

Multivariate interactions in regression

Abhyuday Mandal
School of Industrial and Systems Engineering
Georgia Institute of Technology

Kerby Shedden
Department of Statistics
University of Michigan

Abstract

A “multivariate interaction” in a regression model is a product of two independent variates (linear functions of the regressors) that is an additive component of the regression function $E(Y|X)$. In many cases a substantial portion of the overall pairwise interaction structure in a regression function can be captured by a single multivariate interaction. Due to its parsimonious form, a multivariate interaction can be estimated directly, without the use of sequential variable selection techniques. Inference can be carried out using a straightforward likelihood ratio approach, which yields hypothesis tests of considerably greater power than a comparable regression test when the number of regressor variables is large. We investigate three methods for estimating multivariate interactions. The optimizations required in computing these estimates are neither orthogonalization nor eigenvalue problems, as is the case for other common regression and dimension reduction estimators. We describe an efficient algorithm, present results from a simulation study, and demonstrate the method using U.S. temperature field data and data from a small experiment.

1 Introduction

When the dimensionality p of the regressor vector X in a regression analysis is small, informative interactions can be identified by applying variable selection techniques to the $\binom{p}{2}$ pairwise interactions (e.g. Lindley 1968, Akaike 1973, Mallows 1973, George & McCulloch 1993, Chipman *et al.* 1997). However inference for sequential approaches is not straightforward, and as p grows they become computationally demanding. In this article we consider a parsimonious class of regression functions in which the informative interactions can be factored as a “multivariate interaction” – a product of the form $\theta'X \cdot \eta'X$, where $\theta'X$ and $\eta'X$ are uncorrelated projections of X .

One advantage of using a multivariate interaction to approximate the overall pairwise interaction structure in a regression function is that it may be more interpretable than a sum of variable-pair interactions $\sum c_{ij}X_iX_j$ identified by variable selection techniques. Another key advantage is that the presence of a multivariate interaction can be tested using a straightforward likelihood ratio test. Thus a multivariate interaction provides a direct means to test the presence of pairwise interactions that does not rely on sequential testing. We show that when p is relatively large compared to the sample size n , but $\binom{p}{2} + p < n$, this test may be substantially more powerful than a test based on fitting all pairwise interactions using linear least squares. If $\binom{p}{2} + p \geq n$ then the latter test can not be applied, while the multivariate interaction approach is applicable as long as $p < n/3$.

We will discuss three estimators of a multivariate interaction that result from different estimation principles. The three estimators are related in that they share a common quadratic “kernel”, which is closely related to the objective function in the PHD dimension reduction procedure (Li, 1993). If the regressors are multivariate normal, the three estimators are asymptotically equivalent. However empirical studies show that the three methods may produce very different results when multivariate normality is violated, even if the sample size is large.

The outline of this article is as follows. Section 2 defines the three estimators and establishes their asymptotic equivalence for multivariate normal regressors. Section 3 describes in detail an algorithm for computing two of the three estimators. Section 4 provides simulation results showing that the estimators perform well as point estimates and that the corresponding likelihood ratio tests have good power. Section 5 illustrates the method using temperature data collected at various point in the United States and data from a designed experiment. Section 6 discusses some limitations and points out possibilities for future work.

2 Estimators of multivariate interactions

2.1 Estimation of multivariate interactions via nonlinear regression

Viewed as a nonlinear regression, the problem is to minimize

$$\sum_i (Y_i - \theta' X_i \cdot \eta' X_i)^2 \tag{1}$$

subject to $\text{cov}(\theta' X_i, \eta' X_i) = 0$. Throughout the remainder we will assume without loss of generality that the X_i have been transformed so that $\text{cov}(X_i) = I$, hence the constraint becomes $\theta' \eta = 0$. Moreover, we will assume that the response and regressor variables have all been centered at zero. Since the lengths of θ and η cannot be separately identified from (1), we constrain $\|\theta\| = \|\eta\| = 1$ and introduce a scalar variable λ to represent the overall scale of the multivariate interaction. Minimization of (1) leads to the “nonlinear least squares” (NLS) estimator:

$$(\hat{\theta}, \hat{\eta}, \hat{\lambda}) = \text{argmax}_{\theta, \eta, \lambda} 2\lambda \theta' M_{YXX} \eta - \lambda^2 \sum_i \theta' X_i X_i' \theta \cdot \eta' X_i X_i' \eta / n, \tag{2}$$

where $M_{YXX} = \sum_i Y_i X_i X_i' / n$.

2.2 Estimation of multivariate interactions via Liquid Association

In recent work on gene expression analysis (Li, 2002), the Liquid Association (LA) method

was devised to identify a particular type of interaction in a trivariate problem U, V, W such that the product $(U - EU)(V - EV)$ is strongly associated with W . When such an association exists U, V, W are said to have a “liquid association”. The original motivation was to identify a “scout variable” W such that the magnitude and direction of the interaction between U and V is strongly dependent on W . An example of liquid association would be that when W is large, U and V are positively associated, whereas if W is small, U and V are negatively associated.

