# A Convex Optimization Approach to Modeling Consumer Heterogeneity in Conjoint Estimation

**Theodoros Evgeniou**

Technology Management and Decision Sciences,

INSEAD,

*theodoros.evgeniou@insead.edu*

**Massimiliano Pontil**

University College London,

*pontil@cs.ucl.ac.uk*

**Olivier Toubia**

Marketing Division,

Columbia Business School,

*ot2107@columbia.edu*

# A Convex Optimization Approach to Modeling Consumer Heterogeneity in Conjoint Estimation

**Abstract**

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

**Keywords:** Bayesian Analysis, Data Mining, Econometric Models, Estimation and Other Statistical Techniques, Hierarchical Bayes Analysis, Marketing Research, Regression and Other Statistical Techniques.

# 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 proven 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 towards a population mean (in HB shrinkage is done towards the mean of the first-stage prior on the partworths). In the case of metric data in which closed form expressions are available, we show that the individual-level estimates have the same form. However, while HB samples from a posterior distribution that depends on a set of exogenous parameters (the parameters of the second stage priors), the proposed approach minimizes a convex loss function that depends on a parameter set endogenously

[1] The only exception of which we are aware is an add-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.

(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.

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 in performance 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" 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 datasets. Our work is novel first by its focus on conjoint estimation, second on the loss functions and the convex optimization method used to minimize them, and third on 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 Section 2. In Section 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 Section 4 and two (one for ratings and one for choice) field datasets in Section 5. In Section 6 we illustrate empirically the theoretical differences between our approach and HB outlined in Section 3, and we conclude in Section 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 $\mathbf{X}_i$ 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 details in Appendix C and 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 (RN). 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 $\gamma$ that is typically set using cross-validation:

**Problem 2.1**

$$\min_{\mathbf{w}_i} \frac{1}{\gamma} \sum_{j=1}^{J} \left(y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i\right)^2 + \|\mathbf{w}_i\|^2 \qquad (1)$$
$$\gamma \text{ set by cross} - \text{validation}$$

The loss function $\frac{1}{\gamma}\sum_{j=1}^{J}\left(y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i\right)^2 + \|\mathbf{w}_i\|^2$ is composed of two parts. The first, $\sum_{j=1}^{J}\left(y_{ij} - \mathbf{x}_{ij}\mathbf{w}_i\right)^2$, measures the fit between the estimated utilities and the observed ratings. For a *fixed* $\gamma$, this may be interpreted as the log of the likelihood corresponding

to a normal error term with mean 0 and variance $\gamma$ (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 towards 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 (e.g., smaller). The positive parameter $\gamma$ 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 $-\log(\frac{e^{\mathbf{x}_{ijq*}\mathbf{w}_i}}{\sum_{q=1}^{Q} e^{\mathbf{x}_{ijq}\mathbf{w}_i}})$ (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 (SVM) (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

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 towards 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):

$$\begin{aligned} \min_{\{\mathbf{w}_i\},\mathbf{w}_0,D} \quad & \tfrac{1}{\gamma}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij}-\mathbf{x}_{ij}\mathbf{w}_i\right)^2+\sum_{i=1}^{I}(\mathbf{w}_i-\mathbf{w}_0)^\top D^{-1}(\mathbf{w}_i-\mathbf{w}_0) \\ \text{subject to} \quad & D \text{ is a positive semidefinite matrix scaled to have trace 1} \end{aligned} \tag{2}$$

Let us note that this is *not* the complete method proposed, which includes the estimation of the positive weight $\gamma$ endogenously and is summarized in Section 2.1.5, Problem 2.2. 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 towards 0 as in individual-level RR, it shrinks them towards a vector $\mathbf{w}_0$ (as

will be seen below, the value of $\mathbf{w}_0$ that minimizes the loss function is 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 the appendix 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 $\gamma$ operates the same function as in individual-level RR, namely, achieving a proper trade off between fit and shrinkage. Higher values of $\gamma$ 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 $\infty$ and to only maximize fit.

We consider next the minimization of the loss function (2) given $\gamma$, and in Section 2.1.5 the selection of $\gamma$ using cross-validation. The complete method proposed is summarized in Section 2.1.5, Problem 2.2.

### 2.1.4 Modeling heterogeneity: minimization of the loss function given $\gamma$

For a fixed $\gamma$, 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 (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 Section 3. For a given value of $\gamma$ we use the following iterative method to find the global optimal solution, initializing $D$ to a random positive

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

definite matrix:

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

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

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 $\gamma$ 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 partworths in step 1 (for fixed $\gamma$ and $D$ – see Appendix A for $D$ not invertible) can be written as:

$$\mathbf{w}_i = (X_i^\top X_i + \gamma D^{-1})^{-1} X_i^\top \mathbf{Y}_i + (X_i^\top X_i + \gamma D^{-1})^{-1} \gamma D^{-1} \mathbf{w}_0 \tag{3}$$

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

### 2.1.5 Modeling heterogeneity: setting $\gamma$ using cross-validation

We now describe the estimation of the trade off parameter $\gamma$. Selecting this parameter by minimizing the loss function (2) would be inappropriate as 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 $\gamma$ as follows:

1. Set *Cross-Validation*($\gamma$) $= 0$.

2. For $k = 1$ to $J$:

   (a) Consider the subset of the calibration data $Z^{(-k)} = \bigcup_{i=1}^{I} \{\mathbf{x}_{i1}, \ldots \mathbf{x}_{i(k-1)}, \mathbf{x}_{i(k+1)}, \ldots \mathbf{x}_{iJ}\}$. That is, consider the subset of the calibration data that consists of all questions except the $k^{th}$ 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 $\gamma$ 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.)

[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 questions per respondent. Our particular choice was driven by computational simplicity.

(d) Set *Cross-Validation*($\gamma$) = *Cross-Validation*($\gamma$) + $CV(k)$.

We simply select the parameter $\gamma$ 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 $\gamma$ 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 2.2**

$$\gamma^* = \text{argmin}_\gamma \quad Cross - Validation(\gamma)$$

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

$$\text{subject to} \qquad D \text{ is a positive semidefinite matrix scaled to have trace 1}$$

It is important to note that if $\gamma$ 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

[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.

that would enforce the constraints on $D$. However, *because $\gamma$ is set using cross-validation and the overall estimation method is given by Problem 2.2 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

Choice-based conjoint analysis has become very popular both among practitioners and academics (Carson et al., 1994; Louviere, Hensher, and Swait 2000). We now discuss how to modify the RR-Het method for choice data. 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 solved to estimate the partworths becomes:

**Problem 2.3**

$$\gamma^* = \text{argmin}_\gamma \;\; Cross - Validation(\gamma)$$

$$(\{\mathbf{w}_i^*\}, \mathbf{w}_0^*, D^*) = \text{argmin}_{\{\mathbf{w}_i\},\mathbf{w}_0,D}$$

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

$$\text{subject to} \quad D \text{ is a positive semidefinite matrix scaled to have trace 1}$$

where $\mathbf{x}_{ijq}$ is the $q^{th}$ alternative presented to respondent $i$ in question $j$, and $\mathbf{x}_{ijq_*}$ is the chosen alternative. 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 $\gamma$ is implemented as for RR-Het, with the difference that the cross-validation performance in step (2c) is now measured by the logistic error $-\log\left(e^{\mathbf{x}_{ijq_*}\mathbf{w}_i}/\sum_{q=1}^{Q} e^{\mathbf{x}_{ijq}\mathbf{w}_i}\right)$ on the left out questions. The other major difference from RR-Het is that the minimization of the loss function given $\gamma$ 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.

