 4.1.2 in Amemiya (1985) or theorems 5.1 and 5.2 in Cameron and Trivedi (2005). Amemiya (p. 230) states that identification is synonymous with "the existence of a consistent estimator."

[22] Cameron and Trivedi (2005, p. 127) argue that it is best in such cases to consider the global optimum and apply their theorem 5.1. According to McFadden and Newey (1994, p. 2117), as long as the extremum estimator is consistent and the true parameter is an element of the interior of the parameter space, an extremum estimator will be a root of the first-order conditions asymptotically and hence will be included among the local optima. Amemiya (1985, p. 111) suggests two ways to gain some confidence that a local optimum is a consistent root. First, the solution gives a reasonable value from an economic-theoretic viewpoint. Second, the iteration by which the local optimum was obtained started from a consistent estimator.

[23] For one of the papers, the possibility of local optima is discussed in one of the footnotes.

[24] More specifically, KNITRO generated the exit code of 0 (-100) in 14 (2) instances. According to appendix A in Waltz and Plantenga (2009), the exit code 0 indicates that "the final solution satisfies the termination conditions for verifying optimality," while exit codes -100 to -199 indicate that "a feasible approximate solution was found."

[25] Recently, Judd and Skrainka (2011) show that Monte Carlo integration for evaluating the market share integrals creates ripples in the surface of the objective function that generate spurious local maxima. Skrainka (2011) argues that when instruments are highly collinear, which is often the case for the BLP-type instruments based on product characteristics, the GMM weighting matrix has a high condition number and the outer-loop nonlinear solver finds many local optima.

course, not all of these combinations are meaningful. In the case of the automobile data, there are 20 markets, with the number of products in each market being between 72 and 150, leading to 2,217 product-market combinations. In the case of the cereal data, there are 94 markets with 24 products in each market, leading to 2,256 product-market combinations.

At the product level, we focus on the implication for the own-price elasticities for two products in each data set. The first is the product with the largest observed quantity sold—the top product. The second is the product with the median observed quantity sold—the median product. For each of the two products of interest, our analysis is based on 459 observations for automobiles and 240 observations for cereals. In the former data set, the top product appears in market 2 (1972), and its market share is about 1%. The median product appears in market 16 (1986), and its market share is 0.05%.[26] In the latter data set, the top product is brand 5, a Kellogg's brand, that appears in market 53, and its market share is 45%. The median product is brand 13, a General Mills brand, that appears in market 89, and its market share is 1%.

At the market level, we examine the implications for aggregate elasticity, as well as the change in profits and consumer welfare, for two hypothetical mergers. We calculated the aggregate elasticity by simulating a 1% price increase for all products. We use compensating variation as a measure of change in consumer welfare. For automobiles, we assume GM and Chrysler merge. In the case of cereals, we assume Kellogg's and General Mills merge. As we discussed earlier, we use the approximate solutions for the postmerger prices of equation (13). Similar to the analysis at the product level, we work with 459 observations in the case of automobiles and 240 observations in the case of cereals.

We limit our attention to results implied by those sets of starting values that allowed the optimization algorithms to converge using tight NFP tolerance excluding some additional sets of results. First, we excluded all of the results implied by combinations of starting values and optimization algorithms that gave rise to problematic pre- and postmerger market shares (for example, NaNs). Second, in the case of automobiles (cereals), we excluded results associated with combinations of starting values and optimization algorithms producing objective function values larger than 282.99 (134.92).[27] The process just described led to 459 observations for automobiles and 240 observations for cereals.

According to our findings, the variation across the set of candidate estimates for cereals is much smaller than the variation in the automobile data. This is not surprising given that three of the algorithms nearly always converge at the GMM objective value of 4.56 for cereals. However, it is important to keep in mind that only these three algorithms attained such an objective function value. Therefore, if a researcher were to rely only on parameter estimates for the other eight algorithms, the variation would be more pronounced.

B. Product Level

Panel A of figure 2 provides a histogram of the own-price elasticity (own elasticity, henceforth) for the top automobile product, where we see values as low as -3.48 and as high as -0.93, when we restrict ourselves to only negative elasticities. The mean across the 453 observations for which the elasticity is negative is -2.77, which is similar to the median of -3.03, and the standard deviation is 0.53. Using the absolute value of the mean, the implied coefficient of variation (CoV) equals 0.19. We see substantial variation in the distribution of the own elasticities with a rather prominent spike around -3.2.

The variation is notable even when we limit our attention to the 78 observations for the local optima. Recall from our earlier discussion that we identified seven local optima with objective function values of 157.79, 167.63, 204.56, 211.36, 218.56, 226.59, and 248.14. The implied own elasticities for these seven local optima are -2.30, 2.17, -3.21, -2.59, -2.50, -2.33, and -1.29. The own elasticity for the smallest objective function value of 157.79 is slightly below the 75 percentile, which is -2.34.

The histogram of own elasticities for the median automobile product is available in panel A of figure 3. In this case, the minimum across the 459 observations is smaller than the maximum by almost a factor of 5: -4.91 versus -1.03. Once again, we observe substantial variation in the own-elasticity distribution with a prominent spike in the neighborhood of -3.1. When we exclude own elasticities below -3.28 (5th percentile), the mean is -2.62 and the median is -2.85 across the remaining 437 observations. The standard deviation is 0.6, and the implied CoV is 0.23. Similar to the top product, the variation is substantial when we limit our attention to the observations associated with the seven local optima: -3.22 (157.79), -3.29 (167.63), -3.14 (204.56), -1.60 (211.36), -1.95 (218.56), -2.85 (226.59), and -1.71 (248.14). The own elasticity for the smallest objective function value of 157.79 may be viewed as an outlier given that the 5th percentile is -3.28.

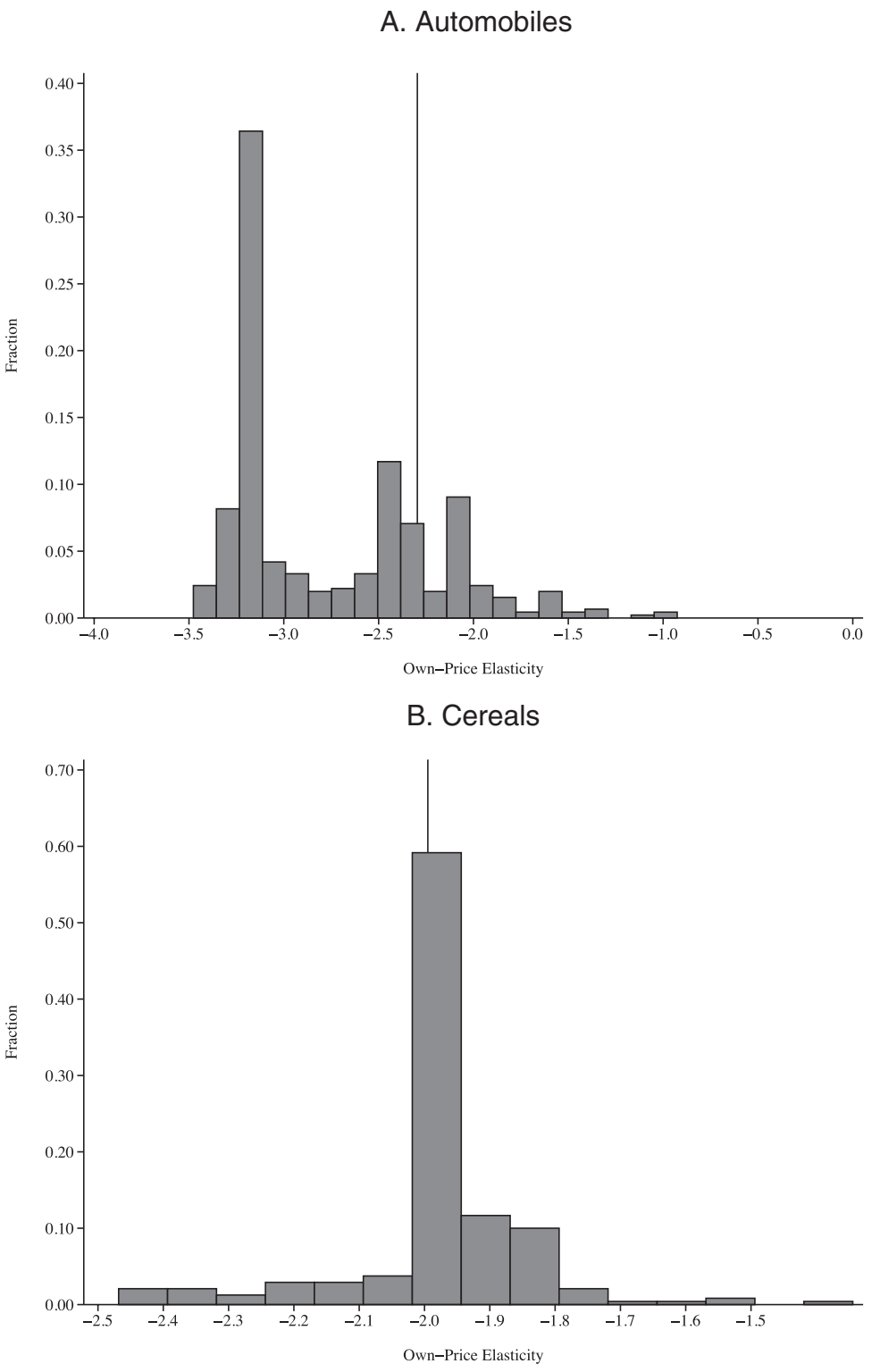
Further support for the variation in own elasticities due to combinations of starting values and algorithms is provided by the histogram in panel A of figure 4. This histogram is based on the own-elasticity CoV for each of the 2,107 automobile product-market pairs across the 459 combinations of optimization algorithms and starting values we considered.[28] The hump of the distribution covers the 0.16 to 0.27 range, and a long right tail is prominent. The mean and median of the CoV distribution are very similar—0.25 and 0.23,