To derive the LA method as presented in Li (2002), let $g(\tilde{W}) = E(\tilde{U}\tilde{V}|\tilde{W})$, where \tilde{U} is the standardization of U (i.e. $\tilde{U} = (U - EU)/\sigma_U$), and similarly \tilde{V} and \tilde{W} are the standardizations of V and W . A large value of g' corresponds to the property of liquid association as described above. Summarizing g' by its expected value, we may apply Stein’s lemma yielding $Eg'(\tilde{W}) = E\tilde{U}\tilde{V}\tilde{W}$. This quantity can be estimated using the moment estimator $\sum_i \tilde{U}_i \tilde{V}_i \tilde{W}_i / n$, where n is the sample size.

The LA method can be applied to the multivariate interaction problem as follows. The regression response variable Y will serve as the scout variable (denoted W above), and $U = \theta'X$ and $V = \eta'X$ will be independent projections of the regressor vector X . Thus we have

$$g(Y) = E((\theta'X) \cdot (\eta'X)|Y),$$

and applying Stein’s lemma yields $Eg'(Y) = \theta' M_{YXX} \eta$, where M_{YXX} is as defined in the previous section. Since X is standardized, it follows that $\theta'X$ and $\eta'X$ are centered, uncorrelated, and have unit variance as long as $\|\theta\| = \|\eta\| = 1$.

In contrast to the original Liquid Association approach, we now have additional parameters θ and η in the problem. In order to produce the strongest relationship between Y and the multivariate interaction $(\theta'X) \cdot (\eta'X)$, we should select θ and η to maximize $Eg'(Y)$,

giving

$$(\hat{\theta}, \hat{\eta}) = \operatorname{argmax}_{\theta, \eta: \theta' \eta = 0} \theta' M_{YXX} \eta / \|\theta\| \|\eta\|. \quad (3)$$

Since 3 is scale-invariant, this method does not directly provide an estimate of the scaling value λ , as defined above, but this can be obtained in a straightforward way, as explained below. Also note that any algorithm that solves (2) can be used to solve 3 by fixing λ at a small constant value.

2.3 Estimation of multivariate interactions via correlation maximization

If a multivariate interaction is present, the maximum correlation between the response values and the fitted interaction values should occur when η and θ are close to their population values. This suggests the following estimator

$$(\hat{\theta}, \hat{\eta}) = \operatorname{argmax}_{\theta, \eta} \operatorname{cor}(Y, \theta' X \cdot \eta' X). \quad (4)$$

If $\operatorname{cor}(\cdot)$ is the Pearson correlation coefficient, (4) can be expressed

$$(\hat{\theta}, \hat{\eta}) = \operatorname{argmax}_{\theta, \eta} \theta' M_{YXX} \eta / \operatorname{SD}(\theta' X \cdot \eta' X), \quad (5)$$

where $\|\theta\| = \|\eta\| = 1$ and $\theta' \eta = 0$.

2.4 Comparison of the three estimators

The estimators (2), (3), and (5) share the kernel term $\theta' M_{YXX} \eta$ and the constraint $\theta' \eta = 0$. However each contains another multiplicative or additive term that does not depend on Y . These terms differ in the three methods. However it can be shown that under a multivariate Gaussian distribution on the regressors, these terms are asymptotically constant (i.e. not functions of θ and η), and hence do not affect the solution. Specifically, in (2), $(\theta' X)^2$ and $(\eta' X)^2$ are independent for Gaussian data, so $\sum_i \theta' X_i' X_i \theta_i \eta' X_i' X_i \eta_i / n \rightarrow 1$.

By this same reasoning, the denominator of (5) becomes 1 in the limit.

3 Optimization

None of the estimators (2), (3), or (5) is an eigenvalue problem, and it is not apparent that any can be transformed into such a problem. For calculation of the NLS estimate (2) we implemented a conjugate-gradient method. We will see that with a trivial modification the same algorithm can be used to calculate the LA estimate (3).

3.1 Reparametrization

Let $f(\theta, \eta)$ be the objective function in (2). Suppose that θ, η is the current point in the optimization, so θ and η are orthogonal and have unit norm. To derive a new search direction, reparameterize the search space in terms of $V \in \mathcal{R}^p$, and $\alpha, \beta \in \mathcal{R}$ using

$$\begin{aligned}\tilde{\theta} &= (\cos \alpha + \cos \beta - \sin \alpha \sin \beta)\theta/2 + (\cos \alpha - \cos \beta - \sin \alpha \sin \beta)\eta/2 - \\ &\quad (\cos \alpha \sin \beta + \sin \alpha)V/(\sqrt{2}\|V\|) \\ \tilde{\eta} &= (\cos \alpha - \cos \beta + \sin \alpha \sin \beta)\theta/2 + (\cos \alpha + \cos \beta + \sin \alpha \sin \beta)\eta/2 + \\ &\quad (\cos \alpha \sin \beta - \sin \alpha)V/(\sqrt{2}\|V\|),\end{aligned}$$

where V is constrained to be perpendicular to both θ and η . In words, $(\tilde{\theta}, \tilde{\eta})$ is related to (θ, η) by a rotation of angle β in the plane spanned by V and $\eta - \theta$, followed by a rotation of angle α in the plane spanned by V and $\eta + \theta$. Note that $\tilde{\theta}$ and $\tilde{\eta}$ remain orthogonal and have unit norm.

