

Hierarchical array priors for ANOVA decompositions

Alexander Volfovsky¹ and Peter D. Hoff^{1,2}

Departments of Statistics¹ and Biostatistics², University of Washington

August 9, 2012

Abstract

ANOVA decompositions are a standard method for describing and estimating heterogeneity among the means of a response variable across levels of multiple categorical factors. In such a decomposition, the complete set of main effects and interaction terms can be viewed as a collection of vectors, matrices and arrays that share various index sets defined by the factor levels. For many types of categorical factors, it is plausible that an ANOVA decomposition exhibits some consistency across orders of effects, in that the levels of a factor that have similar main-effect coefficients may also have similar coefficients in higher-order interaction terms. In such a case, estimation of the higher-order interactions should be improved by borrowing information from the main effects and lower-order interactions. To take advantage of such patterns, this article introduces a class of hierarchical prior distributions for collections of interaction arrays that can adapt to the presence of such interactions. These prior distributions are based on a type of array-variate normal distribution, for which a covariance matrix for each factor is estimated. This prior is able to adapt to potential similarities among the levels of a factor, and incorporate any such information into the estimation of the effects in which the factor appears. In the presence of such similarities, this prior is able to borrow information from well-estimated main effects and lower-order interactions to assist in the estimation of higher-order terms for which data information is limited.

Key Words: array-valued data; Bayesian estimation; cross-classified data; factorial design; MANOVA; penalized regression; tensor; Tucker product; sparse data.

This work was partially supported by NICHD grant 1R01HD067509-01A1.

1 Introduction

Cross-classified data are prevalent in many disciplines, including the social and health sciences. For example, a survey or observational study may record health behaviors of its participants, along with a variety of demographic variables such as age, ethnicity and education level, by which the participants can be classified. A common data analysis goal in such settings is the estimation of the health behavior means for each combination of levels of the demographic factors. In a three-way layout, for example, the goal is to estimate the three-way table of population cell means, where each cell corresponds to a particular combination of factor levels. A standard estimator of the table is provided by the table of sample means, which can alternatively be represented by its ANOVA decomposition into additive effects and two- and three-way interaction terms.

The cell sample means provide an unbiased estimator of the population means, as long as there are observations available for each cell. However, if the cell-specific sample sizes are small then it may be desirable to share information across the cells to reduce the variance of the estimator. Perhaps the simplest and most common method of information sharing is to assume that certain mean contrasts among levels of one set of factors are equivalent across levels of another set of factors, or equivalently, that certain interaction terms in the ANOVA decomposition of population cell means are exactly zero. This is a fairly large modeling assumption, and can often be rejected via plots or standard F -tests. If such assumptions are rejected, it still may be desirable to share information across cell means, although perhaps in a way that does not posit exact relationships among them.

As a concrete example, consider estimating mean macronutrient intake across levels of age, ethnicity and education from the National Health and Nutrition Examination Survey (NHANES). Table 1 summarizes the cell-specific sample sizes for intake of overall carbohydrates as well as two subcategories (sugar and fiber) by age, ethnicity, and education levels for male respondents (more details on these data are provided in Section 5). Studies of carbohydrate intake have been motivated by a frequently cited relationship between carbohydrate intake and health outcomes (Chandalia et al., 2000; Moerman et al., 1993). However, these studies generally report on marginal means of carbohydrate intake across demographic variables, and do not take into account potential non-additivity, or interaction terms, between them (Park et al., 2011; Montonen et al., 2003). A more detailed understanding of the relationship between mean carbohydrate intake and the demographic

Age	Mexican					Hispanic					White					Black				
	P	S	HD	AD	BD	P	S	HD	AD	BD	P	S	HD	AD	BD	P	S	HD	AD	BD
31-40	21	24	23	17	13	12	8	10	11	1	3	37	56	55	56	1	13	31	35	16
41-50	26	10	19	14	6	11	9	10	9	3	10	25	56	57	50	2	25	21	25	17
51-60	29	11	10	14	10	17	6	12	13	11	10	24	46	57	57	3	23	23	24	14
61-70	31	7	5	11	5	19	4	11	6	7	15	23	56	46	54	16	34	20	33	14
71-80	27	2	3	1	3	10	8	5	2	7	61	37	93	72	68	16	10	11	7	12

Table 1: Cross-tabulation of the demographic variables.

variables can be obtained from a MANOVA decomposition of the means array into main-effects, two- and three-way interactions. Evidence for interactions can be assessed with approximate F -tests based on the Pillai trace statistics (Olson, 1976), presented in Table 2.

	approx F	num df	den df	p -value
Education	11.15	15	6102	< 0.01
Ethnicity	18.07	9	6102	< 0.01
Age	21.38	12	6102	< 0.01
Education:Ethnicity	1.67	36	6102	0.01
Education:Age	1.60	48	6102	0.01
Ethnicity:Age	2.05	36	6102	< 0.01
Education:Ethnicity:Age	1.44	144	6102	< 0.01

Table 2: Testing MANOVA variance components via Pillai’s trace statistic.

The F -tests indicate strong evidence that the two- and three-way interactions are not zero. Based on these results, standard practice would be to retain the full model and describe the interaction patterns via various contrasts of sample cell means. However, OLS estimates of these interactions typically have undesirably high mean squared error (MSE), due to the small number of observations for each two- or three-way combination of factors. A variety of penalized least squares procedures have been proposed in order to reduce MSE, such as ridge regression and the lasso. Recent variants of these approaches allow for different penalties on ANOVA terms of different orders, including the ASP method of Beran (2005), and grouped versions of the lasso (Yuan and Lin,

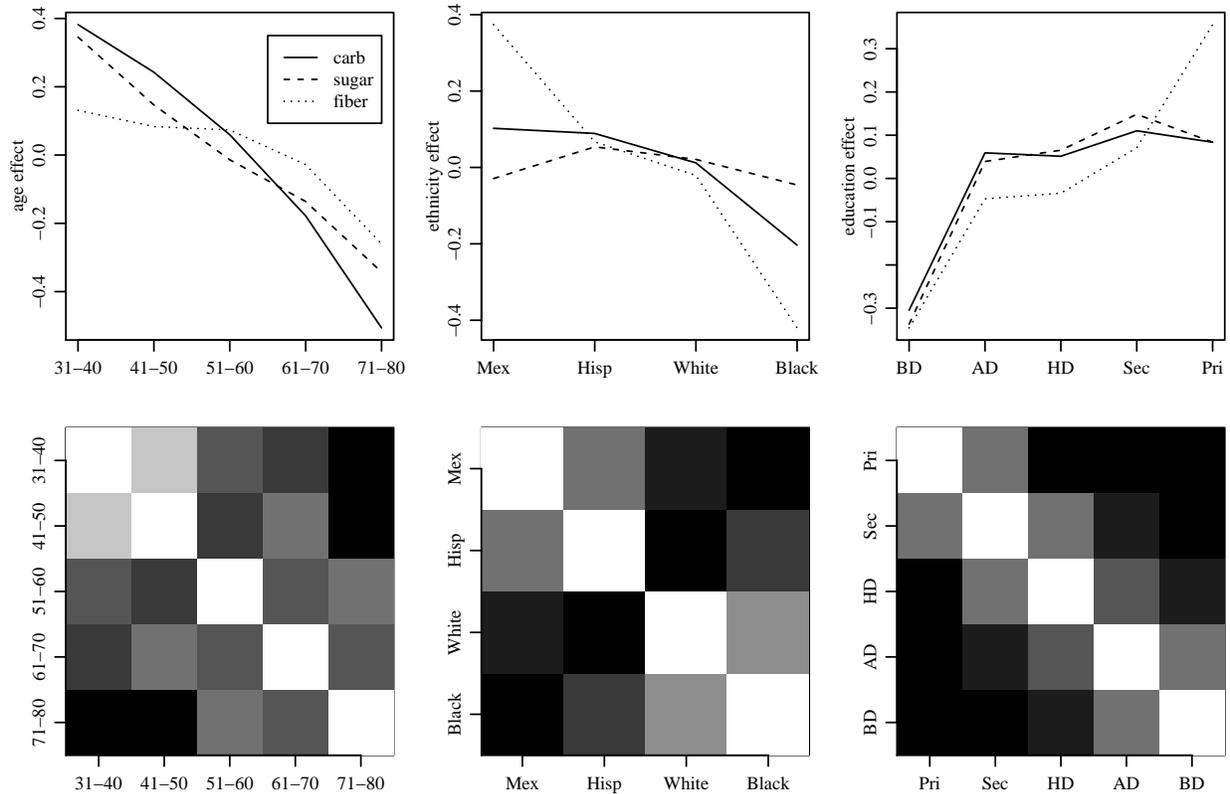


Figure 1: Plots of main effects and interaction correlations for the three outcome variables (carbohydrates, sugar and fiber). The first row of plots gives OLS estimates of the main effects for each factor. The second row of plots gives correlations of effects between levels of each factor, with white representing 1 and black representing -1.