[26] Using the NEWMODV field in the automobile data, the top (median) product is BKRIVE72 (NIPULS84).

[27] For automobiles, 282.99 is the 95th percentile of the objective function value distribution. The 99th percentile is 1,691.29, and the maximum is in excess of 4,700. In the case of cereals, 134.92 is the 75th percentile. The 90th percentile is 314.58, and the maximum is 1,904.78.

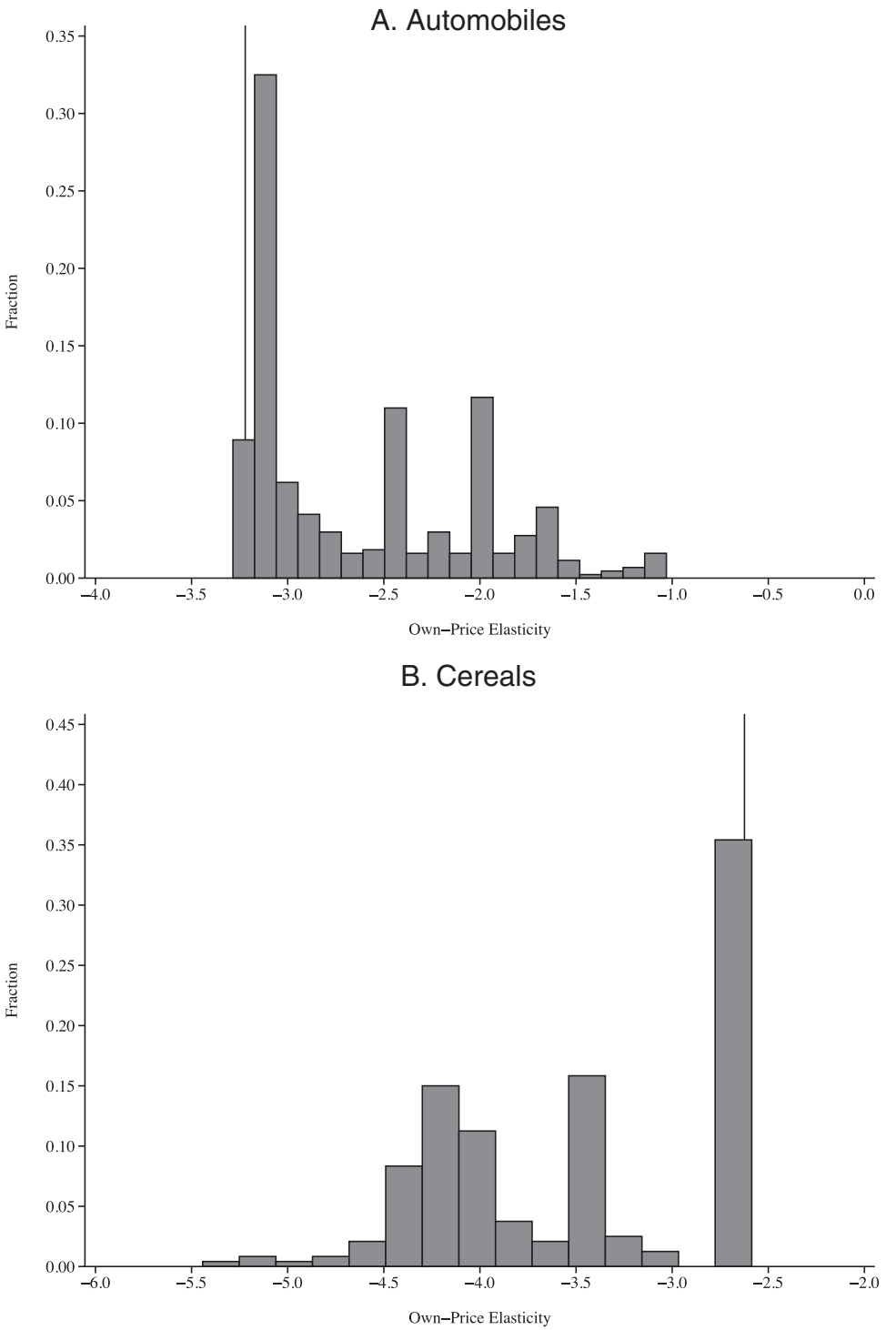
[28] The number of product-market pairs is fewer than 2,217 because we exclude observations exceeding the 95th percentile of the CoV distribution, which is 0.67. Each of 2,217 coefficients of variation is calculated using 459 observations.