# 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 $V = 0$):

$$\begin{array}{rc} \text{Likelihood:} & y_{ij} = \mathbf{x}_{ij}\mathbf{w}_i + \epsilon_{ij} \\ & \epsilon_{ij} \sim N(0, \sigma^2) \\ \text{First-stage prior:} & \mathbf{w}_i \sim N(\mathbf{w}_0, D) \\ \text{Second-stage priors:} & \sigma^2 \sim IG(r_0/2, s_0/2) \\ & D^{-1} \sim W(\eta_0, \eta_0 \times \Delta_0) \end{array}$$

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

$$\begin{array}{rc} \text{Likelihood:} & \text{Prob}(\mathbf{x}_{ijq_*} \text{ is chosen}) = \frac{e^{\mathbf{x}_{ijq_*}\mathbf{w}_i}}{\sum_{q=1}^{Q} e^{\mathbf{x}_{ijq}\mathbf{w}_i}} \\ \text{First-stage prior:} & \mathbf{w}_i \sim N(\mathbf{w}_0, D) \\ \text{Second-stage prior:} & D^{-1} \sim W(\eta_0, \eta_0 \times \Delta_0) \end{array}$$

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.

| **HB** | **RR-Het and LOG-Het** |
|---|---|
| Shrinks towards the mean of the first-stage prior | Shrink towards 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 $\gamma$ |
| The parameters of the second-stage priors are set exogenously | $\gamma$ is determined endogenously using cross-validation |

Table 1: Some characteristics of HB versus RR-Het and LOG-Het.

## 3.1 Similarities

The main similarity between the proposed approach and HB is that they both shrink individual estimates towards a population mean. With metric data, the existence of closed form expressions enables us to clearly identify the individual-specific estimates, the

population means towards 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]

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

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

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

These expressions may 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:

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

$$+[(X_i^\top X_i + \sigma^2 D^{-1})^{-1}(\sigma^2 D^{-1})]\mathbf{w}_0$$

$$= \alpha_{HB}^{(i)}((X_i^\top X_i)^{-1} X_i^\top \mathbf{Y}_i) + (I - \alpha_{HB}^{(i)})\mathbf{w}_0$$

[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).

RR-Het:

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

$$= \alpha_{RR}^{(i)}((X_i^\top X_i)^{-1} X_i^\top \mathbf{Y}_i) + (I - \alpha_{RR}^{(i)})\mathbf{w}_0$$

where $\alpha_{HB}^{(i)} = (X_i^\top X_i + \sigma^2 D^{-1})^{-1}(X_i^\top X_i)$ and $\alpha_{RR}^{(i)} = (X_i^\top X_i + \gamma D^{-1})^{-1}(X_i^\top X_i)$. 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 $\gamma D^{-1}$ 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 $\alpha_{HB}^{(i)}$ over the posterior distributions of $\sigma$ and $D$.

Note that if the parameters $\eta_0$, $\Delta_0$, $r_0$, and $s_0$ in HB were selected to yield a strong prior on $\sigma^2$ and $D$ around the values of $\gamma$ and $D$ obtained by RR-Het estimation, the posterior means provided by HB would converge to the point estimates provided by RR-Het ($\alpha_{HB}^{(i)} \to \alpha_{RR}^{(i)}$). 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 Appendix E 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, $\eta_0$, $\Delta_0$, $r_0$, $s_0$ in the case of metric data and $\eta_0$ and $\Delta_0$ in the case of choice data), the loss functions in RR-Het and LOG-Het are a function of an *endogenous* parameter $\gamma$ (determined from the calibration data using cross-validation). The difference between the way the second-stage priors are set in HB and $\gamma$ 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 Section 6. For example, in the case of metric data, shrinkage is a function of $\sigma^2 D^{-1}$ in HB and $\gamma D^{-1}$ in RR-Het. In HB, the posterior distributions on $\sigma$ 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 $\eta_0$, $\Delta_0$, $r_0$, 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 $\sigma$ and $D$ would clearly lead to maximal estimation accuracy. While such an extreme case may 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 endogenous parameters determined by the minimization of the loss function and by cross-validation: $D$ and $\gamma$ are obtained by solving Problem 2.2 (respectively Problem 2.3).

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 Section 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 comparison may 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 may 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 $\gamma$ is a subset of those achievable by HB by changing $\eta_0$, $\Delta_0$, $r_0$, 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.

[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.

First, Bayesian analysis obeys the likelihood principle (Fisher 1922; Rossi and Allenby 2003; Allenby, Otter and Liu, 2006) 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 may 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, 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 in order 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, *non-hierarchical*, *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 in order to obtain initial point estimates for $\mathbf{w}_0$ and $D$, and then consider the following simple (non hierarchical) model:

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

where $\sigma_0$ 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:

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

Because the full posterior mean of $\mathbf{w}_i$ is given in closed form, there is no need to sample from the posterior in order to obtain point estimates, and the cross-validation error

associated with a given value of $\sigma_0$ may be estimated conveniently fast. Note that varying $\sigma_0$ 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 only available 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).

# 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 Section 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.

[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.

We used a 2 (low vs. high heterogeneity) $\times$ 2 (low vs. high response error) $\times$ 2 (low vs. 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 $\mathbf{w}_0 = [5, 5..., 5]$ 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 $\sigma$ 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 $\sigma$ 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 non-significantly 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.

| 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** |

Table 2: RMSE (lower numbers indicate higher performance) of estimated versus true partworths for the metric-based simulations. 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 non-significantly different from best (at $p < 0.05$) in 7 out of 8 conditions.

## 4.2 Choice-Based Simulations

We extended the simulation setup above to compare choice HB to LOG-Het. As discussed in Section 3, we were unable to test a choice version of the Metric Bayes-CV method. We used again a 2 (low vs. high heterogeneity) $\times$ 2 (low vs. high response error) $\times$ 2 (low vs.

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, Louviere and Anderson (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 $i^{th}$ 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 [$mag$,$mag$,...$mag$] and variance $\sigma^2 = het \times mag$ where the parameter $mag$ controls the amount of response error and the parameter $het$ 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 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]

[9]Because our number of features (10) is 2.5 times the number (4) used by previously published simulations using the same simulation design, we divide the values of typical $mag$ parameters used in previously published simulations (0.5 and 3) by 2.5 in order 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

| 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** |

Table 3: RMSE (lower numbers indicate higher performance) of estimated versus true partworths for the choice simulations. 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.

# 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 four 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

scale.

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

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 non-significantly 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, in order 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 $\sigma$ for all respondents). Third, we do not use demographic variables in the model. We show in Appendix D 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.

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

Table 4: RMSE for holdout questions from the metric field data of Lenk et al., (1996). (Lower numbers indicate higher performance.) 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 non-significantly different from best with both 8 and 16 questions.

## 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 lead to this data set.

The product in this study was carbonated soft drinks. 3 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. 192 respondents were subjected to one of the four 22-question sets (presented in a randomized order). As before, we used 8 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 that, as each question involved 8 products, a random model would achieve a hit rate of 12.5%).

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

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

Table 5: Holdout hit rates (higher numbers indicate higher performance) from the choice field data set. 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 non-significantly different from best in both cases, and significantly better than HB with 16 questions.

# 6 The Relation Between Shrinkage and Estimation Accuracy

Beyond comparing estimation accuracy and predictive ability, we now further explore empirically some of the points raised in Section 3. In particular, we have argued 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 $\gamma$ 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 $\gamma$ 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 ($\gamma$ and $\sigma_0$ 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.

In order to explore the relation between shrinkage and performance, we manually varied the parameters $\gamma$ and $\sigma_0$ 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 only report the graphs based on 16 questions. The graphs based on 8 questions yield similar results and are available from the authors). The parameters $\gamma$ and $\sigma_0$ 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 $\sum_{i=1}^{I} \frac{||\mathbf{w}_i - \mathbf{w}_0||^2}{I}$. Performance, measured by the RMSE of the true vs. 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 $\sigma_0$ (respectively $\gamma$) is varied. The labels "RR-Het" and "Bayes-CV" correspond to the amount of shrinkage and corresponding performance achieved by

the two methods when $\gamma$ and $\sigma_0$ 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 may be viewed as a key difference between our approach and HB.

[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 $\mathbf{w}_0$.