2007; Friedman et al., 2010). Corresponding Bayesian approaches include Bayesian lasso procedures (Yuan and Lin, 2005; Genkin et al., 2007; Park and Casella, 2008) and multilevel hierarchical priors (Pittau et al., 2010; Park et al., 2006).

While these procedures attain a reduced MSE by shrinking linear model coefficient estimates towards zero, they do not generally take full advantage of the structure that is often present in cross-classified datasets. In the data analysis example above, two of the three factors (age and education) are ordinal, with age being a binned version of a continuous predictor. Considering factors such as these more generally, suppose a categorical factor x is a binned version of some underlying continuous or ordinal explanatory variable \tilde{x} (such as income, age, number of children or education level). If the mean of the response variable y is smoothly varying in the underlying

variable \tilde{x} , we would expect that adjacent levels of the factor x would have similar main effects and interaction terms. Similarly, for non-ordinal factors (such as ethnic group or religion) it is possible that two levels represent similar populations, and thus may have similar main-effects and interaction terms as well. We refer to such similarities across the orders of the effects as *order consistent interactions*,

Returning to the NHANES data, Figure 1 summarizes the OLS estimates of the main effects and two-way interactions for the three outcome variables (carbohydrates, sugar and fiber). Not surprisingly, the main effects for the ordinal factors (age and education) are “smooth,” in that the estimated main effect for a given level is generally similar to the effect for an adjacent level. Additionally, some similarities among the ethnic groups appear consistent across the three outcome variables. To assess consistency of such similarities between main effects and two-way interactions, we computed correlations of parameter estimates for these effects between levels of each factor. For example, there are $3 \times 10 = 30$ main-effect and two-way interaction estimates involving each level of age: For each of the three outcome variables, there is 1 main-effect estimate for each age level, 4 estimates from the age by ethnicity interaction and 5 estimates from the age by education interaction. We compute a correlation matrix for the five levels of age based on the resulting 30×5 matrix of parameter estimates, and similarly compute correlations among levels of ethnicity and among levels of education. The second row of Figure 1 gives grayscale plots of these correlation matrices. The results suggest some degree of order consistent interactions: For the ordinal factors, the highest correlations are among adjacent pairs. For the ethnicity factor, the results suggest that on average, the effects for the Mexican category are more similar to the Hispanic category than to the other ethnic categories, as we might expect.

The OLS estimates of the main effects and two-way interactions presented above, along with the fact that two of the three factors are ordinal, suggest the possibility of order consistent interactions among the array of population cell means. More generally, order consistent interactions may be present in a variety of datasets encountered in the social and health sciences, especially those that include ordinal factors, or factors for which some of the levels may represent very similar populations. In this paper, we propose a novel class of hierarchical prior distributions over main effects and interaction arrays that can adapt to the presence of order consistent interactions. The hierarchical prior distribution provides joint estimates of a covariance matrix for each factor, along

with the factor main effects and interactions. Roughly speaking, the covariance matrix for a given factor is estimated from the main effects and interactions in which the factor appears. Conversely, an estimate of a factor’s covariance matrix can assist in the estimation of higher-order interactions, for which data information is limited. We make this idea more formal in the next section, where we construct our prior distribution from a set of related array normal distributions with separable covariance structures (Hoff, 2011), and provide a Markov chain Monte Carlo algorithm for inference under this prior. In Section 3 we provide a simulation study comparing estimation under our proposed prior to some standard estimators. As expected, our approach outperforms others when the data exhibit order consistent interactions. Additionally, for data lacking any interactions, our approach performs comparably to the OLS estimates obtained from the additive model (i.e. the oracle estimator). In Section 4 we extend this methodology to MANOVA models in order to analyze the multivariate NHANES data presented above. In addition to estimates of main effects and interactions, our analysis provides measures of similarity between levels of each of the factors. We conclude in Section 5 with a summary of our approach and a discussion of possible extensions.

2 A hierarchical prior for interaction arrays

In this section we introduce the hierarchical array (HA) prior, and present a Markov chain Monte Carlo (MCMC) algorithm for posterior approximation and parameter estimation. The HA prior is constructed from several semi-conjugate priors, and so the MCMC algorithm can be based on a straightforward Gibbs sampling scheme.

2.1 The hierarchical array prior

For notational convenience we consider the case of three categorical factors, and note that the HA prior generalizes trivially to accommodate a greater number of factors. Suppose the three categorical factors have levels $\{1, \dots, m_1\}$, $\{1, \dots, m_2\}$ and $\{1, \dots, m_3\}$ respectively. The standard ANOVA model for a three-way factorial dataset is

$$\begin{aligned}
 y_{ijkl} &= \mu + a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} + \epsilon_{ijkl} \\
 \{\epsilon_{ijkl}\} &\sim \text{i.i.d. normal}(0, \sigma^2).
 \end{aligned}
 \tag{1}$$

Let a denote the $m_1 \times 1$ vector of main effects for the first factor, (ab) denote the $m_1 \times m_2$ matrix describing the two-way interaction between the first two factors, (abc) denote the $m_1 \times m_2 \times m_3$ three-way interaction array, and let b , c , (ac) , and (bc) be defined similarly. Bayesian inference for this model proceeds by specifying a prior distribution for the ANOVA decomposition $\theta = \{\mu, a, b, c, (ab), (ac), (bc), (abc)\}$ and the error variance σ^2 .

As described in the Introduction, if two levels of a factor represent similar populations, we would expect that coefficients of the decomposition involving these two levels would have similar values. For example, suppose levels i_1 and i_2 of the first factor correspond to similar populations. We might then expect a_{i_1} to be close to a_{i_2} , the vector $\{(ab)_{i_1,j}, j = 1, \dots, m_2\}$ to be close to the vector $\{(ab)_{i_2,j}, j = 1, \dots, m_2\}$, and so on. We represent this potential similarity between levels of the first factor with a covariance matrix Σ_a , and consider a prior distribution on the ANOVA decomposition such that

$$\begin{aligned}\text{Cov}[a] = \text{E}[aa^T] &= \Sigma_a, \\ \text{E}[(ab)(ab)^T] &= k_{ab}\Sigma_a, \\ \text{E}[(ac)(ac)^T] &= k_{ac}\Sigma_a, \\ \text{E}[(abc)_{(1)}(abc)_{(1)}^T] &= k_{abc}\Sigma_a,\end{aligned}$$

where k_{ab} , k_{ac} and k_{abc} are scalars. Here, $(abc)_{(1)}$ is the *matricization* of the array (abc) , which converts the $m_1 \times m_2 \times m_3$ array into an $m_1 \times (m_2 m_3)$ matrix (Kolda and Bader, 2009). To accommodate similar structure for the second and third factors, we propose the following prior covariance model for the main effects and interaction terms:

$$\begin{aligned}\text{Cov}[a] &= \Sigma_a & \text{Cov}[b] &= \Sigma_b & \text{Cov}[c] &= \Sigma_c \\ \text{Cov}[\text{vec}(ab)] &= \Sigma_b \otimes \Sigma_a / \gamma_{ab} & \text{Cov}[\text{vec}(bc)] &= \Sigma_c \otimes \Sigma_b / \gamma_{bc} & \text{Cov}[\text{vec}(ac)] &= \Sigma_c \otimes \Sigma_a / \gamma_{ac} \\ \text{Cov}[\text{vec}(abc)] &= \Sigma_c \otimes \Sigma_b \otimes \Sigma_a / \gamma_{abc},\end{aligned}$$

where “ \otimes ” is the Kronecker product. The covariance matrices Σ_a , Σ_b and Σ_c represent the similarities between the levels of each of the three factors, while the scalars γ_{ab} , γ_{ac} , γ_{bc} , γ_{abc} represent the relative (inverse) magnitudes of the interaction terms as compared to the main effects. Further specifying the priors on the ANOVA decomposition parameters as being mean-zero and Gaussian, the prior on a is then the multivariate normal distribution $N_{m_1}(0, \Sigma_a)$, and the prior on $\text{vec}(ab)$ is

$N_{m_1 m_2}(0, \Sigma_b \otimes \Sigma_a / \gamma_{ab})$. This latter distribution is sometimes referred to as a matrix normal distribution (Dawid, 1981). Similarly, the prior on $\text{vec}(abc)$ is $N_{m_1 m_2 m_3}(0, \Sigma_c \otimes \Sigma_b \otimes \Sigma_a / \gamma_{abc})$, which has been referred to as an array normal distribution (Hoff, 2011).