The search direction is related to the gradient $\partial f(\tilde{\theta}, \tilde{\eta})/\partial V_k$, which can be calculated using the chain rule:

$$\begin{aligned}
\partial f(\tilde{\theta}, \tilde{\eta})/\partial V_k &= \sum_i \partial f/\partial \tilde{\theta}_i \cdot \partial \tilde{\theta}_i/\partial V_k + \partial f/\partial \tilde{\eta}_i \cdot \partial \tilde{\eta}_i/\partial V_k \\
\partial f(\tilde{\theta}, \tilde{\eta})/\partial \alpha &= \sum_i \partial f/\partial \tilde{\theta}_i \cdot \partial \tilde{\theta}_i/\partial \alpha + \partial f/\partial \tilde{\eta}_i \cdot \partial \tilde{\eta}_i/\partial \alpha \\
\partial f(\tilde{\theta}, \tilde{\eta})/\partial \beta &= \sum_i \partial f/\partial \tilde{\theta}_i \cdot \partial \tilde{\theta}_i/\partial \beta + \partial f/\partial \tilde{\eta}_i \cdot \partial \tilde{\eta}_i/\partial \beta
\end{aligned}$$

Calculation of $\partial f/\partial \tilde{\theta}_i$ and $\partial f/\partial \tilde{\eta}_i$ are straightforward. To finish calculating the gradient, we also need

$$\begin{aligned}
\partial \tilde{\theta}_i/\partial V_k &= -(\sin \alpha + \cos \alpha \sin \beta)(\mathcal{I}(i = k)/\|V\| - V_i V_k/\|V\|^3)/\sqrt{2} \\
\partial \tilde{\eta}_i/\partial V_k &= (\cos \alpha \sin \beta - \sin \alpha)(\mathcal{I}(i = k)/\|V\| - V_i V_k/\|V\|^3)/\sqrt{2} \\
\partial \tilde{\theta}_i/\partial \alpha &= -(\sin \alpha + \cos \alpha \sin \beta)\theta_i/2 - (\sin \alpha + \cos \alpha \sin \beta)\eta_i/2 + \\
&\quad (\sin \alpha \sin \beta - \cos \alpha)V_i/(\sqrt{2}\|V\|) \\
\partial \tilde{\eta}_i/\partial \alpha &= (\cos \alpha \sin \beta - \sin \alpha)\theta_i/2 + (\cos \alpha \sin \beta - \sin \alpha)\eta_i/2 - \\
&\quad (\cos \alpha + \sin \alpha \sin \beta)V_i/(\sqrt{2}\|V\|) \\
\partial \tilde{\theta}_i/\partial \beta &= -(\sin \beta + \sin \alpha \cos \beta)\theta_i/2 + (\sin \beta - \sin \alpha \cos \beta)\eta_i/2 - \\
&\quad \cos \alpha \cos \beta V_i/(\sqrt{2}\|V\|) \\
\partial \tilde{\eta}_i/\partial \beta &= (\sin \alpha \cos \beta + \sin \beta)\theta_i/2 + (\sin \alpha \cos \beta - \sin \beta)\eta_i/2 + \\
&\quad \cos \alpha \cos \beta V_i/(\sqrt{2}\|V\|).
\end{aligned}$$

3.2 Alternating optimization procedure for fixed λ

To calculate the solution to (2), we first considered the problem with λ held fixed. This optimizer was then embedded into either a bisection or quadratic interpolation to determine the optimal values of θ , η , and λ (see below). We took this approach after finding that

including λ as a free variable in the optimization along with η and θ frequently caused severe convergence problems.

The optimization of θ and η for fixed λ was performed by alternating between two stages. In the first stage, the Fletcher-Powell conjugate gradient method (Polak, 1971) was used to optimize over V, α, β , using gradients as derived above. This stage was considered converged when the norm of the (V, α, β) gradient was smaller than 10^{-6} . In the second stage, bisection was used to rotate η and θ in the plane spanned by η and θ . This bisection was carried out to 10^{-6} accuracy. Convergence of the overall algorithm was assessed based on KKT conditions. At the solution, the calculated gradient $(\partial f/\partial\eta, \partial f/\partial\theta)$ should lie in the subspace $\mathcal{C} = \langle(\theta, \eta), (\eta, 0), (0, \theta)\rangle$. For convergence, the norm of the gradient vector G projected onto \mathcal{C} was required to be at least 0.999 times the norm of G . This typically occurred in 10-20 alternations between the two optimization stages, which for $p = 10$ regressors took around 10 seconds on a typical machine. Experiments with varying starting values failed to produce any evidence of multiple modes as long as a strict KKT criterion was imposed.