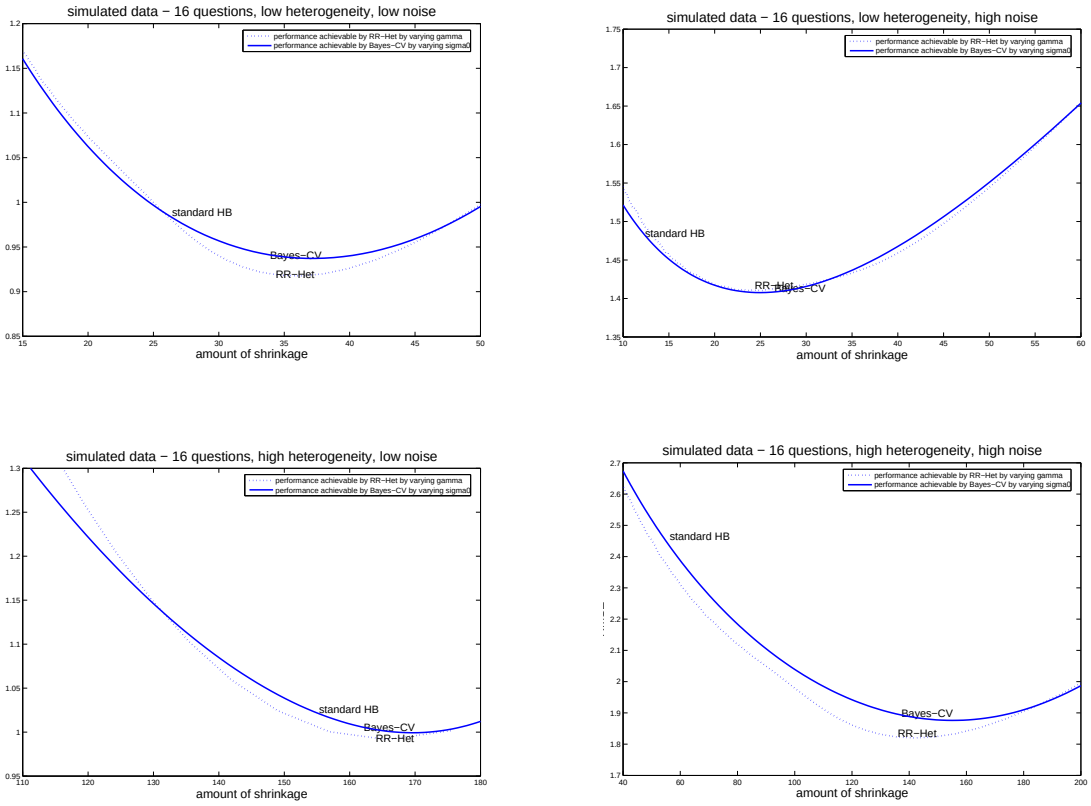


Figure 1: Performance as a function of the amount of shrinkage - metric simulated data. Estimates are based on 16 questions. The amount of shrinkage is measured by $\sum_{i=1}^{I} \frac{||\mathbf{w}_i - \mathbf{w}_0||^2}{I}$. The solid lines represent the amount of shrinkage and corresponding RMSE (estimated versus actual partworths) performance achieved by Metric Bayes-CV as $\sigma_0$ is varied, and the labels "Bayes-CV" represent the amount of shrinkage and performance achieved when $\sigma_0$ is selected using cross-validation. The dotted lines represent the amount of shrinkage and performance achieved by RR-Het as $\gamma$ is varied, and the labels "RR-Het" represent the amount of shrinkage and performance achieved when $\gamma$ is selected using cross-validation. "Standard HB" corresponds to HB with standard second-stage priors, as in Table 2.

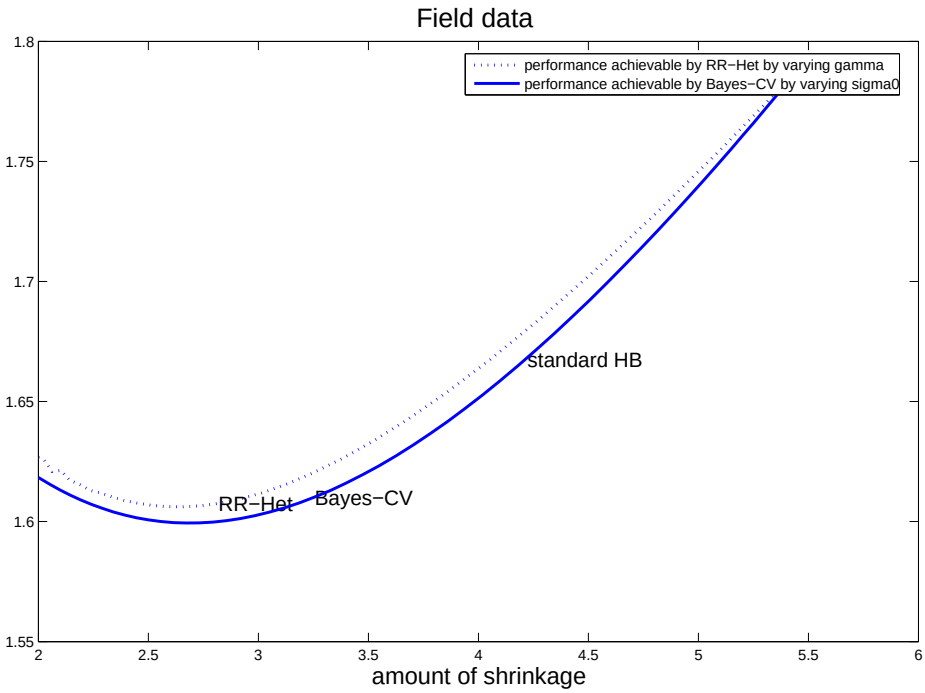


Figure 2: Performance as a function of the amount of shrinkage - field metric data. Estimates are based on 16 questions. The amount of shrinkage is measured by $\sum_{i=1}^{I} \frac{||\mathbf{w}_i - \mathbf{w}_0||^2}{I}$. The solid line represents the amount of shrinkage and corresponding performance (hold-out RMSE) achieved by Metric Bayes-CV as $\sigma_0$ is varied, and the label "Bayes-CV" represents the amount of shrinkage and performance achieved when $\sigma_0$ is selected using cross-validation. The dotted line represents the amount of shrinkage and performance achieved by RR-Het as $\gamma$ is varied, and the label "RR-Het" represents the amount of shrinkage and performance achieved when $\gamma$ is selected using cross-validation. "Standard HB" corresponds to HB with standard second-stage priors, as in Table 4.

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 $\gamma$ and $\sigma_0$. In particular, for the simulations (respectively ratings field data) the RMSE achieved by RR-Het or Metric Bayes-CV when $\gamma$ or $\sigma_0$ is selected using cross-validation is on average only 0.59% (respectively 0.38%) higher than the minimum achievable if the true partworths (respectively holdout ratings) were used to select $\gamma$ and $\sigma_0$. 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

We have proposed a novel convex optimization-based approach for handling consumer heterogeneity in conjoint estimation, and applied it to both metric and choice data. Simulations as well as two empirical data sets suggest that the approach overall outperforms a standard HB implementation. We have also highlighted some important theoretical differences and similarities between our approach and HB. The major difference is that while the amount of shrinkage is a function of a set of exogenous parameters in HB (the parameters of the second-stage priors), it is completely endogenous in our approach. Finally, we also showed empirically that using cross-validation to endogenously select the parameters of the second-stage prior in HB has the *potential* to greatly improve performance.

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. An additional area for future 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. Finally, we have focused in this paper on unimodal representations of heterogeneity. Future research may introduce and model segments of consumers. This may 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. In general, this paper provides only a first step towards the development of optimization and statistical machine learning based methods for modeling heterogeneity. While simulations and field data suggest that this first step is promising, many extensions are possible by modifying the loss function and exploiting the rich literature on convex optimization and statistical machine learning.

# References

[1] Allenby, Greg M., Thomas Otter, and Qing Liu. 2006. Investigating Endogeneity Bias in Conjoint Models. *Working paper, Ohio State University.*

[2] Allenby, Greg M., Peter E. Rossi. 1999. Marketing Models of Consumer Heterogeneity. *Journal of Econometrics*, 89, March/April, p. 57 – 78.

[3] Ando, Rie K. and Tong Zhang. 2005. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. *Journal of Machine Learning Reseearch*, 6, p. 1817–1853.

[4] Arora, Neeraj and Joel Huber. 2001. "Improving parameter estimates And model prediction by aggregate customization in choice experiments", *Journal of Consumer Research*,(September), Vol. 28.

[5] Baxter, Jonathan. 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28, pp. 7–39.

[6] Boyd, Stephen, Lieven Vandenberghe. 2004. Convex Optimization. Cambridge University Press.

[7] Bunch, David S., Jordan J. Louviere, and Don 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.

[8] Carson, Richard T., Jordan J. Louviere, Don A. Anderson, Phipps Arabie, David S. Bunch, David A. Hensher, Richard M. Johnson, Warren F. Kuhfeld, Dan Steinberg, Joffrey Swait, Harry Timmermans, and James B. Wiley. 1994. "Experimental Analysis of Choice", *Marketing Letters*, 5(4), p. 351-367.