In most data analysis situations the similarities between the levels of a given factor and magnitudes of the interactions relative to the main effects will not be known in advance. We therefore consider a hierarchical prior so that Σ_a , Σ_b , Σ_c and the γ -parameters are estimated from the data. Specifically, we use independent inverse-Wishart prior distributions for each covariance matrix, e.g. $\Sigma_a \sim \text{inverse-Wishart}(\eta_{a0}, S_{a0}^{-1})$ and gamma priors for the γ -parameters, e.g. $\gamma_{ab} \sim \text{gamma}(\nu_{ab0}/2, \tau_{ab0}^2/2)$, where η_a , S_a , ν_{ab0} and τ_{ab0}^2 are hyperparameters to be specified (some default choices for these parameters are discussed at the end of this section). This hierarchical prior distribution can be viewed as an adaptive penalty, which allows for sharing of information across main effects and interaction terms: For example, any similarities between levels of the first factor that are consistent across the main effects and two-way interactions will be reflected in the estimate of Σ_a , which in turn assists in the estimation of the three-way interaction for which there is limited data information.

2.2 Posterior approximation

Due to the semi-conjugacy of the HA prior, posterior approximation can be obtained from a straightforward Gibbs sampling scheme. Under this scheme, iterative simulation of parameter values from the corresponding full conditional distributions generates a Markov chain having a stationary distribution equal to the target posterior distribution. For computational simplicity, we consider the case of a balanced dataset in which the sample size in each cell is equal to some common value n , in which case the data can be expressed as an $m_1 \times m_2 \times m_3 \times n$ four-way array Y . A modification of the algorithm to accommodate unbalanced data is discussed in the next subsection.

Derivation of the full conditional distributions of the grand mean μ and the error variance σ^2 are completely standard: Under a $N(\mu_0, \tau_0^2)$ prior for μ , the corresponding full conditional distribution is $N(\mu_1, \tau_1^2)$, where $\tau_1^2 = (1/\tau_0^2 + nm_1 m_2 m_3 / \sigma^2)^{-1}$ and $\mu_1 = \tau_1^2 (\mu_0 / \tau_0^2 + nm_1 m_2 m_3 \bar{r} / \sigma^2)$, where $\bar{r} = \sum_{ijkl} (y_{ijkl} - [a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk}]) / n$. Under an inverse-gamma($\nu_0/2, \nu_0 \sigma_0^2/2$) prior distribution, the full conditional distribution of σ^2 is an inverse-gamma($\nu_1/2, \nu_1 \sigma_1^2/2$) distribution, where $\nu_1 = \nu_0 + nm_1 m_2 m_3$, $\nu_1 \sigma_1^2 = \nu_0 \sigma_0^2 + \sum_{ijkl} (y_{ijkl} - \mu_{ijk})^2$ and $\mu_{ijk} = \mu + a_i + b_j +$

$c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ikj}$. Derivation of the full conditional distributions of parameters other than μ and σ^2 is straightforward, but slightly non-standard due to the use of matrix and array normal prior distributions for the interaction terms. In what follows, we compute the full conditional distributions for a few of these parameters. Full conditional distributions for the remaining parameters can be derived in an analogous fashion.

Full conditionals of a and (abc) : To identify the full conditional distribution of the vector a of main effects for the first factor, let

$$\begin{aligned} r_{ijkl} &= y_{ijkl} - \left(\mu + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} \right) \\ &= a_i + \epsilon_{ijkl}, \end{aligned}$$

i.e., r_{ijkl} is the “residual” obtained by subtracting all effects other than a from from the data. Since $\{\epsilon_{ijkl}\} \sim$ i.i.d. normal($0, \sigma^2$), we have

$$p(Y|\theta, \sigma^2) \propto_a \exp \left\{ -\frac{m_2 m_3 n}{2\sigma^2} (a^T a - 2a^T \bar{r}) \right\},$$

where $\bar{r} = (\bar{r}_1, \dots, \bar{r}_{m_1})$ with $\bar{r}_i = \sum_{jkl} r_{ijkl} / (m_2 m_3 n)$, $\theta = \{a, b, c, (ab), (ac), (bc), (abc)\}$ and “ \propto_a ” means “proportional to as a function of a .” Combining this with the $N_{m_1}(0, \Sigma_a)$ prior distribution for a , we have

$$p(a|Y, \theta_{-a}, \sigma^2) \propto_a \exp \left(-\frac{m_2 m_3 n}{2\sigma^2} [a^T a - 2a^T \bar{r}] - \frac{1}{2} a^T \Sigma_a^{-1} a \right)$$

and so the full conditional distribution of a is multivariate normal with

$$\begin{aligned} \text{Var}[a|Y, \theta_{-a}, \sigma^2] &= (\Sigma_a^{-1} + I m_2 m_3 n / \sigma^2)^{-1} \\ \text{E}[a|Y, \theta_{-a}, \sigma^2] &= (\Sigma_a^{-1} + I m_2 m_3 n / \sigma^2)^{-1} \bar{r} \times (m_2 m_3 n / \sigma^2), \end{aligned}$$

where I is the $m_1 \times m_1$ identity matrix.

Derivation of the full conditional distributions for the interaction terms is similar. For example, to obtain the full conditional distribution of (abc) let r_{ijkl} be the residual obtained after subtracting all other components of θ from the data, so that $r_{ijkl} = (abc)_{ijk} + \epsilon_{ijkl}$. Let \bar{r} be the three-way array of cell means of $\{r_{ijkl}\}$, so that $\bar{r}_{ijk} = \sum_l r_{ijkl} / n$. Combining the likelihood in terms of \bar{r} with the $N_{m_1 m_2 m_3}(0, \Sigma_c \otimes \Sigma_b \otimes \Sigma_a / \gamma_{abc})$ prior density for $\text{vec}(abc)$ gives

$$\begin{aligned} p((abc)|Y, \sigma^2, \Sigma_a, \Sigma_b, \Sigma_c, \gamma_{abc}, \theta_{-(abc)}) &\propto_{(abc)} \exp \left(-\frac{n}{2} \left[\text{vec}(abc)^T \text{vec}(abc) - 2 \text{vec}(abc)^T \text{vec}(\bar{r}) \right] \right) \times \\ &\exp \left(-\frac{1}{2} \text{vec}(abc)^T (\Sigma_c \otimes \Sigma_b \otimes \Sigma_a / \gamma_{abc})^{-1} \text{vec}(abc) \right) \end{aligned}$$

and so $\text{vec}(abc)$ has a multivariate normal distribution with variance and mean given by

$$\begin{aligned}\text{Var}[\text{vec}(abc)|Y, \theta_{-(abc)}, \sigma^2] &= \left((\Sigma_c \otimes \Sigma_b \otimes \Sigma_a / \gamma_{abc})^{-1} + In / \sigma^2 \right)^{-1} \\ \text{E}[\text{vec}(abc)|Y, \theta_{-(abc)}, \sigma^2] &= \left((\Sigma_c \otimes \Sigma_b \otimes \Sigma_a / \gamma_{abc})^{-1} + In / \sigma^2 \right)^{-1} \text{vec}(\bar{r}) \times n / \sigma^2.\end{aligned}$$

Full conditional distributions for the remaining effects can be derived analogously.

