Demand Estimation with High-Dimensional Product Characteristics*

Benjamin J. Gillen
Department of Economics
California Institute of Technology

Hyungsik Roger Moon
Department of Economics
University of Southern California &
Yonsei University

Matthew Shum
Department of Economics
California Institute of Technology

August 25, 2014

Abstract

Structural models of demand founded on the classic work of Berry, Levinsohn, and Pakes (1995) link variation in aggregate market shares for a product to the influence of product attributes on heterogeneous consumer tastes. We consider implementing these models in settings with complicated products where consumer preferences for product attributes are sparse, that is, where a small proportion of a high-dimensional vector of product characteristics influence consumer tastes. We propose a multi-step estimator to efficiently perform uniform inference. Our estimator employs a penalized pre-estimation model specification stage to consistently estimate nonlinear features of the BLP model. We then perform selection via a Triple-LASSO for explanatory controls, treatment selection controls, and instrument selection. After selecting variables, we use an unpenalized GMM estimator for inference. Monte Carlo simulations verify the performance of these estimators.

*We are grateful to David Brownstone, Martin Burda, Garland Durham, Jeremy Fox, Stefan Holderlein, Ivan Jeliazkov, Dale Poirier, Guillame Weisang, Frank Windmeijer, and seminar participants at the Advances in Econometrics Conference on Bayesian Model Comparison and the California Institute of Technology for helpful comments. We owe special thanks to Alexander Charles Smith for important insights early in developing the project.

1 Introduction

A structural model of behavior in which heterogeneous individuals independently make consumption choices according to their own preferences lays the foundation for most empirical investigations into how a product's market share responds to changes in its price. In markets selling complex products, such as cellular phones or automobiles, an individual's preferences could relate to a high-dimension set of attributes describing these products. Despite this complexity, researchers typically assume that the dominant influences on consumers' decisions can be captured by a relatively small set of "important" characteristics implicitly selected as part of the empirical study. As such, a tractable random-utility model for preferences is predicated by a variable selection exercise that defines which observable attributes should be included and which can be ignored. Indeed, in a truly high-dimensional setting where the dimension of the attributes exceeds the number of product-market observations, such a model selection exercise is necessary simply to identify the estimator in finite samples. In the absence of a principled approach to variable selection, the accuracy of counterfactual inference in post-model selection estimators suffers from inefficiency and can be severely distorted by bias.

In this paper, we consider an automated approach to variable selection for a Berry, Levinsohn, and Pakes (1995) (BLP) GMM estimator. In particular, we investigate how variable selection affects post-selection inference for the expected sensitivity of a product's market share to changes in its price. We frame the inference problem as an exercise in estimating the expectation of an endogenous, heterogeneous, treatment effect in the presence of a high-dimensional set of controls and a high-dimensional set of instruments. We propose a set of assumptions to characterize "sparsity" in this context, which ensures that the true model of preference parsimoniously depends on only a few important attributes. We then adopt standard machine learning approaches to variable selection through penalized regression, leveraging the literature on inference in high-dimensional problems to characterize the properties of our estimator.

Driven by computational gains and algorithmic innovations in operations research by computer scientists, researchers are developing new tools for investigating the role of model selection in inference. One cornerstone of these innovations is the LASSO introduced by Tibshirani (1996), which applies penalty to the absolute magnitude of coefficient magnitudes to the sum of squared residuals in a least squares regression model. Subsequent development lead by Fan and Li (2001) and Fan and Peng (2004), proposed an adaptive penalization strategy that relaxes the penalty for "important" coefficients to prevent the penalty from distorting the asymptotic distribution of estimates for selected variables. This adaptive feature allows an estimator to achieve "oracle" efficiency in which the asymptotic distribution for the selected variables matches that of a true model that imposes zero-restrictions as if they were known a priori. Huang et al. (2008) characterized this property for BRIDGE estimators and Caner (2009) introduced a BRIDGE-style estimator for GMM estimators. Other important extensions for these results beyond linear models are provided by Huang et al. (2010), who consider inference in additive nonparametric settings using a grouped LASSO, and Caner and Zhang (2013), who present an elastic net for high-dimensional models identified by a set of moment conditions. The results from Caner and Zhang (2013) are particularly important in the current application, as they allow us to recover the oracle property in a penalized BLP estimator.

As sparsity-based estimation techniques became more broadly applied, new research led by Leeb and Pötscher (2005) began a deeper investigation into the distributional properties of these estimators, identifying contexts in which inference may be distorted despite the oracle property.[1] With Belloni et al. (2012), econometricians started exploring multi-step techniques to address the uniform convergence problems associated with post-variable selec-

[1] Post-model selection estimators suffer from a failure of uniform convergence in parameter space. Leeb and Pötscher (2005) show that this failure occurs in the regions of parameter space where the parameter's value is local to zero, or of order smaller than $n^{-1/2}$, distorting parametric-rate inference against local alternatives. Though the oracle property leads to estimators that perform well pointwise and can be effectively used to test against non-local alternatives, this distortion indicates a practical limitation of oracle efficient estimators. Indeed, a number of papers, including Leeb and Pötscher (2006) and Leeb and Pötscher (2008), cast doubt on the feasibility of uniformly consistent inference based on the oracle property.

tion inference by adopting a semiparametric perspective focused on characterizing specific treatment effects (Belloni et al. (2013a), Belloni et al. (2013b), Farrell (2013)) or characterizing a finite set of parameteric counterfactuals (Kozbur (2013), Hansen and Kozbur (2013)). These techniques allow uniform inference on the effects of interest not by exploiting the oracle property, but rather by adjusting the model selection algorithm to "immunize" inference to selection. By using multiple stages of selection, the sampling distribution of post-selection hypothesis test statistics are robust to the misspecification errors that are inevitable in finite-samples. In linear models, this robustness is established by showing the approximation error between a mis-specified model and a properly-specified model disappears asymptotically. Though our model is inherently nonlinear, the high dimensional aspects of the problem enter through a linear component. As such, given a consistent first-stage estimator for the model, we can perform the variable selection and analyze its effects on the sampling distribution of parameter estimates in a linear context.

