ALAN L. MONTGOMERY and PETER E. ROSSI*

The authors show how price elasticity estimates can be improved in demand systems that involve multiple brands and stores. They treat these demand models in a hierarchical Bayesian framework. Unlike in more standard Bayesian hierarchical treatments, the authors use prior information based on the restrictions imposed by additive utility models. In an additive utility model approach, price elasticities are driven by a general substitution parameter as well as brand-specific expenditure elasticities. The authors employ a differential shrinkage approach in which price elasticities are held closely to the restrictions of the additive utility theory and store-to-store variation is accommodated through differences in expenditure elasticities. Application of these new methods to simulated and real store scanner data shows significant improvements over existing Bayesian and non-Bayesian methods.

Estimating Price Elasticities with Theory-Based Priors

Detailed scanner data on sales response and pricing conditions are now available for virtually every type of consumer packaged good and all major retail formats. The availability of these data permits a systematic approach to studying market structure or brand competition patterns (cf. Allenby 1989), elasticity-based approaches to optimal pricing, and evaluation of promotional profitability. In addition, there is increased interest in using disaggregate data for the purpose of micropricing, in which individual stores or groups of stores charge different prices to exploit differences in consumer price sensitivity (Montgomery 1997). All these analyses rely on the estimation of a demand system and associated price elasticities.

The promise of a quantitative demand-based approach to pricing issues is far from fully being realized because of the difficulty of obtaining reasonable price elasticity estimates. Unrestricted least squares estimates of own- and cross-price elasticities are often of an incorrect sign and unreasonable magnitude, particularly if the analysis is performed at a relatively low level of aggregation, such as the account or store level. One possible solution is to engage in a lengthy specification analysis and data cleanup, which is not guaranteed to remove the problematic estimates. If demand analyses are to be performed for more than one or two units of aggregation or for large and complicated groups of products, it may be impractical to devote a great deal of time to fine-tuning each demand equation.

A traditional solution to these problems involves aggregation, either across geographic units or products. Christen and colleagues (1997) emphasize that aggregation can introduce large biases in the estimates of own-price elasticities and recommend, instead, pooling whenever possible. Although pooling is useful if estimates at a high level of aggregation are desired, this rules out the possibility of using the data to inform pricing decisions at a less aggregate level. For example, the recent trend in category management requires category managers to choose prices so as to maximize total category profits on the account or market level (e.g., Kroger grocery store in Columbus, Ohio, Giant-Eagle in Baltimore); pooled or aggregate analyses are largely irrelevant to this important tactical problem. Aggregation over products to reduce the dimension of the demand system is also difficult because many products in the same category are not perfect substitutes, and the aggregation itself may remove much of the interesting detail (e.g., analysts could form aggregates of national and private-label products, but this would wash out much of the relevant information about relative pricing among different national brands).

One useful approach to the problem of improving price elasticity estimates is to use a statistical procedure in which information is "borrowed," or partially pooled, across the units of analysis (brands or units of aggregation such as stores or chains). In these Bayesian hierarchical approaches, a source of information beyond the data is the prior belief...

*Alan L. Montgomery is an associate professor, Graduate School of Industrial Administration, Carnegie Mellon University (e-mail: alan.montgomery@cmu.edu). Peter E. Rossi is Joseph T. Lewis Professor, Graduate School of Business, University of Chicago (e-mail: peter.rossi@gsb.uchicago.edu). To interact with colleagues on specific articles in this issue, see "Feedback" on the JMR Web site at www.ama.org/pubs/jmr.
that parameters from each demand equation are drawn from some superpopulation of parameters. Blattberg and George (1991) and Montgomery (1997) formulate Bayesian hierarchical models that produce estimates that achieve this effect by "shrinking" them toward a mean of this superpopulation.

Cooper, Klapper, and Inoue (1996) develop another statistical approach to the problem of producing reasonable and stable parameter estimates in high-dimensional models built with scanner data. Cooper, Klapper, and Inoue use a three-way version of factor analysis to reveal a lower-dimensional structure. Although their procedure was designed for application to share models, it could be adapted to find structure in a log-linear demand model such as the one discussed here. This approach relies on the existence of lower-dimensional structure and involves a two-stage estimation process, with an informal process of determining the dimension of the approximate lower-dimensional structure.

In contrast to these purely statistically motivated approaches, we provide a method of exploiting information in the form of restrictions obtained from a demand model derived under the hypothesis of utility maximization. Berndt and Silk (1993) and Allenby (1989) explore imposing exact restrictions from economic theory. In practice, however, the restrictions of economic theory do not hold exactly and are rejected by standard statistical tests. Our view is that economic theory provides a useful source of prior information that should not be imposed exactly but instead through a stochastic prior framework. The data then would be able to modify the prior, with the economic model providing some discipline and resulting in more plausible and reliable estimates.

We develop a new Bayesian approach that uses an additive utility model as the basis for prior information in a hierarchical setting. Our approach combines the benefits of economic theory with statistical shrinkage methods. Because of the nonlinear nature of the theoretical restrictions on the elasticity matrix, we must develop a nonstandard hierarchical model. Our hierarchical approach also features differential shrinkage, in which the prior is tighter on some set of parameters than on others. The standard normal Wishart linear hierarchical model used in the statistics and marketing literature (cf. Gelfand et al. 1990) does not allow for differential shrinkage because the Wishart prior has only one tightness parameter for all elements of the covariance matrix.

Our approach can be viewed as a general approach to shrinkage in demand estimation, which nests many other approaches in the literature as special cases. Our additive utility prior nests Allenby's model, which has subclasses of products with an IIA-style elasticity structure. Blattberg and George (1991) consider standard Bayesian methods that shrink own-price elasticities across brands or stores, whereas Montgomery (1997) considers shrinkage across stores with a full elasticity matrix, including cross-elasticity terms. Again, our approach can be reduced to achieve a close approximation of either Blattberg and George's or Montgomery's approach with appropriate restrictions. We find that, when we apply our approach to scanner data, the restrictions needed to reduce our model to Allenby's or Blattberg and George's are not satisfied. Furthermore, our approach outperforms Montgomery's (1997) in terms of out-of-sample prediction.