[9] Caruana, Rich. 1997. Multi–task learning. *Machine Learning*, 28, p. 41–75.

[10] Chaloner, Kathryn, Isabella Verdinelli. 1995. Bayesian Experimental Design: A Review. *Statistical Science*, 10 (3), p. 273–304.

[11] Cucker, Felipe, Steve Smale. 2002. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1), p. 1–49.

[12] Cui, Dapeng, David Curry. 2005. Prediction in Marketing using the Support Vector Machine. *Marketing Science*, 24(4), p. 595–615.

[13] Efron, Bradley, Robert Tibshirani. 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

[14] Evgeniou, Theodoros, Constantinos Boussios, Giorgos Zacharia. 2005a. Generalized Robust Conjoint Estimation. *Marketing Science*, 24(3).

[15] Evgeniou, Theodoros, Charles Micchelli, and Massimiliano Pontil. 2005b. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6, p. 615–637.

[16] Fisher, Ronald A. 1922. On the Mathematical Foundations of Theoretical Statistics. *Phil. Trans. Royal Soc.*, series A, p. 222–326.

[17] Girosi, Federico, Michael Jones, Tomaso Poggio. 1995. Regularization theory and neural networks architectures. *Neural Computation*, Vol. 7, p. 219–269.

[18] Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman. 2003. The Elements of Statistical Learning. Springer Series in Statistics.

[19] Huber, Joel, and Klaus Zwerina. 1996. "The importance of utility balance in efficient choice designs", *Journal of Marketing Research*, 32 (August), p. 308-317.

[20] Jebara, Tony. 2004. Multi-Task Feature and Kernel Selection for SVMs. In *Proceedings of the twenty-first International Conference on Machine learning.*

[21] Jaakkola Tommi and David Haussler. 1999. Probabilistic kernel regression models. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, San Francisco, Morgan Kaufmann.

[22] Keerthi, S., Duan, K., Shevade, S., and Poo, A. 2005. A Fast Dual Algorithm for Kernel Logistic Regression. *Machine Learning*, Volume 61, Numbers 1-3, November, p. 151-165.

[23] Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, Martin R. Young. 1996. Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. *Marketing Science*, 15(2), p. 173–91.

[24] Louviere, Jordan J., David A. Hensher, and Joffre D. Swait. 2000. Stated Choice Methods: Analysis and Applications, New York, NY, Cambridge University Press.

[25] Micchelli, Charles and Massimiliano Pontil. 2005. On learning vector–valued functions. *Neural Computation*, 17, p. 177-204.

[26] Mika, S., B. Schlkopf, A.J. Smola, K.-R. Mller, M. Scholz, and G. Rtsch. 1999. Kernel PCA and de-noising in feature spaces. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems* 11, pages 536-542. MIT Press,

[27] Minka, Thomas. 2003. A comparison of numerical optimizers for logistic regression. Microsoft research tech report.

[28] Rossi, Peter E., Greg M. Allenby. 1993. A Bayesian Approach to Estimating Household Parameters. *Journal of Marketing Research*, 30(2), p. 171–182.

[29] Rossi, Peter E., Greg M. Allenby. 2003. Bayesian Statistics and Marketing. *Marketing Science*, 22(3), p. 304-328.

[30] Rossi, Peter E., Greg M. Allenby, Robert McCulloch. 2005. Bayesian Statistics and Marketing. John Wiley and Sons.

[31] Shao, Jun. 1993. Linear model selection via cross-validation. *Journal of the American Statistical Association*, 88(422), p. 486–494.

[32] Srinivasan, V. 1998. A Strict Paired Comparison Linear Programming Approach to Nonmetric Conjoint Analysis. *Operations Research: Methods, Models and Applications*, Jay E. Aronson and Stanley Zionts (eds), Westport, CT: Quorum Books, p. 97-111.

[33] Srinivasan, V., Allan D. Shocker. 1973. Linear Programming Techniques for Multidimensional Analysis of Preferences. *Psychometrica*, 38(3), p. 337–369.

[34] Thrun, Sebastian and L. Pratt. 1997. *Learning to Learn.* Kluwer Academic Publishers.

[35] Tikhonov A. N., V. Y. Arsenin. 1977. Solutions of Ill-posed Problems. W. H. Winston, Washington, D.C.

[36] Toubia, Olivier, Duncan I. Simester, John R. Hauser, Ely Dahan. 2003. Fast Polyhedral Adaptive Conjoint Estimation. *Marketing Science*, 22(3), p. 273–303.

[37] Toubia, Olivier, John R. Hauser, Duncan I. Simester. 2004. Polyhedral methods for adaptive choice-based conjoint analysis. *Journal of Marketing Research*, 46 (Feb.), p. 116–131.

[38] Vapnik, Vladimir. 1998. Statistical Learning Theory. New York: Wiley.

[39] Wahba, Grace. 1990. Splines Models for Observational Data. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia.

[40] Zhu, Ji and Hastie, Trevor. 2005. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics* 14(1), p.185-205.

# Appendix

# A Minimization of the RR-Het Loss Function (2) Given $\gamma$

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

We first transform the data as $\tilde{\mathbf{x}}_{ij} = \mathbf{x}_{ij} D^{\frac{1}{2}}$ and define $\tilde{\mathbf{w}}_i = D^{-\frac{1}{2}} \mathbf{w}_i$ and $\tilde{\mathbf{w}}_0 = D^{-\frac{1}{2}} \mathbf{w}_0$ (see the case of a non-invertible $D$ 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

$$\min_{\{\tilde{\mathbf{w}}_i\}, \tilde{\mathbf{w}}_0} \frac{1}{\gamma} \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \tilde{\mathbf{x}}_{ij} \tilde{\mathbf{w}}_i)^2 + \sum_{i=1}^{I} (\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0)^\top (\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_0), \tag{4}$$

and then get the final solution as $\mathbf{w}_i = D^{\frac{1}{2}} \tilde{\mathbf{w}}_i$ and $\mathbf{w}_0 = D^{\frac{1}{2}} \tilde{\mathbf{w}}_0$. This is because $\tilde{\mathbf{x}}_{ij} \tilde{\mathbf{w}}_i = \mathbf{x}_{ij} D^{\frac{1}{2}} D^{-\frac{1}{2}} \mathbf{w}_i = \mathbf{x}_{ij} \mathbf{w}_i$ 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 (4) 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

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

$$\tilde{\mathbf{w}}_i = (\tilde{X}_i^\top \tilde{X}_i + \gamma I_p)^{-1}\tilde{X}_i^\top \mathbf{Y}_i + (\tilde{X}_i^\top \tilde{X}_i + \gamma I_p)^{-1}\gamma \tilde{\mathbf{w}}_0 \tag{5}$$

$$= \hat{\mathbf{w}}_i + \gamma Z_i \tilde{\mathbf{w}}_0,$$

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:

$$\tilde{\mathbf{w}}_0 = \frac{1}{I}\sum_i (\hat{\mathbf{w}}_i + \gamma Z_i \tilde{\mathbf{w}}_0)$$

which implies

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

If the matrix $\left(I_p - \gamma \frac{1}{I_p}\sum_i Z_i\right)$ 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 Section 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 $\mathbf{w}_i = D^{\frac{1}{2}}\tilde{\mathbf{w}}_i$ and $\mathbf{w}_0 = D^{\frac{1}{2}}\tilde{\mathbf{w}}_0$. Finally, to get (3) – which we no not need to compute in practice – we just have to replace $\tilde{X}_i$ with $X_i D^{\frac{1}{2}}$ in (5) and use the fact that $\mathbf{w}_i = D^{\frac{1}{2}}\tilde{\mathbf{w}}_i$.

If $D$ is not invertible we replace $D^{-\frac{1}{2}}$ 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 $D^{\frac{1}{2}}$ (computed using only the non-zero 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 an equation like (3) – which we do not need to compute in practice again – by replacing again $\tilde{X}_i$ with $X_i D^{\frac{1}{2}}$ in (5) and use the fact that $\mathbf{w}_i = D^{\frac{1}{2}}\tilde{\mathbf{w}}_i$.

