Joint Identification of Location and Dispersion Effects in Unreplicated Two-Level Factorials

Andrew J. Henrey
Simon Fraser University, Burnaby, BC, Canada

Thomas M. Loughin[1]
Simon Fraser University, Surrey, BC, Canada

June 29, 2014

[1]Corresponding author: tloughin@sfu.ca

Abstract

Most procedures that have been proposed to identify dispersion effects in unreplicated factorial designs assume that location effects have been identified correctly. Incorrect identification of location effects may impair subsequent identification of dispersion effects. We develop a model for joint identification of location and dispersion effects that can reliably identify active effects of both types. The joint model is estimated using maximum likelihood, and hence effect selection is done using a specially derived information criterion. An exhaustive search through a limited version of the space of possible models is conducted. Both a single-model output and model averaging are considered. The method is shown to be capable of identifying sensible location-dispersion models that are missed by methods that rely on sequential estimation of location and dispersion effects.

Keywords: maximum likelihood, corrected heteroscedastic information criterion, model averaging, exhaustive search, ESMA-CHIC,

[0]This paper has been submitted to Technometrics (`http://amstat.tandfonline.com/loi/tech#.Vi_Nr1XntD9`) and will appear in revised form.

1 Introduction

In many research and quality-improvement endeavors, experiments are run using unreplicated full or fractional factorial designs at 2 levels per factor (generically referred to here as 2^k designs, where k is the number of factors; see Wu and Hamada 2000). These experiments are generally intended to identify factorial effects that influence the mean response. This identification is made difficult by the fact that the natural full model is saturated, but many methods have been proposed to accomplish this goal (see Hamada and Balakrishnan 1998 for a review of methods).

There may furthermore be an interest in estimating process variance and in determining which factorial effects influence dispersion, particularly in quality improvement settings. While unreplicated experiments are obviously ill-suited for variance estimation, efforts have nonetheless been made to try to extract this information from the data. Box and Meyer (1986) developed a seminal procedure to test for "dispersion effects" in data from a 2^k experiment using residuals from a given location model. Several authors have followed this approach by developing improvements or extensions to the Box-Meyer test (e.g., Bergman and Hynén 1997, Wang 1989, Brenneman and Nair 2001, McGrath and Lin 2001a). For a nice review of these procedures see Bursztyn and Steinberg (2006).

All of these procedures use the same basic approach of first fitting a selected location model, and then using residuals from the location model to test for dispersion effects. The resulting tests are well known to be sensitive to the starting location model, so that different location models can yield completely different impressions regarding which dispersion effects are important (Pan 1999, Brenneman and Nair 2001, McGrath and Lin 2001a, 2001b, Pan and Taam 2002). In particular, there is potential for confounding to occur between location and dispersion effects: two active location effects that are excluded from the location model can impart a spurious dispersion effect or can mask a real dispersion effect at their interaction. Thus, it is critical to have the correct model for location before attempting to identify dispersion effects. However, all of the procedures for identifying location models in unreplicated factorial designs are prone to missing small- or moderate-sized real effects (Hamada and Balakrishnan 1998). In Section 3 we show several studies that have each been

analyzed by different authors who identify different "best" models from the same set of data.

In all of these previous analyses, location effects are selected based on one criterion and dispersion effects based on another. There is no direct, objective measure that allows one to compare a model containing one set of location and dispersion effects to a model with a different combination of these effects. The methods in this paper address this shortcoming.

As an alternative to sequential model fitting, we propose in Section 4 to use a joint location and dispersion model for factor screening. This model results in a single likelihood that is used to estimate location and dispersion effects simultaneously. The maximized likelihood is then used for comparing models and selecting effects through an information criterion. Information criteria whose justification is based on asymptotic approximations are of dubious utility in this problem, where the number of potential parameters is roughly twice the number of observations. We therefore develop an exact criterion in Section 4.1 based on the corrected Akaike information criterion (AICc) of Hurvich and Tsai (1998). The form appears somewhat complex, so we simulate the penalty values for different model structures.

The space of all possible models is large, because the presence of dispersion effects can change the ordering of the estimated location effects. When $k \leq 4$, an exhaustive search of the model space is feasible; otherwise we propose using a genetic algorithm to search the space for good-fitting models. Model averaging techniques (Hoeting et al. 1999, Burnham and Anderson 2002) provide a measure of the certainty associated with the importance of each location and dispersion effect and with each model combination. In Section 5 our joint modeling procedure is applied to the examples introduced in Section 3. In each case the procedure provides clear, interpretable results regarding which effects are important and which models are best. In one example, we identify a "best" model that had not previously been detected. Finally, in Section 6 we present results of a small simulation study comparing the joint modeling approach with a combination of popular location- and dispersion-effect identification techniques. The simulations show that the sequential approach has slightly inflated type 1 error rate for identifying dispersion effects. The new procedure experiences a substantially smaller type 1 error rate for dispersion effects *while maintaining uniformly better power*.

2 Previous Approaches

Sequential approaches begin with some method for identifying location effects. They then use the chosen location model to form residuals, which are used for identifying dispersion effects. The methods for identifying the two models are generally completely separate; that is, typical dispersion-effect identification methods assume that the correct location model has been identified, without concern for possible error in the model. We briefly review one procedure for identifying location effects, and three procedures for identifying dispersion effects. These methods are chosen based on performance and apparent popularity.

We present details from the perspective of a full 2^k factorial experiment, although identical results hold for any fractional factorial using $n = 2^k$ runs in the equivalent design. Let $\boldsymbol{W}$ be the $n \times n$ design matrix including all main effects and interactions. We use $+1$ or just $+$ to denote the high level of a factor and -1 or $-$ to represent the low level. Consider a model consisting of p location effects and q dispersion effects, with $p, q = 0, 1, \ldots, n-1$. Denote the corresponding sets of location and dispersion effects by $\mathcal{L}$ and $\mathcal{D}$, respectively, so that a joint location-dispersion model can be represented by $(\mathcal{L}, \mathcal{D})$. Let $\boldsymbol{X} \subseteq \boldsymbol{W}$ be an $n \times (p+1)$ matrix containing the columns corresponding to the effects in $\mathcal{L}$, and $\boldsymbol{U} \subseteq \boldsymbol{W}$ be an $n \times (q+1)$ matrix containing the columns corresponding to the effects in $\mathcal{D}$. Both $\boldsymbol{X}$ and $\boldsymbol{U}$ contain a lead column of ones.

The Lenth test (Lenth 1989) is found by Hamada and Balakrishnan (1998) to be in the class of best methods for identifying location effects in unreplicated 2^k factorials, and it is notable for its simplicity. Assume that responses arise from the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y}$ is the $n \times 1$ vector of responses, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is a vector of parameters corresponding to location effects, and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{0}$ is an $n \times 1$ vector of zeroes and $\boldsymbol{I}$ is an $n \times n$ identity. Note that the Lenth procedure assumes that no dispersion effects are present. This is typical of location-effect identification methods.

Location effects are estimated using ordinary least squares. Because of the orthogonality of the columns in $\boldsymbol{W}$, estimated location effects from the saturated model $\boldsymbol{X} = \boldsymbol{W}$,

$\hat{\beta}_1, \ldots, \hat{\beta}_{n-1}$, are independent. Thus, the estimates for a particular column of $\boldsymbol{W}$ do not depend on which other columns from $\boldsymbol{W}$ are in $\boldsymbol{X}$. With this in mind, Lenth calculates a "pseudo standard error" (PSE) for the effect estimates as follows. Assuming effect sparsity (see, e.g., Wu and Hamada 2000), suppose that all estimated effects of median magnitude and lower are not from active effects. Then the PSE is calculated as

$$\text{PSE} = 1.5 * \text{median}_{\{i:|\hat{\beta}_i|<2.5s_0\}}|\hat{\beta}_i|,$$

where $s_0 = 1.5 * \text{median}_i|\hat{\beta}_i|$. The statistic $t_i = |\hat{\beta}_i|/\text{PSE}$ is used to test $H_0 : \beta_i = 0$. Lenth suggests comparing t_i to a $t_{(n-1)/3}$ distribution, although Loughin (1998) and Ye and Hamada (2000) find better critical values by simulation.

Box and Meyer (1986) use residuals from a chosen location model, assumed known, to identify active dispersion effects. Let $r_1, \ldots, r_n$ be the residuals from any model fit, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Assume that $\boldsymbol{U}$ has only one column in it besides the lead column of ones, and suppose that it corresponds to column d of $\boldsymbol{W}$. Let $d+$ and $d-$ be the sets of observations for which $w_{id} = +1$ and $w_{id} = -1$, respectively. Then the Box-Meyer statistic is

$$F_d = \frac{\sum_{i \in d+} r_i^2}{\sum_{i \in d-} r_i^2}.$$

Although this statistic looks like it should have an F distribution, the residuals in the numerator and denominator are not necessarily independent. Thus, the sampling distribution of F_d is not clear.