Full conditional of Σ_a : The parameters in the ANOVA decomposition whose priors depend on Σ_a are a , (ab) , (ac) and (abc) . For example, the conditional density of (ab) given Σ_a , Σ_b and γ_{ab} can be written as

$$\begin{aligned}p((ab)|\Sigma_a, \Sigma_b, \gamma_{ab}) &= |2\pi\Sigma_b \otimes \Sigma_a / \gamma_{ab}|^{-1/2} \exp(-\text{vec}(ab)^T [\Sigma_b \otimes \Sigma_a / \gamma_{ab}]^{-1} \text{vec}(ab) / 2) \\ &\propto_{\Sigma_a} |\Sigma_a|^{-m_2/2} \text{etr}(-\Sigma_a^{-1} \gamma_{ab} (ab)^T \Sigma_b^{-1} (ab) / 2) \\ &= |\Sigma_a|^{-m_2/2} \text{etr}(-\Sigma_a^{-1} S_{ab} / 2),\end{aligned}$$

where $S_{ab} = \gamma_{ab} (ab)^T \Sigma_b^{-1} (ab)$ and $\text{etr}(A) = \exp\{\text{trace}(A)\}$ for a square matrix A . Similarly, the priors for a , (ac) and (abc) involve the following terms:

$$\begin{aligned}S_a &= aa^T \\ S_{ac} &= \gamma_{ac} (ac) \Sigma_c^{-1} (ac)^T \\ S_{abc} &= \gamma_{abc} (abc)_{(1)} (\Sigma_c \otimes \Sigma_b)^{-1} (abc)_{(1)}^T.\end{aligned}$$

Multiplying together the prior densities for a , (ab) , (ac) and (abc) and the inverse-Wishart(η_{a0}, S_{a0}^{-1}) prior density for Σ_a gives

$$p(\Sigma_a | \theta, \Sigma_b, \Sigma_c, \gamma) \propto |\Sigma_a|^{-(1+m_1+\eta_{a0}+1+m_2+m_3+m_2m_3)/2} \text{etr}(-\Sigma_a^{-1} (S_{a0} + S_a + S_{ab} + S_{ac} + S_{abc}) / 2).$$

It follows that the full conditional distribution of Σ_a is inverse-Wishart(η_{a1}, S_{a1}^{-1}) where $\eta_{a1} = \eta_{a0} + (1 + m_2 + m_3 + m_2m_3)$ and $S_{a1} = S_{a0} + S_a + S_{ab} + S_{ac} + S_{abc}$. The full conditional expectation of Σ_a is therefore $S_{a1} / (\eta_{a1} - m_1 - 1)$, which combines several estimates of the similarities among the levels of the first factor, based the main effects and the interactions.

Full conditional of γ_{abc} : The full conditional distribution of γ_{abc} depends only on the (abc) interaction term. The normal prior for (abc) can be written as

$$p((abc)|\Sigma_a, \Sigma_b, \Sigma_c, \gamma_{abc}) \propto_{\gamma_{abc}} \gamma_{abc}^{m_1 m_2 m_3 / 2} \exp\{-\gamma_{abc} \text{vec}(abc)^T [\Sigma_c \otimes \Sigma_b \otimes \Sigma_a]^{-1} \text{vec}(abc)^T / 2\}$$

Combining this density with a $\text{gamma}(\nu_{abc0}/2, \tau_{abc0}^2/2)$ prior density yields a full conditional for γ_{abc} that is $\text{gamma}(\nu_{abc1}/2, \tau_{abc1}^2/2)$, where

$$\begin{aligned}\nu_{abc1} &= \nu_{abc0} + m_1 m_2 m_3 \\ \tau_{abc1}^2 &= \tau_{abc0}^2 + \text{vec}(abc)^T [\Sigma_c \otimes \Sigma_b \otimes \Sigma_a]^{-1} \text{vec}(abc).\end{aligned}$$

2.3 Balancing unbalanced designs

For most survey data we expect the sample sizes $\{n_{ijk}\}$ to vary across combinations of factors. As a result, the full conditional distributions of the ANOVA decomposition parameters are more difficult to compute. For example, the conditional variance of the three-way interaction $\text{vec}(abc)$ changes from $(\gamma_{abc}(\Sigma_c \otimes \Sigma_b \otimes \Sigma_a)^{-1} + In/\sigma^2)^{-1}$ in the balanced case to $(\gamma_{abc}(\Sigma_c \otimes \Sigma_b \otimes \Sigma_a)^{-1} + D/\sigma^2)^{-1}$ in the general case, where D is a diagonal matrix with diagonal elements $\text{vec}(\{n_{ijk}\})$. Even for moderate numbers of levels of the factors, the matrix inversions required to calculate the full conditional distributions in the unbalanced case can slow down the Markov chain considerably. As an alternative, we propose the following data augmentation procedure to “balance” an unbalanced design. Let \bar{Y}^o be the three-way array of cell means based on the observed data, i.e. $\bar{y}_{ijk}^o = \sum y_{ijkl}/n_{ijk}$. Letting $n = \max(\{n_{ijk}\})$, for each cell ijk with sample size $n_{ijk} < n$ and at each step of the Gibbs sampler, we impute a cell mean based on the “missing” $n - n_{ijk}$ observations as $\bar{y}_{ijk}^m \sim \text{normal}(\mu_{ijk}, \sigma^2/[n_{\max} - n_{ijk}])$, where μ_{ijk} is the population mean for cell ijk based on the current values of the ANOVA decomposition parameters. We then combine \bar{y}_{ijk}^o and \bar{y}_{ijk}^m to form the “full sample” cell mean $\bar{y}_{ijk}^f = (n_{ijk}\bar{y}_{ijk}^o + (n - n_{ijk})\bar{y}_{ijk}^m)/n$. This array of cell means provides the sufficient statistics for a balanced dataset, for which the full conditional distributions derived above can be used.

2.4 Setting hyperparameters

In the absence of detailed prior information about the parameters, we suggest using a modified empirical Bayes approach to hyperparameter selection based on the maximum likelihood estimates (MLEs) of the error variance and mean parameters. Priors for μ and σ^2 can be set as unit information priors (Kass and Wasserman, 1995), whereby hyperparameters are chosen so that the prior means are near the MLEs but the prior variances are set to correspond roughly to only one observation’s worth of information. For the covariance matrices Σ_a , Σ_b and Σ_c , recall that the prior

for the main effect a of the first factor is $N_{m_1}(0, \Sigma_a)$. Based on this, we choose the prior for Σ_a to be inverse-Wishart(ν_{a0}, S_{a0}^{-1}) with $\nu_{a0} = m_1 + 2$ and $S_{a0} = |\hat{a}|^2 I_{m_1} / m_1$, where \hat{a} is the MLE of a . Under this prior, $E[\text{tr}(\Sigma_a)] = |\hat{a}|^2$, and so the scale of the prior matches the empirical estimates. Finally, the γ -parameters can be set analogously, using diffuse gamma priors but centered around values to match the magnitude of the OLS estimates of the interaction terms they correspond to, relative to the magnitude of the main effects. For example, in the next section we use a gamma($\nu_{ab0}/2, \tau_{ab0}^2/2$) prior for γ_{ab} in which $\nu_0 = 1$ and $\tau_{ab0}^2 = |\hat{a}|^2 |\hat{b}|^2 / |(\hat{ab})|^2$, where \hat{a} , \hat{b} and (\hat{ab}) are the OLS estimates.

3 Simulation study

This section presents the results of three simulation studies comparing the HA prior to several competing approaches. The first simulation study uses data generated from a means array that exhibits order consistent interactions. Estimates based on the HA prior outperform standard OLS estimates as well as estimates from a standard Bayesian approach that is similar to the one in Gelman (2005), and is also related to a grouped version of the lasso procedure (Yuan and Lin, 2006). The second simulation study uses data from a means array that exhibits “order inconsistent” interactions, i.e. interactions without consistent similarities in parameter values between levels of a factor. In this case the HA prior still outperforms the OLS and standard Bayes approaches, although not by as much as in the presence of order consistent interactions. The third simulation study uses data from a means array that has an exact additive decomposition, i.e. there are no interactions. The HA prior procedure again outperforms the standard Bayes and OLS approaches, although it does not do as well as OLS and Bayes oracle estimators that assume the correct additive model.

3.1 Data with order consistent interactions

The data in this simulation study is generated from a model where the means array exhibits order consistent interactions. The dimensions of the means array M were chosen to be $m_1 \times m_2 \times m_3 = 15 \times 7 \times 3$, which could represent, for example, the number of categories we might have for age, education level and political affiliation in a cross-classified survey dataset. The means array was generated from a cubic function of three variables that was then binned. Figure 2 plots the mean array across the third factor, demonstrating the nonadditivity present in M . By decomposing M

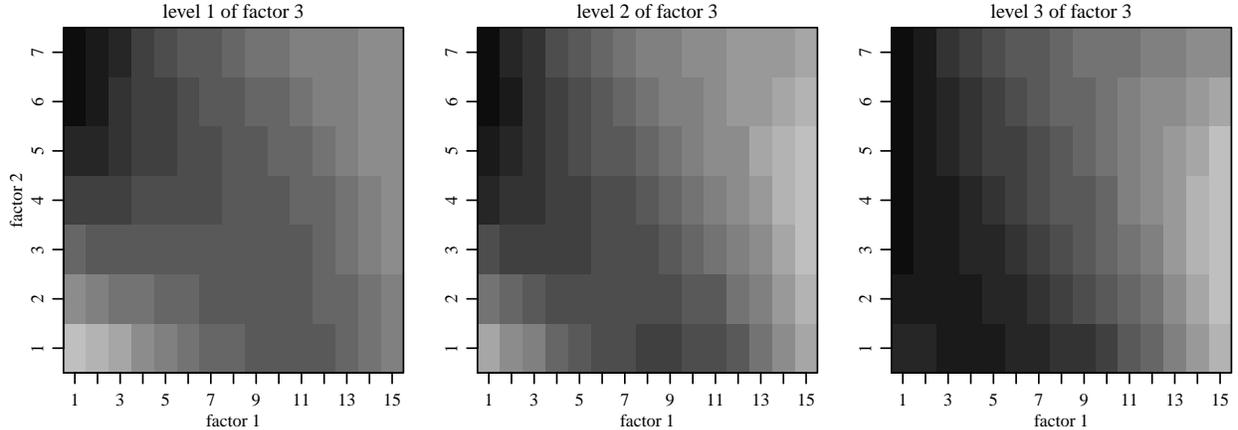


Figure 2: The means array M across levels of the third factor.

