Stepwise paring down variation for identifying influential multifactor interactions

Jing-Shiang Hwang

Academia Sinica, Taiwan

The objective

Identify influential subsets of factor variables $\{X_1, X_2, \ldots, X_m\}$ causally related to a continuous response variable Y

Multifactor ANOVA model

Data: $Y = (y_1, \ldots, y_n)'$, $X_l = (x_{l1}, \ldots, x_{ln})'$, $l = 1, \ldots, m$

$$Y = \mu + \sum_{j=1}^{J} A_j \beta_j + \varepsilon$$

A_j is an $n \times p_j$ matrix corresponding to the jth factor combination and β_j is a coefficient vector of size p_j, $j = 1, \ldots, J = C_1^m + C_2^m + \cdots + C_d^m$.

A_j may consist of a single-factor or two factors or more factors.

A_j is influential if any element of β_j is significantly away from 0.

Possible solutions

- Group lasso
  - Yuan and Lin, JRSSB 2006
- Logic regression
  - Schwender and Ickstadt, Biostatistics 2008
- Bayesian QTL methods
  - Yandell et al., Bioinformatics 2007
- Partition retention
  - Chernoff et al., Annals of Applied Statistics 2009

The proposed approach

1. Exhaustive search for the factor with the largest effect among the *m* single-factor ANOVA models, and compared to a threshold.

   The threshold is determined using a permutation approach.

2. If the factor is identified as influential, the responses are refined by subtracting the mean effects of the identified factor, which are in fact the residuals.

3. Repeat exhaustive search for next single factor with the newly refined responses.

The proposed approach

4. If the estimated effect is no longer larger than the threshold, move to the next stage of exploring sets of factor pairs by fitting all the $m(m\text{-}1)/2$ single term ANOVA models of two-factor combination.

   Repeat steps 1. – 3. for screening pairs.

5. The procedures can be repeated in the next stages to screen for higher order interactions.

The key idea

- The total variation of responses will be pared down stepwise in the screening processes
  - At each run, the total variation of refined responses is reduced significantly, and so is the model error variation in the next run.
  - The remaining influential sets of factors have increased chances of being identified in the next runs.

- Stepwise Paring-down Variation (SPV) algorithm

An illustrative example

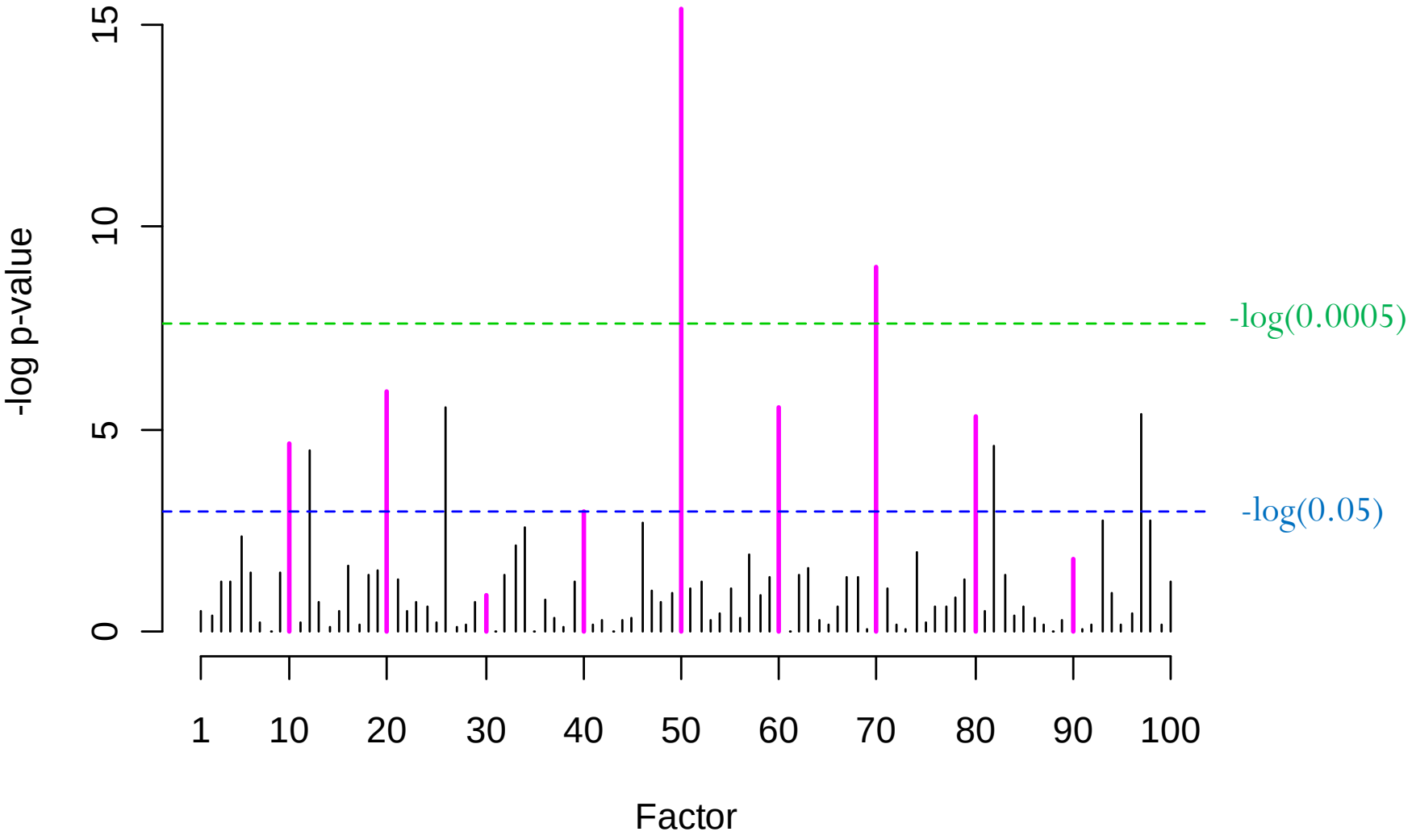
- Generate $m = 100$ independent factors of three levels 0, 1, 2 with equal probabilities for $n = 100$ subjects.
- The responses are affected by the nine factors $X_{10}, X_{20}, \ldots, X_{90}$,