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 only requires $2I$ 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 $\left(\sum_{i=1}^{I}(\mathbf{w}_i - \mathbf{w}_0)(\mathbf{w}_i - \mathbf{w}_0)^\top\right)$, 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, like in the individual-level RR case discussed in Section 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:

$$\begin{array}{ll} \min\limits_{D} & \sum_{i=1}^{I}(\mathbf{w}_i - \mathbf{w}_0)^\top D^{-1}(\mathbf{w}_i - \mathbf{w}_0) \\ \text{subject to} & D \text{ is a positive semidefinite matrix scaled to have trace 1} \end{array}$$

Using a Lagrange multiplier $\rho$ for the trace constraint and taking the derivative with respect to $D$ we have that:

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

$$\Rightarrow D = \frac{1}{2\rho}\left(\sum_{i=1}^{I}(\mathbf{w}_i - \mathbf{w}_0)(\mathbf{w}_i - \mathbf{w}_0)^\top\right)^{\frac{1}{2}} \tag{6}$$

which is positive definite; $\rho$ is simply selected so that $D$ has trace 1.

# 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 $\mathbf{w}_0 = \frac{1}{I}\sum_i \mathbf{w}_i$. Similarly, given $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$ we can solve for $D$ like for RR-Het above - since $D$ appears only in the complexity control. Hence, we only need to solve for $\{\mathbf{w}_i\}$ and $\mathbf{w}_0$ given $D$, and then iterate among the conditional estimations (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 $\tilde{\mathbf{x}}_{ijq} = \mathbf{x}_{ijq}D^{\frac{1}{2}}$ and define $\tilde{\mathbf{w}}_i = D^{-\frac{1}{2}}\mathbf{w}_i$ and $\tilde{\mathbf{w}}_0 = D^{-\frac{1}{2}}\mathbf{w}_0$. Note that

with this transformation we can estimate first $\tilde{\mathbf{w}}_i$ minimizing the modified cost function

$$-\frac{1}{\gamma^*}\sum_{i=1}^{I}\sum_{j=1}^{J}\log\frac{e^{\tilde{\mathbf{x}}_{ijq_*}\tilde{\mathbf{w}}_i}}{\sum_{q=1}^{Q}e^{\tilde{\mathbf{x}}_{ijq}\tilde{\mathbf{w}}_i}}+\sum_{i=1}^{I}(\tilde{\mathbf{w}}_i-\tilde{\mathbf{w}}_0)^\top(\tilde{\mathbf{w}}_i-\tilde{\mathbf{w}}_0) \tag{7}$$

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

Notice that for a fixed $\tilde{\mathbf{w}}_0$, problem (7) is decomposable into $I$ separate sub-problems, 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 (7) 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 only need the gradient and Hessian of the loss function (7). 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

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

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

# C Estimating interactions

The statistical learning approach enables the estimation of highly non-linear models, for example in order to consider all possible attribute interactions of multiple degrees. The use of complexity control and the selection of the trade-off parameter $\gamma$ using methods like cross-validation enables model estimation even when the number of parameters per respondent is very large compared to the number of observations (see for example Vapnik 1998), as has been shown already for individual level conjoint estimation methods (Cui and Curry, 2005; Evgeniou et al., 2005a). Moreover, the estimation of such models with many parameters can be done using the rich theory of kernels and dual optimization methods (Wahba 1990; Vapnik 1998; Hastie et al., 2003). To this purpose we present in the online technical appendix (included at the end of the paper) a dual optimization version of RR-Het and LOG-Het using kernels. We note that one needs to use kernels and this dual optimization approach only when the number of parameters estimated per respondent (i.e., the number of attribute interactions, or other attribute features) is very large, i.e., larger than the number of data $I \times J$ – see online technical appendix – or infinite. Otherwise one can simply expand the vectors $\mathbf{x}_{ij}$ to incorporate all interactions/nonlinearities and use the RR-Het and LOG-Het estimation methods outlined above. For example if we have $p$ binary attributes and consider all $p(p-1)/2$ pairwise interactions, then we can replace $\mathbf{x}_{ij}$ with a $p + p(p-1)/2$ dimensional vector to include these interactions and estimate $p + p(p-1)/2$ parameters per respondent using the RR-Het or LOG-Het methods (or HB).

In order to explore the effectiveness of the statistical learning approach for estimating nonlinear utility functions, we ran choice simulations similar to the ones reported in Section 4.2. We considered the exact same simulation setup as in Section 4.2, except that we added all 45 pairwise attribute interactions. We considered the following 2 scenarios: a) the true partworths for the interactions have half the magnitude of those for the main effects (low interaction level); b) the true partworths for the interactions have the same magnitude as those for the main effects (high interaction level). In both scenarios we set the parameters *mag* and *het* so that we achieve the mean level of response error and heterogeneity used across the 4 conditions in the simulations in Section 4.2 – giving rise to $het = 2$ in both cases, and $mag = 0.2$ and 0.4 in the two scenarios respectively.

We report the results in Table 6. In agreement with previous work on individual-level estimation (Cui and Curry 2005; Evgeniou et al., 2005a), the statistical learning approach is more suitable for estimating models with interactions than HB: LOG-Het significantly outperforms HB both in estimating the interaction coefficients as well as the main effects when interactions are included into the model specification. We also see that when the true model includes a high level of interactions and the number of questions is larger, LOG-Het with a nonlinear specification performs best (marginally significantly, $p = 0.06$) on the estimation of the main effects also compared to models with linear specifications.

We also estimated models with all pairwise attribute interactions for both empirical datasets (resulting in $14+(13\times12)/2 = 92$ parameters per respondent for the metric field dataset and $20+7\times6+7\times7+7\times6 = 153$ parameters per respondent for the choice field

dataset). Including interactions did not improve performance on these datasets. LOG-Het performed not significantly differently with and without interactions on the choice dataset and HB performed significantly worse with interactions than without. On the metric dataset, both RR-Het and HB performed significantly worse with interactions than without. This may be due to the absence of strong interactions and/or to the relatively low number of questions.

| Magnitude of Interactions | Questions | HB | | LOG-Het | |
|---|---|---|---|---|---|
| | | nonlinear specification | linear specification | nonlinear specification | linear specification |
| Low | 8 | 1.2046 | 1.0212 | 1.0216 | **1.0139** |
| | | 1.3254 | – | **1.2041** | – |
| | 16 | 1.0871 | 0.9353 | **0.9241** | **0.9291** |
| | | 1.3033 | – | **1.1571** | – |
| High | 8 | 1.2084 | **1.1131** | **1.1113** | **1.1112** |
| | | 1.3307 | – | **1.1951** | – |
| | 16 | 1.1406 | 1.0368 | **1.0294** | **1.0352** |
| | | 1.2980 | – | **1.1417** | – |

Table 6: RMSE from choice simulations with all pairwise interactions. (Lower numbers indicate better performance.) In each cell the top number is the RMSE (true vs. estimated parameters) for the 10 main effect coefficients and the bottom number is the RMSE for the 45 interaction coefficients. The columns with the header "linear specification" correspond to the case in which only the main effects are estimated, and the columns with the header "nonlinear specification" corresponds to the case in which all main effects and pairwise interactions are estimated. Note that the RMSE is typically larger for the interaction terms because there are 45 interaction coefficients and 10 main effect ones (the RMSE is the average across respondents of the sum of the root mean squared error across coefficients). Bold numbers in each row indicate best or not significantly different from best at the $p < 0.05$ level. In all conditions, LOG-Het significantly outperforms HB in estimating the interaction coefficients as well as the main effects when interactions are included into the model specification.