We apply our economically based hierarchical model to the estimation of a category demand system with 11 competing brands in each of 83 stores. This model yields both improved elasticity estimates and predictive performance superior to what can be achieved with standard Bayesian or classical methods. We obtain more reasonable cross-price elasticity estimates while retaining much of the interesting and potentially valuable store-to-store variation. The out-of-sample predictions from our Bayes estimates outperform pooled and individual store models that employ either restricted or unrestricted least squares estimates, as well as standard Bayesian shrinkage.

**THE BASIC DEMAND MODEL, EXPENDITURE EFFECTS, AND ADDITIVE UTILITY**

All basic pricing decisions and market structure analyses are based on the parameters of a demand system for a group of related products. For example, to determine the optimal price gap between national and store brands of orange juice, the retailer must estimate the own- and cross-price elasticities for all major items in the orange juice category. In many retail categories, there are easily ten major items, with some categories boasting several hundred (e.g., cookies and crackers). Estimation of these price elasticity arrays puts formidable demands on even relatively large scanner data sets. If the retailer wishes to customize prices at the store level, we must estimate price elasticity arrays for each store or group of stores. In other situations, national manufacturers may wish to customize prices at the regional or market level, which poses a similar problem.

Formally, we must estimate a demand system for M products in S stores with T weeks of data. For simplicity and ease of interpretation, we assume a log-log functional form (not only is this the most popular specification in practice, but it appears to fit the data best in our case).

\[
\ln(q_{ist}) = \alpha_s + \beta_i \ln(x_{ist}) + \epsilon_{ist} + \text{diag}(\theta_s) f_{ist} \\
+ \text{diag}(\psi_s) d_{ist} + e_{ist}, \quad e_{ist} \sim N(0, \Sigma_s), \quad s = 1, \ldots, S; \quad t = 1, \ldots, T.
\]

\(
q_{ist} \) and \( p_{ist} \) are vectors of movement and price, \( x_{ist} \) is store expenditures\(^1\) \( (x_{ist} = \sum p_{ist} q_{ist}) \), the i subscript denotes the ith product in the vector, and the i subscript ranges over all products in the store. \( \ln(P_{ist}) = \ln(p_{ist}) \) is a Divisia price index, and \( f_{ist} \) and \( d_{ist} \) are vectors of feature and display variables for store s during week t. Equation 1 departs from other sales response models by including total store expenditures as an independent variable. By including this variable, we measure how demand for items varies with total store expenditures, which enables us to estimate expenditure elasticities.

Estimation of the demand system in Equation 1 will be challenging if either the number of brands or products is large or the unit of aggregation is small. For example, many product categories require at least ten items to characterize market structure properly, which results in a large number of cross-elasticity estimates. Even pooling data across a large number of geographical units (e.g., chains or stores) does not always provide sufficient information to estimate both own- and cross-price elasticities accurately. In situations

\(^1\)Our store expenditure variable is based on the sum of expenditures in 29 categories, which account for approximately one-fifth of all commodity volume.
such as a national manufacturer making marketing decisions at the key account level (Boatright and Rossi 1997) or a retailer making decisions at the store level (Montgomery 1997), we desire inferences for each geographical unit of aggregation (account or store), which multiplies the estimation problem. For each account or store, only a small amount of time series data is available, which limits the price variation. Application of standard least squares techniques at the store or account level will result in many insignificant estimates, often with erroneous signs and unreasonable magnitudes.

The solution to the problems that plague models' fit to scanner data at a low level of aggregation is to bring in various sources of additional information. For example, if we are not interested in the variation across the low-level units, we might decide to pool the data and estimate one common set of parameters. This amounts to imposing the restriction that the price elasticity structure is the same over all units that are pooled, which is an extreme form of prior information. Moreover, even estimates from pooled models can be problematic if there is insufficient relative price variation. This is frequently the case for items that are members of a product line (e.g., different types of Tropicana orange juice that are priced and promoted together, which thereby causes a multicolinearity problem).

**An Economic Interpretation**

The sales response model in Equation 1 is a statistical model that fits the data well. However, we also can interpret this set of regression equations as a demand system. If this category is independent from other categories, Equation 1 can be interpreted as a subset or conditional demand model satisfying all the usual properties of demand models (for example, see Deaton and Muellbauer 1980, Chapter 5). The importance of attaching an economic interpretation to Equation 1 is that the $\mu$ and $E$ parameters can be interpreted as store expenditure and compensated price elasticities, and economic theory can be used to impose structure on the elasticity matrix.

In a standard demand modeling approach, we would specify a form of the utility function and impose the constraints implied by utility maximization subject to a budget constraint. Because the demand system will be applied to aggregate data, we cannot be sure that the restrictions resulting from a particular choice of utility and the assumption of utility maximization will hold exactly. Given the potential aggregation problem, our approach is to specify a reasonable utility function and impose the restrictions of demand theory stochastically through a prior distribution.

**Additive Utility**

In many applications of demand analysis in marketing, researchers investigate the relationship between items in fairly narrowly defined categories of goods. This stands in marked contrast to the literature on applied demand theory in economics, in which highly aggregated groups of goods are investigated, for example, housing, clothing, and food. The products in these narrowly defined categories are all natural substitutes with varying degrees of quality. One very reasonable starting point for a utility specification in this situation is the additive utility function:

\[
\text{utility} = U[\Sigma v_i(q_i)].
\]

Additive utility models form the basis of the popular logit and probit choice models and conjoint analysis. For example, the standard random utility interpretation of the logit and probit models assumes linear utility, which is a special case of additive utility. It is important to emphasize that additive utility can accommodate many common properties of utility functions, including diminishing marginal returns and different degrees of substitutability among goods.

Additive utility implies that the utility gained from one product is unaffected by the utility of other products. For example, there is no interaction in utility from consuming Minute Maid and Tropicana orange juice together. This makes a great deal of sense for products within a category that are typically direct substitutes and not used together. However, additivity may not make sense across products from different categories, such as bread and peanut butter.