FIGURE 2.—OWN-PRICE ELASTICITY HISTOGRAM FOR THE TOP PRODUCT



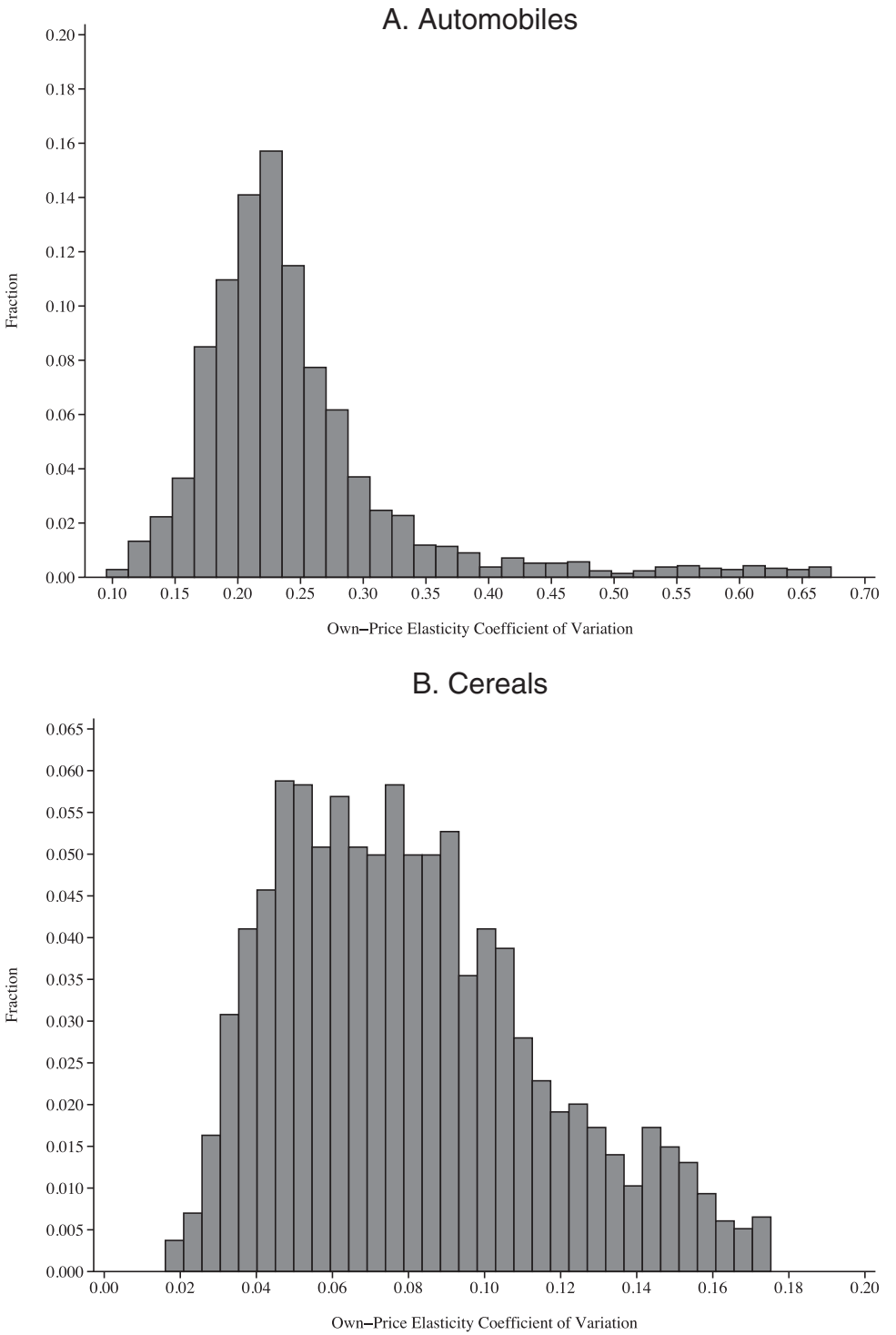
The top product is the product with the largest observed unit market share. For automobiles, the top product is BKRIVE72 in 1972 with a market share of 1%. For cereals, the top product is brand 5 in market 53 with market share of 45%. The histograms are based on 453 (240) observations in the case of automobiles (cereals). Observations exceeding 0 are excluded in the case of automobiles. The number of observations reflects the sets of starting values that allowed the optimization algorithms to converge using tight NFP tolerance, the removal of extraneous observations discussed in section VIIA, and any thresholds discussed here. The vertical lines indicate the value of the own-price elasticity when the objective function value is 157.79 (4.56) for automobiles (cereals).

FIGURE 3.—OWN-PRICE ELASTICITY HISTOGRAM FOR THE MEDIAN PRODUCT



The median product is the product with the median observed unit market share. For automobiles, the median product is NIPULS84 in 1986 with a market share of 0.05%. For cereals, the median product is brand 13 in market 89 with a market share of 1%. The histograms are based on 437 (240) observations in the case of automobiles (cereals). Observations falling below −3.28 are excluded in the case of automobiles. The number of observations reflects the sets of starting values that allowed the optimization algorithms to converge using tight NFP tolerance, the removal of extraneous observations discussed in section VIIA, and any thresholds discussed here. The vertical lines indicate the value of the own-price elasticity when the objective function value is 157.79 (4.56) for automobiles (cereals).

FIGURE 4.—OWN-PRICE ELASTICITY COEFFICIENT-OF-VARIATION HISTOGRAM FOR ALL PRODUCTS



An observation is the ratio of the standard deviation to the absolute value of the mean for a product-market combination. The number of observations used to calculate the two moments is 459 (240) for automobiles (cereals). Coefficients of variation exceeding 0.67 (0.18) for automobiles (cereals) are excluded. These upper bounds are the 95th percentiles of the corresponding distributions. The number of observations reflects the sets of starting values that allowed the optimization algorithms to converge using tight NFP tolerance, the removal of extraneous observations discussed in section VIIA, and any thresholds discussed here.

respectively. Overall, using CoV as a metric, we see substantial variation in the own-elasticity distribution for a large number of product-market pairs due to combinations of algorithms and starting values.

Panel B of figure 2 provides a histogram of the own elasticity for the top cereal product. The own elasticity across the 240 observations is between −2.47 and −1.34 with a mean of −1.98 and a standard deviation of 0.14—the CoV equals 0.07. Although we see notable variation, given the number of routines that terminated at an objective function value of 4.56, the variation is not extreme. The median own elasticity (−1.99) is essentially identical to the own elasticity for the smallest objective function value of 4.56.

We provide a histogram of the own elasticity for the median cereal product in panel B of figure 3. The own elasticity across the 240 observations is between −5.44 and −2.59 with a mean of −3.49 and a standard deviation of 0.73. The implied CoV is 0.21, and the median is −3.47. Similar to the largest product, the variation is notable. The 75th percentile of the distribution, which equals −2.62, is the own elasticity for the smallest objective function value of 4.56.

Finally, the CoV histogram for cereals is available in panel B of figure 4. This histogram is based on the own-elasticity CoV for each of the 2,144 cereal product-market pairs across the 240 combinations of optimization algorithms and starting values we considered.[29] The CoV values range from about 0.02 to around 0.18 with a mean of 0.08 that is indistinguishable from the median. Similar to automobiles, but to a lesser extent, the histogram indicates a nonnegligible amount of variation in the distribution of own elasticities for a large number of product-market pairs.

Product level: Optimization vs. sample variation. Trying to assess whether the variation in economic variables of interest due to the optimization design is different from the variation due to the underlying sample error, we focus on the own elasticities for the automobile products of the previous section.[30] To show the sample variation, we constructed the kernel density of own elasticities implied by 10,000 draws from a $N(\widehat{\theta}_2, \Sigma_{\widehat{\theta}_2})$ distribution. We first used the estimates $\widehat{\theta}_2$ and the variance-covariance matrix $\Sigma_{\widehat{\theta}_2}$ that correspond to the objective function value of 157.79 in table A2 in the online appendix for DIR1-SIM to generate 10,000 sets of parameter estimates. We then calculated the own elasticities implied by each of these sets of parameter estimates. To show the optimization variation, we constructed the kernel density using the 453 (437) own elasticities of the histograms in the bottom panel of figure 2 (figure 3).

For the top automobile product, as panel A of figure 5 indicates, the two densities overlap partially, roughly in the range −3 to −1.5. The density capturing the sample variation is less dispersed. The mean of the own elasticities implied by the 10,000 draws is −2.31, and the standard deviation is 0.27. The mean of the 453 own elasticities capturing the optimization variation is −2.77, and the standard deviation is 0.53. The kernel density capturing the optimization variation exhibits two distinct modes, which correspond to elasticities of −3.21 and −2.47. The first modal own elasticity is very close to the one implied by the parameter estimates for the local optimum at 204.56, while the second one is close to the one implied by the parameter estimates for the local optimum at 218.56.

The partial overlap of the two densities also holds in the case of the median automobile product in panel B of figure 5. Across the 10,000 draws, the mean own elasticity is −3.23, and the standard deviations is 0.54. The mean of the 437 own elasticities capturing the optimization variation is −2.62, and the standard deviation is 0.59. The kernel density capturing the optimization variation has a distinct peak at −3.11 and a plateau around −2.23. The own elasticity of −3.11 is very close to the one implied by the parameter estimates for the local optimum at 204.56, while the own elasticity of −2.23 is similar to the elasticity implied by the parameter estimates for the local optimum at 218.56.

C. *Market Level*

We start our discussion in this section with the histogram of the average aggregate elasticity for automobiles in panel A of figure 6. We calculated the average aggregate elasticity implied by each of the 459 sets of estimates in two steps. In the first step, we calculated a market-specific aggregate elasticity following a price increase of 1% across all products.[31] The average aggregate elasticity ranges from −1.74 to −0.41 with a mean (median) of −1.20 (−1.29) and a standard deviation of 0.24. The ratio of the minimum to the maximum exceeds 4, and the aggregate elasticity exceeds −1 in 105 instances. The variation is still substantial when we exclude the top and bottom 5% of the distribution, that is, values below −1.46 and above −0.65. Limiting our attention to values associated with the local optima, we get the following: −0.57 (157.79), −1.01 (167.63), −1.39 (204.56), −0.82 (211.36), −1.18 (218.56), −0.93 (226.59), and −0.41 (248.14).