Bergman and Hynén (1997) amend the Box-Meyer test by augmenting the location model in such a way that the residuals contained in $d+$ and $d-$ *are* independent. The augmented model consists of the original model plus all effects formed by the interaction of these effects with d. The Bergamn-Hynén test statistic, D_d^{BH}, is structurally identical to F_d, except that it uses the residuals from the augmented location model rather than the original model. Assume that the original location model is correct, so that it contains all active location effects. Also assume that d is the only possible dispersion effect. Then D_d^{BH} has an F distribution with degrees of freedom depending on the the number of effects *not* in the augmented location model. Note that this may be different for each d. Pan (1999) and McGrath and Lin (2001a)

show that this test can have inflated type I error rate or diminished power when the original location model fails to identify location effects of moderate size. Loughin and Malone (2013) describe a testing approach based on D_d^{BH} that provides a measure of safety against this phenomenon.

Harvey (1976) proposes a method for estimating dispersion effects in a general linear regression setting. Let $\boldsymbol{u}_i'$, $i = 1, \ldots, n$ be the ith row of $\boldsymbol{U}$. Harvey uses the model

$$\sigma_i^2 = \exp(\boldsymbol{u}_i'\boldsymbol{\delta}), \tag{1}$$

where $\boldsymbol{\delta}$ is a $(q+1) \times 1$ vector of unknown parameters representing dispersion effects. From this model, he writes $\log r_i^2 = \boldsymbol{u}_i'\boldsymbol{\delta} + v_i$, where $v_i = \log(r_i^2/\sigma_i^2)$, and by analogy with a linear model, uses least squares to estimate $\boldsymbol{\delta}$. In the context of a 2^k factorial design, Brenneman and Nair (2001) show that this results in

$$\hat{\delta}_d = \log\left(\frac{\prod_{i\in d+} r_i^2}{\prod_{i\in d-} r_i^2}\right)^{1/n} = n^{-1}\left(\sum_{i\in d+} \log r_i^2 - \sum_{i\in d-} \log r_i^2\right).$$

Brenneman and Nair (2001) study the bias in several dispersion-effect identification methods. They show that all are biased, with the severity of bias depending on whether an additive or a log-linear model is assumed for the variances. They also show that the bias in the Harvey method is reduced or eliminated when the residuals are computed from the augmented location model used by Bergman and Hynén (1997). They refer to this as the "modified Harvey" method and recommend it for general use because its bias is limited to certain specific cases.

3 Examples

We use three examples from the literature to demonstrate some of the model uncertainty that is inherent in sequential approaches to identifying location and dispersion effects in unreplicated factorials. All three examples are 16-run designs in the 2^{k-l} series, where k is the number of factors and l is the degree of fractionation. These examples have been analyzed multiple times in the literature. Table 1 lists papers in which these examples have

Table 1: Proposed models resulting from different analyses presented in the literature for three different examples described in Section 3. Note that the Lenth-Harvey method may be inappropriate for the Dyestuff example, because discreteness can create residuals of exactly zero.

Example	Authors	Loc. Effects	Disp. Effects
Welding	Box and Meyer (1986)	B, C	C
	Wang (1989)	B, C	C, H, J
	Ferrer and Romero (1993)	B, C	C, J
	Bergman and Hynén (1997)	B, C	C, H, J
	Nelder and Lee (1998***)	B, C, J	C; H or J
	Pan (1999)	B, C, AC, AH, A, H[1]	—
	McGrath and Lin (2001)	B, C	C
	Pan and Taam (2002)	B, C	C; H or J
	Loughin and Malone (2013)	B, C	C
	Lenth-Harvey	B, C	C, H, J
Injection Molding	Montgomery (1990)	A, B, AB	C
	McGrath (2003)	A, B, AB, G, CG	—
	Loughin and Malone (2013)	A, B, AB, G, CG	—
	Lenth-Harvey	A, B, AB, G, CG	E
Dyestuff	Bergman and Hynén (1997)	D	E
	McGrath and Lin (2001)	D	E
	Lenth-Harvey	D, AB	C, BC

been analyzed, along with the results from their different approaches, where "Lenth-Harvey" refers to our re-analysis using the Lenth test for location followed by the modified Harvey test for dispersion. The factor labels used here are those given in first-listed citation for each example.

The first example is the Welding example, a 2^{9-5} design attributed by Box and Meyer (1986) to a technical report by Taguchi and Wu. The general consensus among the previous analyses is that Factors B and C are active location effects, while C is an active dispersion effect. There is some uncertainty regarding whether factors H and/or J might also be dispersion effects, and two authors have found other location effects besides B and C.

Second is the Injection Molding experiment, a 2^{7-3} experiment given in Montgomery (1990). (The data given in Montgomery (1990) contain four centerpoints that have been subsequently ignored by authors analyzing this example. We also ignore these centerpoints.) Montgomery's original analysis identified the interaction triple A, B, and AB as active lo-

cation effects, and C as an active dispersion effect. McGrath and Lin (2001a) recognized that the dispersion effect C lies at the interaction of the next two largest location effects, G and CG. They determined heuristically that the difference between variances at the two levels of C could be largely explained by the product of the fourth and fifth location effects that were missing from Montgomery's model. A more formal analysis in McGrath (2003) concludes that the dispersion effect disappears upon including these two extra effects in the location model. Loughin and Malone (2013) gave a different analysis of these data supporting McGrath's results.

Last, we consider the Dyestuff data first given in Davies (1963) and analyzed for dispersion effects in Bergman and Hynén (1997). The latter authors found a location effect for factor D and a dispersion effect for E, conclusions that were supported by McGrath and Lin (2001a). Although previous analysis results are in agreement for this example, we will see in Section 5 that this does not imply that there is no model uncertainty

All of the analyses cited in Table 1 were performed using sequential analysis approaches. Location effects were selected, then *conditional on the location effects*, dispersion effects were identified. (McGrath (2003) subsequently reconsiders any identified dispersion effects for possible confounding with two location effects, but the dispersion effect must be initially identified using a sequential analysis.) Thus, all of these analyses are susceptible to errors both due to stochastic uncertainty and due to structural propagation of errors from the location analysis into the dispersion analysis, as discussed by Pan (1999), Brenneman and Nair (2001), McGrath and Lin (2001a), (2001b), Pan and Taam (2002). McGrath and Lin (2001a) and Loughin and Malone (2013) show that there is information within the data that can distinguish between two location effects and a dispersion effect. The model-selection procedure needs to be able to weigh the value of each interpretation using some objective measure. The analysis approach proposed in the next section does precisely that, and in addition, allows a direct measure of model uncertainty that is novel among techniques for these data.

4 Effect Selection using a Joint location and dispersion model

Using the same notation from Section 2, a joint location and dispersion model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

where now $\boldsymbol{\epsilon} \sim N(0, \text{Diag}(\exp(\boldsymbol{U}\boldsymbol{\delta})))$, where $\text{Diag}(\cdot)$ makes a diagonal matrix out of a vector. The variance structure in this model is the same one used by Harvey (1976), Cook and Weisberg (1983), Carroll and Ruppert (1988), and others. It has been studied in the context of 2^k factorial models by Wang (1989), who discussed maximum likelihood estimation and derived properties of the estimates, as well as by Nair and Pregibon (1988), Engel and Huele (1996), and others. All previous work, however, has been done under the assumption that $(\mathcal{L}, \mathcal{D})$ is known or has been correctly estimated prior to use of the model. Of course, it is unrealistic to expect that any particular method can achieve this requirement without uncertainty, as the examples in Section 3 show.

Parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$ from (2) are found using maximum likelihood as in Harvey (1976) and Wang (1989). The log-likelihood is

$$l((\boldsymbol{\beta}, \boldsymbol{\delta}); \boldsymbol{Y}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{u}_i'\boldsymbol{\delta} - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{\exp(\boldsymbol{u}_i'\boldsymbol{\delta})} \tag{3}$$

where $\boldsymbol{x}_i'$ is the ith row of $\boldsymbol{X}$ and $\boldsymbol{u}_i'$ is the ith row of $\boldsymbol{U}$. Evaluated at the MLE this becomes

$$l((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}); \boldsymbol{Y}) = -\frac{n}{2}(\log(2\pi\hat{\sigma}_0^2) + 1) \tag{4}$$

where $\sigma_0 = \exp(\delta_0)$ is the variance at the centerpoint of the design space (or when all dispersion effects are inactive). The parameter estimates for $\boldsymbol{\delta}$ have closed form if $q \leq 1$, but otherwise must be estimated using iterative numerical techniques.

