A Convex Optimization Approach to Modeling Consumer Heterogeneity in Conjoint Estimation

Theodoros Evgeniou
Technology Management and Decision Sciences, INSEAD, Bd de Constance, 77300 Fontainebleau, France,
theodoros.evgeniou@insead.edu

Massimiliano Pontil
Department of Computer Science University College London, Malet Place, London WC1E 6BT, United Kingdom,
pontil@cs.ucl.ac.uk

Olivier Toubia
Marketing Division, Columbia Business School, 3022 Broadway, Room 522, New York, New York 10027,
ot2107@columbia.edu

We propose and test a new approach for modeling consumer heterogeneity in conjoint estimation based on convex optimization and statistical machine learning. We develop methods both for metric and choice data. Like hierarchical Bayes (HB), our methods shrink individual-level partworth estimates towards a population mean. However, while HB samples from a posterior distribution that is influenced by exogenous parameters (the parameters of the second-stage priors), we minimize a convex loss function that depends only on endogenous parameters. As a result, the amounts of shrinkage differ between the two approaches, leading to different estimation accuracies. In our comparisons, based on simulations as well as empirical data sets, the new approach overall outperforms standard HB (i.e., with relatively diffuse second-stage priors) both with metric and choice data.

Key words: Bayesian analysis; data mining; econometric models; estimation and other statistical techniques; hierarchical Bayes analysis; marketing research; regression and other statistical techniques
History: This article was received on August 1, 2006, and was with the authors 2 months for 2 revisions; processed by Leonard Lodish.

1. Introduction

A number of optimization-based approaches to conjoint estimation have been proposed in the past. Examples include methods based on linear programming (Srinivasan and Shocker 1973, Srinivasan 1998) or statistical machine learning (Cui and Curry 2005, Evgeniou et al. 2005a), and polyhedral methods (Toubia et al. 2003, Toubia et al. 2004). While these optimization approaches have proved fruitful, they have been exclusively limited to *individual level* estimation and have not modeled heterogeneity.[1] They have therefore underperformed relative to methods such as hierarchical Bayes (HB) (Toubia et al. 2003, Toubia et al. 2004, Evgeniou et al. 2005a).

In this paper we propose and test a new approach to modeling consumer heterogeneity in both metric and

choice-based conjoint estimation using convex optimization and statistical machine learning. We compare our approach with hierarchical Bayes (HB) both theoretically and empirically. Both our methods and HB shrink individual-level partworth estimates toward a population mean (in HB shrinkage is done toward the mean of the first-stage prior on the partworths). However, while HB samples from a posterior distribution that is influenced by a set of exogenous parameters (the parameters of the second stage priors), the proposed approach minimizes a convex loss function that is influenced by a parameter set endogenously (determined from the calibration data) using cross-validation. As a result, the *amounts* of shrinkage differ between HB and our approach. Moreover, we show that the second-stage prior parameters in HB could in theory be set to give rise to HB estimates identical to our estimates, or possibly of higher performance. However, this would require a method for systematically and optimally selecting the second-stage prior parameters in HB. Such selection raises both theoretical and practical issues, which we discuss.

---

[1] The only exception of which we are aware is an ad-hoc heuristic briefly discussed by Toubia et al. (2004), which is impractical because it requires the use of out-of-sample data. In contrast, our goal is to develop a general theoretical framework for modeling heterogeneity.

We use simulations as well as two empirical data sets (one for ratings and one for choice) to compare the performance of our approach to that of a standard HB implementation with relatively diffuse second-stage priors (Allenby and Rossi 1999, Rossi and Allenby 2003). The proposed approach overall outperforms HB with both metric and choice data. We empirically show that the differences between our approach and HB may be linked to differences in the amounts of shrinkage, as suggested by our theoretical comparisons. Moreover, we provide evidence that selecting the parameters of the second-stage priors in HB endogenously (e.g., using cross-validation as in the proposed approach) has the *potential* to greatly improve the predictive performance of HB.

Our approach builds upon and combines ideas from four literatures: statistical learning theory and kernel methods, convex optimization theory, hierarchical Bayes estimation, and the "learning to learn" or "multitask learning" literature in machine learning. "Learning to learn" methods were initially developed mainly using neural networks (Baxter 1997, Caruana 1997, Thrun and Pratt 1997) and recently studied using kernel methods (Jebara 2004, Ando and Zhang 2005, Evgeniou et al. 2005b, Micchelli and Pontil 2005). The central problem addressed by these methods is that of simultaneously estimating regression functions from many different but related data sets. Our work is novel first by its focus on conjoint estimation, second by the particular loss functions and the convex optimization method used to minimize them, and third by the theoretical and empirical comparison with HB.

The paper is organized as follows. We present our approach for metric as well as choice-based conjoint analysis in §2. In §3, we discuss the theoretical similarities and differences between our approach and HB. We then empirically compare the accuracy and predictive performance of our methods with HB using simulations in §4 and two (one for ratings and one for choice) field data sets in §5. In §6 we illustrate empirically the theoretical differences between our approach and HB outlined in §3, and we conclude in §7.

2. Presentation of the Approach

For ease of exposition, we describe the metric version of our approach first and the choice version second.

2.1. Metric Conjoint Estimation Method

2.1.1. Setup and Notation. We assume I consumers (indexed by $i \in \{1, 2, \ldots, I\}$) each rating J profiles (with J possibly different across respondents), represented by row vectors $\mathbf{x}_{ij}$, $j \in \{1, 2, \ldots, J\}$. We assume that the number of partworths is p, i.e., each vector $\mathbf{x}_{ij}$ has p columns. We note with X_i the $J \times p$

design matrix for respondent i (each row of this matrix corresponds to one profile); with $\mathbf{w}_i$ the $p \times 1$ column vector of the partworths for respondent i; and with $\mathbf{Y}_i$ the $J \times 1$ column vector containing the ratings given by respondent i. For simplicity we make the standard assumption of additive utility functions: the utility of the profile $\mathbf{x}_{ij}$ for respondent i is $U_i(\mathbf{x}_{ij}) = \mathbf{x}_{ij}\mathbf{w}_i + \epsilon_{ij}$. It is important to note that the proposed method can be extended to include large numbers of interactions between attributes, using, for example, the kernel approach (Wahba 1990, Vapnik 1998) introduced to marketing by Cui and Curry (2005) and Evgeniou et al. (2005a). We discuss this in detail in the online technical appendix. In agreement with previous research on individual-level conjoint estimation (Cui and Curry 2005, Evgeniou et al. 2005a), the presence of interactions in the model specification enhances the relative performance of our methods compared to HB.

2.1.2. Individual-Level Partworth Estimation Using Statistical Machine Learning: A Brief Review. We build upon a particular individual-level statistical estimation method known as ridge regression (RR), or Regularization Networks. This *individual-level* method (and various extensions, for example to the estimation of general nonlinear functions) has been extensively studied in the statistics and machine learning literatures (see, for example, Tikhonov and Arsenin 1977, Wahba 1990, Girosi et al. 1995, Vapnik 1998, Hastie et al. 2003, and references therein) and more recently in the theoretical mathematics literature (see for example Cucker and Smale 2002).

RR estimates individual-level partworths for respondent i by minimizing a convex loss function with respect to $\mathbf{w}_i$. This loss function is parameterized by a positive weight γ that is typically set using cross-validation.

PROBLEM 1.

$$\min_{\mathbf{w}_i} \frac{1}{\gamma} \sum_{j=1}^{J} (y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i)^2 + \|\mathbf{w}_i\|^2, \qquad (1)$$

$$\gamma \text{ set by cross-validation.} \qquad (2)$$

The loss function $(1/\gamma) \sum_{j=1}^{J} (y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i)^2 + \|\mathbf{w}_i\|^2$ is composed of two parts. The first, $\sum_{j=1}^{J} (y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i)^2$, measures the fit between the estimated utilities and the observed ratings. For a *fixed* γ, this may be interpreted as the log of the likelihood corresponding to a normal error term with mean 0 and variance γ (see, for example, Hastie et al. 2003). The second part, $\mathbf{w}_i^\top \mathbf{w}_i = \|\mathbf{w}_i\|^2$, controls the *shrinkage* (or *complexity*) of the partworth solution $\mathbf{w}_i$ (Vapnik 1998, Cucker and Smale 2002, Hastie et al. 2003). The term "shrinkage" (Hastie et al. 2003) comes from the fact that we effectively "shrink" the partworths toward zero by penalizing deviations from zero ($\|\mathbf{w}_i\|^2$ may be viewed as

the distance between $\mathbf{w}_i$ and 0). The term "complexity control" (Vapnik 1998) comes from the fact that this essentially limits the set of possible estimates, making this set less complex (i.e., smaller). The positive parameter γ defines the trade-off between fit and shrinkage and is typically set using cross-validation (Wahba 1990, Efron and Tibshirani 1993, Shao 1993, Vapnik 1998, Hastie et al. 2003). We will provide a detailed description of cross-validation below, but let us already stress that *cross-validation does not use any out-of-sample data*.

We note that the RR loss function (1) can be generalized by replacing the square error $(y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i)^2$ with other error functions, hence retrieving other individual-based estimation methods—the loss function remains convex as long as the error function is convex. For example, for choice data we will use below the logistic error $-\log(e^{\mathbf{x}_{ijq_*}\mathbf{w}_i}/\sum_{q=1}^{Q} e^{\mathbf{x}_{ijq}\mathbf{w}_i})$ (where $\mathbf{x}_{ijq_*}$ represents the profile chosen by respondent i in choice j, which consists of Q alternatives $\mathbf{x}_{ijq}$, $q \in \{1, \ldots, Q\}$). Using the hinge loss $\theta(y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i)(y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i)$ (where $\theta(x) = 1$ if $x > 0$, and 0 otherwise) leads to the widely used method of Support Vector Machines (Vapnik 1998), introduced to marketing by Cui and Curry (2005) and Evgeniou et al. (2005a). Finally, note that the solution when $\gamma \to 0$ (hence removing the complexity control $\|\mathbf{w}_i\|^2$) converges to the OLS solution $\mathbf{w}_i = (X_i^\top X_i)^{-1}X_i^T Y_i$, where the pseudo-inverse is used instead of the inverse when $(X_i^\top X_i)$ is not invertible (Hastie et al. 2003).