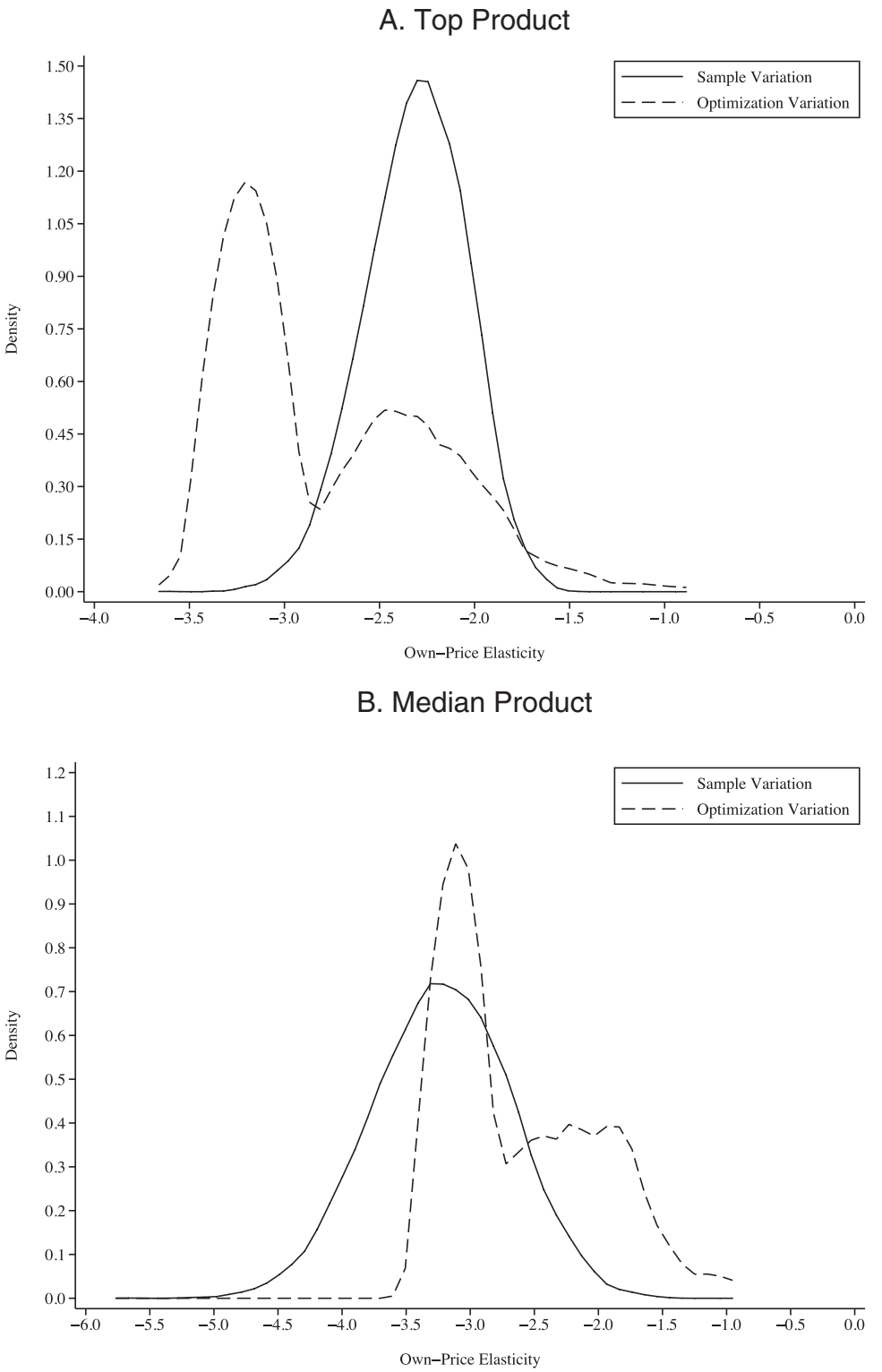
The distribution of the average change in profits due to a hypothetical merger between GM and Chrysler is available in panel A of figure 7. We calculated the average change in profits implied by each of the 459 sets of parameters in two steps. In the first step, we approximated the postmerger prices in each market using equation (13) in section III. In the second step, we calculated the quantity-weighted average change in profits across the twenty markets. Excluding the top 5% and bottom 10% of the distribution, which is equivalent to

[29] The number of product-market pairs is fewer than 2,256 because we exclude observations exceeding the 95th percentile of the CoV distribution, which is 0.175. Each CoV is calculated using 240 observations.

[30] We will refer to the variation due to the optimization design (sample error) as "optimization" ("sample") variation. We use the term *optimization design* to refer to the combination of optimization algorithm, starting values, and NFP tolerance.

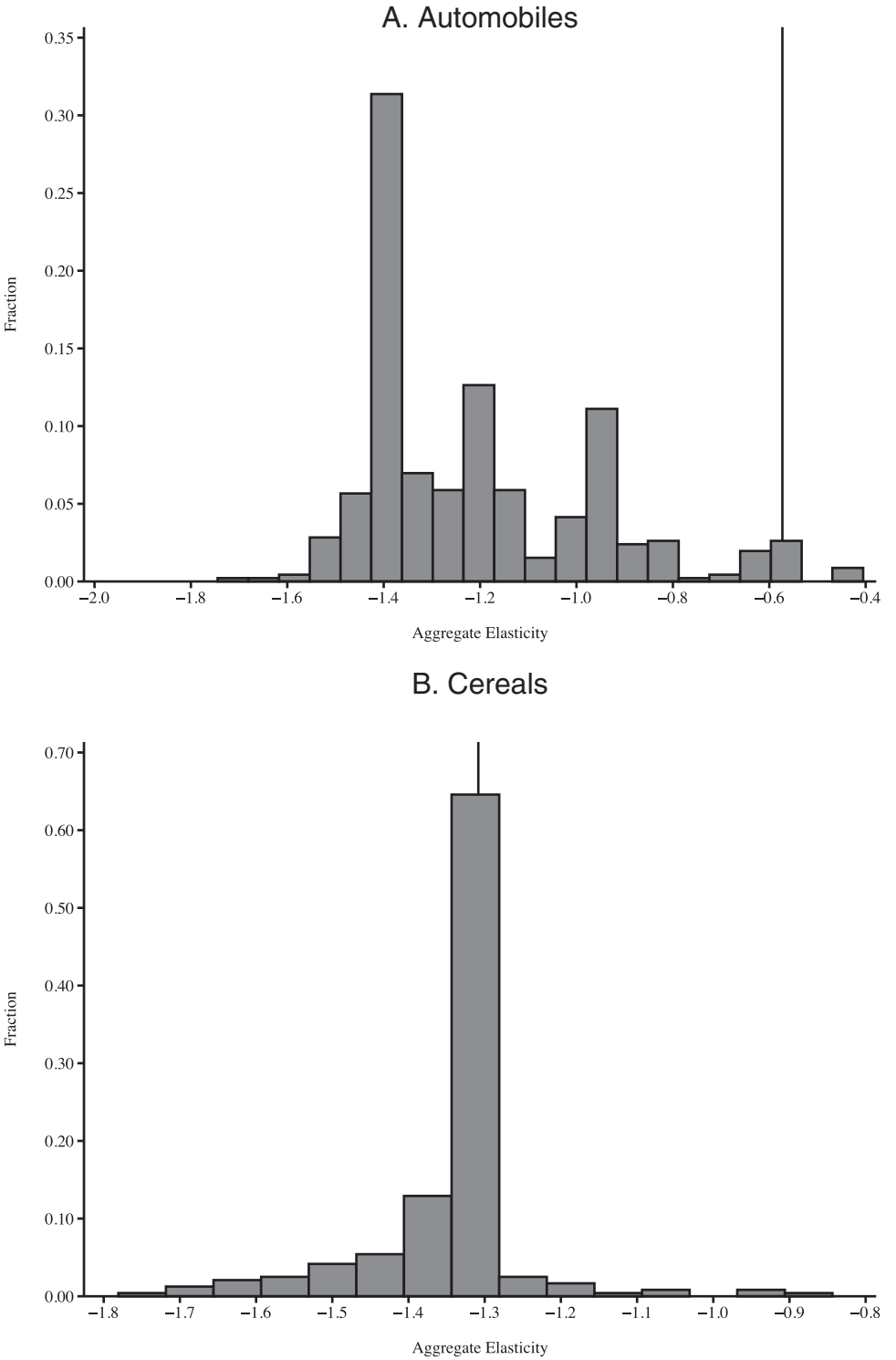
[31] The automobile data contain information about the total market size—that is, including the outside good. We constructed our quantity-weighted average across the twenty markets for each of the 459 sets of estimates using the units associated with all inside goods for all three economic variables of interest.