$$y_i = \sum_{j \in \{10,20,\cdots,90\}} \{0.75 \times I(X_{ij} = 1) - 0.75 \times I(X_{ij} = 0)\} + \varepsilon_i ,$$

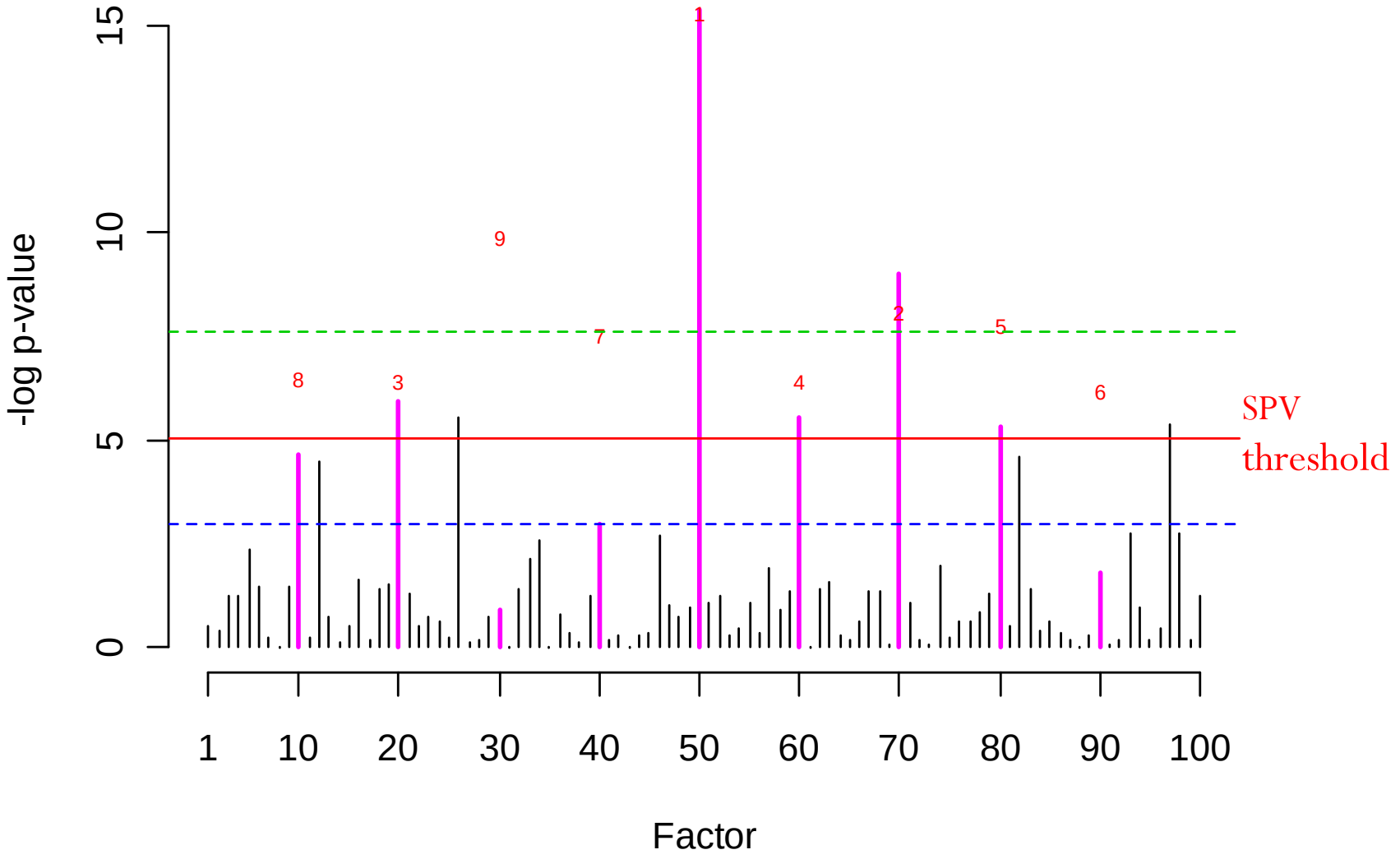
where ε_i is a standard normal distribution.