Though our investigation centers on a specific structural model widely adopted in applied work, the asymptotic properties of our estimator are little different from those analyzed by Belloni et al. (2013b) and Farrell (2013). We follow a multi-step procedure to identify a set of variables to control for observable variation in market shares along with a set of "amelioration" controls. This amelioriation set immunizes the estimator, ensuring uniform convergence in the dimensions of the parameter space for which we wish to perform inference. To select these variables within a nonlinear model, we use a penalized GMM pre-estimator for the BLP objective function that consistently recovers mean utility shocks in the model. We then perform the following three linear-LASSO regressions:

- Explanatory Controls: linear-LASSO of mean utilities on product attributes.
- Treatment Controls: linear-LASSO of prices on product attributes.
- Relevant Instruments: linear-LASSO regression of prices on the selected product attributes, cost shifters, and BLP Instuments.

Finally, we include all variables and instruments which have non-zero attributes from these three LASSO regressions in an unpenalized selected BLP GMM estimator. This approach follows the algorithm laid out in Belloni et al. (2013b), but extends their methodology to a nonlinear GMM setting.

We verify the performance in a set Monte Carlo simulations that require us to confront the computational challenges inherent in estimating structural demand models. We employ the MPEC algorithm from Dube et al. (2012) to optimize the BLP GMM estimator's objective function. An appealing feature of the penalized GMM pre-estimator is that the penalty does not complicate the Hessian of the objective function or the gradients of the equilibrium constraints.

The next section will present the Berry et al. (1995) structural model for demand, the regularity restrictions we impose on the data generating process, and our sparsity assumption on consumer preferences. After that, Section 3 presents the Three-Stage, Triple-LASSO estimation strategy we use to attain uniform inference. The results of our simulation study are reported in section 4, before we close with some commentary on possible refinements for our analysis and directions for further research.

2 The Structural Model, Data Generating Process, and Sparsity

We now detail the structural features of the data generating process and formally state the sampling restrictions we impose on the two. We consider the standard Berry et al. (1995) structural model of demand estimation. The regularity conditions we require are fairly standard for GMM estimators and fully compatible with the structural model. The most novel assumption we introduce relates to our application of sparsity in the consumer preference model.

2.1 BLP Structural Model

The data generating process builds from a set of heterogeneous consumers, with an individual indexed by i, who choose among J products, $j = 1, \ldots, J$ and an outside good $j = 0$ over $t = 1, \ldots, T$ markets. Each product is described by a vector of K characteristics x_{jt} and its price p_{jt}. The latent utility for consumer i from consuming product j in market t is assumed to be linear:

$$u_{ijt} = x'_{jt}\beta_x + p_{jt}\beta_{ip} + \xi_{jt} + \epsilon_{ijt} \tag{1}$$

Implicit in this formulation of utility, we restrict consumers to share uniform tastes for product attributes (β_x) but allow for their sensitivity to prices are be heterogeneous, with $\beta_{ip} \sim N\left(\beta_{0p}, \sigma_p^2\right)$. As usual, the utility formulation also allows for exogenous product-market specific shocks (ξ_{jt}) that provide identifying variation from aggregate market shares. Individual utility's unobserved idiosyncratic component (ϵ_{ijt}) takes the usual Type-I Extreme Value distribution, leading to the logit choice probabilities:

$$Pr\left\{y_{ijt} = j\right\} = \frac{\exp\left\{x'_{jt}\beta_x + p_{jt}\beta_{ip} + \xi_{jt}\right\}}{1 + \sum_{r=1}^{J}\exp\left\{x'_{rt}\beta_x + p_{rt}\beta_{ip} + \xi_{rt}\right\}} \tag{2}$$

To aggregate individual choices into total market shares, we integrate over the individual heterogeneity to translate individual choice probabilities into expected market shares for product j in market t:

$$s_{jt} = \int \frac{\exp\left\{x'_{jt}\beta_x + p_{jt}b + \xi_{jt}\right\}}{1 + \sum_{r=1}^{J}\exp\left\{x'_{rt}\beta_x + p_{rt}b + \xi_{rt}\right\}} dF_{\beta_{ip}}\left(b; \beta_{0p}, \sigma_p^2\right) \tag{3}$$

In practice, the integral is typically calculated using pseudo Monte-Carlo methods, however, naïve approaches to numerical integration can be inefficient. As illustrated in applications that approximating complex likelihoods by Heiss and Winschel (2008) and Heiss (2010), Sparse Grid integration can improve both the computation time and the accuracy of numerical integrals, especially in high-dimensional problems. Recent work by Skrainka and Judd (2011) implements Sparse Grid integration for computing the integral in equation 3, greatly reducing processing time for estimating the model's objective function.

Identifying the parameters of the model, $\theta = \{\beta_x, \beta_{0p}, \sigma_p\}$ requires linking observed market shares to the consumer utility model. A central contribution in Berry et al. (1995) illustrates that, for a given value of θ, a contraction mapping can iteratively solve for the latent product-market shocks ($\xi_{jt}(\theta; X, Z, p, s)$), which can then characterize the mean utility,

$$\delta_{jt}(\theta; X, Z, p, s) = x'_{jt}\beta_x + p_{jt}\beta_{0p} + \xi_{jt}(\theta; X, Z, p, s) \tag{4}$$

Having recovered these mean utility shocks, we can then utilize a set of L exogenous instruments, denoted by z_{jt}, to identify the parameters via the exclusion restriction $E[\xi_{jt}|z_{jt}] = 0$. We let the matrices X, p, s, ξ, δ, and Z collect all observations for the characteristics, price, market shares, utility shocks, mean utilities, and instruments, respectively. We define W as an $L \times L$ weighting matrix and then follow the usual approach of interacting instruments with residuals to define the standard GMM loss function.

$$Q(\theta) = \xi(\theta; X, Z, p, s)' ZWZ'\xi(\theta; X, Z, p, s) \tag{5}$$

The instruments themselves are typically taken to include product attributes, observable cost-shifters for producers (typically embedded in a Bertrand competition model), and the attributes of other products competing in a given market. In a classic reference, Nevo (2000) recommends setting the weighting matrix, W, equal to the inverse of the variance-covariance matrix of the instruments, a common implementation. We will return to discuss instrument selection and weighting after defining sparsity for the model in the next section.

2.2 Sampling Assumptions for the Data Generating Process

We begin by presenting a set of concrete sampling assumptions that are sufficient to ensure the penalized GMM estimator we use in the first stage regression provides a consistent estimator for the parameters in the model. The notation and assumptions here are essentially presenting a simplified set of those maintained in Caner and Zhang (2013), with some adaptations to accommodate traditional variable definitions in the BLP model.