2.1.3. Modeling Heterogeneity: Formulation of the Loss Function. We now extend the RR loss function to model consumer heterogeneity. Individual-level RR estimation does not pool information across respondents and involves minimizing a separate loss function for each respondent. Inspired by HB (Lenk et al. 1996, Allenby and Rossi 1999, Rossi and Allenby 2003, Rossi et al. 2005), we propose modeling heterogeneity and pooling information across respondents by shrinking the individual partworths toward the population mean.

In particular, we consider the following convex optimization problem (if D is not invertible, we replace D^{-1} with the pseudo-inverse of D—see Appendix A for details):

$$\min_{\{\mathbf{w}_i\}, \mathbf{w}_0, D} \quad \frac{1}{\gamma}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i)^2$$
$$+ \sum_{i=1}^{I}(\mathbf{w}_i - \mathbf{w}_0)^\top D^{-1}(\mathbf{w}_i - \mathbf{w}_0),$$

subject to D is a positive semidefinite matrix scaled to have trace 1. (3)

Let us note that this is *not* the complete method proposed, which includes the estimation of the positive

weight γ endogenously and is summarized in §2.1.5, Problem 2. Like the RR loss function (1), this loss function consists of two parts. The first part reflects fit and the second part shrinkage (complexity control). Unlike the individual-level RR loss function (1), the loss function (2) involves solving a single convex optimization problem and estimating all the partworths jointly. Moreover, instead of shrinking the partworths toward 0 as in individual-level RR, it shrinks them toward a vector $\mathbf{w}_0$ (as will be seen below, the value of $\mathbf{w}_0$ that minimizes the loss function is the the population mean) through $(\mathbf{w}_i - \mathbf{w}_0)^\top D^{-1}(\mathbf{w}_i - \mathbf{w}_0)$. Matrix D is related to the covariance matrix of the partworths (see Appendix A for details on the estimation of D based on calibration data), such that the shrinkage penalty is greater for partworths that are distant from the mean $\mathbf{w}_0$ along directions in which there is less variation across respondents. The parameter γ operates the same function as in individual-level RR, namely, achieving a proper trade-off between fit and shrinkage. Higher values of γ result in more homogenous estimates (i.e., more shrinkage). Notice that we scale D by fixing its trace, keeping the problem convex—otherwise, the optimal solution would be to simply set the elements of D to ∞ and to maximize only fit.

We consider next the minimization of the loss function (2) given γ, and in §2.1.5 the selection of γ using cross-validation. The complete method proposed is summarized in §2.1.5, Problem 2.

2.1.4. Modeling Heterogeneity: Minimization of the Loss Function Given γ. For a fixed γ, the loss function (2) is jointly convex with respect to the $\mathbf{w}_i$s, $\mathbf{w}_0$, and matrix D.[2] Hence one can use any convex optimization method (see, for example, Boyd and Vandenberghe 2004) to minimize it.

We choose to solve the first-order conditions directly, which reveals some similarities with HB that will be discussed in §3. For a given value of γ we use the following iterative method to find the global optimal solution, initializing D to a random positive definite matrix:

(1) Solve the first-order conditions for $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$ given γ and D.

(2) Solve the first-order conditions for D given $\{\mathbf{w}_i\}$, $\mathbf{w}_0$, and γ.

In our empirical applications, convergence to a set of parameters ($\{\mathbf{w}_i\}$, $\mathbf{w}_0$, D) that minimizes the loss function (2) (i.e., solves the entire system of first-order conditions) for a given γ was always achieved in fewer than 20 iterations.

We show in Appendix A how to solve the above two steps *in closed form*. We show that the individual

---

[2] This can be seen, for example, from the Hessian, which is positive semidefinite.

partworths in Step 1 (for fixed γ and D – replacing inverses with pseudo-inverses if D is not invertible, as described in Appendix A) can be written as

$$\mathbf{w}_i = (X_i^\top X_i + \gamma D^{-1})^{-1} X_i^\top \mathbf{Y}_i + (X_i^\top X_i + \gamma D^{-1})^{-1} \gamma D^{-1} \mathbf{w}_0,$$
(4)

where the optimal $\mathbf{w}_0$ is shown to be the population mean of the partworths, that is, $\mathbf{w}_0 = (1/I) \sum_i \mathbf{w}_i$. We will see in §3 how this relates to the mean of the conditional posterior in HB.

2.1.5. Modeling Heterogeneity: Setting γ Using Cross-Validation. We now describe the estimation of the trade-off parameter γ. Selecting this parameter by minimizing the loss function (2) would be inappropriate because it would lead to $\gamma = \infty$ and all other parameters equal to 0. Instead, we select this parameter like in individual-level RR, by minimizing the cross-validation error. This standard technique has been empirically validated, and its theoretical properties have been extensively studied (see, for example, Wahba 1990, Efron and Tibshirani 1993, Shao 1993, Vapnik 1998, Hastie et al. 2003, and references therein). It is important to stress that *cross-validation does not require any data beyond the calibration data*. In particular, we measure the cross-validation error corresponding to a given parameter γ as follows:

(1) Set *cross-validation*$(\gamma) = 0$.

(2) For $k = 1$ to J:

(a) Consider the subset of the calibration data

$$Z^{(-k)} = \bigcup_{i=1}^{I} \left\{ \mathbf{x}_{i1}, \ldots, \mathbf{x}_{i(k-1)}, \mathbf{x}_{i(k+1)}, \ldots, \mathbf{x}_{iJ} \right\}.$$

That is, consider the subset of the calibration data that consists of all questions except the kth one for each of the I respondents.[3]

(b) Using only this subset of the calibration data $Z^{(-k)}$, estimate the individual partworths $\{\mathbf{w}_i^{(-k)}\}$, population mean $\mathbf{w}_0^{(-k)}$, and matrix $D^{(-k)}$ for the given γ using the method described in the previous section.

(c) Using the estimated partworths $\{\mathbf{w}_i^{(-k)}\}$, compute the ratings on the I questions (one per respondent) left out from the calibration data $\{\mathbf{x}_{1k}, \mathbf{x}_{2k}, \ldots, \mathbf{x}_{Ik}\}$, and let $CV(k)$ be the sum of squared differences between the estimated and observed ratings for these I calibration questions. (Note that any other performance metric may be used.)

(d) Set *cross-validation*$(\gamma) = $ *cross-validation*$(\gamma) + CV(k)$.

We simply select the parameter γ that minimizes the cross-validation error by using a line search.

The cross-validation error is, effectively, a "simulation" of the out-of-sample error *without* using any out-of-sample data. We refer the reader to the above references for details regarding its theoretical properties, such as its consistency for parameter selection.[4] We will later confirm empirically that selecting γ using cross-validation leads to values very close to optimal (i.e., maximizing estimation accuracy).

To summarize, the proposed method, which we label as RR-Het, is as follows:[5]

PROBLEM 2.

$$\gamma^* = \underset{\gamma}{\arg\min} \quad \text{cross-validation}(\gamma),$$

$$(\{\mathbf{w}_i^*\}, \mathbf{w}_0^*, D^*) = \underset{\{\mathbf{w}_i\}, \mathbf{w}_0, D}{\arg\min} \frac{1}{\gamma^*} \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i)^2$$
$$+ \sum_{i=1}^{I} (\mathbf{w}_i - \mathbf{w}_0)^\top D^{-1} (\mathbf{w}_i - \mathbf{w}_0),$$

subject to D is a positive semidefinite matrix

scaled to have trace 1.

It is important to note that if γ were set exogenously, RR-Het would be equivalent to maximum likelihood estimation (MLE), with the likelihood function proportional to the inverse of the exponential of the loss function (2)—multiplied by an indicator function that would enforce the constraints on D. However, *because γ is set using cross-validation and the overall estimation method is given by Problem 2 and not by the minimization of the loss function* (2), *the comparison of* RR-*Het with MLE is not straightforward*.

2.2. Choice-Based Conjoint Estimation Method

We now show how our approach can be used with choice data. Choice-based conjoint analysis has become very popular, both among practitioners and among academics (Carson et al. 1994, Louviere et al. 2000, Hauser and Toubia 2005). As discussed above, our choice-based method is developed by replacing the square-error loss in RR-Het with the logistic error, hence we call the proposed method LOG-Het. In particular, with choice data, the optimization problem

---

[3] Variations exist. For example, one can remove only one question in total from all I respondents and iterate $I \times J$ times instead of J-leading to the so-called leave-*one*-out cross-validation error—or more than one question per respondent. Our particular choice was driven by computational simplicity.

[4] We say that parameter selection is consistent if the probability of selecting the parameter with optimal out-of-sample performance converges to 1 as the amount of calibration data increases.