Because this model uses a single likelihood for estimating both sets of parameters, model selection criteria such as Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are available (e.g., Burnham and Anderson 2002, Claeskens and Hjort 2008). These criteria are easily calculated from (3) using the general form $IC(r) = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}; \boldsymbol{Y}) + r(p+q+2)$, where r is a chosen penalty coefficient and $p+q+2$ is the number of parameters estimated in model (2). For example, AIC= $IC(2)$, while BIC= $IC(\log n)$.

We use the availability of information criteria as the central focus for our joint location- and dispersion-effect identification method, Exhaustive Search Model Averaging (ESMA) described in detail in the following sections. To outline ESMA, we (1) carry out an exhaustive search of the model space, considering all possible viable combinations of $(\mathcal{L}, \mathcal{D})$; (2) calculate a special form of information criterion on each; and (3) use an approach resembling Bayesian Model Averaging (Hoeting et al. 1999, Burnham and Anderson 2002) to identify those effects that are more or less likely to belong in the model. We can use this approach to quantify the support in the data for any given model.

4.1 Exhaustive search

Our goal is to be able to provide an assessment of the model corresponding to any combination of $(\mathcal{L}, \mathcal{D})$. In principle, this is not difficult to achieve. However, there are several logistical issues that must be addressed in order to complete this task.

First is the sheer size of the problem. In an unreplicated 2^k factorial, there are up to $2^k - 1$ location effects and $2^k - 1$ dispersion effects to consider. When dispersion effects are not considered, the orthogonality and equal standard errors among the location-effect estimates imposes an ordering based on the magnitude of the estimates that does not depend on which other effects are in the model. Thus, only models consisting of the effects with the largest estimated magnitudes need to be considered. There are only 2^k such models to consider.

When dispersion effects are added, the location-effect estimates are no longer orthogonal, and their ordering can change depending on which effects of both types are in the model. Thus, there are nominally $2^{2^{k+1}-2}$ different $(\mathcal{L}, \mathcal{D})$ combinations that can be constructed. For $k = 3$, this is 16,384 models; for $k = 4$, over 1 billion; and for $k = 5$ the number of possible models is on the order of 10^{19}. In most of these models there is no closed-form solution to the likelihood equations, and no obvious way to reduce the computations in the spirit of the leaps-and-bounds algorithm for linear regression (Furnival and Wilson 1974). For $k \geq 5$ no exhaustive search can be run under current computational capacity. In these cases, some kind of alternative search procedure, such as a genetic algorithm (Michalewicz 1998), must be used.

The second complication is that not all $(\mathcal{L}, \mathcal{D})$ combinations lead to viable models. For one thing, $n = 2^k$, so simultaneously estimating all location and dispersion effects is impossible. However, even models with seemingly viable combined size $p + q + 2 < n$ do not always result in valid parameter estimates. Loughin and Rodríguez (2010) observe that sets $\mathcal{D}$ of size $q = 1$ cause the last term in the log-likelihood to factor into two separate sums for independent subsets of the data. A saturated location model can be found for one of the sums with only $p = n/2 - 1$, and fitting this model causes the dispersion effect to go to $\pm\infty$. For example, this occurs for $k = 4$ when $\mathcal{D} = \{A\}$ and $\mathcal{L} = \{B, C, BC, D, BD, CD, BCD\}$, among other cases. Thus models of total size $n/2+2$ can be constructed that lead to infinite likelihoods. Loughin and Rodríguez (2010) also observe that sets $\mathcal{D}$ that consist of an *interaction triple*—three columns from $\boldsymbol{W}$ such that each is the element-wise product of the other two—can combine with certain complementary location models of size $p = n/4 - 1$ to yield a likelihood with multiple monotonically increasing ridges. For example, $\mathcal{D} = \{A, B, AB\}$ with $\mathcal{L} = \{C, D, CD\}$ causes this to occur (as does the same $\mathcal{D}$ with $\mathcal{L} = \{AC, AD, ACD\}$, $\{BC, BD, BCD\}$, or $\{ABC, ABD, ABCD\}$). Again, this means that models of a much smaller total size than n can lead to invalid parameter estimates.

Fortunately, the combinations $(\mathcal{L}, \mathcal{D})$ where this can occur are completely predictable *a priori*. Loughin and Rodríguez (2010) give a series of criteria that determine whether a given location/dispersion model can be fit to a set of data. Carefully enumerating the model space under these restrictions leads to 1,442,837 different joint models in the $k = 4$ problem, which reduces the computational burden considerably.

In our implementation, we further restrict the searchable model space to models satisfying both $p \leq 5$ and $q \leq 5$. This reduces the model space by a significant fraction, and simulations suggest that it has little impact on the results of a search when these restrictions do, in fact, hold. We are not aware of any other procedure that is likely to perform well when a true model does not satisfy some similar sparsity criterion.

4.2 Corrected heteroscedastic information criterion

On each model, an information criterion (IC) is computed. Information criteria are generally based on the Kullback-Leibler (KL) information, which is a measure of the discrepancy

between an estimated model and the true structure that generated the data (Akaike 1973, Konishi and Kitagawa 2008). The KL information consists of two terms, one of which is constant for all models and hence is discarded. The other part is -2 times the expectation of the log-likelihood with respect to the true model, evaluated at the MLE. This expectation cannot be computed because the true model is unknown. The maximized log-likelihood—in this case (4)—can be used as an estimate of the expectation, but it exhibits bias that grows as the size of the model grows (Akaike 1973). Therefore an IC generally consists of -2 times the maximized log likelihood, l, plus a "penalty" term that adjusts for the bias.

Akaike (1973) gives an asymptotic bias adjustment that works for most models. The Akaike Information Criterion (AIC) is $-2l + 2\nu$, where ν is the total number of parameters in the model. However, this penalty term tends to underestimate the bias and in comparisons among models can result in a preference for models that are too large (Burnham and Anderson 2002). Unfortunately, the actual small-sample bias is model-dependent.

Hurvich and Tsai (1989) develop a small-sample estimate of bias for homoscedastic linear and nonlinear regression models, which they use to form the "corrected AIC" (AICc). The penalty in AICc has the form $2n/(n - \nu - 1)$, which is asymptotically equivalent to Akaike's penalty but becomes much larger when the number of parameters in a model is an appreciable fraction of the sample size, a situation that is the norm in modeling data from unreplicated factorials. However, AICc is not fully suitable for our problem because, as noted above, infinite or monotone likelihoods can be produced for models where the total number of parameters is considerably less than $n - 1$. Using AICc on our problem always results in selecting one of the pathological cases of Loughin and Rodríguez (2010) as the best model. Even eliminating these models from consideration is not adequate, because in pilot studies we found that in many data sets simulated from normal distributions, a near-pathological model can be found for which the maximized log-likelihood is much smaller than the penalty adjustment can account for.

We therefore must derive a corrected IC that is appropriate for the heteroscedastic model (2). To do this, we follow the general approach of Hurvich and Tsai (1989). To start, let $(\boldsymbol{\beta}^*, \boldsymbol{\delta}^*)$ represent the true values of the parameters, and for any candidate model $(\mathcal{L}, \mathcal{D})$ let the parameters be $(\boldsymbol{\beta}, \boldsymbol{\delta})$ with MLEs $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}})$. We assume (as do Cavanaugh 1997 and Hurvich

and Tsai 1989) that for any model that is fit, the parameters contained in $(\boldsymbol{\beta}, \boldsymbol{\delta})$ include all of those in $(\boldsymbol{\beta}^*, \boldsymbol{\delta}^*)$. When this is not the case, the bias in $-2l$ may be arbitrarily large in the opposite direction, and hence the model's IC will be much larger than for models that satisfy this assumption. It therefore suffices to consider the true model and fitted model as having parameters of the same dimension, $(p+q+2)$ with $(\boldsymbol{\beta}^*, \boldsymbol{\delta}^*)$ possibly containing some zero values. Finally, let G be the distribution of $\boldsymbol{Y}$.

The quantity that must be estimated from the KL information is

$$-2l^E(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}) \equiv E_G(-2l(\boldsymbol{\beta}, \boldsymbol{\delta}; Y))\Big|_{(\boldsymbol{\beta},\boldsymbol{\delta})=(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})}.$$

The maximized log-likelihood, $-2\hat{l}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}) = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}; \boldsymbol{Y})$, is used to estimate this quantity. Thus, the "penalty" that must be calculated is the bias, $B = -2E_G(l^E(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}) - \hat{l}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}))$.

