# THE STEPWISE RESPONSE REFINEMENT SCREENER (SRRS) AND ITS APPLICATIONS TO ANALYSIS OF FACTORIAL EXPERIMENTS

Frederick Kin Hing Phoa

*Institute of Statistical Science, Academia Sinica, 128 Academia Rd. Sec. 2, Nangang Dist., Taipei 115, Taiwan R.O.C.*

Keywords: Stepwise Response Refinement Screener (SRRS), Modified akaike information criterion, Screening experiment, Nonregular fractional factorial design, Supersaturated design.

Abstract: Two-level supersaturated designs are very useful in the screening experiments and the common goal is to identify sparse but dominant active factors with low cost. Recently, a new analysis procedure called the Stepwise Response Refinement Screener (SRRS) method is proposed to screen important effects. This paper extends this method to the two-level nonregular fractional factorial designs. The applications to several real-life examples suggest that the SRRS method is able to retrieve similar results as the existing methods do. Simulation studies show that compared to existing methods in the literature, the SRRS method performs well in terms of the true model identification rate and the average model size.

## 1 INTRODUCTION

As science and technology have advanced to a higher level nowadays, investigators are becoming more interested in and capable of studying large-scale systems. To address these challenges of expensive experimental costs, research in experimental design has lately focused on the class of supersaturated designs (SSD) for their run-size economy and mathematically novelty. Under the condition of factor sparsity (Box and Meyer, 1986), these experiments aims at correctly identifying the subset of those active factors that have significant impact on the response, so that the whole investigation can be economically proceed via discarding inactive factors prior to the follow-up experiments.

Traditionally, SSDs are employed only for screening main effects, and interactions are discarded due to limited degree of freedom. More refined analysis methods were recently developed and Phoa, Pan and Xu (2009) provides a comprehensive list of recent analysis methods found in the literature. Candes and Tao (2007) proposed the Dantzig selector (DS) and showed that it has some remarkable properties under some conditions. Phoa, Pan and Xu (2009) implemented the DS in practice, introducing a graphical procedure via a profile plot for analysis and an automatic variable selection procedure via a modified Akaike information criterion (*AIC*). Traditionally, *AIC* is used for model selection. For linear models, it is defined as

$$AIC = n\log(RSS/n) + 2p \qquad (1)$$

where $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares and $p$ is the number of parameters in the model. It is known that *AIC* tends to overfit the model when the sample size is small. Phoa, Pan and Xu (2009) imposed a heavy penalty on the model complexity and proposed a new modified *AIC* for the DS method, which is defined as

$$mAIC = n\log(RSS/n) + 2p^2 \qquad (2)$$

The *mAIC* typically chooses a smaller model than *AIC*.

Recently, Phoa (2011) introduce a new variable selection approach via the Stepwise Response Refinement Screener (SRRS). The SRRS chooses the best subset of variables or active factors by two procedures: Factor Screening and Model Searching. This method has shown its superior model selection ability via a comparison to five commonly used methods in the literature, namely SSVS (Chipman et al., 1997), SSVS/IBF (Beattie et al., 2002), SCAD (Li and Lin, 2003), PLSVS (Zhang et al., 2007) and the DS (Phoa et al., 2009a) method. Readers who are interested in the main idea of the SRRS method are referred to Phoa (2011) . This paper aims at extending the SRRS method to the variants of supersaturated experiments. In section 2, we review the notation and the

Kin Hing Phoa F. (2012).
THE STEPWISE RESPONSE REFINEMENT SCREENER (SRRS) AND ITS APPLICATIONS TO ANALYSIS OF FACTORIAL EXPERIMENTS.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 157-161
DOI: 10.5220/0003711101570161

general procedure of SRRS introduced in Phoa (2011) . Section 3 discusses the modifications of the algorithms when the SRRS is used in these extensions. To demonstrate the value of the SRRS method, a simulation study is performed in section 4. The result shows that the SRRS method is powerful for analyzing not only SSDs but also its variant designs. The last section gives some concluding remarks.