Assumption 1 (Sampling Properties for DGP)

1. *$K/T \to 0$ as $T \to \infty$. $\theta_0 \in \Theta_K$, Θ_K is a compact subset of R^K, with a compact limit set $\Theta_\infty := \lim_{T\to\infty} \Theta_K$.*

2. *Our moment conditions apply to the $L \geq K$ instruments, growing so that $L/T \to 0$:*

$$E\left[g_{ljt}(\theta)\right] = E\left[z_{ljt}\xi_{jt}(\theta)\right], l = 1, \ldots, L$$

Each moment condition $g_{ljt}(\theta)$ is continuously differentiable in θ and the vector of moment functions for a given product-market is denoted $g_{jt}(\theta)$.

3. *The following uniform law of large numbers as the number of markets T becomes large:*

$$\sup_{p\in\{1,\ldots,K\}} \sup_{\theta\in\Theta_p} \left[\frac{1}{JT}\sum_{j,t=1}^{J,T} \left|g_{jt}(\theta) - Eg_{jt}(\theta)\right|\right] \to_p 0$$

4. *Define the population moment $m_{lT}(\theta) = E\left[\frac{1}{JT}\sum_{j,t=1}^{J,T} g_{ljt}(\theta)\right]$, then:*

   (a) *$m_{lT}(\theta) \to m_l(\theta)$ uniformly in θ and K,*

   (b) *$m_{lT}(\theta)$ is continuously differentiable and $m_l(\theta)$ is continuous in θ, and,*

   (c) *$m_{lT}(\theta_0) = 0$ and $m_{lT}(\theta) \neq 0, \forall\theta \neq \theta_0$.*

5. *Define the $L \times K$ matrix $\hat{G}_n(\theta) = \frac{\sum \partial g_{jt}(\theta)}{\partial\theta'}$.*

   (a) *As T becomes large, the following uniform law of large numbers holds uniformly in K and a neighborhood $\mathcal{N}$ of θ_0*

$$\frac{\hat{G}_T(\theta)}{T} \to_p G(\theta)$$

   (b) *$G(\theta)$ is continuous in θ, with $G(\theta_0)$ having full column rank K.*

6. *W_T is a positive definite matrix and $\|W_T - W\|_2^2 \to_p 0$, with W also a positive definite and finite matrix.*

7. *The expected value of the outer product of the score is asymptotically defined:*

$$\left\|\Omega^{-1} - \left(\lim_{T\to\infty}(JT)^{-1}\sum_{j,t=1}^{J,T} E\left[g_{jt}(\theta_0)\, g_{jt}(\theta_0)'\right]\right)^{-1}\right\|_2^2 \to 0$$

8. *The minimal and maximal eigenvalues of* $\Sigma = G(\beta_0)' \Omega^{-1} G(\beta_0)$, *denoted* $\underline{e}$ *and* $\overline{e}$, *are bounded between finite constants* $0 < b \leq \underline{e} \leq \overline{e} \leq B < \infty$.

These assumptions are fairly standard in the literature on GMM estimation with some accommodation to reflect the growing number of parameters K and the growing number of instruments L. Importantly for our study, these assumptions allow us to apply Caner and Zhang (2013)'s Consistency Theorem 1 to the penalized GMM estimator we propose in the next section. Caner and Zhang (2013) provide additional discussion of these assumptions and their relation to Newey and Windmeijer (2009)'s investigation of weak instruments and Zou and Zhang (2009)'s study of the adaptive elastic net.

2.3 Sparsity in the Structural Model

With a sufficiently large dimensional set of product attributes, in particular if K is greater than the number of observations JT, the parameter vector θ will not be identified in finite samples. Sparsity provides a reasonable approach to this problem by assuming all but a relatively small number of parameters are restricted to be zero. We follow the customary investigation of the asymptotic distribution for our estimator as the number of markets T becomes large.[2] As such, we are generally considering the number of competing products, J, to be fixed, and the number of markets $T \to \infty$. To reflect the influence of the high-dimensional feature space, we recall Assumption 1.1, which lets the number of characteristics $K \to \infty$, while requiring the ratio $K/T \to 0$. However, we require further restrictions on the number of characteristics assigned that are assigned non-zero influence in consumer utilities:

Assumption 2 (Sparse Consumer Preferences over Product Characteristics) *Consumers' preferences are sparse in the set of product characteristics. That is:*

[2] In a novel analysis of highly competitive markets, Armstrong (2012) characterizes the effect of a large number of products in the market place. As our interest centers on learning about demand behavior for a small number of complex products, the traditional many-market asymptotics provide the relevant limiting behavior for our estimator.

a. *All but* $k_T = o\left(T^{1/4}\right)$ *of the entries in* β_x *are restricted to be equal to zero*

b. *The remaining* k_T *coefficients' true absolute values are all of an order greater than* $T^{-1/2}$.

This formulation of sparsity can be referred to as exact sparsity to indicate that the true data generating process satisfies the zero restrictions exactly. Caner and Zhang (2013)'s consistency proof for the elastic net relies on the exact sparsity assumption, and so we maintain it in its present form.[3] This form of sparsity stands in contrast to the notions of approximate sparsity utilized by Belloni et al. (2012) and Belloni et al. (2013a). The approximate sparsity formulation of the model is naturally motivated as a semi-parametric approximation of the relevant conditional expectations, which we intend to consider the extension to approximate sparsity in future work. In an applied context, the exact sparsity condition is, admittedly, somewhat stronger than we would like from a structural modeling perspective. Exact Sparsity also appears strictly stronger than necessary for our estimation strategy, as the exact sparsity restriction is required only for consistency in the first-stage regression. The inferential properties of the second-stage estimator are established indirectly using robustness through double-selection and optimal instruments rather than a direct application of the oracle property.