Additive utility models result in parsimonious but restrictive matrices of price elasticities:

\[
E_i = \varphi_i \text{diag}(\mu_i) - \varphi_i \mu_i (\lambda_i \cdot w_i)'.
\]

Notice that the cross-elasticity matrix is populated solely by the expenditure elasticities ($\mu$), market shares ($w$), and a general substitution parameter ($\varphi$). This restricted elasticity matrix has $M + 1$ parameters, not including the market shares, as opposed to $M^2$ for the unrestricted form. In the analysis of market share models, several researchers have examined the implications of the IIA property on the price elasticity matrix. It should be noted that the restrictions of additive utility given in Equation 3 do not necessarily have these IIA-based properties.

Allenby (1989) proposes the use of additive utility restrictions in his work on identifying market substructures. In Allenby's view, utility is linear over submarkets of products, which are identified by statistical testing procedures. When a grouping or market substructure is identified, the restrictions from the additive utility over submarkets are imposed exactly. Allenby does not exploit the restrictions imposed by the possibility that additive utility restrictions could be useful within submarkets. In addition, Allenby imposes exact restrictions rather than using additive utility as the basis of a prior.

It has been noted that uncompensated price elasticity matrices are asymmetric (Blattberg and Wisniewski 1989; Kamakura and Russell 1989). In particular, there is empirical evidence that price reductions in higher-price, high-quality brands have more influence on the sales of low-quality, lower-priced brands than vice versa. These asymmetries can be explained by differences in expenditure elasticities and do not require that the compensated price elasticity matrix violate the Slutsky symmetry condition (for application of a nonhomothetic utility model in a choice context, see Allenby and Rossi 1991).

**Endogeneity**

In Equation 1, we have postulated a regression system in which we examine the effects of variation in total grocery expenditure (as measured by our 29-category proxy) and price movements on the quantity of sales. We propose to use restrictions from economic theory as the basis of a prior to help improve estimates of the parameters of this system. This approach is predicated on the assumption that independent variables are exogenous. Because the left-hand side
variable, \( q_{st} \), is included in the expenditure variable \( x_{st} \), and the price index \( P_{st} \), there is a potential endogeneity problem. This should be a small problem because both the expenditure variable and the price index are formed by summing over all items in our 29-category proxy for expenditure. Items in any one category often represent only a tiny fraction of the total store expenditures. However, conclusive proof of this assertion can come only by constructing valid instruments. We completely eliminate these sorts of endogeneity concerns by forming \( x \) and \( P \) from all items except the items modeled in Equation 1. We then simply project the \( x \) and \( P \) with all items on these instruments and rerun our results using these predicted values in a two-stage procedure. This instrumental variables procedure provides identical results to standard one-stage procedures. This is because the variables \( x_{st} \), \( P_{st} \), and \( \ln(x_{st}/P_{st}) \) are correlated almost perfectly with the same variables constructed without the items modeled in Equation 1 (a regression of any of these three variables on the instrument [same variable without category items] produces \( R^2 \)'s exceeding .999). It should be noted, however, that it is still possible that retailers determine price by predicting demand shocks along the lines suggested by Villas-Boas and Winer (1996). If demand shocks are large relative to supply shocks and are predictable, this can introduce some endogeneity into the price variables. We leave this as a subject for further research.

**CONSTRUCTION OF THE PRIOR**

We have shown that the additive utility formulation is an appealing basis of a prior for the compensated elasticity matrix in the demand system of Equation 1. The additive utility specification seems reasonable as a prior because the set of products is similar, and we do not expect to find interaction terms in the utility function. The number of effective free parameters is reduced radically in the additive formulation. Finally, additive utility does not require a homothetic utility structure. An analyst might wish to employ the restrictions implied by additive utility exactly. However, it is likely that in most data sets there will be departures from additivity. In our approach, we use the additive utility as the basis of a prior rather than as an exact restriction.

In standard Bayesian approaches, improved estimates are obtained through the phenomenon of shrinkage, in which similarities across stores and brands are exploited to "shrink" the unrestricted store level estimates toward some common set of parameters. We allow for this sort of shrinkage in our procedure, as well as incorporate the possible benefits of the additive utility structure. This requires a prior specification that will meet the following three criteria:

1. Center the elasticity prior over the restrictions imposed by the additive utility model,
2. Allow for deviations from the additive utility model, and
3. Allow for variation across stores.

A prior meeting these criteria necessarily will take on a different form than those commonly found in the recent literature on Bayesian hierarchical models (cf. Blattberg and George 1991; Gelfand et al. 1990; Montgomery 1997).

**The Prior**

The basic view underlying our prior specification is that differences among stores are driven primarily by differences in expenditure elasticities (\( \mu \)) and the overall substitution parameter (\( \varphi \)). Recall the relationship between the elasticity matrix and these parameters:

\[ E_s = \varphi_s \text{diag}(\mu_s) = \mu_s \cdot w_s. \]

Stores with higher \( |\varphi_s| \) will display larger own- and cross-price elasticities. The expenditure elasticities dictate the extent of asymmetric switching that will occur between high- and low-quality brands. It is possible that the customer population will differ across stores in the extent to which they value and are willing to pay for the higher-quality brands.

With this view, we write our prior on the price elasticity as centered over the value suggested by the theory:

\[ E_s = \text{vec}(E_s), \]

\[ E_s | \mu_s, \varphi_s \sim N(\overline{E}_s, \Delta). \]

The matrix \( \Delta \) determines how close the restrictions of the theory hold. If we set \( \Delta = 0 \), we restrict our elasticity array to that which is implied by the additive utility model, and all differences among stores will be driven by variation in the expenditure elasticities and the overall substitution parameter.