[5] A matlab version of the code, for the metric and choice formats, is available from the authors.

solved to estimate the partworths becomes:

PROBLEM 3.

$$\gamma^* = \arg\min_{\gamma} \; cross\text{-}validation(\gamma),$$

$$(\{\mathbf{w}_i^*\}, \mathbf{w}_0^*, D^*) = \arg\min_{\{\mathbf{w}_i\}, \mathbf{w}_0, D} -\frac{1}{\gamma^*}\sum_{i=1}^{I}\sum_{j=1}^{J}\log\frac{e^{\mathbf{x}_{ijq_*}\mathbf{w}_i}}{\sum_{q=1}^{Q}e^{\mathbf{x}_{ijq}\mathbf{w}_i}}$$
$$+\sum_{i=1}^{I}(\mathbf{w}_i-\mathbf{w}_0)^{\top}D^{-1}(\mathbf{w}_i-\mathbf{w}_0),$$

subject to D is a positive semidefinite matrix scaled to have trace 1,

where $\mathbf{x}_{ijq}$ is the qth alternative presented to respondent i in question j, and $\mathbf{x}_{ijq_*}$ is the chosen alternative. (Question j for respondent i consists of choosing among the Q profiles $(\mathbf{x}_{ijq})_{\{q=1,\dots,Q\}}$.) The parameter J represents the number of choice questions and Q the number of alternatives per question (they do not need to be constant across respondents or questions). Cross-validation for estimating parameter γ is implemented as for RR-Het, with the difference that the cross-validation performance in Step (2c) is now measured by the logistic error $-\log(e^{\mathbf{x}_{ijq_*}\mathbf{w}_i}/\sum_{q=1}^{Q}e^{\mathbf{x}_{ijq}\mathbf{w}_i})$ on the left out questions. The other major difference from RR-Het is that the minimization of the loss function given γ and D may no longer be performed by solving the first order conditions directly. Instead, we use Newton's method (see Appendix B for details and references to other possible estimation methods). As a result, unlike RR-Het, we do *not* have closed-form solutions for the conditional partworth estimates for LOG-Het.

In the previous section, we presented an approach for modeling consumer heterogeneity in conjoint estimation and showed how this approach can be used with both metric and choice data. In the following section, we highlight some theoretical similarities and differences between our approach and hierarchical Bayes.

3. Theoretical Similarities and Differences with HB

We consider the following hierarchical Bayes model for metric data (we assume a standard diffuse prior on $\mathbf{w}_0$, symbolically equivalent to $\mathbf{w}_0 \sim N(0, V^{-1})$ with $V = 0$):

Likelihood: $y_{ij} = \mathbf{x}_{ij}\mathbf{w}_i + \epsilon_{ij}$
$\epsilon_{ij} \sim N(0, \sigma^2);$

First-stage prior: $\mathbf{w}_i \sim N(\mathbf{w}_0, D);$

Second-stage priors: $\sigma^2 \sim IG(r_0/2, s_0/2)$
$D^{-1} \sim W(\eta_0, \eta_0 \times \Delta_0).$

Table 1 Some Characteristics of HB vs. RR-Het and LOG-Het

HB	R-Het and LOG-Het
Shrinks toward the mean of the first-stage prior	Shrink toward the population mean
Samples from posterior distribution	Minimize a convex loss function
Posterior distribution is a function of parameters of the second-stage priors	Loss function is a function of the trade-off parameter γ
The parameters of the second-stage priors are set exogenously	γ is determined endogenously using cross-validation

We consider the following HB model for choice data (again assuming a standard diffuse prior on $\mathbf{w}_0$):

Likelihood: $\text{Prob}(\mathbf{x}_{ijq_*} \text{ is chosen}) = \dfrac{e^{\mathbf{x}_{ijq_*}\mathbf{w}_i}}{\sum_{q=1}^{Q}e^{\mathbf{x}_{ijq}\mathbf{w}_i}};$

First-stage prior: $\mathbf{w}_i \sim N(\mathbf{w}_0, D);$

Second-stage prior: $D^{-1} \sim W(\eta_0, \eta_0 \times \Delta_0).$

Our standard HB implementations, throughout the rest of the paper, follow the literature and use fairly diffuse second-stage priors (see, for example, Allenby and Rossi 1999, Rossi and Allenby 2003): $\eta_0 = p + 3$, $\Delta_0 = I$ for metric and choice HB, and $r_0 = s_0 = 1$ for metric HB.

Table 1 summarizes some key characteristics of HB and the proposed approach.

3.1. Similarities
The main similarity between the proposed approach and HB is that they both shrink individual estimates toward a population mean. With metric data, the existence of closed-form expressions enables us to clearly identify the individual-specific estimates, the population means toward which these estimates are shrunk, and the shrinkage weights. Such explicit comparisons are not readily available with choice data.

In particular, the mean of the conditional posterior distribution of $\mathbf{w}_i$ in metric HB is (see Lenk et al. 1996 for details):[6]

$$E(\mathbf{w}_i \mid \mathbf{w}_0, \sigma, D, data) = (X_i^{\top}X_i + \sigma^2 D^{-1})^{-1}X_i^{\top}\mathbf{Y}_i$$
$$+ (X_i^{\top}X_i + \sigma^2 D^{-1})^{-1}(\sigma^2 D^{-1}\mathbf{w}_0).$$

Compare this expression to the minimizers of the RR-Het loss function (2) given γ and D (see Equation (3)):

$$\mathbf{w}_i = (X_i^{\top}X_i + \gamma D^{-1})^{-1}X_i^{\top}\mathbf{Y}_i$$
$$+ (X_i^{\top}X_i + \gamma D^{-1})^{-1}(\gamma D^{-1}\mathbf{w}_0).$$

---

[6] In Bayesian decision theory, the optimal point estimate corresponding to a quadratic loss function (or to the loss function used to compute RMSE) is the mean of the posterior (Chaloner and Verdinelli 1995, Rossi and Allenby 2003).

These expressions can also be written as follows (if the matrix X_i is not full rank, for the sake of this argument use the pseudo-inverse instead of the inverse):

HB:

$$E(\mathbf{w}_i|\mathbf{w}_0, \sigma, D, data)$$
$$= [(X_i^\top X_i + \sigma^2 D^{-1})^{-1}(X_i^\top X_i)]((X_i^\top X_i)^{-1} X_i^\top \mathbf{Y}_i)$$
$$+ [(X_i^\top X_i + \sigma^2 D^{-1})^{-1}(\sigma^2 D^{-1})]\mathbf{w}_0$$
$$= \alpha_{\text{HB}}^{(i)}((X_i^\top X_i)^{-1} X_i^\top \mathbf{Y}_i) + (I - \alpha_{\text{HB}}^{(i)})\mathbf{w}_0;$$

RR-Het:

$$\mathbf{w}_i = [(X_i^\top X_i + \gamma D^{-1})^{-1}(X_i^\top X_i)]((X_i^\top X_i)^{-1} X_i^\top \mathbf{Y}_i)$$
$$+ [(X_i^\top X_i + \gamma D^{-1})^{-1}(\gamma D^{-1})]\mathbf{w}_0$$
$$= \alpha_{\text{RR}}^{(i)}((X_i^\top X_i)^{-1} X_i^\top \mathbf{Y}_i) + (I - \alpha_{\text{RR}}^{(i)})\mathbf{w}_0,$$

where $\alpha_{\text{HB}}^{(i)} = (X_i^\top X_i + \sigma^2 D^{-1})^{-1}(X_i^\top X_i)$ and $\alpha_{\text{RR}}^{(i)} = (X_i^\top X_i + \gamma D^{-1})^{-1}(X_i^\top X_i)$. These expressions show clearly that the mean of the conditional posterior in HB and the point estimate in RR-Het are both weighted averages between the individual-level OLS estimate $(X_i^\top X_i)^{-1} X_i^\top \mathbf{Y}_i$ and a population mean (in RR-Het, $\mathbf{w}_0$ is equal to the population mean; in HB $\mathbf{w}_0$ is the mean of the first-stage prior on the partworths; and if we assume a diffuse prior on $\mathbf{w}_0$, then the mean of the conditional posterior distribution on $\mathbf{w}_0$ is the population mean). The individual-specific weights (i.e., amounts of shrinkage) are a function of $\sigma^2 D^{-1}$ in HB and of γD^{-1} in RR-Het. The mean of the *full* posterior distribution of $\mathbf{w}_i$ in HB is also a weighted average between the OLS estimate and a population mean, the weights being given by integrating $\alpha_{\text{HB}}^{(i)}$ over the posterior distributions of σ and D.

Note that if the parameters η_0, Δ_0, r_0, and s_0 in HB were selected to yield a strong prior on σ^2 and D around the values of γ and D obtained by RR-Het estimation, the posterior means provided by HB would converge to the point estimates provided by RR-Het ($\alpha_{\text{HB}}^{(i)} \to \alpha_{\text{RR}}^{(i)}$). Hence, in theory the set of point estimates achievable by RR-Het is a subset of those achievable by HB by varying the parameters of the second-stage priors. Therefore, the maximum *potential* performance achievable by HB is at least that achievable by RR-Het. However this does not guarantee higher performance *in practice*. In particular, any poorer performance observed for HB can be attributed to a suboptimal selection of the second-stage prior parameters. We will suggest later that endogenizing the selection of these parameters, although it raises a number of issues that we will discuss, has the potential to improve performance.

3.2. Differences

Two important differences emerge from Table 1. First, HB samples from a posterior distribution while RR-Het and LOG-Het minimize a loss function and hence only produce point estimates. Confidence intervals and hypothesis testing are also possible with RR-Het and LOG-Het using, for example, bootstrapping (Efron and Tibshirani 1993 and references therein). See the online technical appendix for a brief review and an example.