FIGURE 5.—SAMPLE VERSUS OPTIMIZATION VARIATION OF THE OWN-PRICE ELASTICITY FOR AUTOMOBILE PRODUCTS



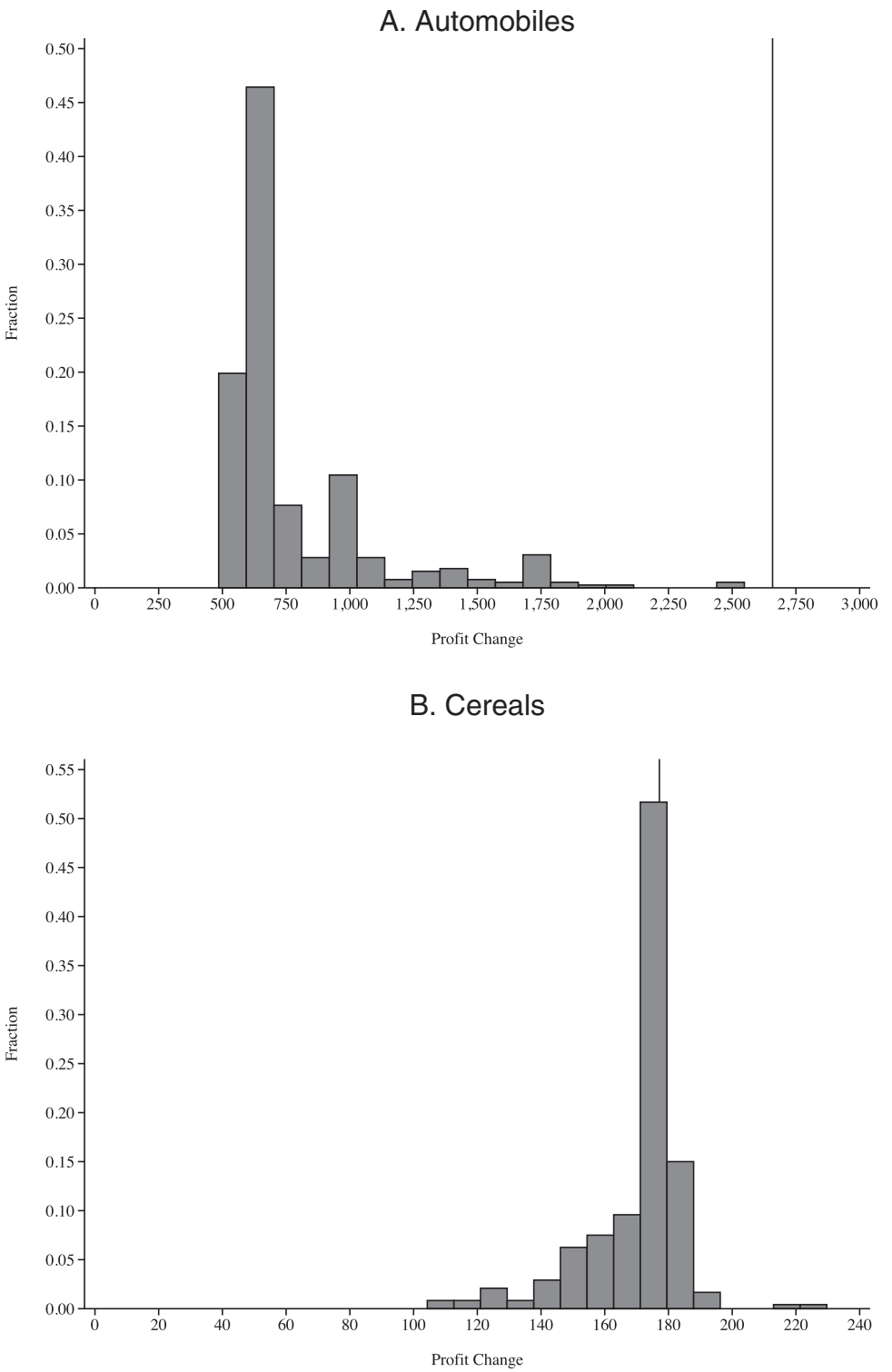
To show the variation due to the underlying sample error, we constructed the kernel density calculating own-price elasticities for 10,000 draws from a $N(\widehat{\theta_2}, \Sigma_{\widehat{\theta_2}})$ distribution. We used the estimates $\widehat{\theta_2}$ and the variance-covariance matrix $\Sigma_{\widehat{\theta_2}}$ that correspond to the objective function value of 157.79 for DIR1-SIM in table A2 in the online appendix. To show the variation due to the optimization design, we constructed the kernel density using the 453 (437) own-price elasticities in the top panel of figure 2 (figure 3).

FIGURE 6.—AVERAGE AGGREGATE ELASTICITY



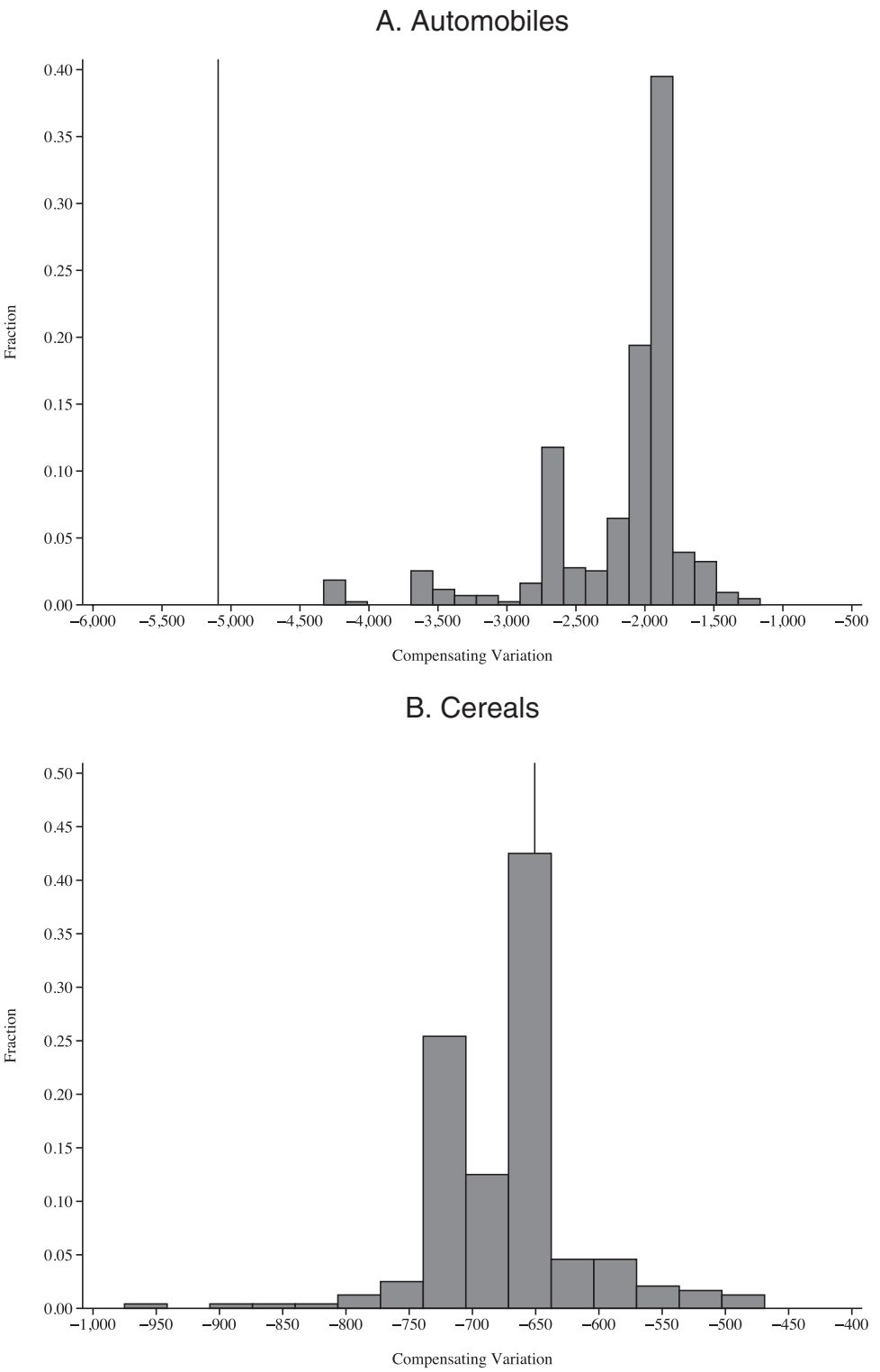
The histogram is based on 459 (240) observations in the case of automobiles (cereals). Each observation corresponds to a weighted average aggregate elasticity across 20 (94) markets using premerger quantities of the inside goods as weights. The number of observations reflects the sets of starting values that allowed the optimization algorithms to converge using tight NFP tolerance, the removal of extraneous observations discussed in section VIIA, and any thresholds discussed here. The details of the aggregate elasticity calculation are available in section VIIC. The vertical lines indicate the value of the average aggregate elasticity when the objective function value is 157.79 (4.56) for automobiles (cereals).