To complete the prior specification, we must specify marginal priors on \( \mu_s \) and \( \varphi_s \), as well as the model intercepts (\( \alpha_s \)), display (\( \theta_s \)), and feature effects (\( \psi_s \)). We stack (\( \mu_s, \theta_s, \psi_s \)) up into a vector and specify a prior on these in the normal form:

\[ \beta_s = \begin{bmatrix} \alpha_s \\ \mu_s \\ \theta_s \\ \psi_s \end{bmatrix} \sim N(\overline{\beta}, \Lambda). \]

This is coupled with a univariate prior on \( \varphi_s \):

\[ \varphi_s \sim N(\gamma, \lambda). \]

**Completing the Prior: Priors on Hyperparameters**

The first stage of our prior is laid out in Equations 4, 5, and 6. This part of the prior centers the price elasticity array over that which would be implied by an additive utility model. In this prior, store variation is driven through differences in the expenditure elasticity parameter vector and the overall substitution parameter. To calibrate this prior, we must decide how large to make the store variation (governed by \( \Lambda \)) and the size of deviations from the additive utility theory that will be tolerated (governed by \( \Delta \)). We also must decide where to center the priors. Rather than leaving this up to the investigator, we introduce a second stage of the prior that will put prior distributions over these hyperparameters. In this framework, the data help us adapt and learn about the central tendencies and variation of the key elasticity parameters. For example, if there is evidence in the data of substantial store variation in \( \mu \) and \( \varphi \), our prior will adapt to a large value of \( \Lambda \), and we will do little shrinkage.

The priors on the hyper parameters are as follows:

\[ \Delta^{-1} \sim W(u_{\Delta}, (u_5 k_5 V_\Delta)^{-1}). \]
Price Elasticities

\[ \Lambda^{-1} = W (u \Lambda (u \Lambda V \Lambda)^{-1}) \]

\[ \bar{\beta} \sim N(\theta, V \theta) \]

\[ \bar{\phi} \sim N(\bar{\phi}, V \phi) \]

\[ \lambda_\phi \sim IG (v_\phi, v_\phi^2) \]

The key priors here are the priors on the covariance matrices given by Equations 7 and 8. The prior parameterization centers these priors over \( V^{-1/k} \). In virtually all work on Bayesian hierarchical models, a Wishart prior is used to govern the extent of adaptive shrinkage. By setting a scale parameter in the prior and the degrees of freedom parameter, this Wishart prior dictates the level of shrinkage. One of the well-known shortcomings of this prior is that there is only one parameter that governs the tightness of the prior. We cannot have a tight prior on some parameters (inducing a great deal of shrinkage) and a loose prior on others. We want the freedom for differential adaptive shrinkage, in which the data could dictate a close adherence to the theoretical restrictions from additive utility while promoting a great deal of store-to-store variation. Our solution to this problem is to calibrate two separate Wishart priors, one for the deviations from theory (Equation 7) and one for the store variation (Equation 8).

Our general philosophy will be to set the prior hyperparameters in Equations 7–11 so as to promote fairly tight adherence to the theory (Equation 4) and allow for a much wider range of store-to-store variation in the overall substitution parameter and expenditure. Because of the conditional structure of the prior and the nonlinearity of mean elasticities as a function of the additive utility parameters, we must take special care in assessing these priors. As illustrated, we observe the marginal prior distributions on key model parameters to ensure that our priors are sufficiently informative to rule out implausible values of parameters (e.g., positive own-price elasticities, negative expenditure elasticities) while still admitting a wide range of possible values.

Our prior can be used in conjunction with a form of the Gibbs sampler to make posterior inferences. Details are available in the Appendix.

Comparison with Other Bayesian Shrinkage Approaches

From a purely statistical point of view, the set of demand systems for M items in each of S stores is a collection of possibly related regression equations. For a collection of 100 stores and ten products, 1000 regression equations would need to be estimated. Standard Bayesian shrinkage approaches would specify a prior over the set of regression coefficients. Bayes estimates from these procedures will exhibit the property known as shrinkage, in which the Bayes estimates will be a compromise between separate least squares estimates for each of the \( M \times S \) regressions and the overall grand mean. The amount of shrinkage is determined both by the evidence in the data, which suggests differences among the brands, and by differences between stores.

As applied, standard Bayesian shrinkage models can be used to shrink across brands (as in Blattberg and George 1991, in a model without cross-elasticities) or across stores (as in Montgomery 1997). In many marketing applications, it might not be reasonable to shrink across brands. Even though all of the items under study may be in the same product category, there are large differences between high-quality national brands and low-quality generic and store brands. In many situations, the higher-quality brands may have own-price elasticities half the size of lower-quality brands. Thus, shrinking all brands to one common own- and cross-price elasticity vector could introduce serious biases. Finally, we want to tell the difference between the situation in which the brands have different price elasticities but there are few differences between stores and the situation in which there are large store differences. The implications for retailer pricing policies are different in these two situations.

The standard Bayesian hierarchical literature offers no clear way of decomposing the variation in the coefficient vectors. Our approach here will be to use a prior that allows for shrinkage across brands that reflects utility theory, as well as an independent level of shrinkage across stores. Our approach nests the standard approaches as special cases. If the prior on the utility restrictions adapts to a loose setting, we will approximate the case of shrinkage across stores closely (as in Montgomery 1997). Shrinkage across brands will be achieved if all brands have the same expenditure elasticities and market shares.

**USING SIMULATED DATA TO UNDERSTAND THE THEORY-BASED SHRINKAGE METHOD**

The nonstandard hierarchical model previously laid out will produce a nonstandard sort of shrinkage procedure. In this section, we use simulated data to understand and illustrate the properties of our theory-based hierarchical model better.

In our theory-based priors, we have two dimensions of variation in the elasticities: variation about the theory and variation from store to store. Our nonstandard model can adapt differentially to each dimension. For example, if the data adhere closely to the restrictions from additive utility but elasticities vary widely from store to store, our procedure will "adapt" and shrink the elasticities close to the theory but allow for a great deal of variation from store to store. However, if the data do not adhere to the restrictions of the additive utility model, we will not do much shrinkage to the theory restrictions, and the posterior of the model will be massed on large values (giving the prior little influence). This avoids the common criticism: "What happens if the prior is strong and wrong?"

We simulate data from a hypothetical sample of stores in two situations: (1) the true price elasticity parameters satisfy the additive utility restrictions and (2) the true elasticities violate additive utility.