into the main, two-way and three-way effects in the same manner as described in Section 2, we can summarize the nonadditivity of M through the magnitudes of the different sums of squares. The magnitudes of the main effects $\|a\|^2/m_1$, $\|b\|^2/m_2$, and $\|c\|^2/m_3$ are 5.267, 0.012, 0.004 respectively. Those of the two-way interactions $\|ab\|^2/(m_1m_2)$, $\|ac\|^2/(m_1m_3)$, and $\|bc\|^2/(m_2m_3)$ are 1.365, 1.312, and 0.384, and the magnitude of the three-way interaction $\|abc\|^2/(m_1m_2m_3)$ is 0.474. For each sample size $n \in \{400, 1000, 5000, 10000\}$, we simulated 50 datasets using the mean array M and independent standard normal errors. In order to make a comparison to OLS possible, we first allocated one observation to each cell of the means array and then distributed the remaining observations uniformly, leading to a complete but potentially unbalanced design. The average number of observations per cell under the sample sizes $\{400, 1000, 5000, 10000\}$ was $\{1.3, 3.2, 15.9, 31.7\}$.

For each simulated dataset we obtained estimates under the HA prior (using the hyperparameter specifications described in Section 2.4), as well as ordinary least squares estimates (OLS) and posterior estimates under a standard Bayesian prior (SB). The SB approach is essentially a simplified version of the HA prior in which the parameter values are conditionally independent given the hyperparameters: $\{a_i\} \sim \text{i.i.d. } N(0, \sigma_a^2)$, $\{(ab)_{ij}\} \sim \text{i.i.d. } N(0, \sigma_{ab}^2)$ and $\{(abc)_{ijk}\} \sim \text{i.i.d. } (0, \sigma_{abc}^2)$ and similarly for all other main effects and interactions. To facilitate comparison to the HA prior, the hyperpriors for these σ^2 -parameters are the same as the hyperpriors for the inverses of the γ -parameters in the HA approach. As a result, this standard Bayes prior can be seen as the limit of a sequence of HA priors where the inverse-Wishart prior distributions for the Σ -matrices converge to point-masses on the identity matrices of the appropriate dimension.

For each simulated dataset, the Gibbs sampler described in Section 2 was run for 11,000 iterations, the first 1,000 of which were dropped to allow for convergence to the stationary distribution. Parameter values were saved every 10th scan, resulting in 1,000 Monte Carlo samples per simulation. Starting values for all the mean effects were set to zero and all variances set to identity matrices of the proper dimensions. We examined the convergence and autocorrelation of the marginal samples of the error variance σ^2 using Geweke’s z -test and the effective sample size. The minimum effective sample size across all simulations was 233 out of the 1000 recorded scans, and the average effective sample size was 895. Geweke’s z -statistic was less than 2 in absolute value in 93, 93, 97, and 95 percent of the Markov chains for the four sample sizes (with the percentages being identical for both Bayesian methods). While the cases in which $|z| > 2$ were not extensively examined, it is assumed that running the chain longer would have yielded improved estimation.

For each simulated data set, the posterior mean estimates \hat{M}_{HA} and \hat{M}_{SB} were obtained by averaging their values across the 1,000 saved iterations of the Gibbs sampler. The average squared error (ASE) of estimation was calculated as $ASE(\hat{M}) = \|\hat{M} - M\|^2 / (m_1 m_2 m_3)$ where M is the means array that generated the data. These values were then compared across the three approaches. The left panel of Figure 3 demonstrates that the SB estimator provided a reduction in ASE when

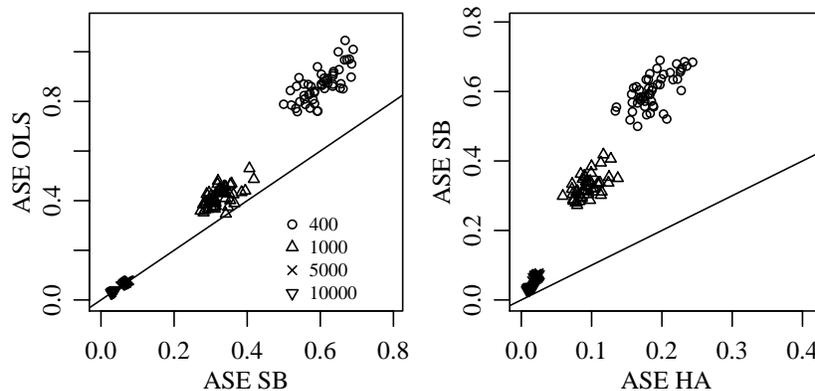


Figure 3: Comparison of ASE for different estimation methods.

compared to the OLS estimator for all data sets with sample sizes 400 and 1000, 96% of the data sets with sample size 5000 and 90% of data sets with sample size 10000. The second panel demonstrates that the HA estimator provides a substantial further reduction in ASE for all data sets. As we would expect, the reduction in ASE is dependent on the sample size and decreases as the sample

size increases.

These results are not surprising: By estimating the variances $\sigma_a^2, \sigma_{ab}^2$, etc. from the data, the SB approach provides adaptive shrinkage and so we expect these SB estimates to outperform the OLS estimates in terms of ASE. However, the SB approach does not use information on the similarity among the levels of an effect, and so its estimation of higher order interactions relies on the limited information available directly in the corresponding sufficient statistics. As such, we expect the SB estimates to perform less well than the HA estimates, which are able to borrow information from well-estimated main effects and low-order interactions to assist in the estimation of higher-order terms for which data information is limited. Additionally, recall that the parameters in the mean array M were generated by binning a third-degree polynomial, and were not generated from array normal distributions, i.e. the HA prior is “incorrect” as a model for M . Even so, the HA prior is able to capture the similarities between adjacent factor levels, resulting in improved estimation. However, we note that not all of the improvement in ASE achieved by the HA prior should be attributed to the identification of order-consistent interactions. The simulation study that follows suggests some of the performance of the HA prior is due to additional parameter shrinkage provided by the inverse-Wishart distributions on the Σ -matrices

3.2 Data with order inconsistent interactions

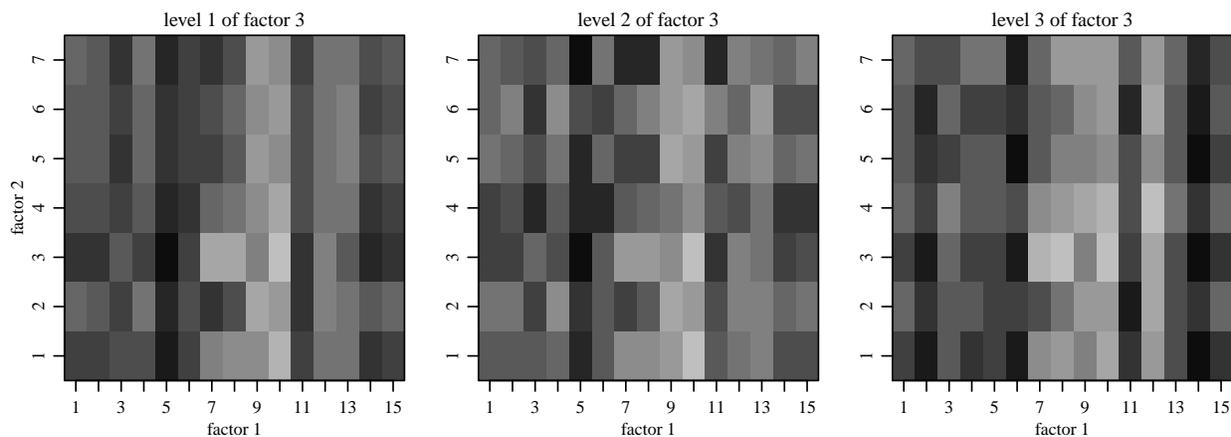


Figure 4: The means array M for the second simulation study, across levels of the third factor.