From (3) it is easy to show that

$$-2\hat{l}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}) = n(\log 2\pi + 1 + \hat{\delta}_0).$$

It can further be shown that

$$-2l^E(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}) = n \log 2\pi + n\hat{\delta}_0 + \mathrm{Tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^*) + (\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

where $\boldsymbol{\Sigma} = \mathrm{Diag}(\exp(\boldsymbol{U}\boldsymbol{\delta}))$, with the exponentiation taken element-wise. Thus,

$$B = E_G[\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^*) + (\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - n]. \tag{5}$$

Unfortunately, except for a few special cases, this does not appear to have a closed-form simplification that does not depend on the true, unknown values of model parameters. The special cases are described in the appendix. We instead perform simulations to estimate the needed expectations, and corroborate these in the cases where a closed-form solution is available.

Before beginning the simulation, we first associate each possible model $(\mathcal{L}, \mathcal{D})$ with a prototype; that is, one model that represents all models whose model matrices are isomorphic under permutation of the rows. For example, it is obvious that the model structure for any single-location-effect model is the same regardless of which location effect is used, so $(\{A\}, \{\emptyset\})$ serves as a prototype for all 15 models of this structure. In model with one

Table 2: CHIC penalty values for various common location-dispersion model prototypes. Subscript is the standard error from the simulation. A subscript of "−" indicates that the value was derived mathematically.

Location	Dispersion model, $\mathcal{D}$				
Model, $\mathcal{L}$	$\emptyset$	A	A,B	A,B,AB	A,B,C
$\emptyset$	4.9_{-}	$10.1_{0.1}$	$17.9_{0.1}$	$42.9_{0.7}$	$31.8_{0.3}$
A	8.0_{-}	12.8_{-}	$25.7_{0.2}$	$54.6_{1.4}$	$58.8_{1.2}$
B	8.0_{-}	$16.9_{0.2}$	$25.7_{0.2}$	$54.6_{1.4}$	$58.8_{1.2}$
A,B	11.6_{-}	$20.0_{0.1}$	$35.3_{0.3}$	$61.0_{2.6}$	$133.3_{5.6}$
A,B,AB	16.0_{-}	$24.1_{0.2}$	$36.3_{0.2}$	64.0_{-}	$190.7_{6.3}$
C	8.0_{-}	$16.9_{0.2}$	$37.7_{0.6}$	$582.4_{148.0}$	$58.8_{1.2}$
A,B,C	16.0_{-}	$32.3_{0.2}$	$80.4_{1.6}$	$644.8_{155.7}$	$332.9_{17.1}$

location effect and one dispersion effect, the dispersion effects is either on the same factor as the location effect or on a different one; hence, $(\{A\},\{A\})$ and $(\{A\},\{B\})$, are the two prototypes for this case.

Next, we need to select a true model under which simulations are to be performed. In keeping with the assumption that the true model does not contain nonzero parameters outside of the candidate model, then the only possible true model that is applicable to any candidate model is $(\mathcal{L},\mathcal{D}) = (\{\emptyset\},\{\emptyset\})$. Thus, the simulation for each prototype consists of generating a large number of data sets, each consisting of 2^k independent standard normal observations; fitting the prototype model to each data set; and computing the sum of terms inside the expectation in (5). The mean of these results is an estimate of the bias adjustment that is needed to complete the information criterion for this heteroscedastic model. We refer to the new criterion as the "corrected heteroscedastic information criterion" (CHIC). Some values of the bias adjustment are shown in Table 2 for certain models when $k = 4$. Exact penalties are available for the cases where $(\mathcal{L},\mathcal{D}) = (\{*\},\{\emptyset\})$, $(\{A\},\{A\})$, or $(\{A,B,AB\},\{A,B,AB\})$. where $*$ here means any location model. Notice some features about these penalties:

1. Dispersion effects are more expensive than location effects in the sense that adding a dispersion effect to a model generally incurs a larger penalty than adding a location effect to the same model.

2. The penalties are *much* larger than the AIC penalty ($2(p+q+2)$), the BIC penalty ($\log(16)(p+q+2)$), and generally also the AICc penalty, especially for larger models

3. The penalties become extremely large before a model becomes saturated in the sense of Loughin and Rodríguez (2010); for example, the estimated penalty for $(\mathcal{L}, \mathcal{D}) = (\{C\}, \{A, B, AB\})$ is 582 (recall that model $(\{C, D, CD\}, \{A, B, AB\})$ is "saturated").

The larger penalties are estimated with decreasing precision because the quantities inside the expectation in (5) are much more variable as models grow closer to saturation. This imprecision is unlikely to be a problem, as the estimated penalties in these cases are huge. Our analyses suggest that such large penalties do their job in keeping near-saturated models from artificially appearing to have excellent fits.

4.3 Model averaging

The last issue that needs to be resolved is how to use CHIC to select effects and/or models. As with other ICs, we seek models with small values of CHIC. However, it is well understood that exhaustive searches often lead to large numbers of models whose IC values are very close to the minimum value (see, e.g., Burnham and Anderson 2002). It is therefore often a matter of random chance which particular model achieves the minimum IC value. Hoeting et al. (1999) give a detailed review of a notion called Bayesian Model Averaging (BMA), wherein BIC values are transformed into approximate posterior probabilities that their respective models are correct, given the data. From the model posterior probabilities, a probability that each parameter is needed in the model can be computed as the sum of probabilities for all models in which the parameter appears.

Burnham and Anderson (2002) discuss extending BMA to any information criterion. The exact same calculations are used, but the results are interpreted as "evidence weights" rather than as probabilities in a Bayesian sense. We apply this general model-averaging construct to the joint location-dispersion models using CHIC follows:

1. Identify the M models to be estimated and calculate CHIC on each model, say CHIC_m, $m = 1, \ldots, M$.

2. Find the minimum CHIC value among all models. Call this value CHIC_0.

3. For each model, compute $\Delta_m = \text{CHIC}_m - \text{CHIC}_0$, and the *model evidence weight*

$$\tau_m = \frac{\exp(-\Delta_m/2)}{\sum_{a=1}^{M} \exp(-\Delta_a/2)}.$$

4. Compute the *effect evidence weight* for effect j as

$$\rho_j = \sum_{m=1}^{M} \tau_m I(\text{Effect } j \text{ is in model } m),$$

where we let $j = 1, \ldots, 2(n-1)$ index the set of all location effects and dispersion effects.

We refer to this procedure as Exhaustive-Search Model Averaging with Corrected Heteroscedastic Information Criterion (ESMA-CHIC).

Notice that the transformation $\text{CHIC} \to \tau$ is a monotone decreasing, so that the smaller a model's CHIC is, the larger its model evidence weight. Furthermore, $\sum_{m=1}^{M} \tau_m = 1$, and if there is a single model whose CHIC is distinctly smaller than the rest, then its evidence weight approaches 1. Thus, evidence weights are interpreted roughly the same as probabilities. Using guidelines suggested by Raftery (1995) for posterior probabilities from BMA, we can interpret evidence weights as offering evidence for a model m that is "weak" when $0.5 \leq \tau_m \leq 0.75$, "positive" when $0.75 < \tau_m \leq 0.95$, "strong" when $0.95 < \tau_m \leq 0.99$, and "very strong" when $\tau_m > 0.99$. Similar interpretation can be applied to each ρ_j. Furthermore, a $\rho_j \approx 0$ can be interpreted as evidence that effect j has practically no influence on the process.

5 Applications

We apply the ESMA-CHIC procedure to the three examples presented in Table 1. The top five models according to CHIC are shown in Tables 3–5, along with the CHIC values, evidence weights, and corresponding ranks for any models suggested by past literature that are not among the top five. Plots of the evidence weights for all effects are shown in Figure 1.

Table 3: Model evidence weights and ranks for top five models and for each model listed in Table 1 for the Welding data.

CHIC-Rank	Location	Dispersion	CHIC	Model Weight
1	B C	C	12.3	0.49
2	B C J "AB"[a]	C	16.0	0.08
3	B C D	C	16.0	0.08
4	B C J AG	C	17.5	0.04
5	B C J	C	17.7	0.03
245	B C	C J	30.0	0.00
409	B C AC AH A H	∅	33.1	0.00
1094	B C J	C J	52.9	0.00
>10,000	B C	C H J	299.0	0.00

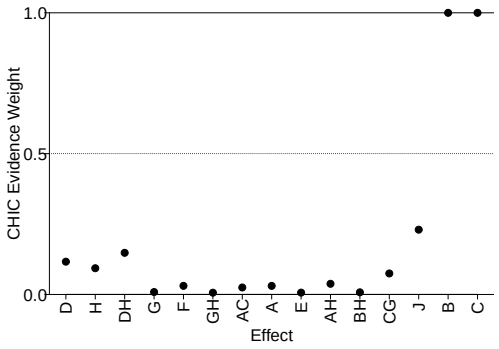
[a]According to Box and Meyer (1986), the experimenters did not believe that any of the aliased effects corresponding to this column of the design matrix, AB=CE=GH, would be active

Table 4: Model evidence weights and ranks for top five models and for each model listed in Table 1 for the Injection Molding data.