We simulate data for a hypothetical sample of 83 stores with three brands observed over 30 weeks (15 weeks for estimation and 15 weeks for predictive validation). To draw the parameters for each store, we use a multivariate normal distribution with a mean that was chosen to satisfy the restrictions of additive utility or not. We also use a fairly diffuse prior on the covariance matrices \( \Lambda \) and \( \Delta \), so that the data are the primary determinants of the posterior on these matrices.

In Table 1, we report the in-sample and out-of-sample mean squared error (MSE) for the additive and nonadditive simulated data. We compare our Bayes estimates to the performance based on the true parameters, least squares esti-
APPLICATION TO RETAIL SCANNER DATA

We apply our methods to store-level scanner data collected from 83 stores in the Dominick's Finer Foods chain in Chicago, Ill. We have 120 weeks of data, which are split in half for the purposes of model validation. We consider products in the refrigerated orange juice category. In Table 2, we list the items under study, average price, and market share. The 11 items represent more than 70% of the dollar volume in this category and cover the range from premium national brands to lower-quality store brands. Our expenditure variable \( x \) is calculated from a subset of 29 store categories with more than 5000 universal product codes. These categories account for more than 25% of total store all commodity volume. We also experiment with more limited definitions of the expenditure variable, such as the total expenditure in the refrigerated juice category or the expenditure only for the 11 brands used here. Using these alternative definitions of expenditure had no material effect on the results.

Results from Standard Classical Approaches

We begin our analysis of the refrigerated orange juice data by considering standard classical procedures. We consider the following procedures:

- Unrestricted least squares on the store level (separate fits for each store),
- Restricted least squares on the store level (additive utility restrictions are imposed exactly),
- Unrestricted pooled models, and
- Restricted pooled models.

In Table 3, we provide summaries of in- and out-of-sample goodness of fit, including in- and out-of-sample MSE, log-likelihood value, Akaike information criterion (AIC), and the Schwarz information criterion (SIC). The 83 stores and 11-dimensional demand system generate a huge number of parameters, which are difficult to estimate even with our relatively long and variable time series of weekly data for each store.

It is immediately clear from Table 3 that pooling is not a good idea. The pooled models fit poorly in sample and have poor out-of-sample performance. This is due to the large degree of heterogeneity in the demand system parameters across stores. At the other extreme, unrestricted store level models seem to overfit the data with excellent in-sample performance but a conspicuous decline in out-of-sample performance.

Table 2

ITEMS, AVERAGE PRICE, AND SHARE

<table>
<thead>
<tr>
<th>Item (ounces)</th>
<th>Average Price</th>
<th>Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropicana Premium 64</td>
<td>2.87</td>
<td>16.1</td>
</tr>
<tr>
<td>Tropicana Premium 96</td>
<td>3.12</td>
<td>10.7</td>
</tr>
<tr>
<td>Florida's Natural 64</td>
<td>2.86</td>
<td>4.0</td>
</tr>
<tr>
<td>Tropicana 64</td>
<td>2.27</td>
<td>15.8</td>
</tr>
<tr>
<td>Minute Maid 64</td>
<td>2.24</td>
<td>16.9</td>
</tr>
<tr>
<td>Minute Maid 96</td>
<td>2.68</td>
<td>5.7</td>
</tr>
<tr>
<td>Citrus Hill 64</td>
<td>2.32</td>
<td>5.1</td>
</tr>
<tr>
<td>Tree Fresh 64</td>
<td>2.18</td>
<td>2.5</td>
</tr>
<tr>
<td>Florida Gold 64</td>
<td>2.07</td>
<td>2.6</td>
</tr>
<tr>
<td>Dominick's 64</td>
<td>1.74</td>
<td>13.6</td>
</tr>
<tr>
<td>Dominick's 128</td>
<td>1.83</td>
<td>6.9</td>
</tr>
</tbody>
</table>

Table 1

PERFORMANCE OF THEORY-BASED BAYES AND OTHER PROCEDURES
(IN-SAMPLE MSE/OUT-OF-SAMPLE MSE)

<table>
<thead>
<tr>
<th></th>
<th>Additive Data</th>
<th>Nonadditive Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>True parameters</td>
<td>.54/.52</td>
<td>.51/.52</td>
</tr>
<tr>
<td>Unrestricted least squares</td>
<td>.31/.93</td>
<td>.30/.86</td>
</tr>
<tr>
<td>Restricted least squares</td>
<td>.40/.70</td>
<td>1.40/.25</td>
</tr>
<tr>
<td>Bayes</td>
<td>.37/.63</td>
<td>.35/.61</td>
</tr>
</tbody>
</table>

Figure 1

DISTRIBUTION OF CROSS-PRICE ELASTICITIES ACROSS PRODUCTS AND STORES: SIMULATED DATA FROM ADDITIVE UTILITY CASE

mates for each store, and additive utility-restricted least squares estimates for each store.

In the case of data simulated with additive utility, the Bayes estimates outperform the unrestricted least squares and perform comparably to the restricted least squares estimates. For data simulated from nonadditive utility, the Bayes estimates have adapted so that the additive utility restrictions are ignored (i.e., \( \Delta \) becomes large). We still get the benefit of shrinkage of estimates across stores, which provides better predictive performance than the unrestricted least squares estimates for each store. In Figure 1, we display the distribution of cross-elasticities across products and stores for data simulated with additive utility holding. We show the large sampling errors in the store-level least squares estimates, which produce approximately 30% negative cross-elasticities. Our Bayes estimates show only a small (less than 10%) fraction of negative cross-elasticities while still exhibiting a lot of variation.

Our analysis of simulated data shows the potential for the proposed theory-based adaptive Bayesian procedure. We obtain improved estimates to the extent that the data adhere to the additive utility restrictions. However, the elasticity estimates will not suffer from appreciable bias if the data do not satisfy these restrictions. We now turn to the analysis of actual scanner data to explore the benefits of our procedure further.
Among all the classical procedures proposed, the restricted store models have the best out-of-sample performance. It seems that the restrictions are useful in cutting down on prediction error, and store-level restricted parameters capture much of the store-to-store variation in demand properties. However, a standard likelihood ratio test of the restrictions would overwhelmingly reject them ($p < .0001$). It may be that the out-of-sample performance of the restricted models is obtained by exploiting the variance-bias trade-off. Ultimately, we want to use these models to predict demand for various pricing policies. The restricted models may have biases that reduce the usefulness of these predictions. Thus, some sort of compromise model that lies close to the restricted store-level models may be the best model for use in practical prediction and pricing policy situations.