FIGURE 7.—AVERAGE CHANGE IN PROFITS ($ MILLIONS)



The histogram is based on 392 (240) observations in the case of automobiles (cereals). For automobiles, values below $485 million (10th percentile) or above $2,548 million (95th percentile) are excluded. Each observation corresponds to a weighted average change in profits across 20 (94) markets using premerger quantities of the inside goods as weights. The number of observations reflects the sets of starting values that allowed the optimization algorithms to converge using tight NFP tolerance, the removal of extraneous observations discussed in section VIIA, and any thresholds discussed here. The details of the compensating variation calculation are available in section VIIC. The vertical lines indicate the value of the average change in profits when the objective function values is 157.79 (4.56) for automobiles (cereals).

FIGURE 8.—AVERAGE MARKET COMPENSATING VARIATION ($ MILLIONS)



The histogram is based on 433 (240) observations in the case of automobiles (cereals). For automobiles, values below −4,328 million (5th percentile) or above 0 are excluded. Each observation corresponds to a weighted average compensating variation across 20 (94) markets using premerger quantities of the inside goods as weights. The number of observations reflects the sets of starting values that allowed the optimization algorithms to converge using tight NFP tolerance, the removal of extraneous observations discussed in section VIIA, and any thresholds discussed here. The details of the compensating variation calculation are available in section VIIC. The vertical lines indicate the value of the average market compensating variation when the objective function values is 157.79 (4.56) for automobiles (cereals).

discarding values below \$485 million and above \$2,548 million, leaves us with 392 observations. Following the truncation of the distribution just described, the mean (median) is \$787 million (\$669 million) and a standard deviation of \$316 million. The change in profits associated with the local optima varies widely: \$2,658 million (157.79), \$1,697 million (167.63), \$668 million (204.56), \$1,427 million (211.36), \$568 million (218.56), \$1,080 million (226.59), and \$3,729 million (248.14).

We provide the histogram of the average market compensating variation (MCV) in panel A of figure 8. We calculated the average MCV implied by each of the 459 sets of parameters in two steps. In the first step, we obtained MCV by multiplying the total market size by the average compensating variation (ACV). We derived ACV using equation (14) in section III and averaging across individuals. In the second step, we calculated the quantity-weighted average MCV across the twenty markets. After dropping the observations in the bottom 5% of the distribution or exceeding 0, which leads to 433 observations, we see values between −\$4,328 million and −\$1,165 million. The mean (median) is −\$2,183 million (−\$1,975 million), and the standard deviation is \$531 million. The change in consumer welfare captured by MCV at the seven local optima is −\$5,093 million (157.79), −\$2,739 million (167.63), −\$1,926 million (204.56), −\$3,550 million (211.36), −\$1,977 million (218.56), −\$2,731 million (226.59), and −\$7,369 million (248.14).

Overall, the distributions of the automobile economic variables of interest at the market level indicate that the values of such variables associated with the smallest objective function value of 157.79 are outliers. In addition, parameter estimates that lead to very similar objective function values can have extremely different economic predictions. For example, when we compare the two local minima at 211.36 and 218.56, the change in profits is larger for the parameter values associated with 211.36 by a factor of 2.5, while the change in consumer welfare is almost twice as large. This underscores our discussion of the potential for a horse race even when the researcher is convinced that the global minimum is the consistent root. One could imagine a situation where, for a slightly different sample, the parameter values associated with the minimum at 218.56 yield a smaller objective value than the parameter values associated with the minimum at 211.36.

The histogram of the average aggregate elasticity for cereals is available in panel B of figure 6. Following the same approach with automobiles, we calculated a quantity-weighted average aggregate elasticity across the 94 markets for each of the 240 sets of parameter estimates. Given that the cereal data do not contain information about the market size, we assumed a total market size (including the outside good) of 250 million servings per day. When expressed in million servings per day, the quantity weight for each market is equal to the share of the inside goods times 250.[32] The average aggregate elasticity is between −1.78 and −0.84, with its distribution having a prominent peak at about −1.3, which is very close to the mean and the median. This is also the value of the average aggregate elasticity if we limit our attention to the minimum with objective function value of 4.56. The standard deviation is 0.11. Given that the 95th percentile is −1.25, we would conclude that the average market is inelastic only for a handful of estimates. The variation is less pronounced if we exclude observations in the top and bottom 5% of the distribution—in this case, the elasticity values are between −1.56 and −1.25.

We also calculated a quantity-weighted average annual change in profits following a hypothetical merger between Kellogg's and General Mills. The histogram of such profit changes is available in panel B of figure 7. We calculated the annual change in profits in each market multiplying the market size by 365. Following the same steps with automobiles, we calculated the quantity-weighted average annual change in profits across the 94 markets for each of the 240 sets of parameter estimates. Similar to the histogram for the average aggregate elasticity, we do not see a single mass point; actually, we observe substantial variation. The average annual change in profits due to the merger ranges from \$104.3 million to \$229.7 million with a mean (median) of \$170.8 million (\$177.1 million) and standard deviation of \$15.7 million. When we limit our attention to the estimates that give rise to the objective function value of 4.56, the change in profits is about \$177.1 million. The variation in profit change when we exclude the top and bottom 5% of the distribution is still substantial: \$138.3 million to \$182.1 million.

Finally, we calculated a quantity-weighted average annual change in consumer welfare for cereals due to the hypothetical merger using MCV; the associated histogram is available in panel B of figure 8. We first calculated a market-specific annual MCV (\$million) as ACV × 250 × 365. We then calculated an average across the 94 markets using the units of the inside goods as weights for each of 240 sets of estimates. As it was the case with the aggregate elasticity and change in profits, we see substantial variation in MCV. The range is from −\$975 million to −\$469 million with a mean (median) of −\$671 million (−\$651 million) and standard deviation of approximately \$60 million. The change in consumer welfare corresponding to the objective function value of 4.56 is almost identical to the median. Even after excluding the top and bottom 5% of the MCV distribution, the MCV exhibits substantial variation, namely, −\$744 million to −\$569 million.