Note that we do not assume the econometrician knows which attributes have zero coefficients. Instead, the objective of the variable selection step is to identify these attributes and impose the correct zero restrictions on the irrelevant attributes. If the post-variable selection estimator almost surely selects the right zero restrictions, then it has the oracle property and

[3]Caner and Zhang (2013) present an adaptive elastic net estimatior in which the number of non-zero coefficients can be allowed to grow at a rate of $T^{1/3}$, so assumption 2.(a) is slightly stronger than it needs to be. However, the sharper 1/4-th root rate of growth in the non-zero parameter vector is consistent with a non-adaptive penalty allowing us to retain consistency in a first-step estimator. Assumption 2.(b) is often made weaker or stated in terms of more primitive objects relating to the penalty of the estimator. However, the presentation of assumption 2.(b) here highlights the importance of the discontinuity that allows us to exclude those variables whose effects are of order less than $T^{-1/2}$.

its asymptotic distribution will be the same as that of an estimator for a model where the econometrician knows the correct zero restrictions a priori.

Given sparsity in consumer preferences, the oracle property is an appealing objective for efficient estimation. However, the variable selection step itself may risk introducing an omitted variable bias by excluding features of the model whose effect on utility may correlate with variation in price. One of the key intuitions on high-dimensional inference for treatment effects in Belloni et al. (2013a) and Belloni et al. (2012) addresses this potential omitted variable bias through an "amelioration" set of variables that augments the controls selected to explain variation in the outcome with additional controls that annihilate endogenous variation in the treatment variable. Belloni et al. (2013b) link this amelioration set to the optimal instruments in a Neyman (1959) c(α) test of composite hypotheses. As our interest centers on the response of market shares to manipulable changes in price, our objective is to immunize our inference on outcomes with respect to variation in price. As such, we focus on a multi-step selection algorithm that selects not only product attributes that explain market share, but also attributes that could be driving causal variation in relative product prices.

Even outside of the sparse setting, the large number of instruments reflected in competing product attributes introduces a many-moment bias to the GMM estimator commonly encountered in dynamic panel models as highlighted in Newey and Windmeijer (2009). One approach to address the many moment issue would be to adopt a continuous updating GMM estimator, for example, with weak instruments Hansen and Kozbur (2013) propose a JIVE estimator for high-dimensional GMM problems. Another approach follows Reynaert and Verboven (2013), who introduce a two-step estimator that uses a consistent pre-estimate for the model to calculate the Chamberlain (1987) optimal instruments for the BLP problem. The Reynaert and Verboven (2013) approach is particularly well-suited to the current application, as the first-step estimator we use for variable selection can also be used to characterize the optimal instruments for those selected variables.

3 The Triple LASSO Estimator

Having established the structural model, sampling properties, and sparsity assumed above, we now describe the algorithm we use to compute our estimator in detail.

3.1 Penalized First-Stage GMM Estimator

Given the assumptions above, a penalized GMM estimator for the BLP objective function can provide a set of consistent estimates for the latent features of the structural model. Denoting the L1 norm of a vector by $\|x\|_1$ we simply perturb the objective function and define the first-stage parameter estimates, implied utility shocks, and mean utilities, respectively as:

$$\begin{aligned}
\tilde{\theta} &= \arg\min_{\{\theta\in\Theta\}} \xi\left(\theta; X, Z, p, s\right)' ZWZ' \xi\left(\theta; X, Z, p, s\right) + \lambda\|\theta\|_1 \\
\tilde{\xi} &= \xi\left(\tilde{\theta}; X, Z, p, s\right) \\
\tilde{\delta} &= X\tilde{\beta}_x + p\tilde{\beta}_{0p} + \tilde{\xi}
\end{aligned} \tag{6}$$

If the penalty term λ goes to zero at rate $1/T$, this penalized GMM estimator will consistently select the important features of the model. We assume this in the following:

Assumption 3 (Asymptotically Diminishing Penalty in First Stage Regression) *The penalty term in the first-stage estimator is asymptotically negligible. In particular,* $\lambda/T \to 0$

As the penalty is not adaptive, i.e., the penalty is not tuned to each parameter based on a pre-pre-estimation step, the resulting estimates for θ will not have the oracle property. In particular, the penalty will lead estimates for the selected attributes to be shrunk towards zero. Nonetheless, the recovered utility shocks $\xi\left(\theta; s\right)$ and mean utilities $\delta\left(\theta; s\right)$ from the estimator retain the parametric rates of consistency necessary for the next selection step.

The L1 LASSO penalty is often coupled with an L2 Ridge penalty term in the elastic net to improve the stability of parameter estimates in the presence of collinearity. We opted for a simpler penalty in this application, as a quadratic penalty for the coefficients would complicate the Hessian for the objective function. By simply adopting the LASSO penalty, the Hessian remains unchanged and the Jacobian is only slightly modified, so the Dube et al. (2012) algorithm remains well-adapted to the problem. The drawback to this approach, however, is that it will require a second step where we estimate a post-selection unpenalized estimator excluding those variables that are not selected in the utility model. We leave the development of a single-step, oracle efficient estimator as a promising candidate for future work.

Nevo (2000) proposes using the inverse of the cross-product of the instruments for the weighting matrix, so $W_T = (Z'Z)^{-1}$. With a large number of instruments in finite-samples, this matrix may be poorly scaled since $Z'Z$ will be rank-deficient. For this reason, we apply an asymptotically negligible regularization the weighting matrix so that $W_T = (Z'Z + \psi I)^{-1}$. A more formal regularization approach could follow Carrasco (2012)'s proposed regularization for the many instruments problem.

3.2 A Triple-LASSO for Attribute and Instrument Selection

We now use the results from the first-stage GMM estimator to identify which variables and instruments to include in a second-stage, unpenalized GMM estimator. In an analogy to standard implementations for Feasible Generalized Least Squares estimators, the consistency of $\tilde{\delta}$ and $\tilde{\xi}$ allows us to perform selection in a relatively simple, linear, auxiliary regression. As such, we characterize and implement this selection step using the LASSO for linear models, though an alternative consistent model selection device would yield similar sampling properties.

In the LASSO formulation, each of these estimators includes some form of L1 penalty denoted by $\lambda_{\{.\}}$ that is allowed to grow with the sample size, but we need impose restrictions so that it has a diminishing impact relative to the sample size asympotically.

Assumption 4 (Asymptotically Diminishing Penalty in Selection)
The penalty terms incorporated in the auxiliary selection estimators are all asymptotically negligible. In particular:

1. $\lambda_x/T \to 0$ *2.* $\lambda_p/T \to 0$ *3.* $\lambda_{zp}/T \to 0$

These penalties present tuning parameters that are difficult to characterize beyond their magnitude relative to the information contained in the sample to ensure that the penalty retains influence relative to sampling error. In our simulation exercises, we selected this value using a block cross-validation technique, which performed well.