3.3 Optimization over λ

To calculate the NLS estimate, an optimal value of λ must be found. Since we can optimize θ and η for a fixed value of λ , to implement an alternating approach we need to devise a method for optimizing λ for fixed values of θ and η . Since λ is a scalar, bisection can be used.

Using the variance decomposition

$$\text{var}Y = \lambda^2 \text{var}(\theta'X)(\eta'X) + E\text{var}(Y|X),$$

and noting that the two projections are uncorrelated and centered, and that Y was standardized, for Gaussian regressors it follows that $\text{var}(\theta'X_i)(\eta'X_i) = 1$, and hence $\lambda^2 = 1 - \sigma^2$. In particular λ should not exceed 1 for Gaussian data. For non-Gaussian data, we found that the

optimal λ value sometimes exceeded 1. In either case, not knowing the value of σ^2 , we started with a bracket of λ values, i.e., a triple $(\lambda_1, \lambda_2, \lambda_3)$ such that $f(\lambda_2) > \max\{f(\lambda_1), f(\lambda_3)\}$, where $f(\lambda)$ is the objective function in (2) restricted to given values of η and θ . A bracket can always be found by shrinking λ_2 until $f(\lambda_2) > 0$, then decreasing λ_1 and increasing λ_3 until a bracket is found.

Upon obtaining the bracket, one alternative is to carry out a bisection over λ to determine the optimal values of λ , η , and θ . A less expensive alternative is to fit a quadratic interpolant to the bracket, and use the maximizer of the interpolant to approximate the optimal λ value. One additional optimization is then run at this λ value to obtain the final estimates. We found in simulations that the values of $\text{corr}(\theta'X \cdot \eta'X, \hat{\theta}'X \cdot \hat{\eta}'X)$ differed by less than 0.01 between the interpolation method and the full bisection method.

4 Simulation study

4.1 Estimation

We carried out a simulation study to evaluate the estimation performance of the method. All combinations of sample size $n = 100, 500$, regressor dimension $p = 10$, and variance ratio $\text{var}(\eta'X \cdot \theta'X)/\text{var}(\epsilon) = 2, 1/2, 1/5$ were used, with 100 replications for each configuration. Two distributions for the regressor (X) variables were used: (i) all X values were simulated as *iid* standard Gaussian draws, and (ii) half of the X values (randomly selected) were *iid* standard Gaussian and the other half were exponentially distributed with mean one. In both cases, the error component ϵ was generated as *iid* Gaussian draws, with the standard deviation set to produce a given variance ratio. Optimization was carried out for both the nonlinear least squares estimator (2) and the LA estimator (3) (calculated using the algorithm for (2) with λ fixed at the small value $1/1000$). Performance was evaluated using the Pearson correlation between the true interaction $(\eta'X) \cdot (\theta'X)$ and the fitted interaction $(\hat{\eta}'X) \cdot (\hat{\theta}'X)$.

The results of this simulation study are shown in Table 1. For sample size 500, excellent agreement ($\rho > 0.8$) between the fitted and population multivariate interaction is found for both the NLS and LA procedures, even when the multivariate interaction only accounts for 1/6 of the response variation (variance ratio 1/5). For sample size 100, correlations exceeding 0.5 are found in all configurations. As expected, higher variance ratios and larger sample sizes produce greater correlations, and both estimation methods perform better when the regressors have a multivariate Gaussian distribution compared to the mixed Gaussian/exponential case.

The NLS method generally outperforms the LA method. The difference in performance is greatest when the sample size is large and the regressors are non-Gaussian, indicating that performance of the LA method rests primarily on its ability to approximate the least squares loss function, for which normality of the regressors is critical. The difference in performance is negligible for Gaussian regressors unless the multivariate interaction accounts for 2/3 of the response variation (variance ratio 2), which is unlikely to occur in practice.

4.2 Omnibus test for pairwise interactions

An important application of the multivariate interaction estimator is testing the presence of pairwise interactions in a regression function. The multivariate interaction can be fit and a likelihood ratio test can be carried out comparing the main effects model to the model with main effects and a multivariate interaction. If the simpler model is rejected, this provides evidence for the presence of pairwise interactions in the regression function.

Assuming normal errors, the likelihood ratio is $L = n(\log(\hat{\sigma}_{\text{me}}^2) - \log(\hat{\sigma}_{\text{mi}}^2))$, where $\hat{\sigma}_{\text{me}}^2$ is the average squared residual for the main effects model and $\hat{\sigma}_{\text{mi}}^2$ is the average squared residual for the model with main effects and a multivariate interaction. We found that the limiting $\chi_{2(p-1)}^2$ distribution gives higher than the nominal type I error level for finite samples, so significance levels were determined by simulating *iid* Gaussian data sets $Y = \epsilon$, and determining the 95th and 99th percentiles of the null distribution of L empirically.