- The nine factors are supposed to contribute the same mean effect.

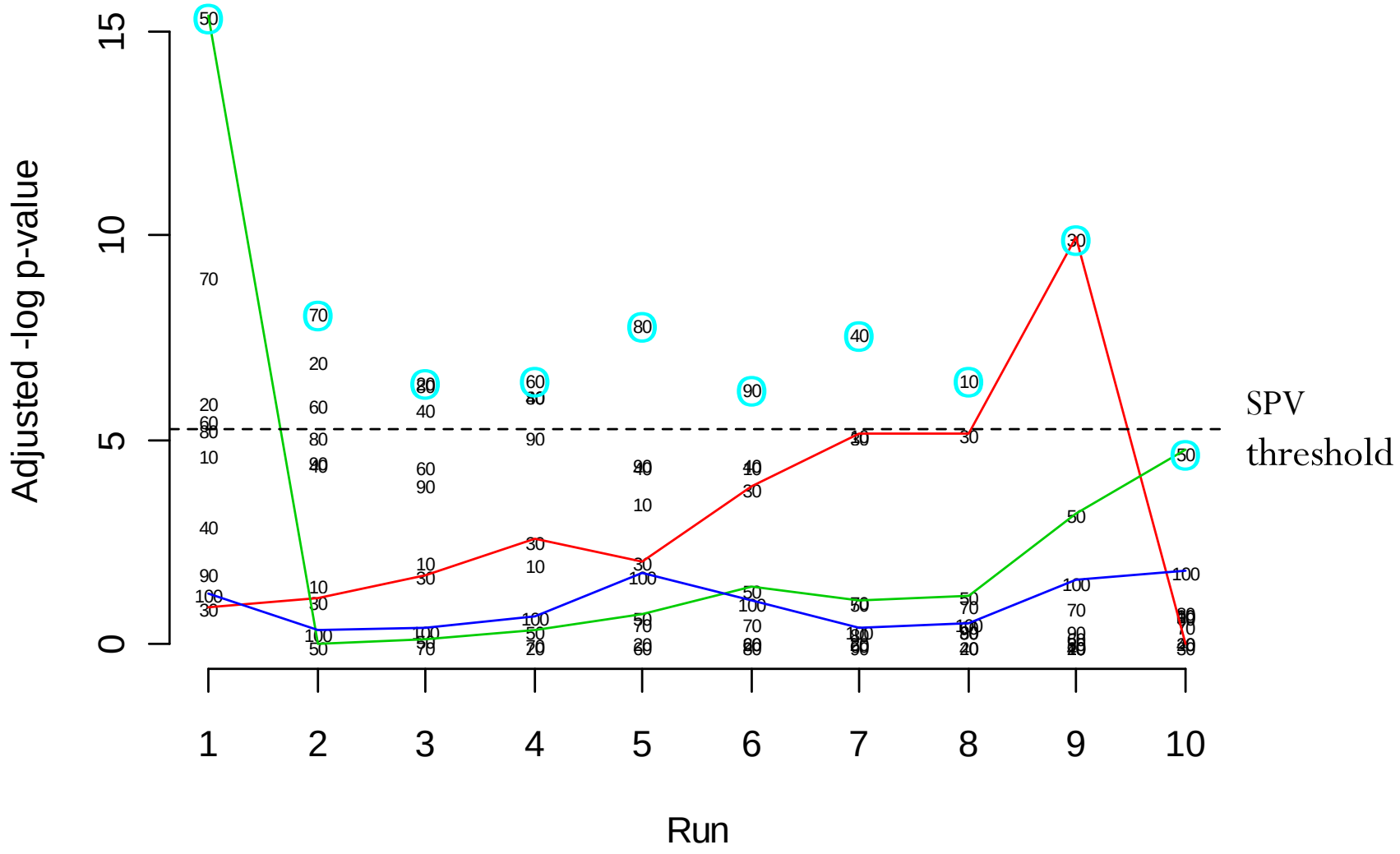
The -log p-values for the 100 single-factor ANOVA models