In this subsection we evaluate the HA approach for populations which exhibit interactions that are order inconsistent. The means array M is constructed by taking the means array from Section

3.1, decomposing it into main effects, two- and three-way interactions, permuting the levels of each factor within each effect, and reconstructing a means array. That is, if $\{a_i : i = 1, \dots, m_1\}$ is the collection of parameters for the first main effect and $\{(ab)_{ij} : i = 1, \dots, m_1, j = 1, \dots, m_2\}$ is the collection of parameters for the two way interaction between the first and second factors in Section 3.1 then $\{a_{\pi_1(i)}\}$ and $\{(ab)_{\pi_2(i)\pi_3(j)}\}$ are the main effect and two-way interaction parameters for the means array in this section, where π_1, π_2 and π_3 are independent permutations. The remaining effects were permuted analogously. Due to this construction, the magnitudes of the main effects, two- and three-way interactions remain the same, but the process becomes less “smooth,” as can be seen in Figure 4.

Again, 50 data sets were generated for each sample size, and estimates \hat{M}_{HA} , \hat{M}_{SB} and \hat{M}_{OLS} were obtained for each data set, where the Bayesian estimates were obtained using the same MCMC approximation procedure as in the previous subsection. Figure 5 compares ASE across the different approaches. The left panel of Figure 5, as with the left panel of Figure 3, demonstrates that the SB estimator provides a reduction in ASE when compared to the OLS estimator. As expected, since neither of these approaches take advantage of the structure of the order consistent interactions, this plot is nearly identical to the corresponding plot in Figure 3.

The second panel demonstrates that the HA estimator provides a further reduction in ASE for all data sets, although this reduction is less substantial than in the presence of order consistent interactions. The lower ASE of the HA estimates may be initially surprising, as there are no order consistent interactions for the HA prior take advantage of. We conjecture that the lower ASE is due to the additional shrinkage on the parameter estimates that the inverse-Wishart priors on the Σ -parameters provide. For example, under both the SB and HA priors we have $\text{Cov}[\text{vec}(ab)] = \Sigma_b \otimes \Sigma_a / \gamma_{ab}$, but under the former the covariance matrices are set to the identity whereas under the latter they have inverse-Wishart distributions.

3.3 Data without interactions

In this subsection we evaluate the HA approach for populations in which interactions are not present. In addition to the SB and OLS estimators, we compare the HA approach to two “oracle” estimators: the additive model least squares estimator (AOLS) and Bayes estimator under the additive model (ASB). The prior used by the ASB approach is the same as the SB prior, but does

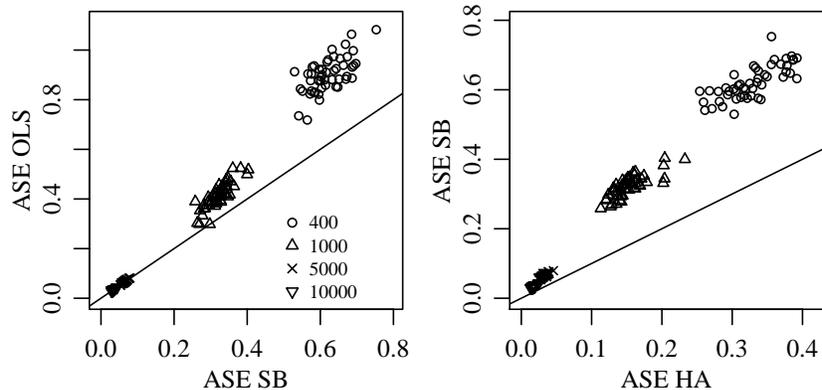


Figure 5: Comparison of ASE for different estimation methods.

not include terms other than main effects in the model.

As before, 50 data sets were generated for each sample size, and estimates \hat{M}_{HA} , \hat{M}_{SB} , \hat{M}_{OLS} , \hat{M}_{ASB} and \hat{M}_{AOLS} were obtained for each data set, where the Bayesian estimates were obtained using the same MCMC approximation procedure as in the previous two subsection. Some results are shown in Figure 6, which compares ASE across the different approaches. In the top row of Figure 6 we see that the performance of the HA estimates is comparable to but not as good as the the oracle least squares and Bayesian estimates in terms of ASE. Specifically, the ASE for the HA estimates is 24.2, 18.6, 20.1 and 17.4 percent higher than for the AOLS estimates for data sets with sample sizes 400, 1000, 5000 and 10000 respectively. Similarly, the ASE for the HA estimates is 25, 19.7, 20.3 and 17.8 percent higher than for the ASB estimates for data sets with sample sizes 400, 1000, 5000 and 10000 respectively. However, the bottom row of Figure 6 shows that the HA prior is superior to the other non-oracle OLS and SB approaches that attempt to estimate the interaction terms.

These results, together with those of the last two subsections, suggest that the HA approach provides a competitive method for fitting means arrays in the presence or absence of interactions. When order consistent interactions are present, the HA approach is able to make use of the similarities across levels of the factors, thereby outperforming approaches that cannot adapt to such patterns. Additionally, the HA approach does not appear to suffer when interactions are not order consistent. In the absence of interactions altogether, the HA approach adapts well, providing estimates similar to those that assume the correct additive model.

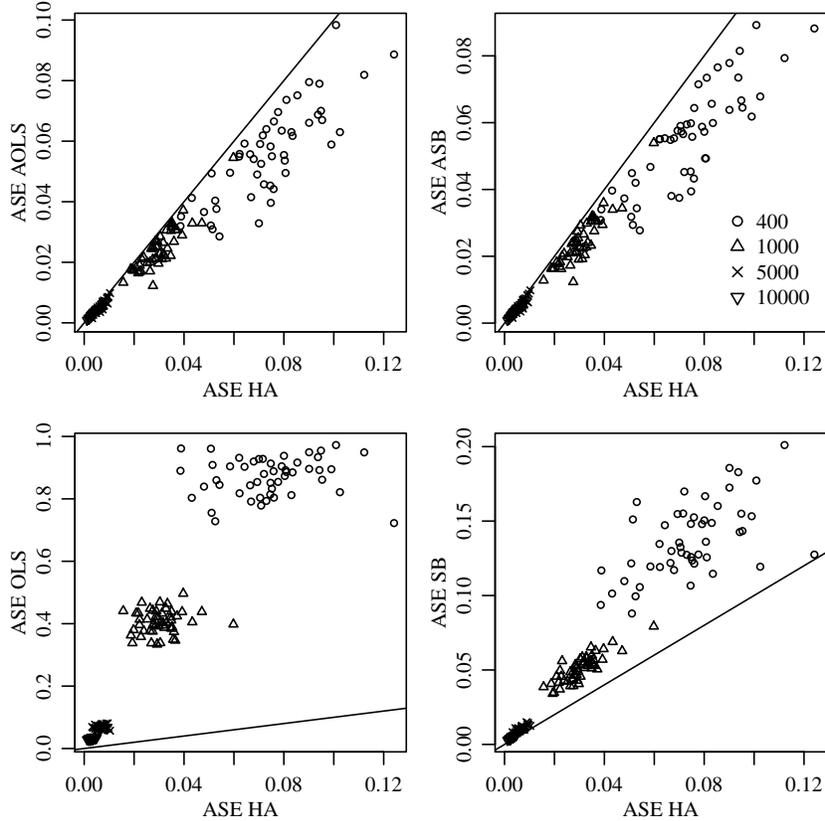


Figure 6: Comparison of ASE for different estimation methods.

4 Analysis of carbohydrate intake

In this section we estimate average carbohydrate, sugar and fiber intake by education, ethnicity and age using the HA procedure described in Section 2. Our estimates are based on data from 2134 males from the US population, obtained from the 2007-2008 NHANES survey. Nutrient intake is self reported on two non-consecutive days. Each day's data include food and beverage intake information from the preceding 24 hour period only, and is calculated using the USDA's Food and Nutrient Database for Dietary Studies 4.1 (USDA, 2010). All intake was measured in grams, and we average the intake over the two days to yield a single measurement per individual. We are interested in relating the intake data to the following demographic variables:

- Age: (31 – 40), (41 – 50), (51 – 60), (61 – 70), (71 – 80).
- Education: Primary (P), Secondary (S), High School diploma (HD), Associates degree (AD), Bachelors degree (BD).