A first question was whether the $Y = \epsilon$ model provides an appropriate null distribution, or whether a model with main effects $Y = X\beta + \epsilon$ should be used for this purpose. Results from the simulation study shown in Table 2 address this question. Data sets were simulated having $n = 500$ observations and $p = 10$ regressors, and in addition to main effects a single pairwise interaction of the form $\gamma X_1 X_2$ was also included as an additive component of the regression function. The error variance was always $\text{var}(\epsilon) = 1$, and the main effects vector β was simulated as *iid* normal values with mean 0 and variance τ^2 . For given values of τ and γ , 1000 data sets were simulated and the likelihood ratio test was applied (based on the $Y = \epsilon$ sampling distribution). Each set of 1000 data sets is characterized in terms of the proportion of variance due to main effects $\text{var}(X'\beta)/\text{var}(Y)$ denoted ME below, and the proportion of variance due to multivariate interactions $\text{var}(X'\eta \cdot X'\theta)/\text{var}(Y)$ denoted MI below. Since β is random these proportions have both a mean and a standard deviation, both of which are reported below.

Focusing first on the $\text{MI} = 0$ case, rejection rates are seen to be very close to the level of the test. This is true for ME values ranging roughly from 3% to 70%, supporting the claim that the sampling distribution of L can be obtained from the $Y = \epsilon$ model without consideration of main effects.

A related question is whether power is affected by the strength of the main effects when interactions are present in the regression function. There are nine configurations in Table 2 with $\text{MI} > 0$. As expected, the power increases as MI increases. However the power is not affected by the strength of the main effects.

Next we asked whether the test based on fitting a multivariate interaction is more powerful than the test based on fitting a regression model containing all $\binom{p}{2}$ pairwise interactions using linear least squares. Table 3 shows the results of a simulation study with $n = 250$ observations and $p = 20$ regressors. If the multivariate interaction explains roughly 12% of the response variation then good power of around 75% (at the 0.05 level) is provided by the

multivariate interaction test. For this same configuration, the test based on fitting all pairs of interactions lags far behind at around 30% power. Performance differences for the level 0.01 test are even greater.

For the results shown in Table 3, the number of pairwise interactions is $\binom{p}{2} = 190$, a substantial proportion of the sample size $n = 250$. As the sample size grows relative to $\binom{p}{2}$, performances of the two tests become more similar. Table 4 shows results for $p = 20$ and $n = 500$. The multivariate interactions test continues to be more powerful, but to a noticeably lesser degree than when $n = 250$. For the case $p = 10$ and $n = 500$, where $\binom{p}{2} = 45$ is an order of magnitude less than the sample size, powers for the two tests were similar.

5.1 Example: US annual mean temperature fields

We obtained daily minimum and maximum temperature values for all days in April, 2001 taken at 1068 stations in the continental United States by the U.S. Geological survey (www.usgs.gov). For each station, we constructed a daily temperature range using the difference between daily maximum and daily minimum temperature. Daily temperature range was then averaged for each station over all days in April, 2001 to produce a monthly range summary. This was used as the response variable for regression analysis with 14 covariates – a 6 dimensional Fourier basis for latitude, a six dimensional Fourier basis for longitude, log elevation, and squared log elevation.

We first fit a model including main effects for the all covariates, which accounted for 45% of variation in the response. Next we fit the multivariate interaction model using both the LA method and the NLS method. These accounted for 51% of variation (LA) and 54% of variation (NLS), and gave likelihood ratio statistics of 273 and 387, respectively. For 1000 simulated Gaussian data sets with $n = 1068$ observations and $p = 14$ regressors, the greatest observed likelihood ratio statistic was 67, indicating very high confidence for the presence of pairwise interactions in the regression function.

Figure 1 contains four scatterplots indicating the way in which the multivariate interaction model fits the residuals in the temperature field after removing main effects. The upper left panel shows the residuals, with the largest 10% denoted by crosses and the smallest 10% (i.e. negative with large magnitude) denoted by boxes. These points are shown with respect to the spatial (latitude and longitude) coordinates of the measurement stations. This plot indicates that the regions where main effects overestimate or underestimate the temperature range cluster into a handful of regions, with the underestimated stations being somewhat more dispersed compared to the overestimated stations.

The upper right panel shows the fitted multivariate interaction $(\hat{\theta}'X) \cdot (\hat{\eta}'X)$ plotted with respect to geographic coordinates. The multivariate interaction corrects several of the faults in the main effects field where many large residuals clustered in small geographic regions. In particular, overestimation of temperature range in southern coastal California, central Colorado, southern Florida, and southern New England are corrected, as is underestimation along the central and northern Pacific coast and in the great plains. The lower panels show the two components $\hat{\theta}'X$ and $\hat{\eta}'X$ of the multivariate interaction.

5.2 Example: a small factorial experiment

A data set from a factorial experiment on the effects of various feeding strategies on the weight of chicks has been discussed in a number of places (e.g. Cox & Snell 1981). Chicks were fed two types of protein (groundnut and soybean) at three different levels, and also were fed two levels of fish solubles. Chicks were raised under all 12 combinations of these three factors, and weights were assessed at 16 weeks. The entire experiment was carried out independently in two “houses”.