## 2 ANALYSIS OF FRACTIONAL FACTORIAL DESIGNS VIA THE SRRS METHODS

Fractional factorial designs (FFDs) are classified into two broad types: Regular FFDs and Nonregular FFDs. Regular FFDs are constructed through defining relations among factors and are described in many textbooks (Wu and Hamada, 2000). These designs have been widely used in scientific researches and industrial processes because they are simple to construct and to analyze. On the other hands, nonregular FFDs such as Plackett and Burman (1946) designs, Quaternary-code designs (Phoa and Xu 2009 , Zhang et. al. 2011 ) and other orthogonal arrays are often used in various screening experiments for their run size economy and flexibility (Wu and Hamada, 2000). Phoa, Xu and Wong (2009) demonstrated the advantages of using nonregular FFDs using two real-life toxicological experiments. Phoa, Wong and Xu (2009) used three real-life chemometrics examples to show the analysis pitfalls when the interactions are assumed to be insignificant without verifications.

In this section, we extend the use of the SRRS method to the analysis of fractional factorial designs (FFDs), including two-level nonregular FFDs and multi-level FFDs.

### 2.1 Modification of the SRRS Method Accompanied for the Analysis of Nonregular Designs

Consider a nonregular FFDs with $k_1$ main effects and $n$ runs, where $n < m$. There are $k_2 = k_1(k_1+1)/2$ interactions between two different main effects. If all two-factor interactions are considered together with all main effects, it is possible that $k_2 > m$, then the design matrix is supersaturated. We express the relationship via a linear regression model $y = X\beta + \varepsilon$ where $y$ is an $n \times 1$ vector of observations, $X$ is an $n \times k$ model matrix for $k = k_1 + k_2$, $\beta$ is a $k \times 1$ vector of unknown parameters, and $\varepsilon$ is an $n \times 1$ vector of random errors. Assume that $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ is a vector of independent normal random variables. In addition, $X$ is assumed to be supersaturated, i.e. $n < k$. We denote $m$ to be the number of potentially important effects (PIEs) and $S_{inf}$ to be the influential set of PIEs found in the process.

Traditionally, the analysis of nonregular FFDs is based on two assumptions: the factor sparsity principle and the effect heredity prinicple. The first assumption has been embedded in the SRRS method, but the second assumption does not. In order to implement the heredity principle into the SRRS method, the two procedures of the SRRS method are slightly modified and presented in the following steps:

I. SRRS (Heredity Prinicple embedded)–Factor Screening:

Step 1. Standardize data so that $y_0$ has mean 0 and columns of $X$ have equal lengths.

Step 2. Compute the marginal correlations $\rho(X_i, y_0)$ for all main effects $X_i$, $i = 1, \ldots, k$. $(*)$

Step 3. Choose $E_0$ such that $|\rho(E_0, y_0)| = \max_{X_i} |\rho(X_i, y_0)|$. Identify $E_0$ as the first PIE and include $E_0$ in $S_{Inf}$.

Step 4. Obtain the estimate $\beta_{E_0}$ by regressing $y_0$ on $E_0$.

Step 5. For the next $m$ PIEs $E_j$ where $j = 1, \ldots, m$, $m < n-2$,

(a) Compute the refined response $y_j = y_{j-1} - E_{j-1}\beta_{E_{j-1}}$.

(b) Compute the marginal correlations $\rho(\{X_i, X_{ij}\}, y_j)$ for all main effects $X_i$, $i = 1, \ldots, k$ and all two-factor interactions $X_{ij}$, $X_j \in S_{Inf}$. $(*)$

(c) Choose $E_j$ such that $|\rho(E_j, y_j)| = \max_{\{X_i, X_{ij}\}} |\rho(\{X_i, X_{ij}\}, y_j)|$. $(*)$

(d) Obtain the estimate $\beta_{E_j}$ by regressing $y_j$ on $E_j$.

(e) Identify $E_j$ as a PIE and include $E_j$ in $S_{Inf}$ if $|\beta_{E_j}| \geq \gamma$, where $\gamma$ is the threshold of noise level.

(f) Repeat (a) to (e) until $E_m$ is not included in $S_{Inf}$.

II. SRRS–Model Searching:

Step 6. Perform all-subset search, with the consideration of the heredity principle, for all $E_j$, from models with one factor to models with $m$ factors, where $m$ is minimum between the ceiling of $n/3$ or the number of $E_j$ in $S_{inf}$. $(*)$

Step 7. Compute *mAIC* for each model and choose the final model with the smallest *mAIC* among all models, and all $E_j$ included in the final model are considered to be significant to the response $y_0$.