VIII. Conclusion

Empirical industrial organization has been increasingly relying on highly nonlinear structural models and probably more so compared to other neighboring fields, such as

[32] The U.S. population was approximately 250 million based on Census figures for July 1990; see http://www.census.gov/popest/data/national/totals/1990s/tables/nat-agesex.txt. The assumption about the total market size is admittedly somewhat arbitrary. However, it is inconsequential for the variation of the results reported here because the market size operates as a scaling factor.

public and labor economics. The reasons for such divergence have been discussed in a rather lively manner in Angrist and Pischke (2010) and Nevo and Whinston (2010). At the same time, a prominent class of such models, the demand models for differentiated products of the Berry et al. (1995) tradition, are becoming increasingly popular in trade, education, housing, health, and environmental economics and shedding light on a wide spectrum of important economic questions.

Nonlinear models are often synonymous with an objective function that is not globally concave or convex. Obtaining parameter estimates and performing inferences is possible, in principle, using a nonlinear search algorithm along with a set of multiple starting values and stopping rules. Both exercises, however, can be particularly challenging when it comes to implementation. In this paper, we document some of the challenges we experienced estimating BLP-type demand models using two widely known data sets, numerous search algorithms, a large number of starting values, and different tolerances of the fixed-point iterations that allow the researcher to infer the structural econometric error.

Our findings point to instances of convergence at points where the first- and second-order optimality conditions fail. We also find that various combinations of optimization algorithms, starting values, and fixed-point iterations may lead to convergence at multiple local optima, as well as to instances of convergence failure. Even on convergence under tight tolerance for the fixed-point iterations, we find substantial variation in the objective function value both within and across optimization algorithms. Although derivative-based algorithms seem to perform better in the case of the cereal data set, there appears to be a tight race between derivative-based and direct-search algorithms for the automobile data set. This variation goes in hand with variation in parameter estimates and translates into variation of the demand models' economic predictions, such as price elasticities, consumer welfare, and firm profits.

In the automobile data set, the range of the own-price elasticity for the product with the highest market share is such that the smallest to the largest values differ by a factor of about 4. The same factor is of similar magnitude in the case of the aggregate elasticity and over 5 for the change in firm profits due to a hypothetical merger. The range of the change in the consumer welfare in our merger exercise is wide too. All of these economic variables of interest exhibit notable variation across multiple optima that we identify through a careful review of first- and second-order optimality conditions.

In the cereal data set, the value of the own-price elasticity of the product with the largest market share exhibits a range such that the smallest and largest values differ by a factor of 2 depending on the combination of optimization algorithm and starting values. At the market level, the range is qualitatively similar for the aggregate elasticity, as well as for the change in consumer welfare and firm profits following a hypothetical merger in the industry.

Drawing from our experience, we offer a number of suggestions to empiricists using (not only) BLP-type demand models. We abstract from the typical robustness checks to the baseline results of an econometric exercise, such as alternative samples, model specifications, and explanatory variables or instruments, to name a few. These important robustness checks receive the proper attention in empirical exercises that are carefully executed. In nonlinear models, there is a second checklist that typically receives less attention, but researchers should go through, *and report*, to feel confident for her results. We not that the typical robustness checks may change the shape of the underlying objective function and, hence, the nature of the optimization problem in hand. In the next paragraphs, we discuss how to navigate through this checklist in a stepwise manner (see also table 4). Goldberg and Hellerstein (2013) provide a very good example of how some of these suggestions can be implemented in empirical work.

TABLE 4.—OPTIMIZATION-DESIGN DETAILS AND DIAGNOSTICS CHECKLIST

Step	Details and Diagnostics
1. Optimization design	Optimization algorithm Starting values Objective function value tolerance Parameter vector tolerance Other optimization settings Fixed-point iteration settings Market share evaluation draws
2. Convergence and local optima	Multiple optima (Y/N) Number of runs converged Algorithm exit code Objective function value Parameter estimates Gradient-based FOC diagnostics Hessian-based SOC diagnostics
3. Implications for economic variables of interest	Variation due to multiple optima, if any, for: • Objective function value • Parameter estimates • Own- and cross-price elasticities: statistics • Other economic variables of interest: statistics

Step 1: Optimization design. Estimate the objective function using multiple optimization algorithms, ideally from different classes, with a large number of starting values. Things to consider include different settings for the algorithm's tunning parameters and stopping rules based on tight tolerances. In the absence of any prior information for starting values, which may come from economic theory, random draws are as good as anything else. Alternatively, use a large number of random draws (10,000) and evaluate the objective function. Among these draws, pick the ones that give rise to, say, the fifty smallest objective function values and use them as the starting values for the nonlinear search. Another possibility is to divide the optimization problem in hand in smaller ones. For example, assume that we can split the parameter vector, say, γ, into two parameter vectors of smaller dimension, say, γ_1 and γ_2. First, fix the elements of γ_2 at some reasonable values and estimate γ_1. With estimates of γ_1 in hand, estimate the elements of γ_2. Iterate this process until the objective function value changes very little. Finally, estimate γ_1 and γ_2 simultaneously.

Consider the possibility of combining optimization algorithms: initiate a random direct search algorithm from the terminal point of a gradient-based algorithm. Analytical gradients and Hessians are highly recommended, but their computation can be demanding in terms of computer code and tedious debugging. Make sure the parameters over which the nonlinear search is performed are of similar magnitudes because improper scaling may affect the progress of the optimization algorithm (numerical issues); there are tips and tricks about preconditioning and scaling in most optimization books and manuals that accompany optimization software. It is a good idea to plot the objective function value against each of the parameters over an interval around the final estimate, while holding the remaining ones constant, and make sure that you pass the "ocular" identification test. Use tight tolerance and potentially multiple starting values for the NFP iterations if employed. Section VIA offers some guidance for this step. The accuracy in the evaluation of the market share integrals using Monte Carlo (MC) simulation improves as $\sqrt{R}$, where R is the number of draws—the empiricist must increase the number of draws by 100 to gain a digit in accuracy. Finally, recent work has identified alternatives to (MC) simulation for the evaluation of market share integrals.

Step 2: Convergence and local optima. Once estimation is complete, collect the parameter estimates, objective function values, and the algorithms' exit codes across the various combinations of the optimization design in step 1. Based on the algorithms' exit codes and documentation, first discard those combinations for which there is a clear indication that the algorithm did not converge. It is important to maintain and review carefully log files recording the algorithm's progress during the nonlinear search and make sure that they follow their rate of convergence. Second, examine whether the objective function values and the parameter estimates that pass the convergence test correspond to local minima by examining the associated gradient vectors and Hessian matrices using diagnostics similar to the ones we discussed in section VIB. Most of the practitioners tend to focus only on the global minimum despite the well-documented literature that any of the plausible from economic-theoretic viewpoint local optima can be a consistent root. As we discussed in section VIC, it is often difficult to find the global minimum of a criterion function aside from the case of a globally convex criterion function. We suggest that in the presence of multiple local minima, the results of all other optima should be reported. If the researcher can eliminate some set of results on economic, or statistical grounds, this discussion should also be included.

Step 3: Implications for economic variables of interest. Summarize the variation in the economic variables of interest implied by the underlying sample error and the nature of the underlying objective function if needed. Regarding the sample error, each economic variable of interest is a function of the estimated parameters, and its standard error can be obtained using bootstrap (see, for example, figure 5). Regarding the nature of the underlying objective function, we suggest the following "fire-detection" test: after step 2, construct a histogram of the objective function values; if there is no a single mass point ("smoke"), look for variation in the variables of interest focusing on local optima ("fire"). In the case of the BLP-type demand models, own- and cross-price elasticities can serve as a natural starting point for documenting such variation (see our discussion in section VIIB). After all, the variation in substitution patterns implied by the demand system should translate into variation in economic variables of interest that are application specific, such as profits, or consumer welfare. Even when the values of the objective function differ by a small amount, variables of economic interest can vary considerably.

REFERENCES

Ackerberg, D., and M. Rysman, "Unobserved Product Differentiation in Discrete-Choice Models: Estimating Price Elasticities and Welfare Effects," *Rand Journal of Economics* 36 (2005), 771–788.

Allon, G., A. Federgruen, and M. Pierson, "Price Competition Under Multinomial Logit Demand Function with Random Coefficients," Northwestern University Kellogg School of Management, Columbia University Graduate School of Business, and Harvard Business School, technical report (2011).

Amemiya, T., *Advanced Econometrics* (Cambridge, MA: Harvard University Press, 1985).