We fit and subtracted main effects, treating level of protein as a linear (1 df) variable, then we estimated a multivariate interaction using the LA approach. The response variable was log-transformed, as in previous analyses of these data. Using likelihood ratios for inference, the multivariate interactions were significant with p-values 0.005 and 0.047 in the two houses,

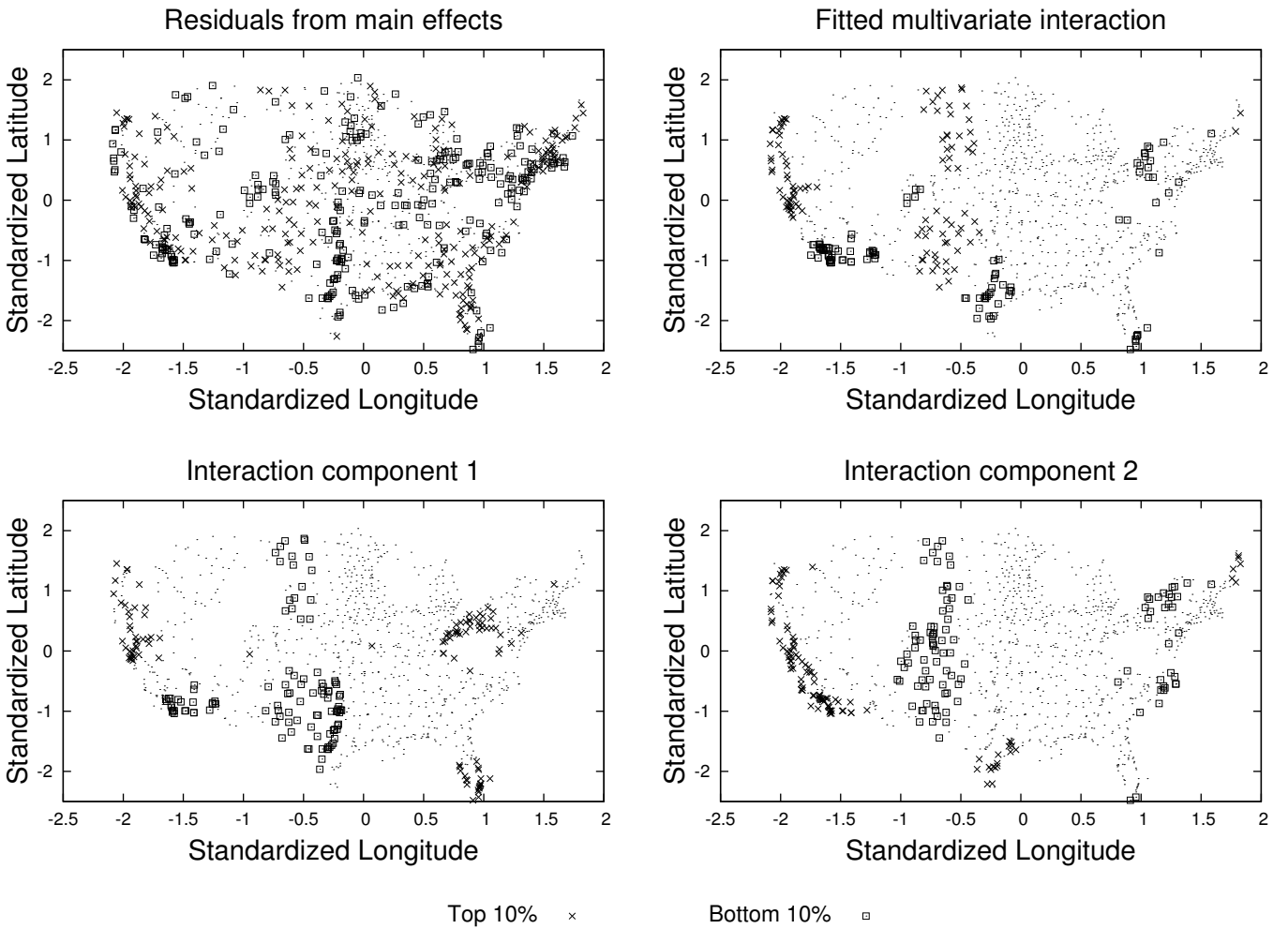


Figure 1: Multivariate interaction for U.S. temperature field. Upper left panel: residuals from main effects, upper right panel: fitted multivariate interaction, lower panels: components of the multivariate interaction. Each plot shows the upper 10% of values plotted as crosses and the lower 10% of values plotted as boxes. These points are plotted with respect to the geographic (latitude and longitude) coordinates of the stations.

respectively. Denoting the variables as T (type of protein), L (level of protein), and F (level of fish solubles), the fitted interactions for house 1 and house 2 are:

$$\begin{aligned} & (0.035T - 0.565L - 0.144F) \cdot (0.957T - 0.014L + 0.288F) \quad \text{house 1} \\ & (-0.006T - 0.489L + 0.141F) \cdot (0.873T + 0.125L + 0.472F) \quad \text{house 2.} \end{aligned}$$