Second, while the posterior in HB is a function of a set of *exogenous* parameters (the parameters of the second-stage priors, η_0, Δ_0, r_0, s_0 in the case of metric data and η_0 and Δ_0 in the case of choice data), the loss functions in RR-Het and LOG-Het are a function of an *endogenous* parameter γ (determined from the calibration data using cross-validation). The difference between the way the second-stage priors are set in HB and γ is set in RR-Het and LOG-Het translates into differences in the way the amount of shrinkage is determined, as will be confirmed empirically in §6. For example, in the case of metric data, shrinkage is a function of $\sigma^2 D^{-1}$ in HB and γD^{-1} in RR-Het. In HB, the posterior distributions on σ and D are influenced both by the data and by the second-stage priors $\sigma^2 \sim IG(r_0/2, s_0/2)$ and $D^{-1} \sim W(\eta_0, \eta_0 \times \Delta_0)$. The exogenous parameters η_0, Δ_0, r_0, and s_0 are often selected to induce fairly diffuse and uninformative second-stage priors. Other values could yield different second-stage priors, resulting in different amounts of shrinkage. For example, strong priors around the "true" values of σ and D would clearly lead to maximal estimation accuracy. While such an extreme case can be studied hypothetically using simulations, in field settings where the truth is unknown, one typically has to revert to fairly diffuse second-stage priors. On the other hand, in RR-Het (respectively LOG-Het), the amount of shrinkage is a function of endogeneous parameters determined by the minimization of the loss function and by cross-validation: D and γ are obtained by solving Problem 2 (respectively, Problem 3).

It is important to note that this second difference is not intrinsic, and that the second-stage prior parameters in HB could be set in practice endogenously—for example, using cross-validation as well. The systematic incorporation of cross-validation in a Bayesian framework raises several issues and is beyond the scope of this paper. We discuss these issues briefly in the next section and demonstrate the *potential* of this approach empirically in §4.

3.3. Using Cross-Validation to Select the Parameters of the Second-Stage Priors in HB

Our empirical comparisons will suggest that our approach usually significantly outperforms a standard HB implementation (with fairly diffuse second-stage priors). However, such a comparison might be perceived as unfair because the posterior in HB is

a function of exogenous parameters, while the loss function in our approach is a function of an endogenous parameter set using cross-validation.[7] It seems reasonable to hypothesize that selecting the parameters of the second-stage priors in HB using cross-validation could yield a performance level comparable to RR-Het and LOG-Het. For example, we have shown above that the set of point estimates achievable by RR-Het by changing γ is a subset of those achievable by HB by changing η_0, Δ_0, r_0, and s_0. However, let us first note that the fact that the set of point estimates achievable by RR-Het is a subset of those achievable by HB *does not* guarantee that endogenously selecting the second-stage priors will improve performance relative to RR-Het. For example, because the number of parameters of the second-stage priors in HB is much larger than the number of parameters set using cross-validation in RR-Het or LOG-Het ($p^2 + 3$ versus 1 in the metric case and $p^2 + 1$ versus 1 in the choice case), there is a risk of overfitting.

Moreover, at least three potential issues arise regarding the use of cross-validation to select the parameters of the second-stage priors in HB.

First, Bayesian analysis obeys the likelihood principle (Fisher 1922, Rossi and Allenby 2003, Liu et al. 2007), which states that all the information relevant for inference is contained in the likelihood function. It is not clear whether cross-validation satisfies this principle, as it appears that the data are used both to set some parameters and to make some inference based on these parameters. It might be possible to construct an alternative HB specification that would include cross-validation, i.e., cross-validation and estimation would be part of the same comprehensive model and the likelihood principle would be satisfied (to the best of our knowledge, this is an open problem). At this point we are agnostic on whether cross-validation can be justified in a Bayesian framework. Our goal in this paper is only to explore whether it has the potential to improve the predictive performance of HB and not to justify its use theoretically, which we leave for future research.

Second, a practical issue arises due to the number of parameters of the second-stage priors in HB. Indeed, in the case of metric data the number of parameters is $p^2 + 3$, and in the case of choice data it is $p^2 + 1$. Setting the values of all these parameters directly using cross-validation in a hierarchical Bayes framework would be intractable in most practical applications given the set of candidate values.

Third, another practical issue arises from the fact that the computation of the cross-validation error

associated with each set of values of the second-stage prior parameters usually requires sampling from the corresponding posterior distribution to obtain point estimates. This is, again, a computational issue given the set of candidate parameter values.

We hope that future research will address these two practical issues. In this paper we are able to assess the *potential* of using cross-validation in Bayesian estimation by considering a simpler, *nonhierarchical*, *metric* model with only one hyperparameter (therefore avoiding the first practical issue) and by taking advantage of the existence of closed form expressions for the posterior means in the *metric* case (therefore avoiding the second practical issue).

In particular, we first run metric HB with standard second-stage priors to obtain initial point estimates for $\mathbf{w}_0$ and D, and we then consider the following simple (nonhierarchical) model:

$$\text{Likelihood:} \quad y_{ij} = \mathbf{x}_{ij}\mathbf{w}_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim N(0, \sigma_0^2),$$
$$\text{First-stage prior:} \quad \mathbf{w}_i \sim N(\mathbf{w}_0, D),$$

where σ_0 is a parameter set using cross-validation. This specification is a special case of the metric HB specification in which $\Delta_0 = D$, $\eta_0 \to \infty$, $s_0 = r_0 \times \sigma_0$, and $r_0 \to \infty$. The full posterior mean of $\mathbf{w}_i$ has the same expression as the conditional mean in the general model:

$$E(\mathbf{w}_i \mid data) = (X_i^\top X_i + \sigma_0^2 D^{-1})^{-1} X_i^\top \mathbf{Y}_i$$
$$+ (X_i^\top X_i + \sigma_0^2 D^{-1})^{-1}(\sigma_0^2 D^{-1}\mathbf{w}_0).$$

Because the full posterior mean of $\mathbf{w}_i$ is given in closed form, there is no need to sample from the posterior to obtain point estimates, and the cross-validation error associated with a given value of σ_0 can be estimated conveniently fast. Note that varying σ_0 directly varies the amount of shrinkage characterized by $\sigma_0^2 D^{-1}$. Note also that unlike in RR-Het, $\mathbf{w}_0$ and D are fixed here. We label this model Metric Bayes-CV.[8]

Unfortunately, such closed-form expressions are available only for metric data and not for choice data. Hence, we are unable to test an equivalent model for choice (note that the second practical problem would remain even if we were able to address the first).

In the previous section, we showed that although both our approach and hierarchical Bayes shrink individual-level estimates toward a population mean, the amount of shrinkage is partly exogenous in HB while it is completely endogenous in our approach. Endogenizing the amount of shrinkage in HB (by endogenizing the selection of the second-stage prior

---

[7] Note, however, that our approach does *not* use any additional data compared to HB: all methods only use the calibration data and use the *same* calibration data.

[8] This model is in the spirit of the empirical Bayes approach of Rossi and Allenby (1993), to the extent that $\mathbf{w}_0$ and D are based on a preliminary analysis of the data. However, Rossi and Allenby do not use cross-validation.

parameters) raises some theoretical and practical issues. Despite these issues, we explore the *potential* of such modification with a simple, nonhierarchical, metric Bayesian model. In the following two sections, we compare the estimation accuracy and predictive performance of our approach to that of hierarchical Bayes.

4. Simulation Experiments

We first compare our approach with HB both for metric and choice data using simulations. We compare the methods using two field data sets (one for ratings and one for choice) in §5.

4.1. Metric-Based Simulations

We compare RR-Het to the following methods:

(1) A standard HB implementation using typical values for the parameters of the second-stage priors (resulting in fairly diffuse second-stage priors): $\eta_0 = p + 3$, $\Delta_0 = I$, $r_0 = s_0 = 1$.

(2) The Metric Bayes-CV method described above.

We used a 2 (low versus high heterogeneity) $\times$ 2 (low versus high response error) $\times$ 2 (low versus high number of questions) simulation design. We simulated ratings-based conjoint questionnaires with 10 binary features (plus an intercept). The true partworths were drawn from $\mathbf{w}_i \sim N(\mathbf{w}_0, \sigma_w \times I)$ where $\mathbf{w}_0 = [5, 5, \ldots, 5]$ and $\sigma_w = 2$ in the low-heterogeneity case and $\sigma_w = 4$ in the high-heterogeneity case. The profiles were obtained from an orthogonal and balanced design with 16 profiles, and the ratings were equal to $y_{ij} = \mathbf{x}_{ij}\mathbf{w}_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma_e)$ with $\sigma_e = 2$ in the low-response error case and $\sigma_e = 4$ in the high-response error case. In the low-number-of-questions conditions, 8 profiles were drawn randomly without replacement from the orthogonal design for each simulated respondent. In the high-number-of-questions conditions, all 16 profiles were rated by each simulated respondent. We simulated 5 sets of 100 respondents in each condition, estimation being performed separately for each set. Our performance metric was the root mean square error (RMSE) between the estimated and true partworths.

We note that the model used to generate the data follows the distributional assumptions of HB. If strong second-stage priors around the true values of σ and D were used, then we would clearly expect HB to perform best. We focus on a more realistic and practical setting in which no prior information on the values of σ and D is available to either method.

Table 2 reports the average RMSE across respondents in each magnitude $\times$ heterogeneity $\times$ number of questions cell.

We see the following:

(1) RR-Het performs significantly better than standard HB in 7 out of 8 conditions. Overall, it is best or nonsignificantly different from best (at $p < 0.05$) in 7 out of 8 conditions.

Table 2 RMSE (Lower Numbers Indicate Higher Performance) of Estimated vs. True Partworths for the Metric-Based Simulations

Heterogeneity	Response error	Questions	Standard HB	Metric Bayes-CV	RR-Het
Low	Low	8	1.502	**1.453**	**1.459**
		16	0.989	0.941	**0.920**
Low	High	8	1.751	**1.736**	1.861
		16	1.485	**1.414**	**1.417**
High	Low	8	3.189	2.479	**2.358**
		16	1.026	1.005	**0.993**
High	High	8	3.363	2.909	**2.839**
		16	2.465	1.898	**1.834**

Notes. Bold numbers in each row indicate best or not significantly different from best at the $p < 0.05$ level. The proposed method, RR-Het, is significantly better than standard HB in 7 out of 8 conditions. It is overall best or nonsignificantly different from best (at $p < 0.05$) in 7 out of 8 conditions.