3.2.1 Explanatory Control Variables

A primary consideration in selecting attributes for demand estimation is to include those features that drive mean utility. We identify these attributes by a LASSO regression of the recovered mean utilities, $\tilde{\delta}$, on product attributes.

$$\hat{\gamma}_x = \arg\min \sum_{jt} \left(\tilde{\delta}_{jt} - x'_{jt}\gamma_x\right)^2 - \lambda_x \|\gamma_x\|_1 \tag{7}$$

Note that, even without the penalty term, the LASSO regression would asymptotically estimate a zero parameter for those characteristics that are irrelevant to demand estimation.

3.2.2 Treatment-Selection Control Variables

In addition to those attributes which explain mean utilities, we also include a set of variables to control for attribute-driven variation in price. These treatment-selected attributes are identified by the penalized LASSO regression:

$$\hat{\gamma}_p = \arg\min \sum_{jt} \left(p_{jt} - x'_{jt}\gamma_p\right)^2 - \lambda_p \|\gamma_p\|_1 \tag{8}$$

This selection step incorporates the double LASSO proposed in Belloni et al. (2013a). As noted in Assumption 4.2, the penalty term λ_p becomes small relative to the sample size and selection is consistent.

3.2.3 The Optimal Instrument Attributes

The Chamberlain (1987) optimal instruments for the BLP estimator are discussed in Reynaert and Verboven (2013). As is usual in IV estimators, the optimal instruments for the exogenous control variables are the control variables themselves and the optimal instrument for the endogenous variable is its exogenous conditional expectation. To identify this set of instruments, we include the selected attributes identified in the previous two sections, along with a set of instruments for explaining variation in price and its impact on utility.

$$\hat{\zeta}_p = \arg\min \sum_{jt} \left(p_{jt} - z'_{jt}\zeta_p\right)^2 - \lambda_{zp}\|\zeta_p\|_1 \tag{9}$$

We could further refine our instruments by following Reynaert and Verboven (2013)'s characterization of the optimal instrument for the variance in consumers' response to price as the expected derivative of consumer utility with respect to σ_p. We conjecture that the optimal instrument for σ_p could be constructed from the first-stage penalized regression estimates, but have yet to establish this result formally. This extension is not trivial, as it requires additional smoothness assumptions to ensure the expectation of these derivatives can be consistently estimated by our first-stage model.

3.3 The Unpenalized Post-Triple Selection Estimator

Our final estimates for the parameters in the demand system come from an unpenalized GMM estimator of the BLP model using the selected attributes and instruments described above. The attributes associated with the parameters whose estimates in either of the two control selection regressions 7 and 8 identify the set of product attributes to include in the demand model. These attributes, along with the instruments that have non-zero coefficients in regression 9, then form the set of selected instruments with which to identify that model.

Define X^* and Z^* as the respective matrices of selected controls and instruments and let $\Theta^* \subset \Theta$ as the subset of parameter space where the unselected attributes' corresponding coefficients are fixed at zero. Then then Post-Triple LASSO BLP estimator is given by:

$$\theta^* = \arg\min_{\theta\in\Theta^*} Q^*(\theta) \tag{10}$$

$$Q^*(\theta) = \xi\left(\theta; X^*, Z^*, p, s\right)' Z^* W Z^{*\prime} \xi\left(\theta; X^*, Z^*, p, s\right)$$

The post-triple LASSO BLP estimator satisfies the conditions for uniformly consistent inference on β_{0p} established in Belloni et al. (2013b). Uniform consistency and uniform asymptotic normality for σ_0 is not assured, however, as inference could be affected by an Andrews (2001) problem in a region of parameter space near the boundary value of $\sigma_0 = 0$.

4 Monte Carlo Simulations

We verify the proposed estimator's performance in a number of simulation experiments. Here, we describe our simulation experiments, present the set of estimators we evaluate, and characterize the finite-sample distributions of the estimates for the price-related parameters β_{0p} and σ_p^2.

4.1 The Simulated Data Generating Process

To describe the data generating process used in our simulations, we begin by generating product attributes that are jointly normally distributed with a fixed covariance matrix as follows:

$$X_{jt} \sim N\left(0, \Sigma_X\right); \quad \Sigma_{X,\{i,j\}} = 0.5^{|i-j|}/16 \tag{11}$$

The scaling of the covariance matrix allows the variation in utilities due to product attributes in our utility model to roughly match the scale of variation in prices. In our simulations, only the first four attributes of X_{jt} are important to the model. All other attributes vary independently of consumer demand behavior and variation in price. A representative consumer's utility is then given by:

$$\begin{aligned} u_{ijt} &= [4, 4, 2, 2, 0, \ldots, 0] x_{jt} - \beta_{ip} p_{jt} + \xi_{jt} + \epsilon_{ijt} \\ \epsilon_{ijt} &\sim \text{Type I Extreme Value} \\ \xi_{jt} &\sim N\left(0, 1/4\right); \ \beta_{ip} \sim N\left(5, 0.1\right) \end{aligned} \tag{12}$$

Price is endogenous to the system, correlated with the important elements of X_{jt} and ξ_{jt}, but its effect is identified by some exogenous variation in ν_{jt} that is identifiable by its projection onto a single exogenous instrument z_{jt}.

$$p_{jt} = \xi_{jt} + [1, 1, 0, 0, 0, \ldots, 0]x_{jt} + \nu_{jt};\ \nu_{jt} \sim U\,[0, 1] \tag{13}$$
$$z_{jt} = \nu_{jt} + [1, 1, 1, 1, 0, \ldots, 0]x_{jt} + \eta_{jt};\ \eta_{jt} \sim N\,(0, 1)$$

We assume all shocks in all equations are i.i.d. across products and markets. We simulate 20 markets with 10 products, making an expected number of 100 observations, and evaluate the estimators with $K \in \{5, 20, 50, 100, 200\}$.

4.2 The Fitted Estimators

As a benchmark, we evaluate the sampling distribution of the post-Triple LASSO BLP estimator relative to that of an estimator that drops all variables with a zero coefficient from the model. The Oracle Benchmark Estimator is estimated using an unpenalized BLP estimator optimized using the MPEC algorithm imposing the zero restrictions on attributes as if they were known a priori. By exploiting this information to maximum effect for perfect variable selection, the Oracle Benchmark Estimator provides the best possible estimate in the model.