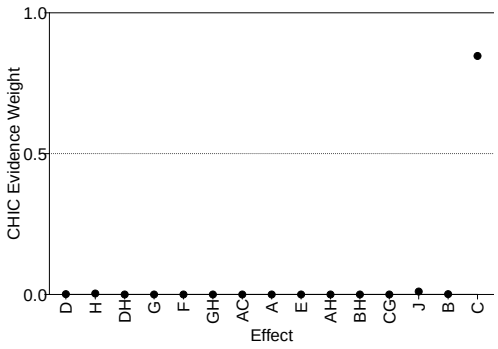
CHIC-Rank	Location	Dispersion	CHIC	Model Weight
1	A B AB G CG	∅	72.3	1.00
2	A B AB CG	∅	85.6	0.00
3	A B AB G	∅	87.9	0.00
4	A B AB	∅	90.3	0.00
5	A B AB BC CG	∅	90.5	0.00
277	A B AB	C	116.1	0.00

Table 5: Model evidence weights and ranks for top five models and for each model listed in Table 1 for the Dyestuff data.

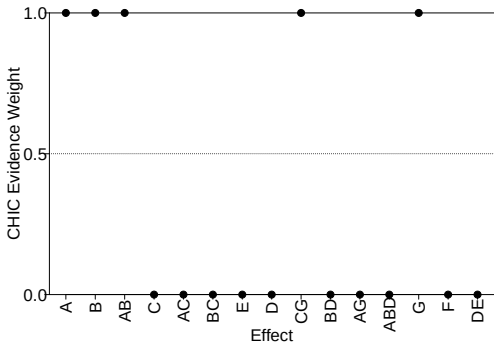
CHIC-Rank	Location	Dispersion	CHIC	Model Weight
1	C D AB CD	C	117.5	0.33
2	C D CD	∅	119.5	0.12
3	D CD	∅	121.3	0.05
4	C D BC CD	∅	121.7	0.04
5	C D CD BE	∅	122.3	0.03
15	D	E	124.9	0.01



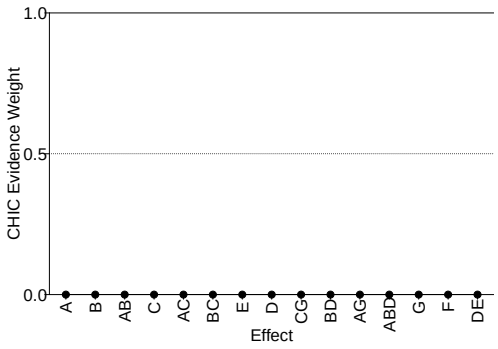
(a) Welding Location Effects



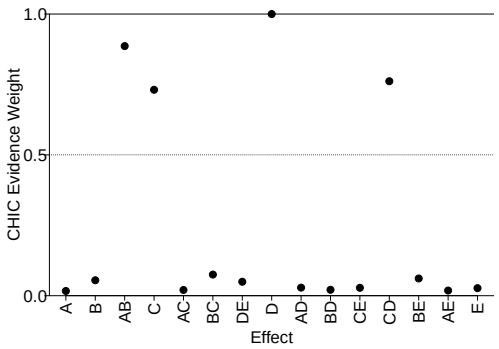
(b) Welding Dispersion Effects



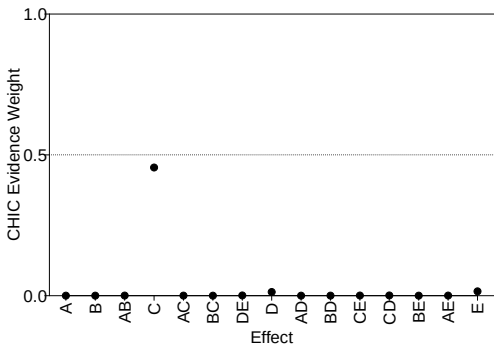
(c) Injection Molding Location Effects



(d) Injection Molding Dispersion Effects



(e) Dyestuff Location Effects



(f) Dyestuff Dispersion Effects

Figure 1: Plots of effects evidence weights for three example data sets: Welding (top), Injection Molding (middle), and Dyestuff (bottom)

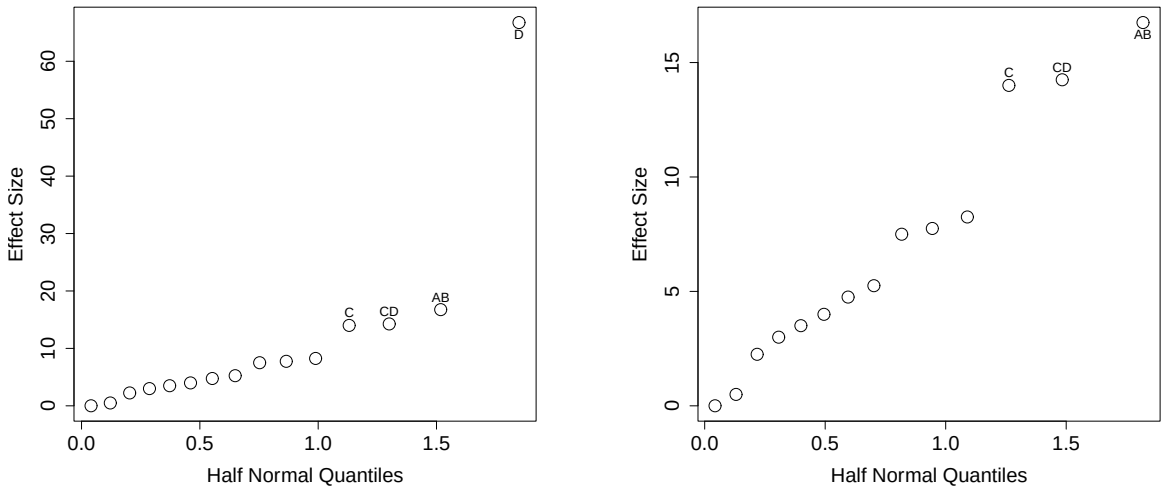
For the Welding example, Table 3 shows that the top model is $(\{B, C\}, \{C\})$ as suggested by Box and Meyer (1986), McGrath and Lin (2001) and Loughin and Malone (2013). This model has 0.49 evidence weight, so the data are not conclusive in their support for this model. However, no other model is a particular challenge for the best; this model is favored by at least 6:1 over any other model. The other top models all contain $(\{B, C\}, \{C\})$, along with a variety of additional location effects. Models with different dispersion effects do not fare well at all. Based on these results, it is no surprise that Figure 1a shows location effects B and C with evidence weights that round to 1 and none others that are particularly strong. Most have very little support from the data. Figure 1b shows that dispersion effect C has evidence weight 0.85. All other dispersion effects have negligible weight.

In the Injection Molding example, Table 4 shows that the data are conclusive in their preference for the model with five location effects and no dispersion effect, agreeing with McGrath (2003) and Loughin and Malone (2013). There is essentially no support for the alternative interpretation of a dispersion effect on C rather than location effects on G and CG. This is noteworthy, as it is the first time that these two competing models have been compared objectively using a single criterion. It serves to highlight one of the fundamental advantages of the ESMA-CHIC approach. Figures 1c and 1d re-express these results by showing evidence weights of 1 on location effects A, B, AB, G, and CG, and zero for all other effects.

The Dyestuff example represents another interesting problem. Past analyses using different methods by Bergman and Hynén (1997) and McGrath and Lin (2001) arrive at the same conclusion of a location effect on D and a dispersion effect on E. However, ESMA-CHIC shows that there is considerable uncertainty regarding what model best represents the data, with only 0.33 evidence weight on the top model, $(\{C, D, AB, CD\}, \{C\})$. All five of the top models contain location effects D and CD, and four contain C, but AB appears only in the first among the top five. The previously identified model, $(\{D\}, \{E\})$ is ranked 15th with an evidence weight of only 0.01. According to the data, it is extremely unlikely that this model describes the true process adequately.