Andrews, D., "A Stopping Rule for the Computation of Generalized Methods of Moments Estimators," *Econometrica* 65 (1997), 913–931.

Angrist, J., and J. Pischke, The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics, *Journal of Economic Perspectives* 24 (2010), 3–30.

Armantier, O., and O. Richard, "Domestic Airline Alliances and Consumer Welfare," *Rand Journal of Economics* 39 (2008), 875–904.

Audet, C., and J. Dennis, Mesh Adaptive Direct Search "Algorithms for Constrained Optimization," *SIAM Journal on Optimization* 17 (2006), 188–217.

Bayer, P., F. Ferreira, and R. McMillan, "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy* 115 (2007), 588–638.

Berry, S., "Estimating Discrete-Choice Models of Product Differentiation," *Rand Journal of Economics* 25 (1994), 242–262.

Berry, S., J. Levinsohn, and A. Pakes, "Automobile Prices in Market Equilibrium," *Econometrica* 63 (1995), 841–890.

——— "Voluntary Export Restraints on Automobiles. Evaluating a Trade Policy," *American Economic Review* 89 (1999), 400–430.

——— "Differentiated Products Demand Systems from a Combination of Micro and Macro Date: The New Car Market" *Journal of Political Economy* 112 (2004), 68–105.

Berry, S., O. Linton, and A. Pakes, "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems," *Review of Economic Studies* 71 (2004), 613–654.

Bonnet, C., and P. Dubois, "Inference on Vertical Contracts between Manufacturers and Retailers Allowing for Nonlinear Pricing and Resale Price Maintenance," *Rand Journal of Economics* 41 (2010), 139–164.

Burke, J., A. Lewis, and M. Overton, "The Speed of Shor's R-algoritm," *Journal of Numerical Analysis* 28 (2008), 711–720.

Cameron, A., and P. Trivedi, *Microeconometrics: Methods and Applications* (Cambridge: Cambridge University Press, 2005).

Chu, C., "The Effect of Satellite Entry on Cable Television Prices and Product Quality," *Rand Journal of Economics* 41 (2010), 730–764.

Copeland, A., W. Dunn, and G. Hall, "Inventories and the Automobile Market," *Rand Journal of Economics* 42 (2011), 121–149.

Davidson, J., *Econometric Theory* (Malden, MA: Blackwell, 2000).
Davis, P., "Spatial Competition in Retail Markets: Movie Theaters," *RandJournal of Economics* 37 (2006), 964–982.
De Haan, L., "Estimation of the Minimum of a Function Using Order Statistics," *Journal of the American Statistical Association* 76 (1981), 467–469.
Dorsey, R., and W. Mayer, "Genetic Algorithms for Estimation Problems with Multiple Optima, Nondifferentiability, and Other Irregular Features," *Journal of Business and Economic Statistics* 13 (1995), 53–66.
Dube, J., J. Fox, and C. Su, "Improving the Numerical Performance of BLP Static and Dynamic Discrete Choice Random Coefficients Demand Estimation," *Econometrica* 80 (2012), 2231–2267.
Furlong, K., "Quantifying the Benefits of New Products: Hybrid Vehicles," working paper (2012).
Goeree, M., "Limited Information and Advetising in U.S. Personal Computer Industry," *Econometrica* 76 (2008), 1017–1074.
Goffe, W., G. Ferrier, and J. Rogers, "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics* 60 (1994), 65–69.
Goldberg, P., and R. Hellerstein, "A Structural Approach to Identifying the Sources of Local-Currency Price Stability," *Review of Economic Studies* 80 (2013), 175–210.
Goolsbee, A., and A. Petrin, "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica* 72 (2004), 351–381.
Iizuka, T., "Experts Agency Problems: Evidence from the Prescription Drug Market in Japan," *Rand Journal of Economics* 38 (2007), 844–862.
Jiang, R., P. Manchandab, and P. Rossi, "Bayesian Analysis of Random Coefficient Logit Models Using Aggregate Data," *Journal of Econometrics* 149 (2009), 136–148.
Judd, K., *Numerical Methods in Economics* (Cambridge, MA: MIT Press, 1998).
Judd, K., and B. Skrainka, "High Performance Quadrature Rules: How Numerical Integration Affects a Popular Model of Product Differentiation," CEMMAP working paper CWP03/11 (2011).
Kappel, F., and A. Kuntsevich, "An Implementation of Shor's r-Algorithm," *Computational Optimization and Applications* 15 (2000), 193–205.
Lagarias, J., J. Reeds, and M. Wright, "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," *SIAM Journal on Optimization* 9 (1998), 112–147.
McCullough, B., and H. Vinod, "Verifying the Solution from a Nonlinear Solver: A Case Study," American Economic Review 93 (2003), 873–892.
McFadden, D., "Econometric Models of Probabilistic Choice," in C. Manski and D. McFadden (eds.), *In Structural Analysis of Discrete Data* (Cambridge, MA: MIT Press, 1981).
McFadden, D., and W. Newey, "Large Sample Estimation and Hypothesis Testing," in R. Engle and D. McFadden (eds.), *Handbook of Econometrics* Vol. 4 (New York: Elsevier, 1994).
Nakamura, E., and D. Zerom, "Accounting for Incomplete Pass-Through," *Review of Economic Studies* 77 (2010), 1192–1230.
Nevo, A., "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry, University of California," Berkeley Competition Policy Center working paper CPC 99-02 (1997).
———- "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry," *Rand Journal of Economics* 31 (2000a), 395–421.
———- "A Practitioner's Guide to Estimation of Random Coefficients Logit Models of Demand," *Journal of Economics and Management Strategy* 9 (2000b), 513–548.
——— "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica* 69 (2001), 307–342.
——— "New Products, Quality Changes, and Welfare Measures from Estimated Demand Systems," this REVIEW 85 (2003), 266–275.
Nevo, A., and M. Whinston, "Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference," *Journal of Economic Perspectives* 24 (2010), 69–82.
Pakes, A. and D. Pollard, "Simulation and the Asymptotics of Optimization Estimators," *Econometrica* 57 (1989), 1027–1057.
Petrin, A., "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy* 110 (2010), 705–729.
Quandt, R., "Computational Problems and Methods," in Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, vol. 1 (New York: Elsevier, 1983).
Reiss, P., and F. Wolak, "Structural Econometric Modeling: Rationales and Examples from Industrial Organization," in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, vol. 6A (New York: Elsevier, 2007).
Rekkas, M., "The Impact of Campaign Spending on Votes in Multiparty Elections," this REVIEW 89 (2007), 573–585.
Rust, J., "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica* 55 (1987), 999–1033.
Ruud, P., *An Introduction to Classical Econometric Theory* (New York: Oxford University Press, 2000).
Simon, C., and L. Blume, *Mathematics for Economists* (New York: Norton, 1994).
Skrainka, B., "A Large Scale Study of the Small Sample Performance of Random Coefficient Models of Demand," University College London Centre for Microdata Methods and Practice technical report (2011).
Small, K., and H. Rosen, "Applied Welfare Economics with Discrete Choice Models," *Econometrica* 49 (1985), 105–130.
Song, M., "Measuring Consumer Welfare in the CPU Market: An Application of the Pure-Characteristics Demand Model," *Rand Journal of Economics* 38 (2007), 429–446.
Su, C., and K. Judd, "Constrained Optimization Approaches to Estimation of Structural Models," *Econometrica* 80 (2012), 2213–2230.
Torczon, V., "On the Convergence of Pattern Search Algorithms," *SIAM Journal on Optimization* 7 (1997), 1–25.
Veall, M., "Testing for a Global Maximum in An Econometric Context," *Econometrica*, 58 (1990), 1459–1465.
Villas-Boas, S., "Vertical Relationships between Manufacturers and Retailers: Inference with Limited Date," *Review of Economics Studies* 74 (2007), 625–652.
——— "An Empirical Investicgation of the Welfare Effects of Banning Wholesale Price Discrimination," *Rand Journal of Economics* 40 (2009), 20–46.
Waltz, R., and T. D. Plantenga, *KNITRO User's Manual Version 6* (Evanston, IL: Ziena Optimization, 2009).