The adjusted –log p-values obtained from the first 9 runs of SPV algorithm



The adjusted –log p-values for the 9 causal factors and an unrelated factor X_{100} in the first 10 runs of SPV algorithm



Why the SPV algorithm works

Suppose that only two single factors, say X_j and X_k, have effects on the responses

$$y_i = \mu + \sum_{a=1}^{g_j} \alpha_a I(X_{ij} = a) + \sum_{b=1}^{g_k} \beta_b I(X_{ik} = b) + \varepsilon_i,$$

where μ is a constant, α_a is the effect of the X_j variable with constraint $\sum \alpha_a = 0$, β_b is the effect of the X_k variable with constraint $\sum \beta_b = 0$, and $\varepsilon_i \sim N(0, \sigma^2)$ is the random error component.

Sum of squares decomposition

- Let the average of all responses under the *c*th level of variable X_l be denoted by $\bar{y}_{l_c} = \sum_{i=1}^{n} \frac{y_i I(X_{il} = c)}{n_{l_c}}$, where $n_{l_c} = \sum_{i=1}^{n} I(X_{il} = c)$ is the size of the corresponding category.
- The total sum of squares may be written as

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [\sum_{a=1}^{g_j} \bar{y}_{j_a} I(X_{ij} = a) - \bar{y}]^2 + \sum_{i=1}^{n} [\sum_{b=1}^{g_k} \bar{y}_{k_b} I(X_{ik} = b) - \bar{y}]^2$$
$$+ \sum_{i=1}^{n} [y_i - \sum_{a=1}^{g_j} \bar{y}_{j_a} I(X_{ij} = a) - \sum_{b=1}^{g_k} \bar{y}_{k_b} I(X_{ik} = b) + \bar{y}]^2.$$

- This is usually written as $SS_T = SS_{X_j} + SS_{X_k} + SS_E$.

The *F* statistic

- In the first run of searching for the most influential factor among *m* factors, we fit a single-term ANOVA model $y_i = \mu + \sum_{c=1}^{g_l} \lambda_c I(X_{il} = c) + \varepsilon_i$ to the response variable and factor X_l.
- The test statistic is

$$\mathrm{F}^*(X_l) = \frac{SS_{X_l}/(g_l - 1)}{(SS_T - SS_{X_l})/(n - g_l + 1)} = \frac{SS_{X_l}/(g_l - 1)}{SS_{E|X_l}/(n - g_l + 1)} = \frac{MSS_{X_l}}{MSS_{E|X_l}}.$$