Effect evidence weights in Figures 1e and 1f are similarly uncertain regarding what effects "belong" in the model. Among location effects, clearly D has "very strong" evidence in its

Figure 2: Half-normal plot of location effects for the Dyestuff example



(a) All effects

(b) All but the largest effect

favor (weight = 1), AB and CD both have "positive" evidence (weights 0.89 and 0.76, respectively), while C has "weak" evidence to support it (weight = 0.73). No other location effects have any appreciable support from the data. Among dispersion effects, none surpass the threshold for "weak" evidence, although C is very close with weight 0.46. Note that dispersion effect E has negligible weight, and hence is not considered to be a part of any serious explanation for these data.

How does such uncertainty come to pass, in particular with such a discrepancy between the results of ESMA-CHIC and past analyses? Part of the answer can be seen from a half-normal plot of the location effects for this example, as shown in Figure 2. The plot on the left shows that location effect D stands out as an "obviously" active effect, while those for AB, CD, and C are somewhat less obvious. Subjective assessment of this plot could either include or exclude these three effects from a model. However, the choice becomes much clearer when we exclude the obvious outlier and rescale the plot, as shown in Figure 2b. Now the three uncertain location effects are seen to stand out a bit more clearly from the line formed by the rest of the points, corroborating their evidence weights.

The belief that E should be a dispersion effect appears to be another example of the location-dispersion confounding issue touched on in Section 1 and discussed in more detail

in Pan (1999), McGrath and Lin (2001a, 2001b), and Loughin and Malone (2013). Notice that the two largest "uncertain" location effects, AB and CD, form an interaction triple with E. Thus, it is no surprise that failure to identify AB and CD as location effects should result in possible spurious identification of a dispersion effect at E. Once again we see the immense potential offered by this new ability to compare different combinations of location effects and dispersion effects using a single criterion.

6 Analyses of Simulated Data

To complement the results obtained on the three examples in the previous section, we assess the performance of ESMA-CHIC using simulated data where the true models are known. We generate data sets from model (2) for a 2^4 design using selected combinations $(\mathcal{L}, \mathcal{D})$. We perform analyses using either ESMA-CHIC or a two-step procedure consisting of the Lenth location-effect test followed by the modified Harvey dispersion-effect test (Lenth 1989, Brenneman and Nair, 2001). The Lenth test is chosen due to its simplicity and its reputation for reasonably good performance, even in the presence of dispersion effects (Wu and Hamada 2000, Hamada and Balakrishnan 1998, Zhang 2010, MS Thesis). The modified Harvey method is chosen because it is very similar in structure to the popular Bergman-Hynén test, but it is specifically designed to be compatible with our loglinear model for dispersion effects (Brenneman and Nair 2001). We simulate 100 data sets for ESMA-CHIC and 5000 for Lenth-Harvey to achieve a balance between simulation error and run time. We summarize the simulation results by computing the average power and type I error rate for detecting location effects and the same measures for detecting dispersion effects. For ESMA-CHIC, effects in the model with the smallest CHIC value are declared active. Note that ESMA-CHIC is not calibrated to achieve any specific Type I error rate, so there is no specific expectation for its performance in this measure. Both parts of the The Lenth-Harvey test are conducted using a nominal 0.05 level, with the critical values for the Lenth test taken from Loughin (1998).

We use all combinations of six true location models and five true dispersion models, for a total of 30 different models. We choose $\mathcal{L} = \{\emptyset\}, \{A\}, \{B\}, \{A, B\}, \{A, B, C\}$, and

$\{A, B, AB, C, AC, D\}$ because these represent five fairly typical location-model structures and one somewhat large model. Notice that the last location model has six active effects, which is more than the five that our implementation of ESMA-CHIC is designed to detect. Thus, we have a check on the price we pay for assuming more stringent effect sparsity than is actually present. We pair these location models with $\mathcal{D} = \{\emptyset\}$, $\{A\}$, $\{A, B\}$, $\{A, B, AB\}$, $\{A, B, C\}$.

Each active location effect is set to a level that would be detected with approximately 50% by a Wald test using the known variance of the contrast under the true dispersion model (McGrath 2003). This level is re-calibrated for each different location model, so that power for all Lenth tests remains at approximately 50% regardless of the model. Each active dispersion effect is set to a standard-deviation ratio of 5:1; i.e., the errors at the + level of the effect are multiplied by $\sqrt{5}$, while those at the lower level are divided by $\sqrt{5}$.

6.1 Results

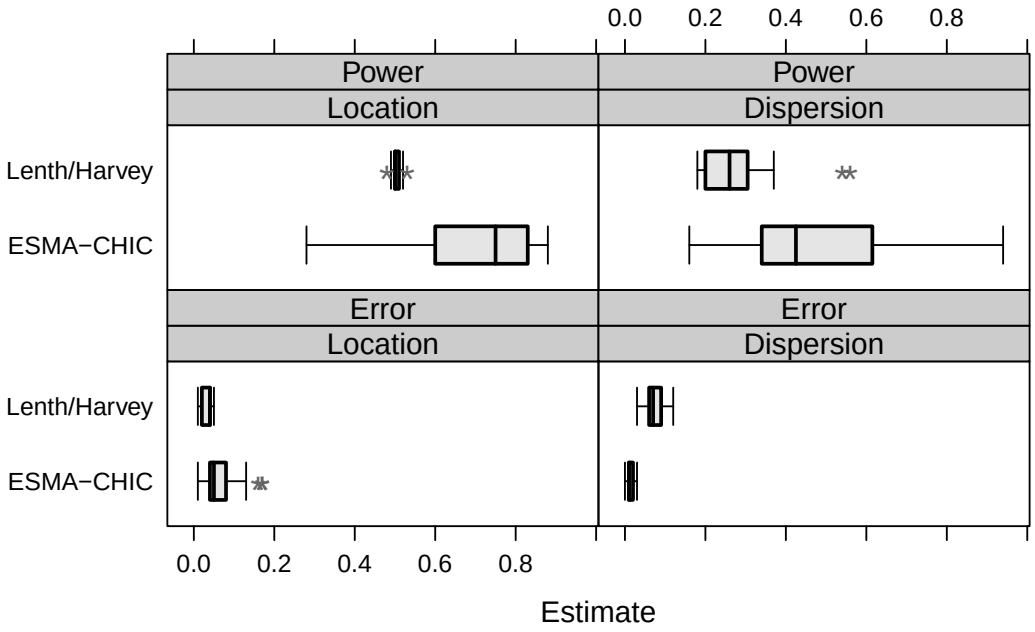
Results of the simulations are shown in Table 6 and depicted in Figure 3. Comparing location-effect detection for the two methods, note that the Lenth test has been calibrated to maintain both its error rate and its power. The simulations show that this is largely achieved, although the Lenth-test error rate does decrease as model size increases. By comparison, ESMA-CHIC has much more model-dependent error rates and power for detecting location effects. On average, error rates are slightly larger than those for the Lenth test, although when there are no dispersion effects they are considerably larger. When the true location model is larger than we assume, ESMA-CHIC has error rates that drop to around 1–2%.

Power for detecting location effects mirrors the error rates. ESMA-CHIC detection rates average close to 75%, although they are much lower when the model is too large. In many cases, the power is quite high—over 80%—even when the observed error rate is at or below 0.05. In contrast, the Lenth test becomes rather conservative as the size of the location model grows and when there are dispersion effects. The re-calibration of the sizes of the active effects allows its power to remain at 50% by increasing the sizes of the location effects in the larger dispersion models. ESMA-CHIC seems to have less difficulty detecting these large location effects than the Lenth test does.

Table 6: Estimated power ("Pow") and Type I Error rate ("Err") for detecting location effects ("L") and dispersion effects ("D") using the ESMA-CHIC procedure (100 simulations) and the Lenth/modified-Harvey test (5000 simulations).