A standard, Unpenalized BLP estimator provides another reference point for our proposed estimator's performance. In practice, this estimator is only identified for finite samples when $K < T$, so we can't estimate it for all of our simulation specifications. However, this standard provides a useful device for identifying the loss of efficiency by not exploiting sparsity in those samples for which it is identified.

We also consider two estimators where model selection is not done robustly. The first non-robust estimator comes from the output of our first-stage LASSO-Penalized GMM Estimator. This estimator will not have the oracle property, as the penalty is shrinking its coefficients toward zero and, in so doing, introducing a finite-sample bias. However, we'd also expect a failure of uniform inference, characterized by a breakdown in the normality of the estimator's sampling distribution.

The second non-robust estimator performs only a single-round of attribute selection, selecting the model that best explains mean utility. We call this estimator the Post-Single LASSO BLP estimator. This estimator will also suffer from a failure of uniformly consistent inference, delivering a non-normal sampling distribution for the parameter estimates.

4.3 The Sample Distribution of Estimators

We consider the simulated mean, bias, standard error, and root mean square error for each of the five estimators from the previous subsection as reported in Table 1. We find the performance lines up nicely with the theoretical properties. For low-dimensional problems, the unpenalized GMM estimator, penalized GMM estimator, and the Post-Triple LASSO estimator all perform quite well, delivering estimates that compare favorably to the Oracle estimator. In these cases, the only striking result comes from the Post-Single LASSO estimator, which is badly distorted by a selection-induced omitted variable bias. As the number of nuisance attributes becomes large, the importance of consistent selection and an unpenalized second-stage regression become apparent. After $K = 100$, the unpenalized GMM estimator is not identified in finite samples, giving rise to highly unstable parameter estimates. Further, the penalized GMM estimator begins to suffer from its bias towards zero, even though its root mean square error is still quite reasonable until $K = 200$. Importantly, the Post-Triple LASSO estimator's performance doesn't deteriorate that badly relative to the Oracle Benchmark even when the number of parameters is twice the expected number of observations.

Next, we evaluate the LASSO selection procedure itself to ensure the estimator is including the correct attributes in the model and check the frequency with which incorrect attributes are included. Table 2 reports these statistics, illustrating that the Penalized GMM and Post-Triple LASSO estimators effectively include the important attributes in the model. It also suggests the presence of an omitted variable bias in the Post-Single LASSO estimator, which regularly omits at least one important variable. We also see that the Triple-LASSO is fairly effective at filtering out irrelevant attributes, including on average only 1-2 additional attributes beyond the important controls. Finally, we note that, although the Penalized

Table 1: Sampling Properties of Estimated Price Parameters

The table reports simulated sampling Bias, Standard Error, and Root Mean Square Error for five different approaches to model selection for demand estimation. The data generating process is presented in section 4.1 and the presented results are computed from a set of 500 simulation runs for $J = 10$ products competing in $T = 20$ markets, giving an expected count of 100 observations. The Oracle Benchmark provides the best infeasible estimator, based on perfect model selection a priori. The Unpenalized GMM estimator is the usual BLP estimator based on equation 5. The Penalized GMM estimator comes from the first-stage LASSO-BLP estimation characterized by equation 6. The Post-Single LASSO estimator includes only those variables selected via 7. The Post-Triple LASSO uses our full algorithm of triple-selection followed by unpenalized estimation of the selected model, as characterized by equation 10.

	Simulated Bias for β_{0p}					Simulated Bias for σ_0				
	$K = 5$	$K = 20$	$K = 50$	$K = 100$	$K = 200$	$K = 5$	$K = 20$	$K = 50$	$K = 100$	$K = 200$
Oracle Benchmark	0.00	0.04	(0.07)	0.06	0.16	0.05	0.05	0.06	0.08	0.08
Unpenalized GMM	0.03	0.12	0.08	-	-	0.00	0.00	0.00	-	-
Penalized GMM	0.04	0.12	0.08	0.41	1.18	0.04	0.04	0.04	0.02	0.03
Post-Single LASSO	0.65	0.51	0.32	0.18	0.93	0.07	0.07	0.08	0.08	0.08
Post-Triple LASSO	0.06	0.12	(0.04)	0.07	0.12	0.05	0.05	0.06	0.08	0.08

	Simulated Standard Error for β_{0p}					Simulated Standard Error for σ_0				
	$K = 5$	$K = 20$	$K = 50$	$K = 100$	$K = 200$	$K = 5$	$K = 20$	$K = 50$	$K = 100$	$K = 200$
Oracle Benchmark	0.46	0.42	0.65	0.29	0.65	0.04	0.04	0.03	0.02	0.01
Unpenalized GMM	0.68	0.33	0.93	-	-	0.00	0.00	0.00	-	-
Penalized GMM	0.68	0.33	0.90	0.44	0.62	0.04	0.04	0.04	0.08	0.06
Post-Single LASSO	1.68	1.20	1.21	1.84	1.98	0.02	0.02	0.02	0.01	0.00
Post-Triple LASSO	0.61	0.83	0.53	0.62	0.71	0.04	0.04	0.03	0.02	0.01

	Simulated Mean Square Error for β_{0p}					Simulated Mean Square Error for σ_0				
	$K = 5$	$K = 20$	$K = 50$	$K = 100$	$K = 200$	$K = 5$	$K = 20$	$K = 50$	$K = 100$	$K = 200$
Oracle Benchmark	0.46	0.42	0.65	0.29	0.67	0.06	0.06	0.07	0.08	0.08
Unpenalized GMM	0.68	0.35	0.93	-	-	0.00	0.00	0.00	-	-
Penalized GMM	0.69	0.35	0.90	0.60	1.33	0.06	0.06	0.05	0.08	0.07
Post-Single LASSO	1.80	1.30	1.25	1.85	2.19	0.08	0.08	0.08	0.08	0.08
Post-Triple LASSO	0.61	0.84	0.54	0.62	0.72	0.06	0.06	0.07	0.08	0.08

Table 2: Model Selection Accuracy and Efficiency

The table reports on the effectiveness of our proposed model selection algorithms. Panel A reports the frequency with which the model selected the four "important" attributes assigned non-zero coefficients in the simulated data generating process. Panel B reports the average number of additional, "null" attributes are included in the selected model. The data generating process and estimators match those reported in Table 1.