# D Adding Covariates

Our approach can be extended to include demographic covariates that affect the individual partworths, using standard convex optimization techniques. In particular, we consider the following loss function:

$$\begin{array}{ll} \min\limits_{\{\mathbf{w}_i\},\boldsymbol{\Theta},D} & \frac{1}{\gamma}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij}-\mathbf{x}_{ij}\mathbf{w}_i\right)^2+\sum_{i=1}^{I}(\mathbf{w}_i-\boldsymbol{\Theta}\mathbf{z}_i)^\top D^{-1}(\mathbf{w}_i-\boldsymbol{\Theta}\mathbf{z}_i) \\ \text{subject to} & D \text{ is a positive semidefinite matrix scaled to have trace 1} \end{array}$$

where $\boldsymbol{\Theta}$ is a $p \times r$ matrix of regression coefficients and $\mathbf{z}_i$ is a $r$-dimensional vector of covariates for respondent $i$. Note that the above loss function remains jointly convex with respect to $\{\mathbf{w}_i\}$, $\boldsymbol{\Theta}$, and $D$. Also, for a given $\boldsymbol{\Theta}$ and $D$ we can solve for $\{\mathbf{w}_i\}$ as before (it is easy to see that $\mathbf{w}_i = (X_i^\top X_i + \gamma D^{-1})^{-1}X_i^\top \mathbf{Y}_i + (X_i^\top X_i + \gamma D^{-1})^{-1}\gamma\boldsymbol{\Theta}\mathbf{z}_i)$. Similarly, for a given $\boldsymbol{\Theta}$ and $\{\mathbf{w}_i\}$ we solve for $D$ as before. Hence we only show here how to solve for $\boldsymbol{\Theta}$ given $\{\mathbf{w}_i\}$ and $D$. We assume for simplicity that the matrix $D$ is invertible. If not, we replace $D^{-1}$ with the pseudo-inverse as before.

Taking the derivative with respect to $\boldsymbol{\Theta}$ we see that:

$$\sum_{i=1}^{I}(\mathbf{w}_i\mathbf{z}_i^\top)^\top = \sum_{i=1}^{I}(\mathbf{z}_i\mathbf{z}_i^\top)\boldsymbol{\Theta}^\top$$

from which we solve for $\boldsymbol{\Theta}$ as:

$$vec(\boldsymbol{\Theta}) = (ZZ^\top \otimes D^{-1})^{-1}(Z \otimes D^{-1})vec(\mathbf{W})$$

where $vec(\mathbf{\Theta})$ is a column vector with the columns of matrix $\mathbf{\Theta}$ stacked on each other, $\otimes$ is the Kronecker product, and $vec(\mathbf{W})$ is a column vector obtained by stacking the $\mathbf{w}_i$'s. Note that we still have closed form solutions for $\{\mathbf{w}_i\}$ and $\mathbf{\Theta}$, and the estimation of the partworths $\{\mathbf{w}_i\}$ is still decomposed across the individuals.

We tested this method on Lenk et al. (1996)'s dataset, using the demographic variables collected by Lenk et al. (1996) as covariates (see Table 2 in Lenk et al. for a description of these variables). The performance results are reported in Table 7. Notice that the results are qualitatively similar to those reported in Section 5. RR-Het with covariates is significantly better than HB with covariates for both 8 and 16 questions. Metric Bayes-CV is also significantly better than HB. Note also that the addition of covariates did not improve the predictive performance of HB or RR-Het. We report the estimated $\mathbf{\Theta}$'s from HB and RR-Het in Table 8, along with the confidence intervals (see Appendix E).

| Questions | Standard HB with covariates | Metric Bayes-CV with covariates | RR-Het with covariates |
|---|---|---|---|
| 8 | 1.945 | 1.916 | **1.845** |
| 16 | 1.673 | **1.607** | **1.609** |

Table 7: Holdout questions RMSE (lower numbers indicate better performance) from the field data of Lenk et al., (1996) when covariates are used. 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 non-significantly different from best with both 8 and 16 questions.

# E Confidence intervals

A number of bootstrapping methods can be used to generate confidence intervals for RR-Het or LOG-Het (see for example Efron and Tibshirani 1993 and references therein). These methods rely on the following principle: estimate the parameters (i.e., $\{\mathbf{w}_i\}, \mathbf{w}_0, D$) multiple times (i.e., 1000 times - see Efron and Tibshirani 1993), each time using a different random bootstrap subsample of the data, and then compute confidence intervals from these estimates. The bootstrap samples are obtained by sampling the original data with replacement and are of the same size as the original data ($I \times J$ in our case). The estimates across these bootstrap repetitions are then used to produce confidence intervals. Various methods can be used for this purpose, i.e., from the simple one of computing the standard deviation of the estimates across the bootstrap samples, to more "advanced" ones such as ABC-bootstrap or $BC_a$ bootstrap estimates (Efron and Tibshirani 1993). We refer the reader to the rich literature on bootstrap confidence intervals for further information.

As an illustration, we computed confidence intervals for the estimate of $\mathbf{\Theta}$ obtained from Lenk et al.'s dataset, using the simplest bootstrap confidence intervals method: compute the standard deviation of the estimated parameters across 1000 bootstrap samples. For HB, we obtained confidence intervals from the standard deviation of the posterior distribution of each parameter, estimated as the standard deviation of the parameter values sampled (also 1000 times) from the posterior. We report the results in Table 8. Note that the standard errors estimated from the two methods are comparable. For example,

the average standard deviations are 0.992 for RR-Het and 0.1051 for HB, the number of parameters for which the estimate is more than one standard deviation away from zero is 32 for RR-Het and 33 for HB, and for 91 out of 98 parameters HB and RR-Het agree on whether the parameter is more than one standard deviation away from 0.

| Intercept | Female | Years | Own | Tech | Apply | Expert |
|---|---|---|---|---|---|---|
| **3.55 3.54** | -0.12 -0.11 | **-0.07 -0.07** | 0.01 0.02 | **-0.28 -0.28** | **0.15 0.15** | **0.15 0.15** |
| (0.50 0.25) | (0.24 0.11) | (0.05 0.02) | (0.31 0.15) | (0.25 0.12) | (0.07 0.04) | (0.06 0.03) |
| -0.13 -0.12 | **0.28 0.27** | -0.00 -0.00 | -0.06 -0.05 | 0.00 -0.00 | -0.00 -0.01 | **0.03** 0.03 |
| (0.22 0.25) | (0.11 0.11) | (0.02 0.02) | (0.14 0.14) | (0.11 0.11) | (0.03 0.03) | (0.03 0.03) |
| **0.51 0.49** | -0.10 -0.10 | -0.01 -0.01 | **0.21 0.21** | **0.16 0.16** | **0.06 0.06** | **-0.07 -0.06** |
| (0.21 0.24) | (0.11 0.11) | (0.02 0.02) | (0.13 0.14) | (0.10 0.11) | (0.03 0.03) | (0.03 0.03) |
| 0.18 0.19 | **-0.15 -0.15** | -0.01 -0.01 | -0.02 -0.02 | 0.09 0.10 | 0.01 0.01 | 0.01 0.01 |
| (0.22 0.25) | (0.10 0.11) | (0.02 0.02) | (0.13 0.14) | (0.10 0.11) | (0.03 0.03) | (0.03 0.03) |
| -0.14 -0.16 | -0.10 -0.10 | **-0.02 -0.02** | 0.07 0.06 | 0.09 0.09 | 0.04 **0.04** | **0.06 0.06** |
| (0.27 0.24) | (0.13 0.11) | (0.02 0.02) | (0.16 0.14) | (0.13 0.11) | (0.04 0.04) | (0.03 0.03) |
| -0.07 -0.07 | -0.10 -0.09 | -0.02 -0.02 | 0.02 0.02 | -0.04 -0.04 | 0.02 0.02 | **0.04 0.04** |
| (0.24 0.24) | (0.11 0.11) | (0.02 0.02) | (0.14 0.14) | (0.11 0.11) | (0.03 0.03) | (0.03 0.03) |
| **0.65 0.65** | **-0.19 -0.18** | -0.01 -0.01 | 0.02 0.02 | -0.07 -0.07 | 0.01 0.01 | -0.01 -0.01 |
| (0.25 0.23) | (0.12 0.12) | (0.02 0.02) | (0.16 0.14) | (0.12 0.12) | (0.04 0.03) | (0.03 0.03) |
| **-0.31 -0.31** | -0.03 -0.03 | -0.00 -0.01 | 0.09 0.10 | 0.01 0.00 | **-0.04 -0.03** | **0.05 0.05** |
| (0.22 0.24) | (0.11 0.11) | (0.02 0.02) | (0.13 0.14) | (0.11 0.11) | (0.03 0.03) | (0.03 0.03) |
| 0.15 0.17 | -0.04 -0.05 | 0.00 0.00 | 0.02 0.02 | **-0.11** -0.11 | -0.00 0.00 | -0.01 -0.02 |
| (0.20 0.23) | (0.09 0.11) | (0.02 0.02) | (0.12 0.15) | (0.10 0.11) | (0.03 0.04) | (0.03 0.03) |
| 0.10 0.10 | 0.03 0.03 | **0.02** 0.02 | 0.12 0.11 | **-0.16 -0.16** | -0.01 -0.01 | -0.02 -0.02 |
| (0.20 0.23) | (0.10 0.12) | (0.02 0.02) | (0.12 0.13) | (0.10 0.11) | (0.03 0.04) | (0.02 0.03) |
| -0.05 -0.04 | **0.11** 0.10 | **0.04 0.04** | -0.02 -0.01 | -0.00 -0.00 | 0.02 0.02 | -0.01 -0.01 |
| (0.21 0.25) | (0.10 0.11) | (0.02 0.02) | (0.12 0.13) | (0.09 0.11) | (0.03 0.03) | (0.03 0.03) |
| **0.38 0.37** | -0.03 -0.03 | **-0.04 -0.03** | -0.02 -0.02 | 0.09 0.08 | 0.00 0.00 | -0.00 -0.00 |
| (0.22 0.25) | (0.10 0.11) | (0.02 0.02) | (0.12 0.14) | (0.10 0.11) | (0.03 0.03) | (0.03 0.03) |
| 0.06 0.04 | 0.05 0.05 | **0.03 0.03** | **-0.17 -0.17** | -0.03 -0.04 | -0.00 -0.00 | 0.01 0.01 |
| (0.22 0.25) | (0.10 0.11) | (0.02 0.02) | (0.13 0.14) | (0.10 0.11) | (0.03 0.03) | (0.03 0.03) |
| **-1.51 -1.49** | **0.37 0.37** | **0.03 0.03** | -0.19 **-0.19** | -0.06 -0.06 | 0.01 0.01 | 0.03 **0.03** |
| (0.34 0.25) | (0.17 0.11) | (0.03 0.02) | (0.21 0.14) | (0.17 0.11) | (0.05 0.03) | (0.04 0.03) |