- If X_l is not one of the two causal factors, the ratio is

$$\frac{E\{MSS_{X_l}\}}{E\{MSS_{E|X_l}\}} = \frac{\sigma^2 + \frac{\sum_{c=1}^{g_l} n_{l_c} \lambda_c^2}{g_l - 1}}{\sigma^2 + \frac{\sum_{a=1}^{g_j} n_{j_a} \alpha_a^2}{g_j - 1} + \frac{\sum_{b=1}^{g_k} n_{k_b} \beta_b^2}{g_k - 1}} < 1 < \text{(threshold* >1)}$$

Hardly identify both two factors

- When X_j or X_k is considered in the single-term ANOVA model,

$$\frac{E\{MSS_{X_j}\}}{E\{MSS_{E|X_j}\}} = \frac{\sigma^2 + \frac{\sum_{a=1}^{g_j} n_{j_a} \alpha_a^2}{g_j - 1}}{\sigma^2 + \frac{\sum_{b=1}^{g_k} n_{k_b} \beta_b^2}{g_k - 1}} \quad \text{and} \quad \frac{E\{MSS_{X_k}\}}{E\{MSS_{E|X_k}\}} = \frac{\sigma^2 + \frac{\sum_{b=1}^{g_k} n_{k_b} \beta_b^2}{g_k - 1}}{\sigma^2 + \frac{\sum_{a=1}^{g_j} n_{j_a} \alpha_a^2}{g_j - 1}}$$

- Whether the p-values of the F^* statistics for factor X_j and X_k are significant depend not only on mean effect size of each factor but mean effect size of the other factor.

Increased chance for the other factor

- If X_j is identified, the total sum of squares is pared down to

$$\sum_{i=1}^{n}[y_i^{(1)}]^2 = \sum_{i=1}^{n}[\sum_{b=1}^{g_k}\bar{y}_{k_b} I(X_{ik}=b) - \bar{y}]^2 + \sum_{i=1}^{n}[y_i^{(1)} - \sum_{b=1}^{g_k}\bar{y}_{k_b} I(X_{ik}=b) + \bar{y}]^2.$$

In the second run, we expect that variable X_k has an increased chance of being identified as an influential factor because the updated ratio is larger than the previous one, that is,

$$\frac{E\{MSS_{X_k}\}}{E\{MSS_{E|X_k}\}} = \frac{\sigma^2 + \frac{\sum_{b=1}^{g_k} n_{k_b}\beta_b^2}{g_k - 1}}{\sigma^2} > \frac{\sigma^2 + \frac{\sum_{b=1}^{g_k} n_{k_b}\beta_b^2}{g_k - 1}}{\sigma^2 + \frac{\sum_{a=1}^{g_j} n_{j_a}\alpha_a^2}{g_j - 1}}$$

- Therefore, we expect a larger –log p-value for factor X_k now.

Little chance for noisy factors

- If factor X_k is successfully identified, the newly refined responses are formed by removing the factor effects.
- Since the ratio between the mean sum of squares for any factor and error variance is close to one, there is little chance of any single variable to be found from fitting a one-single factor ANOVA model on the refined responses.

Simulation studies

- The aim of these simulation studies is to evaluate the power of the SPV algorithm and the false discovery rate compared to competing methods available in R packages.
- The simulation schemes are slightly modified from those in the related literature so that the direct comparison of the simulation results is relatively fair.

Example 1: Compared with the variable selection approach of group lasso (Yuan and Lin, JRSSB, 2006)

- Generated 120-dimensional multivariate normal variables $Z_1,\ldots,Z_{120}$ with mean vector **0** and covariance matrix of element $\Sigma_{ij}=0.5^{|i-j|}$.
- Each Z_j was transformed to three levels 0, 1 or 2 if it was smaller than $\Phi^{-1}(1/3)$, larger than $\Phi^{-1}(2/3)$, or in between, respectively.
- *A* was a set of 20 numbers randomly selected from $1,2,\ldots,120.$

$$Y=\sum_{j\in A}\{\alpha_j I(Z_j=0)+\beta_j I(Z_j=1)\}+\varepsilon\,,$$

where α_j and β_j were generated uniformly from $[-1.25,-0.75]\cup[0.75,1.25]$, the error was a standard normal.