(2) Metric Bayes-CV performs significantly better than standard HB in all 8 conditions (these significance tests are *not* reported in the table). This suggests that selecting the parameters of the second-stage priors in HB using cross-validation has the potential to greatly improve predictive ability.

4.2. Choice-Based Simulations

We extended the simulation setup above to compare choice HB to LOG-Het. As discussed in §3, we were unable to test a choice version of the Metric Bayes-CV method. We used again a 2 (low versus high heterogeneity) $\times$ 2 (low versus high response error) $\times$ 2 (low versus high number of questions) design, assumed 10 binary features, and used 8 and 16 as our low and high numbers of questions. We assumed two profiles per choice set and derived our orthogonal design by applying the shifting method of Bunch et al. (1994) (see also Huber and Zwerina 1996, Arora and Huber 2001) to the orthogonal design used for the metric simulations (if X_i is the ith row of the effects-coded orthogonal design, then choice i is between X_i and $1 - X_i$). Following the tradition of choice-based conjoint simulations (Arora and Huber 2001, Evgeniou et al. 2005a, Toubia et al. 2004), we drew the true partworths from normal distributions with mean $[mag, mag, \ldots, mag]$ and variance $\sigma^2 = het \times mag$ where the parameter *mag* controls the amount of response error and the paramater *het* the amount of heterogeneity. We set the parameters *mag* and *het* to capture the range of response error and heterogeneity used in the previous simulations in the aforementioned studies. In particular, we set *mag* = 1.2 and 0.2, respectively in the low and high response error conditions,[9] and *het* = 1 and 3, respectively, in

---

[9] Because our number of features (10) is 2.5 times the number (4) used by previously published simulations using the same

Table 3 RMSE (Lower Numbers Indicate Higher Performance) of Estimated vs. True Partworths for the Choice Simulations

Het	Response error	Quest	HB	LOG-Het
Low	Low	8	**0.5933**	0.6235
		16	**0.4486**	0.4600
Low	High	8	0.9740	**0.9609**
		16	0.8050	**0.7946**
High	Low	8	0.7389	**0.7289**
		16	0.4970	**0.4827**
High	High	8	0.9152	**0.9013**
		16	0.6935	**0.6878**

Notes. LOG-Het is the proposed method, HB is hierarchical Bayes. Bold numbers in each row indicate best or not significantly different from best at the $p < 0.05$ level. LOG-Het performs significantly better than HB in 6 out of 8 conditions.

Table 4 RMSE for Holdout Questions from the Metric Field Data of Lenk et al. (1996) (Lower Numbers Indicate Higher Performance)

Questions	Standard HB	Metric Bayes-CV	RR-Het
8	1.905	1.851	**1.794**
16	1.667	**1.610**	**1.608**

Notes. Bold numbers in each row indicate best or not significantly different from best at the $p < 0.05$ level. Both RR-Het and metric Bayes-CV perform significantly better than standard HB with both 8 and 16 questions. RR-Het performs overall best or nonsignificantly different from best with both 8 and 16 questions.

the low and high heterogeneity conditions. We used logistic probabilities to simulate the answers to the choice questions. We measure performance using the RMSE between the true and estimated partworths, normalized to have a norm of 1.

The results of the simulations (based on 5 sets of 100 respondents) are summarized in Table 3. We see that the proposed method LOG-Het performs significantly better than standard HB in 6 out of 8 conditions.[10] HB outperforms LOG-Het only in the case of low response error and low heterogeneity.

5. Comparisons Based on Field Data

5.1. Comparison of the Metric-Based Methods Using Field Data

We compared RR-Het, HB, and Metric Bayes-CV on a field data set used in a previously published paper (Lenk et al. 1996).[11] The data come from a ratings-based conjoint study on computers, with 180 consumers rating 20 profiles each. The first 16 profiles form an orthogonal and balanced design and are used for calibration; the last 4 are holdouts used for validation. The independent variables are 13 binary attributes and an intercept (see Table 2 in Lenk et al. 1996 for a description). The dependent variable is a rating on an 11-point scale (0 to 10). We measured performance using the root mean square error (RMSE) between the observed and predicted holdout ratings. We estimated the partworths using 8 (randomly selected) and 16 questions.

We report the results in Table 4. Both RR-Het and Metric Bayes-CV perform significantly better than

standard HB with both 8 and 16 questions. RR-Het performs overall best or nonsignificantly different from best with both 8 and 16 questions. This further confirms the potential of RR-Het, as well as the potential of using cross-validation in Bayesian estimation. Note that our numbers are comparable but not equal to the ones reported by Lenk et al. for the following reasons. First, to perform significance tests, we compute the RMSE for each respondent and report the averages across respondents, as opposed to computing an aggregate metric as in Lenk et al. Second, we assume homoskedasticity (same σ for all respondents). Third, we do not use demographic variables in the model. We show in the online technical appendix how RR-Het can be extended to include such covariates, and compare the performance of this extension to that of HB with covariates and Metric Bayes-CV with covariates. The same conclusions apply.

5.2. Comparison of the Choice-Based Methods Using Field Data

We compared LOG-Het to HB on an empirical conjoint data set kindly made available to us by Research International.[12] Note that we were not involved in the design of the conjoint study that led to this data set.

The product in this study was carbonated soft drinks. Three attributes were included: brand (6 levels), size (7 levels), and price (7 levels), for a total of 20 partworths per respondent. A pseudo-orthogonal design was first generated with 76 choice tasks, each involving 8 alternatives. This design was divided into 4 subsets of 18 questions plus 4 additional questions. There were 192 respondents subjected to 1 of the 4 22-question sets (presented in a randomized order). We used 8 (randomly selected from the first 16) or 16 questions to estimate the models, and the last 6 as holdouts.

We compare performance in Table 5. LOG-Het is not significantly different from HB with 8 questions and significantly better with 16 questions. As a reference, a homogeneous estimate obtained by logistic regression achieved a hit-rate error of 21.7% (note

---

simulation design, we divide the values of typical *mag* parameters used in previously published simulations (0.5 and 3) by 2.5 to make the overall utilities, and hence the level of response error, comparable.

[10] Note that the numbers from Tables 2 and 3 are not comparable because they are not on the same scale.

[11] We thank Peter Lenk for kindly sharing this data set with us.

[12] The data are proprietary but are available from the authors and Research International upon request.

Table 5 Holdout Hit Rates (Higher Numbers Indicate Higher Performance) from the Choice Field Data Set

Questions	Standard HB (%)	LOG-Het (%)
8	**48.37**	47.76
16	51.04	**52.34**

Notes. LOG-Het is the proposed method, HB is hierarchical Bayes. Bold numbers in each row indicate best or not significantly different from best at the $p < 0.05$ level. LOG-Het performs overall best or nonsignificantly different from best in both cases, and significantly better than HB with 16 questions.

that, as each question involved 8 products, a random model would achieve a hit rate of 12.5%).

The empirical comparisons reported in the previous two sections indicate that our approach, overall, outperforms standard hierarchical Bayes (with relatively diffuse second-stage priors) and show that endogenizing the selection of the second-stage prior parameters in hierarchical Bayes has the potential to greatly improve estimation accuracy and predictive performance. In the following section, we further explore the relation between the amount of shrinkage and performance, and we assess the validity of using cross-validation for parameter selection.

6. The Relation Between Shrinkage and Estimation Accuracy

We have argued in §3 that RR-Het and LOG-Het differ from HB in the approach used to determine the parameters on which the posterior distribution (respectively, the loss function) depend (parameters of the second-stage priors exogenous in HB versus γ endogenously estimated using cross-validation in RR-Het and LOG-Het), and that these differences translate into differences in the amounts of shrinkage performed by the estimators. We have also argued, based on past literature, that cross-validation is an effective way of selecting the parameter γ on which the RR-Het and LOG-Het loss functions depend, and hypothesized that it could be an effective way of selecting the second-stage prior parameters on which the HB posterior distribution depends. This raises the following two sets of questions, which we address empirically:

1. What is the relation between the amount of shrinkage and performance? Are differences in performance between methods systematically coupled with differences in the amount of shrinkage?

2. Does cross-validation in RR-Het, LOG-Het, and Metric Bayes-CV yield parameter values (γ and σ_0 respectively) close to the ones that maximize estimation accuracy?

We addressed these questions both with metric and choice data. We report the case of metric data here because of the availability of Metric Bayes-CV. The

conclusions with choice data are identical—details and graphs are available from the authors.

To explore the relation between shrinkage and performance, we manually varied the parameters γ and σ_0 in RR-Het and Metric Bayes-CV and assessed the corresponding performance. See Figure 1 for the simulations and Figure 2 for the field data (we report only the graphs based on 16 questions. The graphs based on 8 questions yield similar results and are available from the authors). The parameters γ and σ_0 are not on the same scale; however, there is a one-to-one mapping between each of these parameters and the amount of shrinkage. Hence, we report the realized amount of shrinkage on the x-axis, measured by $\sum_{i=1}^{I} \|\mathbf{w}_i - \mathbf{w}_0\|^2 / I$. Performance, measured by the RMSE of the true versus estimated partworths for the simulations and by the holdout RMSE for the field data (as in Tables 2 and 4), is reported on the y-axis. The solid and dotted curves represent the amount of shrinkage and the corresponding performance achieved respectively by Metric Bayes-CV and RR-Het as σ_0 (respectively γ) is varied. The labels "RR-Het" and "Bayes-CV" correspond to the amount of shrinkage and corresponding performance achieved by the two methods when γ and σ_0 are selected using cross-validation (i.e., they correspond to the numbers reported in Tables 2 and 4).[13] We also report the amount of shrinkage and performance achieved by standard HB.