Table 8: The $14 \times 7$ parameters of the covariates matrix $\boldsymbol{\Theta}$ for the field data. In each cell the first number is the estimate from HB and the second from RR-Het with covariates. In parenthesis we report the standard deviations, first number for HB and second for RR-Het. We report in bold the parameters which are more than one standard deviation away from 0 (32 for RR-Het and 33 for HB).

# F Online Technical Appendix: Estimation of General Nonlinear Utility Functions Using Kernels

We present a novel dual optimization method for solving RR-Het and LOG-Het using kernels (Vapnik 1998; Hastie et al., 2003; Wahba 1990). One needs to consider this dual method when the number of parameters estimated per respondent is very large (i.e., larger than $I \times J$, the size of the kernel matrix – see below) or infinite. Otherwise we can use the primal optimization method outlined in Appendices A and B, for example after expanding the data $\mathbf{x}_{ij}$ to include any attribute interactions – i.e. if we have $p$ binary attributes and consider all $p(p-1)/2$ pairwise interactions, then we can replace $\mathbf{x}_{ij}$ with a $p+p(p-1)/2$ dimensional vector to include these interactions and estimate $p+p(p-1)/2$ parameters per respondent using the RR-Het or LOG-Het (primal) estimation methods developed in Appendices A and B. In this particular example we will need to consider matrices of size $(p + p(p-1)/2) \times (p + p(p-1)/2)$ for each individual, which can be smaller than the size $IJ \times IJ$ of the kernel matrix below.

We begin with noting that for a fixed matrix $D$ and $\mathbf{w}_0$ both RR-Het and LOG-Het are decomposed across the $I$ respondents independently. For RR-Het, if we define $\hat{y}_{ij} = y_{ij} - \mathbf{x}_{ij}\mathbf{w}_0$ (see below how to estimate $\mathbf{x}_{ij}\mathbf{w}_0$ using dual parameters) and $\hat{\mathbf{w}}_i = \mathbf{w}_i - \mathbf{w}_0$, we can estimate $\hat{\mathbf{w}}_i$ by solving the individual level RR $\hat{\mathbf{w}}_i = \text{argmin} \sum_j (\hat{y}_{ij} - \mathbf{x}_{ij}\hat{\mathbf{w}}_i)^2 + \gamma \hat{\mathbf{w}}_i^\top D^{-1} \hat{\mathbf{w}}_i$. Similarly, for LOG-Het we can estimate $\hat{\mathbf{w}}_i$ by solving the individual level LOG-Het $\hat{\mathbf{w}}_i = \text{argmin} \sum_j -\log(\frac{e^{\mathbf{x}_{ijq_*}\mathbf{w}_0} e^{\mathbf{x}_{ijq_*}\hat{\mathbf{w}}_i}}{\sum_{q=1}^{Q} e^{\mathbf{x}_{ijq}\mathbf{w}_0} e^{\mathbf{x}_{ijq}\hat{\mathbf{w}}_i}}) + \gamma \hat{\mathbf{w}}_i^\top D^{-1} \hat{\mathbf{w}}_i$. A standard result in the literature (see

for example Vapnik 1998; Hastie et al., 2003; Wahba 1990) is that the minimizing vector $\hat{\mathbf{w}}_i$ of the individual level RR-Het or LOG-Het can be written as

$$\hat{\mathbf{w}}_i = \sum_{j=1}^{J} \alpha_{ij} D \mathbf{x}_{ij}^\top \tag{8}$$

where the coefficients $\alpha_{ij}$ are *dual* variables of the individual-level optimization problem for RR-Het (or LOG-Het) above. To estimate the $\alpha_{ij}$'s we replace $\hat{\mathbf{w}}_i$ with (8) and solve the resulting problem. For example for RR-Het we solve the optimization problem for respondent $i$ (for fixed $D$ and $\mathbf{w}_0$):

$$\min_{\alpha_{ij}} \frac{1}{\gamma} \sum_{j=1}^{J} \left( \hat{y}_{ij} - \sum_{k=1}^{J} \alpha_{ik} K_D(\mathbf{x}_{ij}, \mathbf{x}_{ik}) \right)^2 + \sum_{j,k=1}^{J} \alpha_{ij} \alpha_{ik} K_D(\mathbf{x}_{ij}, \mathbf{x}_{ik}) \tag{9}$$

(we similarly express $\mathbf{x}_{ijq} \hat{\mathbf{w}}_i$ using the dual parameters $\alpha_{ij}$ in LOG-Het), where we define the so called *kernel function* (Wahba 1990; Vapnik 1998)

$$K_D(\mathbf{x}, \mathbf{z}) := \mathbf{x} D \mathbf{z}^\top$$

and we write

$$\mathbf{x} \hat{\mathbf{w}}_i = \sum_{j=1}^{J} \alpha_{ij} \mathbf{x} D \mathbf{x}_{ij}^\top = \sum_{j=1}^{J} \alpha_{ij} K_D(\mathbf{x}, \mathbf{x}_{ij}). \tag{10}$$

With some abuse of notation, we note with $K_D$ the $IJ \times IJ$ *kernel matrix* with elements $(K_D)_{st} = K_D(\mathbf{x}_s, \mathbf{x}_t)$ where $s$ and $t$ correspond to some $ij$'s (taking all pairs of data across all respondents). Note that *we only need this kernel matrix, and not matrix $D$, in order*

*to estimate the dual parameters* $\alpha_{ij}$.

To consider attribute interactions and general nonlinear utility functions a standard approach (Wahba 1990; Hastie et al., 2003; Vapnik 1998) is to define a $d$-dimensional ($d$ much larger than $p$) feature space $\phi(\mathbf{x})$ that, for example, includes all attribute interactions (for example if we have $p$ binary attributes and consider all $p(p-1)/2$ pairwise interactions, then $\phi(\mathbf{x})$ is a $p + p(p-1)/2$ dimensional vector – or we can include higher order interactions and increase the dimensionality of $\phi(\mathbf{x})$ possibly making it infinite to consider any function). We then estimate a high dimensional $\mathbf{w}_i$ for each task such that the utility of a product $\mathbf{x}$ is $\phi(\mathbf{x})\mathbf{w}_i$. The key observation is that to estimate such a high dimensional (possibly infinite) $\mathbf{w}_i$ we still need to estimate only a finite number of parameters, namely the $I \times J$ dual parameters $\alpha_{ij}$, as long as we can compute the kernel matrix $K_D$ using $K_D(\mathbf{x}, \mathbf{z}) := \phi(\mathbf{x})D\phi(\mathbf{z})^\top$ (where matrix $D$ now has dimensionality equal to that of $\phi(\mathbf{x})$, possibly infinite). The dual derivations above still hold, replacing everywhere $\mathbf{x}$ with $\phi(\mathbf{x})$.