True Model		Lenth/modified-Harvey				ESMA-CHIC			
Location	Dispersion	PowL	ErrL	PowD	ErrD	PowL	ErrL	PowD	ErrD
∅	∅		0.05		0.06		0.17		0.02
A	∅	0.50	0.04		0.06	0.74	0.17		0.01
B	∅	0.51	0.04		0.06	0.73	0.16		0.01
A,B	∅	0.49	0.04		0.06	0.76	0.13		0.02
A,B,C	∅	0.48	0.03		0.07	0.79	0.11		0.00
A,B,AB,C,AC,D	∅	0.50	0.01		0.10	0.54	0.01		0.02
∅	A		0.05	0.54	0.06		0.08	0.73	0.02
A	A	0.51	0.04	0.56	0.10	0.71	0.07	0.78	0.01
B	A	0.50	0.04	0.28	0.06	0.86	0.10	0.50	0.01
A,B	A	0.50	0.04	0.27	0.07	0.83	0.07	0.66	0.01
A,B,C	A	0.50	0.03	0.20	0.07	0.85	0.09	0.25	0.01
A,B,AB,C,AC,D	A	0.51	0.01	0.32	0.10	0.30	0.02	0.94	0.00
∅	A,B		0.04	0.37	0.06		0.04	0.70	0.02
A	A,B	0.51	0.04	0.29	0.07	0.83	0.05	0.55	0.01
B	A,B	0.50	0.04	0.30	0.07	0.86	0.05	0.60	0.01
A,B	A,B	0.51	0.03	0.19	0.09	0.77	0.06	0.45	0.01
A,B,C	A,B	0.51	0.03	0.18	0.08	0.71	0.06	0.18	0.03
A,B,AB,C,AC,D	A,B	0.51	0.01	0.24	0.11	0.28	0.02	0.49	0.00
∅	A,B,C		0.02	0.26	0.05		0.01	0.63	0.02
A	A,B,C	0.50	0.02	0.20	0.07	0.88	0.04	0.40	0.01
B	A,B,C	0.50	0.02	0.20	0.07	0.87	0.05	0.39	0.01
A,B	A,B,C	0.50	0.02	0.18	0.09	0.75	0.05	0.37	0.01
A,B,C	A,B,C	0.50	0.02	0.20	0.09	0.67	0.06	0.16	0.03
A,B,AB,C,AC,D	A,B,C	0.50	0.01	0.18	0.12	0.28	0.02	0.30	0.01
∅	A,B,AB		0.04	0.34	0.03		0.05	0.48	0.02
A	A,B,AB	0.52	0.05	0.25	0.04	0.76	0.04	0.38	0.01
B	A,B,AB	0.50	0.05	0.25	0.04	0.79	0.05	0.36	0.01
A,B	A,B,AB	0.49	0.04	0.31	0.04	0.45	0.04	0.35	0.01
A,B,C	A,B,AB	0.53	0.04	0.26	0.08	0.60	0.05	0.33	0.00
A,B,AB,C,AC,D	A,B,AB	0.49	0.01	0.26	0.11	0.29	0.02	0.33	0.00

Figure 3: Estimated power and type I Error rates for for detecting location effects and dispersion effects using the ESMA-CHIC procedure (100 simulations) and the Lenth/modified-Harvey test (5000 simulations).



Following the Lenth test with the modified Harvey test results in slightly larger-than-nominal error rates in most cases, with the greatest inflation occurring with the very large location model. In that model, there are many location effects that may be missed by the Lenth test, which creates more opportunities for creation of spurious dispersion effects to be detected. On the other hand, ESMA-CHIC has very low error rates for dispersion effects. It tends to be very conservative in declaring effects active. Nonetheless, its power is *uniformly* better than the modified Harvey test, in some cases by more than double. This is true despite the latter test's larger error rate.

7 Discussion and Conclusions

In this paper we have developed the first fully automated analysis procedure for 2^k factorial designs that can identify both location and dispersion effects in a single step. This is a critical advance, as it finally provides an objective approach to choosing between models where location-dispersion confounding may take place. Evidence of its effectiveness in this

regard comes from the simulations, where it competently detects moderately sized active effects and avoids detecting spurious effects of both types with very reasonable frequency.

The Lenth-Harvey approach that we used as its competitor was chosen deliberately to be a competitive alternative. In particular, the Harvey test assumes a normal distribution with loglinear dispersion effects, which is the same model that was used for the simulations. Thus, we gave our new method no "home-field advantage." The main flaw that we anticipated with the Harvey test in this context was its performance in the wake of the inability to correctly glean the location model with complete certainty through the Lenth test. Extended simulation results show that this issue is only one aspect of the whole difficulty with using a sequential testing scheme like Lenth/modified Harvey. In particular, several factors influence whether dispersion effects are spuriously identified by this procedure. For example, the presence of two real dispersion effects induces a spurious dispersion effect on their interaction (noted, e.g., by Brenneman and Nair 2001). This effect is apparent in the simulations from $\{A, B\}, \{A, B\}$, where the combination of location and dispersion confounding causes the AB dispersion effect to be falsely detected roughly 40% of the time—more often than either of the two active dispersion effects! This does not happen with ESMA-CHIC, because the model with $\mathcal{D} = \{A, B\}$ can be compared directly to the model with $\mathcal{D} = \{AB\}$ using a single criterion.

Another huge advantage of the ESMA-CHIC procedure is its ability to provide assessments of uncertainty regarding the importance of the model parameters. Although not true Bayesian posterior probabilities, the evidence weights derived from the CHIC nonetheless convey useful information about the relative importance of effects and models. In addition to providing listing of top models, the procedure affords an analyst who has a model in mind prior to analysis the opportunity to examine the evidence in its favor relative to other models. This should have great appeal to engineers and other application specialists, who typically know something about the processes that they are investigating and might like an objective assessment of their prior beliefs. Furthermore, the model-averaging aspect also allows one to find more realistic measures of uncertainty on the actual parameter estimates through unconditional variance estimates computed from the multitude of models that have been fit (Burnham and Anderson 2002). These variances account for the model-selection

uncertainty as well as typical within-model sampling variability.

Unfortunately, an exhaustive search of the model space is a computationally intensive process. As previously stated, we have shortened our computation time considerably (by about 60%) by imposing maxima of five location and five dispersion effects on any model fit. Even with these restrictions, analyzing one data set took an average of about 25 minutes in R on a Quad-core 2.40GHz processor with 32GB of RAM, using 3 of the 4 cores. Some of the code is written in C to improve runtime. There are two main sources for potential further improvement in runtime. First, the ESMA-CHIC algorithm would be easily parallelized: the majority of the computation is done on individual models, and this can be distributed easily to as many cores as are available. Second, one could simply choose to evaluate only models with a suitably low penalty term, as models with several location and dispersion effects have exceedingly large penalties and are not likely to be chosen among the best models.

An argument against fully automated model-selection procedures is that they often pay no attention to whether combinations of variables make practical sense together. For example, our algorithm for ESMA-CHIC makes no use of effect heredity or hierarchy. Under the circumstances, however, we do not consider this to be a particular weakness. The model-averaging aspect of ESMA-CHIC allows an analyst to peruse the top models for those that do make sense. If no practical models are highly supported by the data, then this may be a signal either that something was wrong with the experiment or that something unexpected is driving the responses. In either case, we are not aware of any other objective analysis method would be able to provide a sensible model for the analyst when faced with such data, and indeed would provide far less information regarding the extent to which "sensible" models are *not* supported by the data.

A final criticism is that our procedure is based on a parametric model, which means that it may be sensitive to mis-specification of that model. We have not assessed this sensitivity, but instead can point out that practically every other location- or dispersion-testing procedure is based on a variant of the same model (a notable exception is the permutation test of Loughin and Noble 1997). Furthermore, its basis in a parametric model means that there is nothing that limits the procedure to problems in the 2^k series. It could conceivably be developed to apply to other design series or even to much more general regression problems. Of course,

this would require deriving or estimating new CHIC penalty values.

Appendix: Special cases where CHIC Penalty can be calculated exactly

When there are no dispersion effects, the denominator in the last term in the log-likelihood (3) is a constant that factors out of the sum, leading to closed-form solutions for $\boldsymbol{\beta}$. The CHIC values coincide with AICc values in this case.

When there is a single dispersion effect, say A, the denominator takes two values according to the level of A for observation i. If the location model is also A, then the sum can be split into separate sums according to the level of A:

$$\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{\exp(\boldsymbol{u}_i'\boldsymbol{\delta})} = \sum_{i\in A+} \frac{(y_i - (\beta_0 + \beta_A))^2}{\exp(\delta_0 + \delta_A)} + \sum_{i\in A-} \frac{(y_i - (\beta_0 - \beta_A))^2}{\exp(\delta_0 - \delta_A)},$$

where $A+ = \{i : x_{iA} = +1\}$ and $A- = \{i : x_{iA} = -1\}$. Noting that the middle term of (3) reduces to $(n/2)\sigma_0^2$, then the likelihood is maximized by separately minimizing each of the above sums. This results in $\hat{\beta}_0 + \hat{\beta}_1 = \bar{y}_{A+}$, $\hat{\beta}_0 - \hat{\beta}_1 = \bar{y}_{A-}$, $\exp(\hat{\delta}_0 + \hat{\delta}_A) = \hat{\sigma}^2_{A+}$, and $\exp(\hat{\delta}_0 - \hat{\delta}_A) = \hat{\sigma}^2_{A-}$, where $\bar{y}_S$ is the sample mean over $i \in S$ and $\hat{\sigma}^2_S$ is $(|S| - 1)/|S|$ times the sample variance over $i \in S$.