Expanding these trinomial products yields

$$\begin{aligned} & 0.033T^2 + 0.008L^2 - 0.041F^2 - 0.541TL - 0.127TF - 0.161LF \quad \text{house 1} \\ & -0.005T^2 - 0.061L^2 + 0.066F^2 - 0.428TL + 0.120TF - 0.214LF \quad \text{house 2.} \end{aligned}$$

Previous analysis (Cox & Snell 1981) found the TL (type of protein by level of protein) interaction to be significant at the 0.01 level and the LF (level of protein by level of fish solubles) interaction to be “suggestive” at the 0.05 level. The multivariate interaction is similarly dominated by the TL interaction, with the LF interaction being the second most important contributing term for both houses. Conveniently, we are able to give a significance level that encompasses both interactions. In addition, our significance level also conveys that pure quadratic terms are not very predictive. The third (TF) interaction is not far behind the LF interaction in magnitude, but is discounted since the direction of the effect is not consistent in the two houses.

6 Discussion and future work

A multivariate interaction provides a low-dimensional parameterization of a limited range of first order interaction structures that can be tractably estimated. Since the dimensionality of the multivariate interaction is only of order $2p$, estimation variability is low even for medium-sized data sets, allowing powerful hypothesis tests to be carried out. In addition, a multivariate interaction allows models to be built that capture a substantial portion of the overall pairwise interaction structure with little risk of overfitting the data.

An important antecedent for the work presented here is the PHD method of Li (1992). While PHD is presented in the more general context of regression with an unknown link function, it may also be applied when a linear link is assumed. In that case, PHD is seen to estimate a multivariate “perfect square” which is an additive term in the regression function. Such a perfect square is more constrained than a multivariate interaction as it is represented by only p coefficients. Although the multivariate interaction parameterizes a larger range of structures, it is not able to represent a perfect square, as the two components of the multivariate interaction are constrained to be orthogonal. This orthogonality constraint could be removed, yielding a $2p - 1$ dimensional parameterization. By analogy with classical multivariate analysis techniques such as canonical correlation analysis, we have imposed the orthogonality constraint here in the hope of making the two projections in the interaction more interpretable. In future work it may be valuable to investigate the procedure in which this constraint is relaxed.

As an inferential tool, the multivariate interaction provides a powerful procedure for testing the presence of interactions by eliminating the need to use sequential procedures. A particularly favorable case is when a single variable pair interaction dominates. Rather than searching through all $\binom{p}{2}$ variable pairs to identify the dominant interaction, a multivariate interaction identifies it in a single fit. We recognize that if the interaction structure is more complex, with multiple variable pairs contributing, it is not certain that the overall effect of pairwise interactions will be closely approximated by a multivariate interaction. We have found in a number of examples, however, that multivariate interactions continue to explain a substantial portion of the overall variance due to pairwise interactions. Finally, it is not unreasonable in certain cases to hypothesize that a multivariate interaction might be an exact description of the underlying mechanism. In this case the fitted multivariate interaction represents the optimal frame for viewing the first order interaction structure of the regression function, just as Linear Discriminant Analysis identifies the optimal frame for

viewing the regressors in certain classification problems.

An important aspect of this work is that the least squares regression function (2) can be approximated by a more tractable expression (3) if the variables follow a multivariate Gaussian distribution reasonably closely. This generally leads to around a 10-fold decrease in computation time. We emphasize that in a real data set with up to $p = 50$ variables either algorithm could be used. However in simulation studies for power in which thousands of fits must take place, there is a substantial practical benefit to using the LA method. Moreover limited simulations with the NLS method showed little improvement over the LA method when using Gaussian data. The simulation results for point estimation, shown in Table 1, leave open the possibility that the NLS method may yield better power for non-Gaussian variables.

Acknowledgment:

Support of Pfizer Global Research and Development, Ann Arbor Laboratories and NSF grant DMS-0305996 is gratefully acknowledged.

References:

Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Ed. B.N. Petrov and F. Csaki, pp. 267-81. Budapest: Akademia Kiado.

Chipman H, Hamada M and Wu CFJ, (1997) A Bayesian variable selection approach for analyzing designed experiments with complex aliasing, *Technometrics* **39**:372.

Cox DR and Snell EJ (1981). *Applied Statistics: Principles and Examples*. Chapman-Hall, London.

George EI and McCulloch RE (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Society* **88**:881-889.

Li KC (2002). Genome-wide coexpression dynamics: Theory and application. *PNAS* **99**:16875-16880.

Li KC (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's Lemma. *Journal of the American Statistical Association*, **87**:1025-1040.

Lindley DV (1968). The choice of variables in regression. *Journal of the Royal Statistical Society B* **30**:31.

Mallows CL (1973), Some Comments on Cp. *Technometrics* **15**:661.