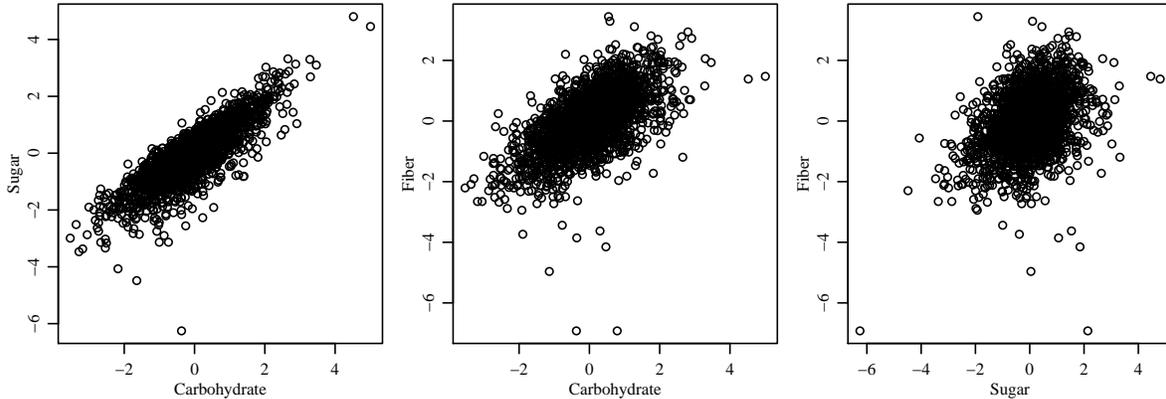


Figure 7: Two-way plots of the transformed data.

- Ethnicity: Mexican (not Hispanic), Hispanic, white (not Hispanic) and black (not Hispanic).

Sample sizes for age-education-ethnicity combinations were presented in Table 1 in Section 1. Of the 2234 male respondents within the above demographic groups, 100 were missing their nutrient intake information for both days, with similar rates of missingness across the demographic variables. For the purpose of this example analysis, we treat this information as missing at random.

The data on the original scale are somewhat skewed and show heteroscedasticity across the demographic variables. Since different variances across groups can lead to bias in the sums of squares estimates, making F -tests for interactions anti-conservative (Miller and Brown, 1997), stabilizing the variance is desirable. Figure 7 provides two-way scatterplots of the response variables after applying a quarter power transformation to each variable, which we found stabilized the variances across the groups better than either a log or square-root transformation. Additionally, we centered and scaled each response variable to have mean zero and variance one.

4.1 MANOVA model and parameter estimation

As presented in Table 1 of Section 1, F -tests indicate evidence for the presence of interactions in the array of population cell means. However, 12% of all age-education-ethnicity categories have sample sizes less than 5, and so we are concerned with overfitting of the OLS estimates. As an alternative, we extend the HA methodology described in Section 2 to accommodate a MANOVA model. Our MANOVA model has the same form as the ANOVA model given by Equation (1), except that each effect listed there is a three-dimensional vector corresponding to the separate effects for each of the

three response variables. Additionally, the error terms now have a multivariate normal distribution with zero-mean and unknown covariance matrix Σ_y .

We extend the hierarchical array prior discussed above to accommodate the p -variate MANOVA model as follows: Our prior for the $m_1 \times p$ matrix a of main effects for the first factor is $\text{vec}(a) \sim N_{m_1 p}(0, I \otimes \Sigma_a)$, where I is the $p \times p$ identity matrix. Our prior for the $m_1 \times m_2 \times p$ array (ab) of two-way interaction terms is given by $\text{vec}(ab) \sim N_{m_1 m_2 p}(0, \Gamma_{ab}^{-1} \otimes \Sigma_b \otimes \Sigma_a)$. Here, Γ_{ab} is a $p \times p$ diagonal matrix whose terms determine the scale of the two-way interactions for each of the p response variables. Similarly, our prior for the four-way array (abc) of three-way interaction terms is $\text{vec}(abc) \sim N_{m_1 m_2 m_3 p}(0, \Gamma_{abc}^{-1} \otimes \Sigma_c \otimes \Sigma_b \otimes \Sigma_a)$. Priors for other main effects and interaction terms are defined similarly. The hyperpriors for each diagonal entry of Γ are independent gamma distributions, chosen as in Section 2.4 so that the prior magnitude of the effects for each response is centered around the sum of squares of the effect from the OLS decomposition.

A Gibbs sampling scheme similar to the one outlined in Section 2 was iterated 200,000 times with parameter values saved every 10 scans, resulting in 20,000 simulated values of the means array M and the covariance matrices $\{\Sigma_{\text{eth}}, \Sigma_{\text{age}}, \Sigma_{\text{edu}}\}$ for posterior analysis. Mixing of the Markov chain for M was good: Figure 8 shows MCMC samples of 4 out of 320 entries of M (chosen so that their trace plots were visually distinct). The autocorrelation across the saved scans was low, with the lag-10 autocorrelation for the thinned chain being less than 0.14 in absolute value for each element of M (97.3% of entries have lag-10 autocorrelation less than 0.07 in absolute value) and effective sample sizes between 1929 and 13520. The mixing for the elements of the covariance matrices $\Sigma_{\text{eth}}, \Sigma_{\text{age}}, \Sigma_{\text{edu}}$ was not as good as that of the means array M : The maximum absolute value of lag-10 autocorrelation of the saved scans for the three rescaled covariance matrices was 0.18, 0.12, and 0.19 respectively. The effective sample sizes for the elements of the covariance matrices were at least 1684.

4.2 Posterior inference for M and the Σ s

We obtain a Monte Carlo approximation to $\hat{M} = E[M|Y]$ by averaging over the saved scans of the Gibbs sampler. Figure 9 provides information on the shrinkage and regularization of the estimates due to the HA procedure, as compared to OLS. The first panel plots the difference between the OLS and Bayes estimates of the cell means versus cell-specific sample sizes. For small sample sizes,

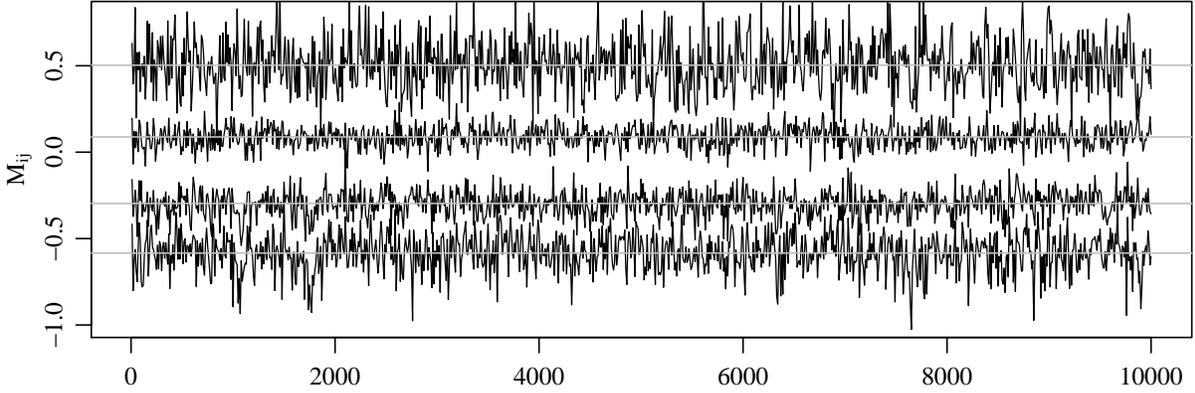


Figure 8: MCMC samples of 4 out of 320 entries of the means array M .

the Bayes estimate for a given cell is affected by the data from related cells, and can generally be quite different from the OLS estimate (the cell sample mean). For cells with large sample sizes the difference between the two estimates is generally small. The second panel of the figure plots the OLS estimates of the cell means for carbohydrate intake of black survey participants across age and education levels. Note that there appears to be a general trend of decreasing intake with increasing age and education level, although the OLS estimates themselves are not consistently ordered in this way. In contrast, these trends are much more apparent in the Bayes estimates plotted in the third panel. The HA prior allows the parameter estimates to be close to additive, while not enforcing strict additivity in this situation where we have evidence of non-additivity via the F -tests.