We use this result to evaluate the terms in the bias, (5). First,

$$(\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \sum_{i\in A+} \frac{((\beta_0^* + \beta_A^*) - \bar{y}_{A+})^2}{\hat{\sigma}^2_{A+}} + \sum_{i\in A-} \frac{((\beta_0^* - \beta_A^*) - \bar{y}_{A-})^2}{\hat{\sigma}^2_{A-}}.$$

Since $\bar{y}_{A+} \sim N((\beta_0^* + \beta_A^*), \exp(\delta_0^* + \delta_A^*))$ and $(n/2)\hat{\sigma}^2_{A+}/\exp(\delta_0^* + \delta_A^*) \sim \chi^2_{(n/2)-1}$, we have

$$\frac{(n/2)(\bar{y}_{A+} - (\beta_0^* + \beta_A^*))^2/\exp(\delta_0^* + \delta_A^*)}{(n/2)\hat{\sigma}^2_{A+}/((n/2) - 1)\exp(\delta_0^* + \delta_A^*)} \sim F_{1,(n/2)-1}.$$

The expectation of this statistic is $((n/2) - 1)/((n/2) - 3)$. Thus,

$$E\left(\sum_{i\in A+} \frac{((\beta_0^* + \beta_A^*) - \bar{y}_{A+})^2}{\hat{\sigma}^2_{A+}}\right) = \frac{n/2}{(n/2) - 3} = \frac{n}{n-6}.$$

Hence, $E((\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}})) = 2n/(n-6)$..

Next,

$$\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^*) = (n/2)\exp(\delta_0^* + \delta_A^*)/\hat{\sigma}^2_{A+} + (n/2)\exp(\delta_0^* - \delta_A^*)/\hat{\sigma}^2_{A-}.$$

Since $(n/2)\hat{\sigma}^2_{A+}/\exp(\delta_0^* + \delta_A^*) \sim \chi^2_{(n/2)-1}$, we have

$$E(\exp(\delta_0^* + \delta_A^*)/(n/2)\hat{\sigma}^2_{A+}) = ((n/2) - 3)^{-1}.$$

Thus,

$$E((n/2)\exp(\delta_0^* + \delta_A^*)/\hat{\sigma}_{A+}^2) = (n/2)^2/((n/2) - 3) = n^2/2(n-6),$$

and hence $E(\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^*) = n^2/(n-6)$.

Combining terms, we have $B = n^2/(n-6) + 2n/(n-6) - n = 8n/(n-6)$. When $n = 16$, this yields $B = 12.8$.

Similar calculations for $\mathcal{L} = \mathcal{D} = \{A, B, AB\}$ are based on factoring the last term in (3) into four independent terms. This yields $B = n^2/(n-12) + 4n/(n-12) - n = 16n/(n-12)$. When $n = 16$ we have $B = 64$.

References

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In *Second international symposium on information theory*, Petrov, B. N. and Csaki, F, editors. Academiai Kiado, Budapest, 267–281.

Bergman, B. and Hynén, A. (1997). "Dispersion Effects from Unreplicated Designs in the 2^{k-p} Series." *Technometrics*, 39, pp. 191–198.

Box, G.E. and Meyer, R.D. (1986). "Dispersion Effects from Fractional Designs." *Technometrics*, 28, pp. 19–27.

Brenneman, W.A. and Nair, V.N. (2001). "Methods for Identifying Dispersion Effects in Unreplicated Factorial Experiments: A Critical Analysis and Proposed Strategies." *Technometrics*, 43, pp. 388–405.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.)*. Springer, New York.

Bursztyn, D. and Steinberg, D.M. (2006). "Screening Experiments for Dispersion Effects." In *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, Dean, A. and Lewis, S., editors. Springer, New York, NY, pp. 21–47

Carroll, R. J. and Ruppert, D. (1988) *Transformation and Weighting in Regression*. Chapman and Hall, New York.

Cavanaugh, J. E. (1997). "Unifying the Derivations for Akaike and Corrected Akaike Information Criteria." *Statistics and Probability Letters*, 33, 201–208.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Cook, R.D. and Weisberg, S. (1983). "Diagnostics for Heteroscedasticity in Regression." *Biometrika*, 70, pp. 1–10.

Davies, O. L. (ed.) (1963), *The Design and Analysis of Industrial Experiments*, Oliver and Boyd, London.

Engel, J and Huele, A. F. (1996). "A Generalized Linear Modeling Approach to Robust Design." *Technometrics*, 38, 365–373

Ferrer, A.J. and Romero, R. (1993). "Small Samples Estimation of Dispersion Effects from Unreplicated Data." *Communications in Statistics: Computation and Simulation*, 22, pp. 975–995.

Furnival, G. M. and Wilson, R. W., Jr. (1974). "Regressions by Leaps and Bounds." *Technometrics*, 16, 499-511.

Hamada, M. and Balakrishnan, N. (1998). "Analyzing Unreplicated Factorial Experiments: A Review with Some New Proposals." *Statistica Sinica*, 8, pp. 1–41.

Harvey, A. C. (1976). "Regression Models with Multiplicative Heteroscedasticity" *Econometrica*, 44, 461–465.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). "Bayesian Model Averaging: A Tutorial (with Discussion)." *Statistical Science*, 14, 382–401.

Hurvich, C. M. and Tsai, C.-L. (1989). "Regression and Time Series Model Selection in Small Samples." *Biometrika*, 76, 297–307.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer, New York.

Lenth, R.V. (1989). "Quick and Easy Analysis of Unreplicated Factorials." *Technometrics*, 31, pp. 469–473.

Loughin, T. M. (1998). "Calibration of the Lenth Test for Unreplicated Factorial Designs." *Journal of Quality Technology*, 30, 171–175.

Loughin. T. M. and Malone, C. J. (2013). "An Adaptation of the BergmanHynén Test for Dispersion Effects in Unreplicated Two-Level Factorial Designs When the Location Model May Be Incorrect." *Journal of Quality Technology*, **45**, 350–359.

Loughin, T. M. and Noble, W. (1997). "A Permutation Test for Effects in an Unreplicated Factorial Design." *Technometrics*, **39**, 180–190.

Loughin, T. M. and Rodríguez, J. E. (2011). "Computational Issues with Fitting Joint Location/Dispersion Models in Unreplicated 2^k Factorials." *Computational Statistics and Data Analysis*, 55, 491–497.

McGrath, R.N. (2003). "Separating Location and Dispersion Effects in Unreplicated Fractional Factorial Designs." *Journal of Quality Technology*, 35, pp. 306–316.

McGrath, R.N. and Lin, D.K. (2001a). "Confounding of Location and Dispersion Effects in Unreplicated Fractional Factorials." *Journal of Quality Technology*, 33, pp. 129–139.

McGrath, R.N. and Lin, D.K. (2001b). "Testing Multiple Dispersion Effects in Unreplicated Fractional Factorial Designs." *Technometrics*, 43, pp. 06–414.

Michalewicz, Z. (1998). *Genetic algorithms + Data Structures = Evolution Programs*. Springer, New York.

Montgomery, D.C. (1990). "Using Fractional Factorial Designs for Robust Process Development." *Quality Engineering*, 3, pp. 193–205.

Nair, V.N. and Pregibon, D. (1988). "Analyzing Dispersion Effects from Replicated Factorial Experiments." *Technometrics*, 30, pp. 247–257.

Nelder, J.A. and Lee, Y. (1991). "Generalized Linear Models for the Analysis of Taguchi-Type Experiments." *Applied Stochastic Models and Data Analysis*, 7, pp. 107–120.

Pan, G. (1999). "The Impact of Unidentified Location Effects on Dispersion Effect Identification from Unreplicated Factorial Designs." *Technometrics*, 41, pp. 313–326.

Pan, G and Taam, W (2002). "On Generalized Linear Model Method for Detecting Dispersion Effects in Unreplicated Factorial Designs." *Journal of Statistical Computation and Simulation*, 72, pp. 431–450.

Raftery, A. E. (1995). "Bayesian Model Selection in Social Research." *Sociological Methodology*, 25, 111–163.

Wang, P.C. (1989). "Test for Dispersion Effects from Orthogonal Arrays." *Computational Statistics and Data Analysis*, 8, pp. 109–117.

Wu, C. F. J. and Hamada, M. S. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley and Sons, Hoboken, NJ.

Ye, K., and Hamada, M. S. (2000). "Critical Values of the Lenth Method for Unreplicated Factorial Designs," *Journal of Quality Technology*, 32, 57–66.

Zhang, Y. (2010). "A Comparison of Location Effect Identification Methods for Unreplicated Fractional Factorials in the Presence of Dispersion Effects." MSc. Thesis, Simon Fraser University, Burnaby, BC, Canada.