**Results from the Bayesian Approach**

As emphasized previously, our Bayesian model has the capability to adapt to models very close to the restricted/unrestricted/pooled/unpooled models considered in Table 3. There are two dimensions along which shrinkage can occur: (1) toward the restrictions of the additive utility model and (2) across stores to the pooled solution. The data themselves and the choice of prior hyperparameters dictate how much and what type of shrinkage will occur. These shrinkage dimensions correspond in our Bayesian model to the key covariance matrices, $\Delta$ and $\Lambda$. If $\Delta$ is small, then we are allowing little variation from the restrictions of additive utility theory. If $\Lambda$ is small, then we are allowing little variation across stores. For example, if we set both $\Delta$ and $\Lambda$ to matrices of large numbers, posterior means from the Bayesian hierarchical model would approximate the unrestricted store level estimates closely. Of course, we do not set the elements of $\Delta$ and $\Lambda$ directly. Instead, we specify priors on these matrices. These priors and the data dictate how much shrinkage actually occurs. The key decision that must be made is to specify the location of these priors on $\Delta$ and $\Lambda$.

The priors on $\Delta$ and $\Lambda$ have a degree of freedom parameter, $v$, and a scaling parameter, $k$:

$$\Delta^{1} - W(v_{\Delta}, (v_{\Delta}k_{\Delta}V_{\Delta}^{-1})^{-1}); \Lambda^{1} - W(v_{\Lambda}, (v_{\Lambda}k_{\Lambda}V_{\Lambda}^{-1})^{-1}).$$

Setting a high degree-of-freedom value (relative to the sample size) will tighten the prior around the location set by $k$ and $V$ (note: $E[\Delta^{1}] = (k_{\Delta}V_{\Delta}^{-1})$, and $E[\Lambda^{1}] = (k_{\Lambda}V_{\Lambda}^{-1})$). By appropriate choice of prior parameters, we can move the Bayes estimates quite close to any of the restricted/unrestricted/pooled/individual models previously discussed. However, the advantage of the hierarchical Bayes approach is that we can learn from the data the direction and amount of shrinkage.

We first consider a very diffuse prior setting, $v_{\Delta} = \text{dim}(\Delta) + 3$ and $v_{\Lambda} = \text{dim}(\Lambda) + 3$. These are barely proper Wishart priors of the sort often used in the hierarchical Bayes literature. For these prior settings, our model will adapt fully with minimal prior influence. We compute Bayes estimates of all elasticity parameters and then evaluate in- and out-of-sample performance, as in Table 3. We achieve an in-sample MSE of .211 and an out-of-sample MSE of .296. These numbers dominate all the classical procedures, which shows that a flexible Bayes model captures much of the advantages of various restricted or pooled approaches without sacrificing flexibility.

**Comparison with Standard Bayesian Shrinkage**

In addition, our theory-based prior outperforms the standard Bayesian shrinkage approach outlined by Blattberg and George (1991) or Montgomery (1997). Applying standard shrinkage across brands and stores, as suggested by Blattberg and George (1991), yields an in-sample MSE of .203 and an out-of-sample MSE of .319. Shrinkage only across stores, as suggested by Montgomery (1997), yields an in-sample MSE of .198 and an out-of-sample MSE of .312. Thus, even a very diffuse prior in our theory-based shrinkage approach yields an improvement of at least 5% in predictive performance over standard Bayesian adaptive shrinkage methods.

**Prior Sensitivity**

Although many researchers might find the use of diffuse but trainable priors attractive, our own view is that informative theory-based priors are plausible. One problem that may arise with informative theory-based priors is that they might require a lot of data to overwhelm the prior if the data do not support the prior. To explore this possibility, we conducted an extensive analysis of tighter priors, evaluating each on the basis of predictive performance. To achieve a tighter prior, we set $v_{\Delta} = \text{dim}(\Delta) + 5 \times 83$ and $v_{\Lambda} = \text{dim}(\Lambda) + 5 \times 83$, because 83 is effectively the sample size for the estimation of these covariance matrices. We experimented with a variety of values for $k_{\Delta}$ and $k_{\Lambda}$, which determine how much shrinkage occurs. In Table 4, we present these results.

### Table 3
**COMPARISON TO NONBAYESIAN PROCEDURES**

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>MSE</th>
<th>Log-Likelihood</th>
<th>AIC</th>
<th>SIC</th>
<th>Predictive MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrestricted store</td>
<td>12,865</td>
<td>.170</td>
<td>49560.9</td>
<td>−73391.9</td>
<td>40778.2</td>
<td>.379</td>
</tr>
<tr>
<td>Restricted store</td>
<td>2,905</td>
<td>.247</td>
<td>24798.1</td>
<td>−43786.1</td>
<td>−18005.8</td>
<td>.318</td>
</tr>
<tr>
<td>Unrestricted pooled</td>
<td>155</td>
<td>.314</td>
<td>20850.9</td>
<td>−41391.8</td>
<td>−40016.3</td>
<td>.385</td>
</tr>
<tr>
<td>Restricted pooled</td>
<td>35</td>
<td>.358</td>
<td>11598.5</td>
<td>−23126.9</td>
<td>−22816.4</td>
<td>.402</td>
</tr>
</tbody>
</table>

Notes: MSE = mean squared error, AIC = Akaike information criterion, and SIC = Schwarz information criterion.