Panel A: Frequency of Selecting Important Attributes

	$K = 5$	$K = 20$	$K = 50$	$K = 100$	$K = 200$
Oracle Benchmark	100%	100%	100%	100%	100%
Unpenalized GMM	100%	100%	100%	100%	100%
Penalized GMM	100%	100%	100%	100%	100%
Post-Single LASSO	75%	75%	75%	75%	39%
Post-Triple LASSO	100%	100%	100%	100%	88%

Panel B: Average Number of Null Attributes Selected

	$K = 5$	$K = 20$	$K = 50$	$K = 100$	$K = 200$
Oracle Benchmark	-	-	-	-	-
Unpenalized GMM	3.76	17.78	46.75	97.15	198.87
Penalized GMM	3.76	17.77	46.73	93.81	192.13
Post-Single LASSO	0.02	0.01	0.02	0.03	1.09
Post-Triple LASSO	0.86	1.15	1.25	1.34	2.24

GMM estimator is influencing parameter values, it does assign a number of null attributes significant coefficients that are not selected by the Triple-LASSO algorithm. This lack of sharp selection suggests a call for more aggressive penalization, but this enhanced penalty would further bias the estimates. We conjecture that these issues could be jointly addressed by deriving an adaptively penalized estimator for this problem, a possibility we leave for further research.

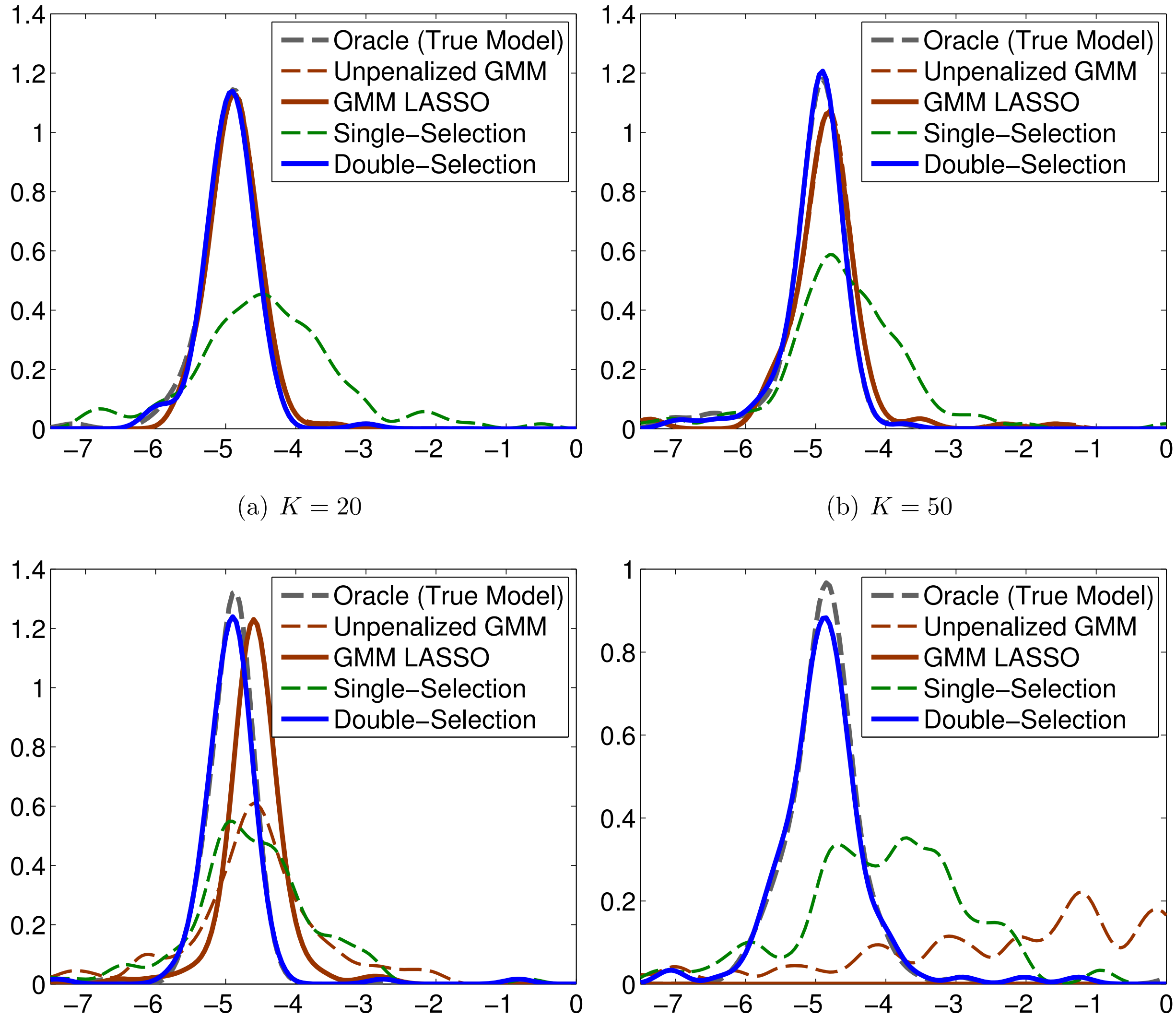
Finally, we consider the full sampling distribution of the price-related parameters. Figure 1 presents kernel density plots of the estimates for β_{0p} for the proposed estimators. We again see the pattern indicated in table 1. In a low-dimensional context, all the estimators except the variable-omitting Post-Single LASSO estimator perform relatively well. As the number of null features grows, the bias becomes more severe for the penalized GMM estimator because the penalty term grows relative to the magnitude of the GMM objective function. We also see the unpenalized GMM estimator begin to lose efficiency with a relatively large number of parameters. However, even with a high dimensional parameter space, the Post-Triple LASSO estimator performs comparably to the Oracle benchmark.

5 Today's Conclusions and Tomorrow's Directions

Our analysis provides a proposed method for selecting attributes from a high-dimensional feature space for estimating consumer demand models. We adopt sufficiently strong conditions to show that our estimator achieves uniform, $\sqrt{n}$-consistent inference. Further, we have verified this performance and the importance of robust selection in our simulation study.