Figures 1 and 2 illustrate the existence of a U-shaped relationship between the amount of shrinkage and performance. Moreover, they confirm that differences in performance between the different methods are systematically coupled with differences in the amount of shrinkage: the smaller the difference in the amount of shrinkage, the smaller the difference in performance. This confirms that the approach used to determine the amount of shrinkage can be viewed as a key difference between our approach and HB.

Finally, Figures 1 and 2 also suggest that the amount of shrinkage and performance achieved by RR-Het and Metric Bayes-CV when selecting parameters using cross-validation is close to the bottom of the corresponding curves, i.e., it is close to what would be achieved if the true partworths (or holdout ratings) were used to calibrate the parameters γ and σ_0. In particular, for the simulations (respectively ratings field data) the RMSE achieved by RR-Het or Metric Bayes-CV when γ or σ_0 is selected using cross-validation is on average only 0.59% (respectively, 0.38%)

---

[13] For each set of simulated respondents, the labels "Bayes-CV" and "RR-Het" lie exactly on the corresponding curves. However, this does not necessarily hold for our graphs because they are based on *averages* across the five sets of simulated respondents. Note also that the differences between the two curves are due to differences in D and $\mathbf{w}_0$.

Figure 1   Performance as a Function of the Amount of Shrinkage—Metric Simulated Data
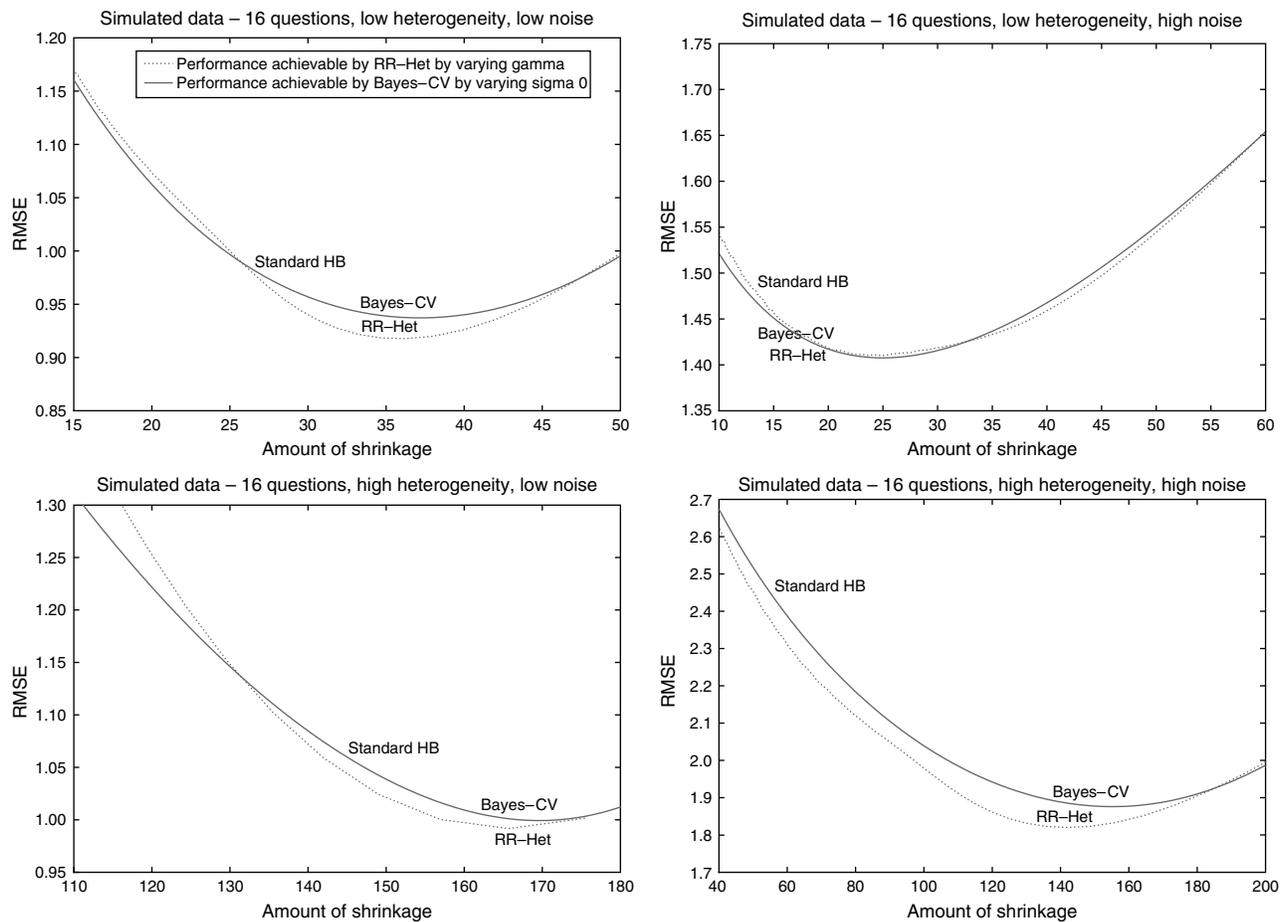
Notes. Estimates are based on 16 questions. The amount of shrinkage is measured by $\sum_{i=1}^{I} \|\mathbf{w}_i - \mathbf{w}_0\|^2 / I$. The solid lines represent the amount of shrinkage and corresponding RMSE (estimated versus actual partworths) performance achieved by metric Bayes-CV as σ_0 is varied, and the labels "Bayes-CV" represent the amount of shrinkage and performance achieved when σ_0 is selected using cross-validation. The dotted lines represent the amount of shrinkage and performance achieved by RR-Het as γ is varied, and the labels "RR-Het" represent the amount of shrinkage and performance achieved when γ is selected using cross-validation. "Standard HB" corresponds to HB with standard second-stage priors, as in Table 2.

higher than the minimum achievable if the true partworths (respectively, holdout ratings) were used to select γ and σ_0. This confirms that cross-validation is an effective method for parameter selection, both for RR-Het and Metric Bayes-CV, and hence potentially for all the second-stage prior parameters in HB.

7.   Conclusions and Future Research

Our main results can be summarized as follows.

• We have proposed a novel approach for handling consumer heterogeneity in conjoint estimation based on convex optimization and machine learning, and we applied it to both metric and choice data (§2).

• This approach shares some similarities with hierarchical Bayes. However, one of the major differences is that while the amount of shrinkage is influenced by a set of exogenous parameters in HB (the parameters of the second-stage priors), it is completely endogenous in our approach (§3).

• Simulations, as well as two empirical data sets, suggest that the approach overall outperforms a standard HB implementation (with relatively diffuse second-stage priors) (§§4 and 5).

• Selecting the second-stage prior parameters in HB endogenously, like in our approach, raises some practical and theoretical issues. However, we show the *potential* of this modification with a simple, metric, nonhierarchical model (§§3, 4, and 5).

• There exists a U-shaped relation between amount of shrinkage and performance, and differences in performance can be traced to differences in the amounts of shrinkage. Selecting some of the shrinkage parameters using cross-validation gives rise to an amount of shrinkage that is close to optimal (§6).

The experimental results suggest that an important and challenging area for future research is to develop systematic and computationally efficient ways of selecting the parameters of the second-stage priors in HB more optimally. A second area for future

Figure 2   Performance as a Function of the Amount of Shrinkage—Field Metric Data
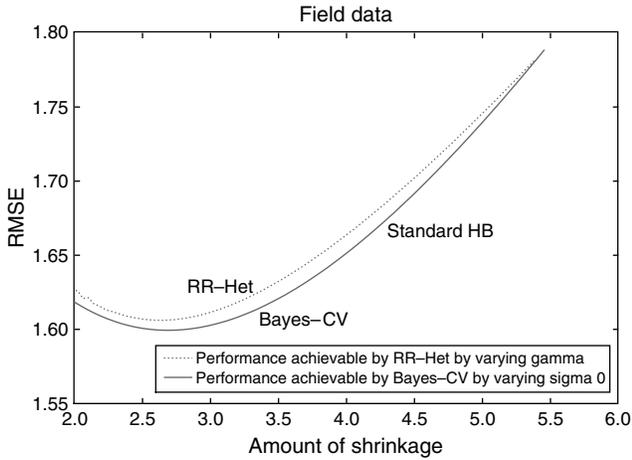
Notes. Estimates are based on 16 questions. The amount of shrinkage is measured by $\sum_{i=1}^{I} \|\mathbf{w}_i - \mathbf{w}_0\|^2 / I$. The solid line represents the amount of shrinkage and corresponding performance (holdout RMSE) achieved by metric Bayes-CV as σ_0 is varied, and the label "Bayes-CV" represents the amount of shrinkage and performance achieved when σ_0 is selected using cross-validation. The dotted line represents the amount of shrinkage and performance achieved by RR-Het as γ is varied, and the label "RR-Het" represents the amount of shrinkage and performance achieved when γ is selected using cross-validation. "Standard HB" corresponds to HB with standard second-stage priors, as in Table 4.

research would be to explore the use of population based complexity/shrinkage control in other individual level optimization based methods (e.g., Srinivasan and Shocker 1973, Srinivasan 1998, Toubia et al. 2003, Toubia et al. 2004), for estimation and possibly as well for adaptive questionnaire design. Third, in this paper we have focused on unimodal representations of heterogeneity. Future research could introduce and model segments of consumers. This could be achieved by modifying the form of the complexity control in loss function (2) to reflect, for example, the existence of multiple clusters of respondents. Finally, optimization and statistical learning methods could be used to capture and model other phenomena beyond consumer heterogeneity. For example, our methods could be extended to capture recently researched learning phenomena in conjoint analysis (Liechty et al. 2005, Bradlow et al. 2004). Another potential area of application is modeling the formation of consideration sets (Gilbride and Allenby 2004, 2006; Jedidi and Kohli 2005; Hauser et al. 2006).