Here is a table showing the effect of prior scaling on parameter sensitivity:

### Table 4
**PRIOR SCALING PARAMETER SENSITIVITY ANALYSIS**

<table>
<thead>
<tr>
<th>$k_{\Delta}k_{\Lambda}$</th>
<th>.0001</th>
<th>1</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0001</td>
<td>2511.318</td>
<td>2151.297</td>
<td>.1821337</td>
</tr>
<tr>
<td>1</td>
<td>2151.294</td>
<td>2141.292</td>
<td>.1811335</td>
</tr>
<tr>
<td>4900</td>
<td>2098.301</td>
<td>2081.298</td>
<td>.1771350</td>
</tr>
</tbody>
</table>

Notes: In-sample MSE/Out-of-sample MSE.
The four corners of Table 4 correspond to prior settings, much in the spirit of each of the four classical approaches. Small values of \( k_A \) restrict deviations from the theory, and small values of \( k_A \) restrict store-to-store parameter variation.

The main message from Table 4 is that priors that are "loose" on the theory have poor out-of-sample performance (see the right-hand column). The best setting in Table 4 is the middle cell, with the k's set to 1. This prior setting performs comparably to the adaptable diffuse priors previously reported.

In summary, a range of priors, from fairly diffuse to tighter, in Table 4 can produce comparable out-of-sample performance because of the adaptive nature of the shrinkage process. Our view is that the tighter prior is preferred because we believe that the theoretical model is plausible. However, some may be uncomfortable with this position and want further data-based guidance. To provide this, we compute posterior probabilities for the various prior settings using Newton and Raftery's (1994) approach. These results overwhelmingly support a model with a very tight prior on the theory and only a modest amount of shrinkage across stores (\( k_A = .0001, k_A = 4900 \) has highest posterior probability of all the settings in Table 4).

Exploring the Posterior

To this point, we have concentrated on various summary statistics of model performance without much exploration of the posterior and prior distributions of various key parameters. In this subsection, we examine various posterior and prior marginal distributions for the key price elasticity, expenditure elasticity, and overall substitution parameters for our preferred prior setting (k's = 1 and vs set to dimension + \( 5 \times 83 \)).

We first examine the implied priors over key model parameters to ensure that our priors are sufficiently diffuse to enable the data to play the primary role in making inferences. In Figure 2, we show the prior marginal distribution of the expenditure elasticities and overall substitution parameter. Economic theory suggests that expenditure elasticities should be distributed around 1 with a negative overall substitution parameter. The prior marginal distributions of own- and cross-price elasticities are shown in Figure 3. Our priors cover a broad range of plausible values. Because we are satisfied that the priors are reasonably diffuse without including implausible values, we turn to a brief description of various key features of the posterior.

The pronounced variation in key parameters is evident in Figures 4, 5, and 6, which display various marginal posteriors for a spectrum of stores. To produce each figure, we sort the stores from smallest to largest posterior mean for each parameter and display boxplots for the marginal posteriors for the 5, 15, 25, 35, 45, 55, 65, 75, 85, and 95 percentiles of the store distribution. Figure 4 shows large variation in the overall level of substitutability across stores. One interpretation of Figure 4 is that the customer base in some stores regards this set of brands as more substitutable than others; that is, effective quality distinctions are smaller for certain store subpopulations. In Figure 5, the expenditure elasticity
Price Elasticities

for Citrus Hill is plotted for various stores, again displaying large variation. Differences across stores in the level of substitutability and in the expenditure elasticities drive differences in uncompensated elasticities, as is demonstrated in Figure 5, which displays price elasticities for Tropicana 64-ounce juice.

In Figures 7 and 8, we provide the distributions of various elasticity estimates across stores and products for three estimation procedures: unrestricted least squares applied store by store, restricted least squares applied store by store, and our theory-based Bayes estimates. Figure 7 shows the distribution of own-price elasticities; note that there is one estimate for each of the 83 stores and the ten products, equal to 830 estimates plotted in the form of a boxplot. All the own-price elasticities are of the correct sign. What is most striking is the difference in variation between the restricted store estimates and either the Bayes or individual least squares estimates. The restricted store estimates have much smaller variation across stores and products. With the additive utility restrictions, own-price elasticity variation can be driven only by changes in expenditure elasticities, overall substitution, and market shares (see Equation 3).

With the superior performance of the Bayes estimates in out-of-sample prediction, it seems clear that the additive utility restrictions are a bit too severe and that this restricts the variation in own-price elasticities. Figure 5 shows the distribution of cross-price elasticities for the same three estimation procedures. The Bayes estimates have fewer nega-
tive cross-price elasticities, as might be expected. However, there is still a good deal of information in the data that suggests some cross-price elasticities are negative. If the researcher believed this was completely implausible, the prior settings could be adjusted to tighten down on the theory, providing more shrinkage to positive cross-price elasticities that are implied by the additive utility structure. As discussed previously, we used fairly diffuse priors that adapt to the information in the data. This data set illustrates the importance of using the theory as the basis of a prior rather than imposing it on the data. There is value in the additive utility theory when it is applied in a nondogmatic fashion.

**CONCLUSIONS**

We show how demand estimates can be improved when they occur across multiple brands and stores. We treat these demand models in a hierarchical Bayesian framework. Unlike more standard Bayesian hierarchical treatments, we use prior information based on the restrictions imposed by additive utility models. In an additive utility model approach, price elasticities are driven by a general substitution parameter, as well as brand-specific expenditure elasticities. We employ a differential shrinkage approach in which price elasticities are held closely to the restrictions of the additive utility theory but store-to-store variation is accommodated through differences in expenditure elasticities. Our differential shrinkage approach is adaptive. For example, if the data lend support to the restrictions of additive utility theory but still exhibit substantial variation across stores, the approach proposed here will "train" itself to a prior that is tight around the restrictions but still allows for variation across the stores.

We apply our approach to store-level scanner data for a large grocery chain with 11 brands in the refrigerated orange juice category. Our method produces estimates that outperform standard demand estimates produced by unrestricted or restricted (by the additive utility theory) least squares. We also outperform various pooled estimates in both in- and out-of-sample predictive performance. Finally, our method outperforms standard Bayesian shrinkage approaches, as judged by out-of-sample predictive performance.