We now show how to iteratively estimate the kernel matrix $K_D$ and the dual parameters $\alpha_{ij}$ *without* estimating matrix $D$. This will be similar in spirit to the (primal) method discussed in Appendix A, but unlike that method we now iterate between $\alpha_{ij}$ and $K_D$ instead of $\{\mathbf{w}_i\}$ and $D$. We call this the *dual RR-Het (or LOG-Het)* method. This dual method consists of the following steps.

**Initialization of $\{\alpha_{ij}\}$, matrix $K_D$, and $\mathbf{x}_{ij}\mathbf{w}_0$:** We initialize the $\alpha_{ij}$'s with the in-

dividual level estimates, that is, setting $D$ to the identity matrix and $\mathbf{w}_0$ to a vector of zeros. To do so we only need to solve a standard individual level RR (or kernel logistic regression for LOG-Het) with kernel $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})\phi(\mathbf{z})^\top$ – for example this can be computed using the polynomial kernel $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}\mathbf{z}^\top)^2$ to incorporate all pairwise attribute interactions (see Vapnik 1998; Wahba 1990; Cui and Curry 2005; Evgeniou et al., 2005). Having the dual parameters $\alpha_{ij}$ we initialize $\mathbf{x}_{ij}\mathbf{w}_0^\top$ as the average of the $I$ numbers $\mathbf{x}_{ij}\hat{\mathbf{w}}_i$ using (10). We initialize matrix $K_D$ using the kernel $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})\phi(\mathbf{z})^\top$.

**Estimation of $\{\mathbf{x}_{ij}\mathbf{w}_0\}$ given $\{\alpha_{ij}\}$ and matrix $K_D$:** At every iteration we can estimate $\mathbf{x}_{ij}\mathbf{w}_0$ using the previous $\mathbf{x}_{ij}\mathbf{w}_0$ (initialized to 0) and the new $\mathbf{x}_{ij}\hat{\mathbf{w}}_i$ computed using the $\alpha_{ij}$'s from (10) as $\mathbf{x}_{ij}\mathbf{w}_{0,\text{new}} = \frac{1}{I}\sum_{i=1}^{I}(\mathbf{x}_{ij}\hat{\mathbf{w}}_i + \mathbf{x}_{ij}\mathbf{w}_0)$.

**Estimation of the dual parameters $\{\alpha_{ij}\}$ given matrix $K_D$ and $\{\mathbf{x}_{ij}\mathbf{w}_0\}$:** For a given kernel matrix $K_D$ we can estimate the dual parameters $\{\alpha_{ij}\}$ using standard dual estimation methods for RR-Het (Vapnik 1998; Hastie et al, 2003) or LOG-het (Minka 2003; Keerthi et al., 2005; Zhu and Hastie 2005; Jaakkola and Haussler 1999). We refer the reader to the rich literature on dual methods for RR-Het and LOG-Het for more information on dual estimation methods, and to Vapnik (1998), Hastie et al., (2003), Wahba (1990) for more information on the definition and use of kernels in general. We focus here on the iterative estimation of the kernel matrix $K_D$ given $\{\alpha_{ij}\}$ which we discuss next.

**Estimation of the matrix $K_D$ given $\{\alpha_{ij}\}$ and $\{\mathbf{x}_{ij}\mathbf{w}_0\}$:** To iteratively estimate $K_D$ given $\{\alpha_{ij}\}$ we use an observation standard in kernel methods used for example for kernel Principal Component Analysis (Mika et al., 1999). Let $W$ be the matrix with columns $\hat{\mathbf{w}}_i = (\mathbf{w}_i - \mathbf{w}_0)$. Matrices $WW^\top$ and $W^\top W$ have the same eigenvalues, $\sigma_1^2 \geq \cdots \geq \sigma_R^2 > 0$, where the $\sigma_r > 0$ are the (non-zero) singular values and $R \leq \min(d, I)$ with $d$ the dimensionality of $\phi(\mathbf{x})$ – typically $d \geq I$ (with $d$ possibly infinite) when we use the dual method. Thus we can use the $I \times I$ matrix $W^\top W$ to compute the singular values $\sigma_r$ instead of the (possibly infinite dimensional) matrix $WW^\top$ used to estimate matrix $D$ in the update equation of matrix $D$ (see Appendix A and Equation (13) below). For this purpose, we observe that

$$(W^\top W)_{i\ell} = \mathbf{w}_i^\top \mathbf{w}_\ell = \sum_{j,k=1}^{J} \alpha_{ij}\alpha_{\ell k} K_{D^2}(\mathbf{x}_{ij}, \mathbf{x}_{\ell k}) \tag{11}$$

where $K_{D^2}(\mathbf{x}, \mathbf{z}) := \phi(\mathbf{x})DD^\top\phi(\mathbf{z})^\top$. Moreover, if we denote by $\mathbf{u}_r \in \mathcal{R}^d$ and $\mathbf{v}_r \in \mathcal{R}^I$ the eigenvectors of $WW^\top$ and $W^\top W$ respectively, they are related through

$$\sigma_r \mathbf{u}_r = W\mathbf{v}_r \quad \text{and} \quad \sigma_r \mathbf{v}_r = W^\top \mathbf{u}_r. \tag{12}$$

We have (see Appendix A) that the new matrix $D$ is given by

$$D_{\text{new}} = \frac{(WW^\top)^{\frac{1}{2}}}{\text{Trace}(WW^\top)^{\frac{1}{2}}} = \beta \sum_r \sigma_r \mathbf{u}_r \mathbf{u}_r^\top, \tag{13}$$

where $\beta = (\sum_{r=1}^{R} \sigma_r)^{-1}$. Hence, we have that

$$K_{D_{\text{new}}}(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) D_{\text{new}} \phi(\mathbf{z})^\top = \beta \sum_r \sigma_r (\phi(\mathbf{x})\mathbf{u}_r)(\phi(\mathbf{z})\mathbf{u}_r). \tag{14}$$

In order to compute the right hand side of the above equation we use (12) to get

$$\phi(\mathbf{x})\mathbf{u}_r = \frac{1}{\sigma_r}\phi(\mathbf{x})W\mathbf{v}_r = \frac{1}{\sigma_r}\sum_{i=1}^{I}\phi(\mathbf{x})\hat{\mathbf{w}}_i\mathbf{v}_{ri} = \frac{1}{\sigma_r}\sum_{i=1}^{I}\sum_{j=1}^{J}\mathbf{v}_{ri}\alpha_{ij}K_D(\mathbf{x}_{ij}, \mathbf{x}) \tag{15}$$

where $\mathbf{v}_{ri}$ is the $i$'s element of the $I$-dimensional eigenvector $\mathbf{v}_r$. Finally, we get the new kernel matrix $K_{D_{\text{new}}}$ by combining equations (14) and (15) as:

$$K_{D_{\text{new}}} = \beta K_D A \Sigma^{-1} A^\top K_D, \tag{16}$$

where $A$ is the $IJ \times R$ matrix defined by $A_{ij,r} = \alpha_{ij}\mathbf{v}_{ri}$ and $\Sigma$ is the $R \times R$ diagonal matrix with elements $\Sigma_{rr} = \sigma_r$. Note that in order to alternate the minimization with respect to the dual parameters $\{\alpha_{ij}\}$ and $K_D$ we need to also update the matrix $K_{D^2}$ used in (11) which is similarly given by

$$K_{D^2} = \beta^2 K_D A A^\top K_D.$$

We note that if the dimensionality $d$ of $\phi(\mathbf{x})$ is finite, as is typically the case in practice (e.g., we incorporate all possible interactions among attributes), then we can compute the final $\mathbf{w}_i$'s using the dual parameters $\{\alpha_{ij}\}$ and matrix $D$ from (13) after each iteration.

If $d$ is infinite, we cannot estimate the $\mathbf{w}_i$'s. In that case, to estimate the utility of a new product $\mathbf{z}$ we use (10). To this purpose we need to compute $K_D(\mathbf{x}_{ij}, \mathbf{z})$. This can be done recursively using (16) and all the parameters $\beta$, $\Sigma$, $A$ computed across all the iterations.