Acknowledgments

Appendix A. Minimization of the RR-Het Loss Function (2) Given γ

A.1.   Estimating $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$ Given D

We first transform the data as $\tilde{\mathbf{x}}_{ij} = \mathbf{x}_{ij} D^{1/2}$ and define $\tilde{\mathbf{w}}_i = D^{-1/2}\mathbf{w}_i$ and $\tilde{\mathbf{w}}_0 = D^{-1/2}\mathbf{w}_0$ (see the case of a noninvertible D below). Note that with this transformation we can estimate first the $\tilde{\mathbf{w}}_i$'s and $\tilde{\mathbf{w}}_0$ using the transformed data $\tilde{\mathbf{x}}_{ij}$ and the modified cost function

$$\min_{\{\tilde{\mathbf{w}}_i\}, \tilde{\mathbf{w}}_0} \frac{1}{\gamma} \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \tilde{\mathbf{x}}_{ij} \tilde{\mathbf{w}}_i)^2 + \sum_{i=1}^{I} (\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0)^{\top} (\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0), \quad \text{(A.1)}$$

and then get the final solution as $\mathbf{w}_i = D^{1/2}\tilde{\mathbf{w}}_i$ and $\mathbf{w}_0 = D^{1/2}\tilde{\mathbf{w}}_0$. This is because $\tilde{\mathbf{x}}_{ij}\tilde{\mathbf{w}}_i = \mathbf{x}_{ij} D^{1/2} D^{-1/2}\mathbf{w}_i = \mathbf{x}_{ij}\mathbf{w}_i$ and $(\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0)^{\top}(\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0) = (\mathbf{w}_i - \mathbf{w}_0)^{\top}D^{-1}(\mathbf{w}_i - \mathbf{w}_0)$. With this transformation we never compute the inverse of matrix D.

Note that (A.1) is jointly convex with respect to the pair of variables $\{\tilde{\mathbf{w}}_i\}$ and $\tilde{\mathbf{w}}_0$. Taking the derivative with respect to $\tilde{\mathbf{w}}_0$ we see that

$$\tilde{\mathbf{w}}_0 = \frac{1}{I} \sum_{i=1}^{I} \tilde{\mathbf{w}}_i.$$

Taking the derivative with respect to $\tilde{\mathbf{w}}_i$ we have that

$$\frac{2}{\gamma} \tilde{X}_i^{\top} \tilde{X}_i \tilde{\mathbf{w}}_i - \frac{2}{\gamma} \tilde{X}_i^{\top} \mathbf{Y}_i + 2(\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0) = 0$$

$$\Rightarrow \tilde{\mathbf{w}}_i = (\tilde{X}_i^{\top} \tilde{X}_i + \gamma I_p)^{-1} \tilde{X}_i^{\top} \mathbf{Y}_i + (\tilde{X}_i^{\top} \tilde{X}_i + \gamma I_p)^{-1} \gamma \tilde{\mathbf{w}}_0$$

$$= \hat{\mathbf{w}}_i + \gamma Z_i \tilde{\mathbf{w}}_0, \quad \text{(A.2)}$$

where I_p is the p-dimensional identity matrix, $\tilde{X}_i$ is the matrix with rows $\tilde{\mathbf{x}}_{ij}$, $\hat{\mathbf{w}}_i = (\tilde{X}_i^{\top} \tilde{X}_i + \gamma I_p)^{-1} \tilde{X}_i^{\top} \mathbf{Y}_i$ and $Z_i = (\tilde{X}_i^{\top} \tilde{X}_i + \gamma I_p)^{-1}$. Finally, substituting $\tilde{\mathbf{w}}_i$ into the equation for $\tilde{\mathbf{w}}_0$ we get

$$\tilde{\mathbf{w}}_0 = \frac{1}{I} \sum_{i=1}^{I} (\hat{\mathbf{w}}_i + \gamma Z_i \tilde{\mathbf{w}}_0),$$

which implies

$$\tilde{\mathbf{w}}_0 = \left(I_p - \gamma \frac{1}{I} \sum_{i=1}^{I} Z_i\right)^{-1} \frac{1}{I} \sum_{i=1}^{I} \hat{\mathbf{w}}_i.$$

If the matrix $(I_p - \gamma(1/I_p) \sum_i Z_i)$ is not invertible, we follow the individual RR literature and take its pseudo-inverse. It can be shown, like in the individual-level RR case discussed in §2.1.2, that using the pseudo-inverse is equivalent to adding to the loss function (2) an extra term $\delta \mathbf{w}_0^{\top} D^{-1} \mathbf{w}_0$ with $\delta \to 0$.

Having estimated $\tilde{\mathbf{w}}_i$ and $\tilde{\mathbf{w}}_0$ we then get $\mathbf{w}_i = D^{1/2}\tilde{\mathbf{w}}_i$ and $\mathbf{w}_0 = D^{1/2}\tilde{\mathbf{w}}_0$. Finally, to get (4)—which we no not need to compute in practice—we just have to replace $\tilde{X}_i$ with $X_i D^{1/2}$ in (A.2) and use the fact that $\mathbf{w}_i = D^{1/2}\tilde{\mathbf{w}}_i$.

If D is not invertible, we replace $D^{-1/2}$ with the square root of the pseudo-inverse of D and follow the exact same computations above—note that we never have to compute D^{-1}. In this case, the projections on $D^{1/2}$ (computed using only the nonzero eigenvalues of D) above also ensure that $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$ are in the range of D – otherwise notice

that the complexity control can be set to 0 by simply considering $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$ in the null space of D. We can also get (4)—which we do not need to compute in practice again—with all inverses being pseudo-inverses by replacing again $\tilde{X}_i$ with $X_i D^{1/2}$ in (A.2) and use the fact that $\mathbf{w}_i = D^{1/2}\tilde{\mathbf{w}}_i$.

Note that we have closed-form solutions for both $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$. Moreover, the estimation of the partworths $\mathbf{w}_i$ is decomposed across the individuals and requires only $2I$ inversions of p-dimensional (small) matrices.

A.2. Estimating D Given $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$

We assume for simplicity that the covariance of the $\mathbf{w}_i$s, and hence the matrix $(\sum_{i=1}^{I}(\mathbf{w}_i - \mathbf{w}_0)(\mathbf{w}_i - \mathbf{w}_0)^\top)$, has full rank (which is typically the case in practice when we have many respondents). If the covariance matrix is not full rank, we replace the inverse of the solution D below with the pseudo-inverse. It can be shown as in the individual-level RR case discussed in §2.1.2, that using the pseudo-inverse is equivalent to adding to the loss function (2) the term $\epsilon \text{Trace}(D^{-1})$ with $\epsilon \to 0$, keeping the loss function convex.

Given $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$ we solve

$$\min_D \quad \sum_{i=1}^{I}(\mathbf{w}_i - \mathbf{w}_0)^\top D^{-1}(\mathbf{w}_i - \mathbf{w}_0),$$

subject to $\quad D$ is a positive semidefinite matrix scaled to have trace 1.

Using a Lagrange multiplier ρ for the trace constraint and taking the derivative with respect to D we have that

$$-\frac{1}{2}D^{-1}\left(\sum_{i=1}^{I}(\mathbf{w}_i - \mathbf{w}_0)(\mathbf{w}_i - \mathbf{w}_0)^\top\right)D^{-1} + \rho I = 0$$

$$\Rightarrow \quad D = \frac{1}{2\rho}\left(\sum_{i=1}^{I}(\mathbf{w}_i - \mathbf{w}_0)(\mathbf{w}_i - \mathbf{w}_0)^\top\right)^{1/2} \quad \text{(A.3)}$$

which is positive definite; ρ is simply selected so that D has trace 1.

Appendix B. Newton's Method for LOG-Het

Notice that for given $\{\mathbf{w}_i\}$ and D, assuming D is invertible (otherwise, as for RR-Het, use the pseudo-inverse of D) we get as before that $\mathbf{w}_0 = (1/I)\sum_i \mathbf{w}_i$. Similarly, given $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$ we can solve for D like for RR-Het above—because D appears only in the complexity control. Hence, we need only to show how to solve for $\{\mathbf{w}_i\}$ given D and $\mathbf{w}_0$, and then iterate among the conditional estimations like for RR-Het (in all our experiments, fewer than 20 iterations were required for convergence). As for RR-Het above, to avoid computing the inverse of D, we first transform the data as $\tilde{\mathbf{x}}_{ijq} = \mathbf{x}_{ijq} D^{1/2}$ and define $\tilde{\mathbf{w}}_i = D^{-1/2}\mathbf{w}_i$ and $\tilde{\mathbf{w}}_0 = D^{-1/2}\mathbf{w}_0$. Note that with this transformation we can estimate first $\tilde{\mathbf{w}}_i$ minimizing the modified cost function

$$-\frac{1}{\gamma^*}\sum_{i=1}^{I}\sum_{j=1}^{J}\log\frac{e^{\tilde{\mathbf{x}}_{ijq_*}\tilde{\mathbf{w}}_i}}{\sum_{q=1}^{Q}e^{\tilde{\mathbf{x}}_{ijq}\tilde{\mathbf{w}}_i}} + \sum_{i=1}^{I}(\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0)^\top(\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0) \quad \text{(B.1)}$$

and then get the final solution as $\mathbf{w}_i = D^{1/2}\tilde{\mathbf{w}}_i$.