Polak E (1971). *Computational Methods in Optimization*. Academic Press, New York.

n	LA			NLS		
	VR=2	VR=1/2	VR=1/5	VR=2	VR=1/2	VR=1/5
100	0.86(0.05)	0.74(0.11)	0.58(0.16)	0.94(0.06)	0.76(0.10)	0.55(0.15)
500	0.96(0.02)	0.94(0.02)	0.88(0.05)	0.99(0.05)	0.97(0.01)	0.91(0.05)
100	0.80(0.10)	0.69(0.14)	0.56(0.17)	0.89(0.09)	0.70(0.11)	0.51(0.14)
500	0.88(0.05)	0.86(0.06)	0.83(0.08)	0.99(0.003)	0.96(0.05)	0.90(0.06)

Table 1: Simulation study for estimation. Column 1: sample size, columns 2-4: results of Liquid Association (LA) estimation, columns 5-7: results of nonlinear least squares (NLS) estimation. Each entry is the mean(standard deviation) over 100 replicates of the correlation between $\hat{\eta}'X \cdot \hat{\theta}'X$ and $\eta'X \cdot \theta'X$. The upper part of the table shows the results for Gaussian regressors, the lower part shows the results for a mixture of Gaussian and exponential regressors. The regressor dimension was $p = 10$ in all cases.

Level		Regression components	
0.05	0.01	ME	MI
0.035	0.010	0.037(0.017)	0
0.094	0.024	0.037(0.016)	0.010(0.001)
0.501	0.278	0.037(0.016)	0.037(0.006)
0.928	0.838	0.035(0.015)	0.080(0.011)
0.050	0.008	0.373(0.102)	0
0.121	0.020	0.359(0.106)	0.006(0.001)
0.506	0.292	0.359(0.102)	0.025(0.005)
0.945	0.848	0.353(0.104)	0.050(0.010)
0.058	0.004	0.684(0.110)	0
0.115	0.021	0.686(0.003)	0.003(0.001)
0.489	0.229	0.674(0.108)	0.013(0.004)
0.951	0.850	0.669(0.103)	0.027(0.009)

Table 2: Invariance of type I/II error to strength of main effects. A multivariate interaction was fit to data simulated from the model $Y = X'\beta + \gamma X_1 X_2 + \epsilon$, and a likelihood ratio test was used to test for interactions. For 15 different model specifications with $p = 10$ regressors and $n = 500$ observations, 1000 data sets were simulated. The proportions of data sets for which the null hypothesis was rejected at significance levels 0.05 and 0.01 are shown in columns 1 and 2. Columns 3 and 4 show the mean(standard deviation) for $\text{var}(X'\beta)/\text{var}(Y)$ (column 3) and $\text{var}(\gamma X_1 X_2)/\text{var}(Y)$ (column 4) for each of the model specifications.

Level	Multivariate interaction		All pairs		ME	MI
	0.05	0.01	0.05	0.01	Variance Ratios	
	0.049	0.015	0.058	0.018	0.542(0.091)	0
	0.064	0.014	0.045	0.008	0.541(0.091)	0.005(0.001)
	0.105	0.028	0.059	0.020	0.529(0.086)	0.018(0.005)
	0.125	0.043	0.086	0.023	0.530(0.087)	0.028(0.007)
	0.211	0.050	0.105	0.026	0.519(0.090)	0.040(0.010)
	0.270	0.124	0.116	0.036	0.514(0.093)	0.054(0.013)
	0.408	0.236	0.129	0.045	0.506(0.090)	0.068(0.017)
	0.551	0.315	0.220	0.058	0.501(0.088)	0.085(0.020)
	0.656	0.461	0.296	0.088	0.493(0.090)	0.101(0.024)
	0.787	0.668	0.312	0.096	0.486(0.092)	0.121(0.028)

Table 3: Comparison of power using multivariate interactions (columns 1 and 2) to power using a regression fit of all pairwise interactions (columns 3 and 4). For seven model configurations with $n = 250$ observations and $p = 20$ regressors, 1000 data sets were simulated and powers were determined empirically. The proportion of variance due to main effects and to multivariate interactions are given in columns 5 and 6 respectively.

Level	Multivariate interaction		All pairs		ME	MI
	0.05	0.01	0.05	0.01	Variance Ratios	
	0.053	0.009	0.070	0.021	0.541(0.091)	0
	0.069	0.018	0.103	0.031	0.537(0.085)	0.005(0.001)
	0.194	0.066	0.221	0.080	0.527(0.084)	0.018(0.004)
	0.298	0.150	0.305	0.105	0.529(0.028)	0.028(0.006)
	0.623	0.372	0.412	0.187	0.524(0.084)	0.039(0.008)
	0.813	0.631	0.632	0.338	0.514(0.086)	0.053(0.011)
	0.952	0.883	0.768	0.541	0.511(0.085)	0.067(0.014)

Table 4: Comparison of power using multivariate interactions (columns 1 and 2) to power using a regression fit of all pairwise interactions (columns 3 and 4). For seven model configurations with $n = 500$ observations and $p = 20$ regressors, 1000 data sets were simulated and powers were determined empirically. The proportion of variance due to main effects and to multivariate interactions are given in columns 5 and 6 respectively.