The first modification is in Step 2. Due to the heredity principle, two-factor interactions are never be selected as the first PIE, so only the marginal correlations of all main effects are compared for selecting the first PIE. The second and third modifications are in Step 5. During the search of the $j^{th}$ PIE, not all two-factor interactions are considered in the comparison of marginal correlation. According to the heredity principle, a two-factor interaction $X_{ij}$ is considered in Step 5(b) if and only if either $X_i$ or $X_j$ or both parents main effects have been included in $S_{Inf}$ in the previous searches. Therefore, the modifications in Step 5 take away a subset of two-factor interactions that none of their corresponding parent main effects have been PIEs. The last modification is in Step 6. The reduced models built in this step must follow the heredity principle in order to avoid the situation that some significant two-factor interactions are included in the reduced model but none of their parent main effects have been included.

## 2.2 Two Illustrating Examples

We illustrate the analysis of nonregular FFDs via the SRRS method step by step using the following two examples. The Factor Screening procedures is terminated via the noise threshold in the first example and via the maximum number of PIEs in the second example.

**Example 1.** Consider the cast fatigue experiment (Wu and Hamada 2000 , section 7.1), a real data set consisting of seven two-level factors. The design matrix and the response are found in Wu and Hamada (2000) . When all two-factor interactions are considered to be as important as the main effects, the design matrix consists of 21 additional interactions and is supersaturated.

In the Factor Screening procedure, the first PIE being identified is $F$ and its absolute marginal correlation to $y_0$ is the highest among all main effects (0.6672). A regression model between $y_0$ and $F$ is built and the magnitude of the slope estimate $|\beta_F| = 0.4576$. Then we set the threshold $\gamma = 0.04$, about 10% of $\beta_F$.

To search for the second PIE, the new response $y_1$ is refined by subtracting $F\beta_F$ from $y_0$. Then among all main effects and all the two-factor interactions that consist of $F$, $FG$ (the interaction between main effects $F$ and $G$) has the highest absolute marginal correlation (0.8980) to $y_1$ and so it is identified as the second PIE. A regression model between $y_1$ and $FG$, $F$ is built and the magnitude of the slope estimate $|\beta_{FG}| = 0.4588 > \gamma$. This means $FG$ is important enough to be included in the influential set $S_{Inf}$ together with $F$.

Table 1: Factor Screening of Cast Fatigue Experiment Data.

| m | PIE | Marginal Correlation | $\lvert\beta\rvert$ | Continue or Stop |
|---|---|---|---|---|
| 0 | $F$ | 0.6672 | 0.4576 | Continue |
| 1 | $FG$ | −0.8980 | 0.4588 | Continue |
| 2 | $D$ | −0.4677 | 0.1183 | Continue |
| 3 | $EF$ | −0.6336 | 0.1442 | Continue |
| 4 | $C$ | 0.5032 | 0.0758 | Continue |
| 5 | $E$ | −0.5817 | 0.0785 | Continue |
| 6 | $AE$ | −0.7667 | 0.1482 | Continue |
|  | $AE$ | −0.6835 | 0 | Stop |

PIEs in $S_{Inf}$ after Factor Screening:
$C, D, E, F, AE, EF, FG$

The procedure continues to search for the next five PIEs. Table 1 shows every step of the process of Factor Screening. Note that in the last step, the absolute magnitude of the slope estimate of $AE$ is close to 0, so the search stops and seven PIEs are identified in the Factor Screening procedure.

Since there are 12 observations in the data, the maximum number of active factors is suggested to be 4. There are totally 98 reduced models up to four-factors models that are constructed from seven PIEs, but only 49 of them fulfill the heredity principle. A comparison of the *mAIC*s of these 49 reduced models shows that the two-effects model with $F$ and $FG$ has the lowest $mAIC = -27.82$. Thus the SRRS method suggests that $F$ and $FG$ have significant impacts to the response $y_0$. This result is also recommended by Wu and Hamada (2000, Section 8.4) and the Dantzig selector (DS) method in Phoa, Pan and Xu (2009) .