Notice that for a fixed $\tilde{\mathbf{w}}_0$, problem (B.1) is decomposable into I separate subproblems, one for each respondent, each

of them being a standard (widely studied) regularized kernel logistic regression problem (Jaakkola and Haussler 1999, Hastie et al. 2003, Keerthi et al. 2005, Minka 2003, Zhu and Hastie 2005). We can solve (B.1) for $\tilde{\mathbf{w}}_i$ using various standard methods used for logistic regression (e.g., see Minka 2003). We use here a standard Newton's method implemented based on the matlab code of Minka (2003) available at http://research.microsoft.com/~minka/papers/logreg/. For this purpose we need only the gradient and Hessian of the loss function (B.1). These are given as

$$G = \sum_{j=1}^{J}\left(\tilde{\mathbf{x}}_{ijq_*}^\top - \frac{\sum_{q=1}^{Q}e^{\tilde{\mathbf{x}}_{ijq}\tilde{\mathbf{w}}_i}\tilde{\mathbf{x}}_{ijq}^\top}{\sum_{q=1}^{Q}e^{\tilde{\mathbf{x}}_{ijq}\tilde{\mathbf{w}}_i}}\right) + \gamma(\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0)$$

for the gradient and

$$H = \sum_{j=1}^{J}\sum_{q=1}^{Q}\left(-\frac{e^{\tilde{\mathbf{x}}_{ijq}\tilde{\mathbf{w}}_i}\tilde{\mathbf{x}}_{ijq}^\top\tilde{\mathbf{x}}_{ijq}}{\sum_{q'=1}^{Q}e^{\tilde{\mathbf{x}}_{ijq'}\tilde{\mathbf{w}}_i}}\right.$$
$$\left. + \frac{e^{\tilde{\mathbf{x}}_{ijq}\tilde{\mathbf{w}}_i}\tilde{\mathbf{x}}_{ijq}^\top\left(\sum_{q'=1}^{Q}e^{\tilde{\mathbf{x}}_{ijq'}\tilde{\mathbf{w}}_i}\tilde{\mathbf{x}}_{ijq'}\right)}{\left(\sum_{q'=1}^{Q}e^{\tilde{\mathbf{x}}_{ijq'}\tilde{\mathbf{w}}_i}\right)^2}\right) + \gamma I_p$$

for the Hessian. At each Newton step the new $\tilde{\mathbf{w}}_i$ (for each respondent i independently) is given by $\tilde{\mathbf{w}}_i^{\text{new}} = \tilde{\mathbf{w}}_i^{\text{old}} - H^{-1}G$.

References

Allenby, G. M., P. E. Rossi. 1999. Marketing models of consumer heterogeneity. *J. Econometrics* **89**(March/April) 57–78.

Ando, R. K., T. Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Machine Learn. Res.* **6** 1817–1853.

Arora, N., J. Huber. 2001. Improving parameter estimates and model prediction by aggregate customization in choice experiments. *J. Consumer Res.* **28**(September).

Baxter, J. 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learn.* **28** 7–39.

Boyd, S., L. Vandenberghe. 2004. *Convex Optimization.* Cambridge University Press, Oxford, UK.

Bradlow, E., Y. Hu, T.-H. Ho. 2004. A learning-based model for imputing missing levels in partial conjoint profiles. *J. Marketing Res.* **41**(4) 369–381.

Bunch, D. S., J. J. Louviere, D. Anderson. 1994. A comparison of experimental design strategies for multinominal logit models: The case of generic attributes. Working paper, Graduate School of Management, University of California at Davis.

Carson, R. T., J. J. Louviere, D. A. Anderson, P. Arabie, D. S. Bunch, D. A. Hensher, R. M. Johnson, W. F. Kuhfeld, D. Steinberg, J. Swait, H. Timmermans, J. B. Wiley. 1994. Experimental analysis of choice. *Marketing Lett.* **5**(4) 351–367.

Caruana, R. 1997. Multi-task learning. *Machine Learn.* **28** 41–75.

Chaloner, K., I. Verdinelli. 1995. Bayesian experimental design: A review. *Statistical Sci.* **10**(3) 273–304.

Cucker, F., S. Smale. 2002. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.* **39**(1) 1–49.

Cui, D., D. Curry. 2005. Prediction in marketing using the support vector machine. *Marketing Sci.* **24**(4) 595–615.

Efron, B., R. Tibshirani. 1993. *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Evgeniou, T., C. Boussios, G. Zacharia. 2005. Generalized robust conjoint estimation. *Marketing Sci.* **24**(3) 415–429.

Evgeniou, T., C. Micchelli, M. Pontil. 2005b. Learning multiple tasks with kernel methods. *J. Machine Learn. Res.* **6** 615–637.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc., Ser. A* 222–326.

Gilbride, T., G. M. Allenby. 2004. A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Sci.* **23**(3) 391–406.

Gilbride, T., G. M. Allenby. 2006. Estimating heterogenous EBA and economic screening rule choice models. *Marketing Sci.* **25**(5) 494–509.

Girosi, F., M. Jones, T. Poggio. 1995. Regularization theory and neural networks architectures. *Neural Comput.* **7** 219–269.

Hastie, T., R. Tibshirani, J. H. Friedman. 2003. *The Elements of Statistical Learning*. Springer Series in Statistics.

Hauser, J., O. Toubia. 2005. The impact of utility balance and endogeneity in conjoint analysis. *Marketing Sci.* **24**(3) 498–507.

Hauser, J., G. J. Tellis, A. Griffin. 2005. Research on innovation: A review and agenda for marketing science. *Marketing Sci.* **25**(6) 687–717.

Huber, J., K. Zwerina. 1996. The importance of utility balance in efficient choice designs. *J. Marketing Res.* **32**(August) 308–317.

Jebara, T. 2004. Multi-task feature and kernel selection for SVMs. *Proc. Twenty-First Internat. Conf. Machine Learning*, Banff, Alberta, Canada, 55–63.

Jaakkola, T., D. Haussler. 1999. Probabilistic kernel regression models. *Proc. Seventh Internat. Workshop on Artificial Intelligence and Statist.* Morgan Kaufmann, San Francisco, CA.

Jedidi, K., R. Kohli. 2005. Probabilistic subset-conjunctive models for heterogenous consumers. *J. Marketing Res.* **42** 483–494.

Keerthi, S., K. Duan, S. Shevade, A. Poo. 2005. A fast dual algorithm for kernel logistic regression. *Machine Learn.* **61**(1–3, November) 151–165.

Lenk, P. J., W. S. DeSarbo, P. E. Green, M. R. Young. 1996. Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Sci.* **15**(2) 173–91.

Liechty, J. C., D. K. H. Fong, W. S. DeSarbo. 2005. Dynamic models incorporating individual heterogeneity: Utility evolution in conjoint analysis. *Marketing Sci.* **24**(2) 285–293.

Liu, Q., T. Otter, G. M. Allenby. 2007. Investigating endogeneity bias in marketing. *Marketing Sci.* **26**(5) 642–650.

Louviere, J. J., D. A. Hensher, J. D. Swait. 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge University Press, New York.

Micchelli, C., M. Pontil. 2005. On learning vector–valued functions. *Neural Comput.* **17** 177–204.

Mika, S., B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, G. Rätsch. 1999. Kernel PCA and de-noising in feature spaces. M. S. Kearns, S. A. Solla, D. A. Cohn, eds. *Advances in Neural Information Processing Systems*, Vol. 11. MIT Press, Cambridge, MA, 536–542.

Minka, T. 2003. A comparison of numerical optimizers for logistic regression. Microsoft research tech report.

Rossi, P. E., G. M. Allenby. 1993. A Bayesian approach to estimating household parameters. *J. Marketing Res.* **30**(2) 171–182.

Rossi, P. E., G. M. Allenby. 2003. Bayesian statistics and marketing. *Marketing Sci.* **22**(3) 304–328.

Rossi, P. E., G. M. Allenby, R. McCulloch. 2005. *Bayesian Statistics and Marketing*. John Wiley and Sons, New York.

Shao, J. 1993. Linear model selection via cross-validation. *J. Amer. Statist. Assoc.* **88**(422) 486–494.

Srinivasan, V. 1998. A strict paired comparison linear programming approach to nonmetric conjoint analysis. J. E. Aronson, S. Zionts, eds. *Operations Research: Methods, Models and Applications*. Quorum Books, Westport, CT, 97–111.

Srinivasan, V., A. D. Shocker. 1973. Linear programming techniques for multidimensional analysis of preferences. *Psychometrica* **38**(3) 337–369.

Thrun, S., L. Pratt. 1997. *Learning to Learn*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Tikhonov, A. N., V. Y. Arsenin. 1977. *Solutions of Ill-Posed Problems*. W. H. Winston, Washington, D.C.

Toubia, O., J. R. Hauser, D. I. Simester. 2004. Polyhedral methods for adaptive choice-based conjoint analysis. *J. Marketing Res.* **46**(February) 116–131.

Toubia, O., D. I. Simester, J. R. Hauser, E. Dahan. 2003. Fast polyhedral adaptive conjoint estimation. *Marketing Sci.* **22**(3) 273–303.

Vapnik, V. 1998. *Statistical Learning Theory*. Wiley, New York.

Wahba, G. 1990. *Splines Models for Observational Data*, Vol. 59. *Series in Applied Mathematics*. SIAM, Philadelphia, PA.

Zhu, J., T. Hastie. 2005. Kernel logistic regression and the import vector machine. *J. Computational Graphical Statist.* **14**(1) 185–205.