**APPENDIX**

**Gibbs Sampler for Our Model**

Our basic log-log demand model can be written in the standard notation for the seemingly unrelated regression (SUR) model:

\[ y_s \sim N(X_s \delta_s, \Sigma_s \Theta_s) \]

Here, we have stacked the vector of demand quantities for each brand and allowed the demand equations errors to be correlated contemporaneously:

\[
q = \begin{bmatrix}
q_{1s} \\
q_{2s} \\
\vdots \\
q_{Ms}
\end{bmatrix},
\Sigma = \begin{bmatrix}
\text{ln}(q_{11}) \\
\text{ln}(q_{21}) \\
\vdots \\
\text{ln}(q_{Mt})
\end{bmatrix},
X = \begin{bmatrix}
X_{1s} \\
X_{2s} \\
\vdots \\
X_{Ms}
\end{bmatrix},
\]

\[
X_{ls} = \begin{bmatrix}
1 & \text{ln}(x_{1s}/p_{1s}) & \text{ln}(p_{1s}) & \ldots & \text{ln}(p_{M_{1s}}) & f_{1ls} & d_{1ls} \\
1 & \text{ln}(x_{2s}/p_{2s}) & \text{ln}(p_{2s}) & \ldots & \text{ln}(p_{M_{2s}}) & f_{12s} & d_{12s} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
1 & \text{ln}(x_{Ms}/p_{Ms}) & \text{ln}(p_{Ms}) & \ldots & \text{ln}(p_{M_{Ms}}) & f_{1Ms} & d_{1Ms}
\end{bmatrix}
\]

\[
\delta_s = \begin{bmatrix}
\delta_{ls} \\
\delta_{xs} \\
\vdots \\
\delta_{Ms}
\end{bmatrix},
\Sigma_s = \begin{bmatrix}
\sigma_{ls}^2 & \ldots & \text{c(t)}(\phi_{ls}) & \ldots & \sigma_{Ms}^2
\end{bmatrix},
\Theta_s = \begin{bmatrix}
\phi_{ls} & \ldots & \theta_{ls} & \ldots & \phi_{Ms}
\end{bmatrix},
\]

where \( \epsilon_{ls} \) is the \( l \)th row of \( E_{ls} \).

We can combine this likelihood with the prior laid out in Equations 4–11 to form the joint posterior for all of the model parameters:

\[
\{\epsilon_{ls} \}, \{\beta_{ls} \}, \{\Sigma_s \}, \{\phi_{ls} \}, \Delta \mid \{\epsilon_{ls} \}, \{X_{ls} \}.
\]

Bayes inferences for price elasticities for a particular store are derived from the marginal posterior distribution of \( \epsilon_{ls} \).
For example, the posterior mean is used frequently as a Bayes estimate. The nonlinear structure of the additive utility restrictions, coupled with the SUR model framework, make it impossible to perform the necessary integrals to derive the marginal distribution analytically. For this reason, we develop a simulation-based method to approximate the marginal posterior to any desired degree of accuracy.

We develop a specialized version of the Gibbs sampler to handle this model (Gibbs samplers have been used extensively in the analysis of linear hierarchical regression models). Gibbs sampling is a subset of the group of methods called Markov chain Monte Carlo. The basis of all of these methods is the idea that a Markov chain that has the posterior as its equilibrium distribution can be constructed. By "running," or simulating, this chain, we can simulate effectively from the posterior distribution. Gibbs sampling is based on a Markov chain, which is constructed by successively sampling from a set of conditional posterior distributions. We take full advantage of the conditional nature of our prior specification to construct conditional posterior distributions that can be sampled using standard normal Wishart theory. The full set of conditionals is as follows:

(A1a) \( \beta_{ls} \mid \epsilon_{ls}, \Sigma_s, \gamma_{ls}, X_{ls} \);

(A1b) \( \epsilon_{ls} \mid \beta_{ls}, \Sigma_s, \gamma_{ls}, X_{ls} \);

(A2) \( \phi_{ls} \mid \beta_{ls}, \Delta \);

(A3) \( \Sigma_s \mid \beta_{ls}, \epsilon_{ls}, \gamma_{ls}, X_{ls} \);

(A4) \( \Delta \mid \{\epsilon_{ls} \}, \{\beta_{ls} \}, \{\phi_{ls} \} \);

(A5) \( \Lambda \mid \{\beta_{ls} \}, \bar{\beta} \);

(A6) \( \bar{\beta} \mid \{\beta_{ls} \}, \Lambda \);
(A7) \[ \overline{\varphi} \mid \{\varphi_s\}, \lambda_{\varphi}; \]

and

(A8) \[ \lambda_{\varphi} \mid \{\varphi_s\}, \overline{\varphi}. \]

We break the conditional posterior of the regression coefficients into two parts. The joint conditional posterior of \((\beta_s, \varepsilon_s)\) is not in a standard form because of the nonlinearity of the mean of \(\varepsilon_s\) in terms of \((\mu_s, \phi_s)\). Breaking into two conditionals maintains a conditionally linear structure and enables standard normal regression Bayes results to be used. To determine how this is done, we recall that, given the price elasticities, we can write the model as a standard SUR:

\[
[\ln(q_{its}) - \Sigma e_{is}\ln(p_{jts})] = \alpha_{ts} + \mu_{is}\ln(x_{ts}/P_{ts}) + \theta_{ts}d_{its} + \psi_{its}d_{its} + \varepsilon_{its}.
\]

Thus, \(\beta_s \mid \varepsilon_s, \Sigma_{sv}, y_s, X_s\) is a normal distribution from the standard Bayes theory of linear models. Likewise, \(\varepsilon_s \mid \beta_s, \Sigma_{sv}, y_s, X_s\) is also a normal distribution with a mean that is highly nonlinear in \((\mu_s, \phi_s)\). This is found by writing the model as follows:

\[
[\ln(q_{its}) - [\alpha_{ts} + \mu_{is}\ln(x_{ts}/P_{ts}) + \theta_{ts}d_{its} + \psi_{its}d_{its}]]
\]

\[
\quad = \Sigma e_{is}\ln(p_{jts}) + \varepsilon_{its}
\]

The rest of the conditionals are straightforward and based on standard normal Wishart theory (details and code are available on request from the authors).

REFERENCES