**Example 2.** Consider the high-performance liquid chromatography (HPLC) experiment (Vander-Heyden et al., 1999), a real data set consisting of eight two-level factors. The design matrix and the response are found in Phoa, Wong and Xu (2009) . When all two-factor interactions are considered to be as important as the main effects, the design matrix consists of 28 additional interactions and is supersaturated.

In the Factor Screening procedure, the first PIE being identified is $E$ and its absolute marginal correlation to $y_0$ is the highest among all main effects (0.5019). A regression model between $y_0$ and $E$ is built and the magnitude of the slope estimate $|\beta_F| = 0.5583$. Then we set the threshold $\gamma = 0.05$, about 10% of $\beta_E$.

To search for the second PIE, the new response $y_1$

Table 2: Factor Screening of HPLC Experiment Data.

| m | PIE | Marginal Correlation | $\|\beta\|$ | Continue or Stop |
|---|---|---|---|---|
| 0 | $E$ | $-0.5019$ | 0.5583 | Continue |
| 1 | $EF$ | 0.8055 | 0.7750 | Continue |
| 2 | $F$ | 0.7747 | 0.4417 | Continue |
| 3 | $H$ | $-0.7396$ | 0.3000 | Continue |
| 4 | $FH$ | 0.5897 | 0.1625 | Continue |
| 5 | $A$ | 0.6922 | 0.1389 | Continue |
| 6 | $FI$ | $-0.5295$ | 0.0893 | Continue |
| 7 | $EI$ | 0.5713 | 0.0836 | Continue |
| 8 | $AF$ | $-0.6587$ | 0.0792 | Continue |
|  | $EF$ | 0.6951 | 0.0667 | Continue |

PIEs in $S_{Inf}$ after Factor Screening:
$A, E, F, H, AF, EF, EI, FH, FI$

is refined by subtracting $E\beta_E$ from $y_0$. Then among all main effects and all the two-factor interactions that consist of $E, EF$ (the interaction between main effects $E$ and $F$) has the highest absolute marginal correlation (0.8055) to $y_1$ and so it is identified as the second PIE. A regression model between $y_1$ and $EF, E$ is built and the magnitude of the slope estimate $|\beta_{EF}| = 0.7750 > \gamma$. This means $EF$ is important enough to be included in the influential set $S_{Inf}$ together with $E$.

The procedure continues to search for the next eight PIEs. Table 2 shows every step of the process of Factor Screening. Note that in the last step, although the absolute magnitude of the slope estimate of $EF$ is $0.0667 > \gamma$, the $m < n-2$ criterion stops the search and nine PIEs are identified in the Factor Screening procedure.

Since there are 12 observations in the data, the maximum number of active factors is suggested to be 4. With nine PIEs found in the previous step, there are totally 255 reduced models up to four-factors models, but only 102 of them fulfill the heredity principle. A comparison of the *mAIC*s of these 102 reduced models shows that the three-effects model with $E$, $F$ and $EF$ has the lowest $mAIC = -6.48$. Thus the SRRS method suggests that $E$, $F$ and $EF$ have significant impacts to the response $y_0$.

Phoa, Wong and Xu (2009) previously analyzed the same data and concluded that an additional effect $H$ was also significant to the response. The *mAIC* of the model consisting of $E$, $F$, $H$ and $EF$ is $-3.95$, which is slightly higher than our suggested model. The increase of *mAIC* when $H$ is added comes from the heavy penalty to the number of factors in the model. If other penalty terms are used, results may be different. For example, the original *AIC* favors the addition of $H$. Therefore, $H$ may be barely significant and some follow-up experiments are suggested to investigate the significance of $H$ to the response.

Table 3: Summary of Simulation Results in Example 3.

| Case |  | I | II | III | IV |
|---|---|---|---|---|---|
| Min | TMIR | 94% | 47% | 5% | 0% |
|  | Size | 1.00 | 1.85 | 2.05 | 1.06 |
| 1st Q. | TMIR | 97% | 97% | 44% | 15% |
|  | Size | 1.01 | 2.01 | 3.00 | 2.42 |
| Median | TMIR | 98% | 97% | 96% | 53% |
|  | Size | 1.02 | 2.02 | 3.00 | 3.30 |
| 3rd Q. | TMIR | 99% | 99% | 99% | 88% |
|  | Size | 1.03 | 2.03 | 3.01 | 3.76 |
| Max | TMIR | 100% | 100% | 100% | 99% |
|  | Size | 1.06 | 2.05 | 3.04 | 3.98 |

## 3 SIMULATION STUDIES

In order to judge the value of the SRRS method, we randomly generate some models and evaluate the performance of the SRRS method.