The first row of Figure 10 provides the estimates of the main effects from the HA procedure. The second row of Figure 10 summarizes covariance matrices $\{\Sigma_{\text{eth}}, \Sigma_{\text{age}}, \Sigma_{\text{edu}}\}$ via the posterior mean estimates of the correlation matrices $\{C_{d,ij}\} = \{\Sigma_{d,ij}/\sqrt{\Sigma_{d,ii}\Sigma_{d,jj}}\}$ for $d \in \{\text{eth}, \text{age}, \text{edu}\}$. In this figure, the diagonal elements are all 1, and darker colors represent a greater departure from one. The range of the estimated correlations was -0.34 to 0.42 for age categories, -0.30 to 0.35 for ethnic groups, and -0.37 to 0.38 for educational categories. For the two ordered categorical variables, age and education, we see that closer categories are generally more positively correlated than ones that are further apart. While the ethnicity variable is not ordered, its correlation matrix informs us of which categories are more similar in terms of these response variables. The middle panel of the second row of Figure 10 confirms the order-consistent interactions we observed in Figure 1: Mexican survey participants are more similar to Hispanic participants in terms of carbohydrate

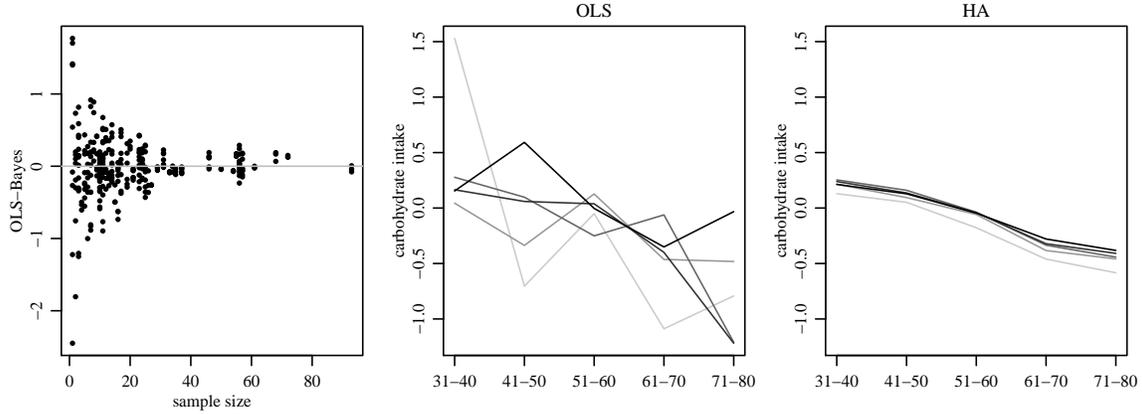


Figure 9: Shrinkage and regularization plots. The first panel plots the difference between the OLS and HA estimates of a cell-mean against the cell-specific sample sizes. The second and third panels plot estimated cell-means for black survey participants across age and education levels, where lighter shades represent higher levels of education.

intake patterns than to white or black participants.

5 Discussion

This article has presented a novel hierarchical Bayes method for parameter estimation of cross-classified data under ANOVA and MANOVA models. Unlike least-squares estimation, a Bayesian approach provides for regularized estimates of the potentially large number of parameters in a MANOVA model. Unlike the non-hierarchical Bayesian approach, the hierarchical approach provides a data-driven method of regularization, and unlike the standard hierarchical Bayes, the hierarchical array prior can identify similarities among categories and share this information across interaction effects to assist in the estimation of higher-order terms for which data information is limited. In a simulation study the HA approach was able to detect interactions when they were present, and to estimate the means array better than a full least squares or standard Bayesian approaches (in terms of mean squared error). When the true means array was completely additive, the HA prior was able to adapt to this smaller model better than the other full model estimation approaches under consideration.

Generalizations of the HA prior are applicable to any model whose parameters consist of vectors, matrices and arrays for which some of the index sets are shared. This includes generalized linear

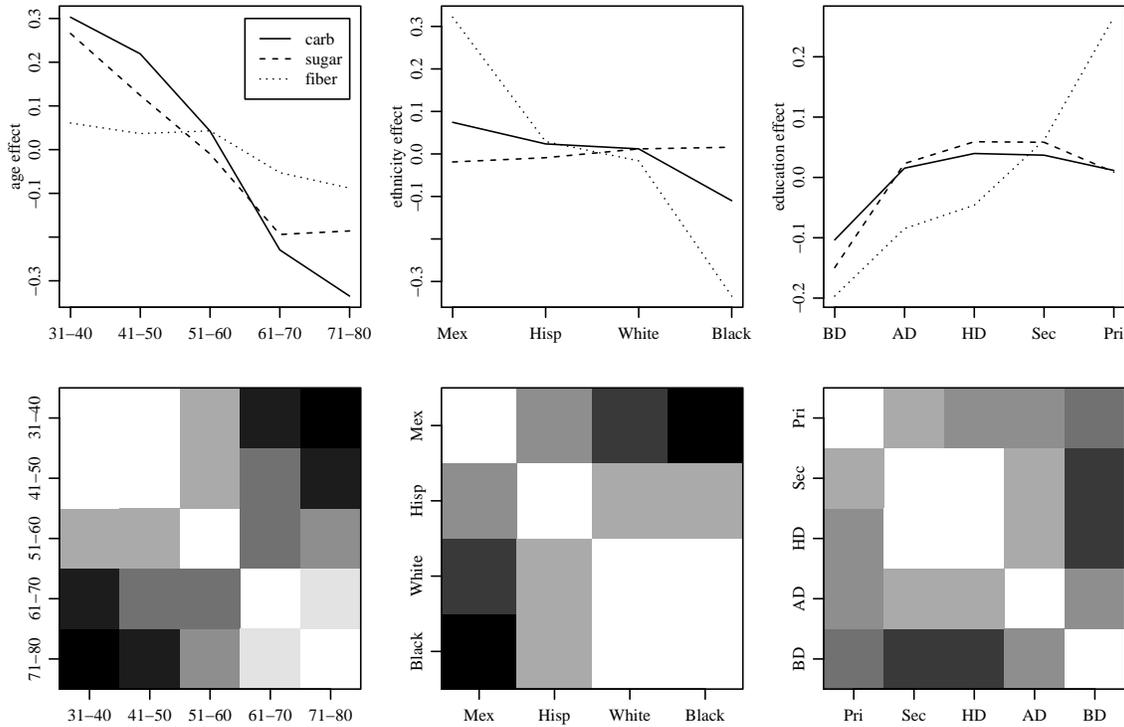


Figure 10: Plots of main effects and interaction correlations for the three outcome variables (carbohydrates, sugar and fiber). The first row of plots gives HA estimates of the main effects for each factor. The second row of plots gives correlations of effects between levels of each factor, with white representing 1 and black representing -1.

models with categorical factors, as well as ANCOVA models that involve interactions between continuous and categorical explanatory variables. As an example of the latter case, suppose we are interested in estimating the linear relationship between an outcome and a set of explanatory variables for every combination of three categorical factors. The regression parameters then consist of an $m_1 \times m_2 \times m_3 \times p$ array, where m_1, m_2, m_3 are the numbers of factor levels and p is the number of continuous regressors. The usual ANCOVA decomposition can be used to parametrize this array in terms of main effects and interactions arrays, for which a hierarchical array prior may be used.

Computer code and data for the results in Sections 3 and 4 are available at the authors' websites:

www.stat.washington.edu/~volff, www.stat.washington.edu/~hoff

References

- Beran, R. (2005). ASP fits to multi-way layouts. *Annals of the Institute of Statistical Mathematics*, 57(2):201–220.
- Chandalia, M., Garg, A., Lutjohann, D., von Bergmann, K., Grundy, S., and Brinkley, L. (2000). Beneficial effects of high dietary fiber intake in patients with type 2 diabetes mellitus. *New England Journal of Medicine*, 342(19):1392–1398.
- Dawid, A. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *Arxiv preprint arXiv:1001.0736*.
- Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, 33(1):1–53.
- Genkin, A., Lewis, D., and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304.
- Hoff, P. (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196.
- Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934.
- Kolda, T. and Bader, B. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455.
- Miller, R. and Brown, B. (1997). *Beyond ANOVA: basics of applied statistics*. Chapman & Hall/CRC.
- Moerman, C., De Mesquita, H., and Runia, S. (1993). Dietary sugar intake in the aetiology of biliary tract cancer. *International journal of epidemiology*, 22(2):207–214.

- Montonen, J., Knekt, P., Järvinen, R., Aromaa, A., and Reunanen, A. (2003). Whole-grain and fiber intake and the incidence of type 2 diabetes. *The American journal of clinical nutrition*, 77(3):622–629.
- Olson, C. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83(4):579.
- Park, D., Gelman, A., and Bafumi, J. (2006). State level opinions from national surveys: Post-stratification using multilevel logistic regression. *Public Opinion in State Politics*, pages 209–28.
- Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.
- Park, Y., Subar, A., Hollenbeck, A., and Schatzkin, A. (2011). Dietary fiber intake and mortality in the NIH-AARP Diet and Health Study. *Archives of internal medicine*, 171(12):1061.
- Pittau, M., Zelli, R., and Gelman, A. (2010). Economic disparities and life satisfaction in European regions. *Social indicators research*, 96(2):339–361.
- USDA (2010). *Food and Nutrient Database for Dietary Studies 4.1*. U.S. Department of Agriculture, Agricultural Research Service, Food Surveys Research Group, Beltsville, MD.
- Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.
- Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.