Though we illustrated the estimator properties through simulation, our results have implications for any exercise in estimating demand models. The techniques are directly applicable to analyzing complex products with many features, such as cellular phones, hybrid automobiles, or home theater systems. Additionally, these methods can be applied to settings where we may not know the shape of the utility function and wish to use a sieve to approximate a non-linear specification of individual utility. The results confirm the intuition that controls in a well-designed empirical studies of market demand ought to include those attributes with a substantial influence on either individual choices or prices across products and should ignore other attributes as irrelevant. Further, the results regarding instrument selection for identifying the price elasticity of demand must satisfy some minimal degree of relevance, otherwise it's better to drop those instruments that are too weak.

Though our estimator has appealing finite-sample properties, the conditions under which it was derived are stronger than strictly necessary. A natural next step would more closely inspect the features of our first-stage estimator that may allow for a more efficient estimation strategy. We are particularly interested in an adaptive penalization strategy that controls the bias introduced by the LASSO estimator by penalizing each parameter based on the results of a first-stage estimator. This approach would extend the adaptive elastic net proposed by Caner and Zhang (2013), however it would require extending their results to a setting with many irrelevant instruments. There are some promising innovations in the work by Fan



(a) $K = 20$ (b) $K = 50$

(c) $K = 100$ (d) $K = 200$

Figure 1: Distribution of Simulated Price Coefficient β_{0p}

This figure plots kernel density estimates of the simulated sampling distribution for our demand estimation models as we increase the number of irrelevant attributes in the model. The data generating process and estimators match those reported in Table 1.

and Liao (2012), Hansen and Kozbur (2013), and Cheng and Liao (2013) for models with an exponential number of irrelevant instruments that could provide theoretical foundations for the exercise. This exercise may be more amenable to an empirical likelihood approach, following the specification of Conlon (2013) for the demand estimation problem and the asymptotic analysis of Tang and Leng (2010).

Another potential refinement of our approach would consider directly modeling the optimal instruments for the random coefficients. As noted by Belloni et al. (2013b), there is a tight link between the robustness properties of a Post-Double LASSO and the optimal instruments identifying hypothesis tests for the relevant parameters. Using our first-stage regression results, we could apply the algorithms discussed in Reynaert and Verboven (2013) to include only the optimal instruments for our problem rather than the full set of instruments selected in the third iteration of the LASSO.

Finally, it would be interesting to explore the feasibility of selecting the attributes for which consumers display heterogeneous preferences. In our current exercise, the coefficients for product attributes were fixed in the population and the price effect was known a priori to be heterogeneous. Can we effectively test for multiple random coefficients and identify which attributes should be assigned these random coefficients? How does selection interact with unobserved heterogeneous effects? Evaluating this question is complicated by the non-negativity of the variance parameter for the random coefficients, which introduces an Andrews (2001) problem.

At a foundational level, we would also like to motivate the sparsity assumption from a model of consumer behavior featuring either limited attention or satisficing. This structural foundation could be more readily established in the context of an approximate formulation of sparsity. That is, even though consumers focus on only a few attributes, we would rather not assume they jointly coordinate on which attributes to consider. It would be more natural to directly model the approximation error induced by model selection to the consumers' true utility representation. Such a formulation would also lead to a model of consumer demand with complex products that more closely reflects the empirical environment.

References

Donald W. K. Andrews. Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69(3):683–734, 2001.

Timothy Armstrong. Large market asymptotics for differentiated product demand estimators with economic models of supply. Technical report, Working Paper, 2012.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 2013a.

Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Honest confidence regions for logistic regression with a large number of controls. *arXiv preprint arXiv:1304.3969*, 2013b.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995. ISSN 00129682. doi: 10.2307/2171802. URL http://dx.doi.org/10.2307/2171802.

Mehmet Caner. Lasso-type gmm estimator. *Econometric Theory*, 25(01):270–290, 2009.

Mehmet Caner and Hao Helen Zhang. Adaptive elastic net for generalized methods of moments. *Journal of Business & Economic Statistics*, (just-accepted), 2013.

Marine Carrasco. A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398, 2012.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

Xu Cheng and Zhipeng Liao. Select the valid and relevant moments: An information based lasso for gmm with many moments. 2013.

Christopher T Conlon. The empirical likelihood mpec approach to demand estimation. *Available at SSRN*, 2013.

Jean-Pierre Dube, Jeremy Fox, and Che-Lin Su. Improving the Numerical Performance of BLP Static and Dynamic Discrete Choice Random Coefficients Demand Estimation. *Econometrica*, 2012. forthcoming.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Yuan Liao. Endogeneity in ultrahigh dimension. *Working Paper*, 2012.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *arXiv preprint arXiv:1309.4686*, 2013.

Christian Hansen and Damian Kozbur. Instrumental variables estimation with many weak instruments using regularized jive. *Working Paper*, 2013.

Florian Heiss. The panel probit model: Adaptive integration on sparse grids. In William Greene and R. Carter Hill, editors, *Maximum Simulated Likelihood Methods and Applications*, volume 26 of *Advances in Econometrics*, chapter 6, pages 41–64. Emerald Group Publishing Limited, 2010.

Florian Heiss and Viktor Winschel. Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, 144:62–80, 2008.

Jian Huang, Joel L Horowitz, and Shuangge Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, 36(2):587–613, 2008.

Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282, 2010.

Damian Kozbur. Inference in additively separable models with a high dimensional component. *Working Paper*, 2013.

Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01):21–59, 2005.

Hannes Leeb and Benedikt M Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, pages 2554–2591, 2006.

Hannes Leeb and Benedikt M Pötscher. Sparse estimators and the oracle property, or the return of hodges estimator. *Journal of Econometrics*, 142(1):201–211, 2008.

Aviv Nevo. A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy*, 9(4):513–548, 2000.

Whitney K Newey and Frank Windmeijer. Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719, 2009.

Jerzy Neyman. Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics*, 57:213, 1959.

Mathias Reynaert and Frank Verboven. Improving the performance of random coefficients demand models: the role of optimal instruments. *Journal of Econometrics*, 2013.

Benjamin S Skrainka and Kenneth L Judd. High performance quadrature rules: How numerical integration affects a popular model of product differentiation. *Available at SSRN 1870703*, 2011.

Cheng Yong Tang and Chenlei Leng. Penalized high-dimensional empirical likelihood. *Biometrika*, 97(4):905–920, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.