**Example 3.** In this example, we generate data from the same linear model as in Example 1. Since there are only 12 observations in the data, the maximum possible number of active factors is 4. Therefore, we consider four cases for *beta*. There are $i$ active factors for case $i$, $1 \leq i \leq 4$. For each case, we generate 500 models where the selection of active factors is random without replacement, the signs of the active factors are randomly selected from either positive or negative, and the magnitudes are randomly selected from 2 to 10 with replacement. For each model, we generate data 100 times and obtain the True Model Identified Rate (TMIR) and the average model size. In the simulations we fix $\gamma = 1$, which is approximately equal to 10% of $\max|\beta_i|$. Table 3 gives the summary statistics of these two quantities among 500 models.

The SRRS method is very effective in identifying 1, 2 and 3 active factors; the TMIR in these cases are at least 96% in average true model identified rate and only a few cases that have average model sizes slightly higher than the true numbers of active factors. The performance of the method decreases in identifying 4 active factors. It is mainly because of the limit posted on the allowed number of active factors, which leads to a slightly underfitting situation.

## 4 CONCLUDING REMARKS

The Stepwise Response Refinement Screener (SRRS) method has shown its satisfactory performance on screening the supersaturated designs in Phoa (2011) . In this paper, we modify the SRRS method in order to adapt for analyzing the nonregular FFDs with the consideration of interactions. Under the validity of the factor sparsity and effect heredity assumptions, the calculations needed to carry out the analysis are simple and easily performed with little computation time. Simulation suggests that the SRRS method performs well in most of the cases, except when it is on the line of maximum number of allowed active factors. In addition, we cannot ensure that thsi method works well in every case as its fundamental theorems are still under investigation. Sometimes it may still possible to reach misleading conclusion, so it is highly recommended that once the suggested set of significant factors is found, a follow-up experiment is needed for validating the results.

## REFERENCES

Beattie, S., Fong, D., and Lin, D. (2002). A two-stage bayesian model selection strategy for supersaturated designs. *Technometrics*, 44:55–63.

Box, G. and Meyer, R. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, 28:11–18.

Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when $p$ is much larger than $n$. *Annals of Statistics*, 35:2313–2351.

Chipman, H., Hamada, H., and Wu, C. (1997). A bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics*, 39:372–381.

Li, R. and Lin, D. (2003). Analysis methods for supersaturated design: some comparisons. *Journal of Data Science*, 1:249–260.

Phoa, F. (2011). The stepwise response refinement screener (srrs). *(in review)*.

Phoa, F., Pan, Y., and Xu, H. (2009a). Analysis of supersaturated designs via the dantzig selector. *Journal of Statistical Planning and Inference*, 139:2362–2372.

Phoa, F., Wong, W., and Xu, H. (2009b). The need of considering the interactions in the analysis of screening designs. *Journal of Chemometrics*, 23:545–553.

Phoa, F. and Xu, H. (2009). Quater-fraction factorial design constructed via quaternary codes. *Annals of Statistics*, 37:2561–2581.

Phoa, F., Xu, H., and Wong, W. (2009c). The use of nonregular fractional factorial designs in combination toxicity studies. *Food and Chemical Toxicology*, 47:2183–2188.

Plackett, R. and Burman, J. (1946). The design of optimum multifactorial experiments. *Biometrika*, 33:305–325.

Vander-Heyden, Y., Jimidar, M., Hund, E., Niemeijer, N., Peeters, R., Smeyers-Verbeke, J., D.L., M., and Hoogmartens, J. (1999). Determination of system suitability limits with a robustness test. *Journal of Chromatography A*, 845:145–154.

Wu, C. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization.* Wiley, New York.

Zhang, Q., Zhang, R., and Liu, M. (2007). A method for screening active effects in supersaturated designs. *Journal of Statistical Planning and Inference*, 137:235–248.

Zhang, R., Phoa, F., Mukerjee, R., and Xu, H. (2011). A trigonometric approach to quaternary code designs with application to one-eighth and one- sixteenth fractions. *Annals of Statistics*, 39:931–955.