The average number of correctly and falsely identified factors based on 200 simulated datasets from the model of 20 causal factors

n	Correct model (%)			Number of correct factors			Number of false factors		
	$Glasso_{0.5}$	$Glasso_{0.25}$	SPV_0	$Glasso_{0.5}$	$Glasso_{0.25}$	SPV_0	$Glasso_{0.5}$	$Glasso_{0.25}$	SPV_0
200	0	0	17	11.52	18.41	18.32	2.15	8.70	0.67
300	0	2	55	13.04	19.31	19.93	0.65	4.28	0.53

Example 2: Two- and three-factor interactions using the model of Meier, et al. (JRSSB, 2008)

- Generated 15 correlated factors of three levels denoted by $Z_1, \ldots, Z_{15}$.
- The response *Y* was affected by 12 terms.
- The corresponding binary design matrices are denoted by $A[Z_j] \in \mathbf{R}^{n\times 3}$, $A[Z_1 : Z_j] \in \mathbf{R}^{n\times 9}$ and $A[Z_1 : Z_5 : Z_6] \in \mathbf{R}^{n\times 27}$.
- The simulation model is

$$Y = \sum_{j=1}^{6} A[Z_j]\beta_{1j} + \sum_{j=2}^{6} A[Z_1 : Z_j]\beta_{2j} + A[Z_1 : Z_5 : Z_6]\beta_3 + \varepsilon$$

where the coefficient vectors β_{1j}, β_{2j} and β_3 were from standard normal and adjusted to a zero sum for each coefficient vector.

The average number of correctly and falsely identified factors based on 200 simulated datasets from the model of 12 causal terms

n	Correct model (%)			Number of correct terms			Number of false terms		
	$\text{Glasso}_{0.2}$	$\text{Glasso}_{0.1}$	SPV_0	$\text{Glasso}_{0.2}$	$\text{Glasso}_{0.1}$	SPV_0	$\text{Glasso}_{0.2}$	$\text{Glasso}_{0.1}$	SPV_0
200	0	0	4	7.38	8.11	9.85	3.31	15.59	1.33
400	0	0	16	7.41	8.42	11.25	1.22	7.46	1.28

Example 3: Compares the SPV algorithm to the logic regression method available in the R package "logicFS"

- Used the second simulation of Schwender and Ickstadt (Biostatistics, 2008) but replaced the binary outcome with a continuous trait.
- Data of 1,000 observations and 50 SNPs were generated, where each SNP exhibited a minor allele frequency of 0.25.
- The *Y* was generated from a normal distribution with variance one and mean equal to $\sum_{j=1}^{5} \beta L_j$ where the five logic expressions were

  $L_1 = S_1 \wedge S_2^C$, $L_2 = S_3 \wedge S_4^C$, $L_3 = S_5^C \wedge S_6^C$, $L_4 = S_7^C \wedge S_8^C \wedge S_9^C$ and $L_5 = S_{10}^C \wedge S_{11}^C \wedge S_{12}$.

- S_j = "SNP *j* is not of the homozygous reference genotype".

The average numbers of correctly and falsely identified terms based on 100 simulated datasets from the model of 5 causal terms

β	Number of correct terms			Number of false terms		
	logicFS	SPV_0	SPV_2	logicFS	SPV_0	SPV_2
1	1.97	3.90	3.65	3.01	1.07	0.23
2.5	2.46	4.90	4.70	2.49	1.17	0.04

Example 4: For comparison with the partition retention method (Chernoff et al., AOAS, 2009)

- 1,000 binary factor variables, denoted by $X_1, \ldots, X_{1000}$.
- The response $Y \sim N(\mu, \sigma^2)$, $\mu = \max(\mu_1, \mu_2) + 0.1(\mu_1 + \mu_2)$ and $\sigma = \max(\sigma_1, \sigma_2)$

  with $\mu_1 = 4X_1X_2X_3$, $\mu_2 = 6X_4X_5X_6X_7$, $\sigma_1 = 1 + X_1X_2X_3$ and $\sigma_2 = 1 + 2X_4X_5X_6X_7$.

- $X_j \sim Bin(1, p_j)$ p_j = .4, .5, .6, .35, .45, .55 and .65 for the first 7 variables.
- The probabilities for the remaining 993 binary variables were randomly uniformly selected in the range of 0.4 to 0.6.
- Sample size n = 400

Simulation result

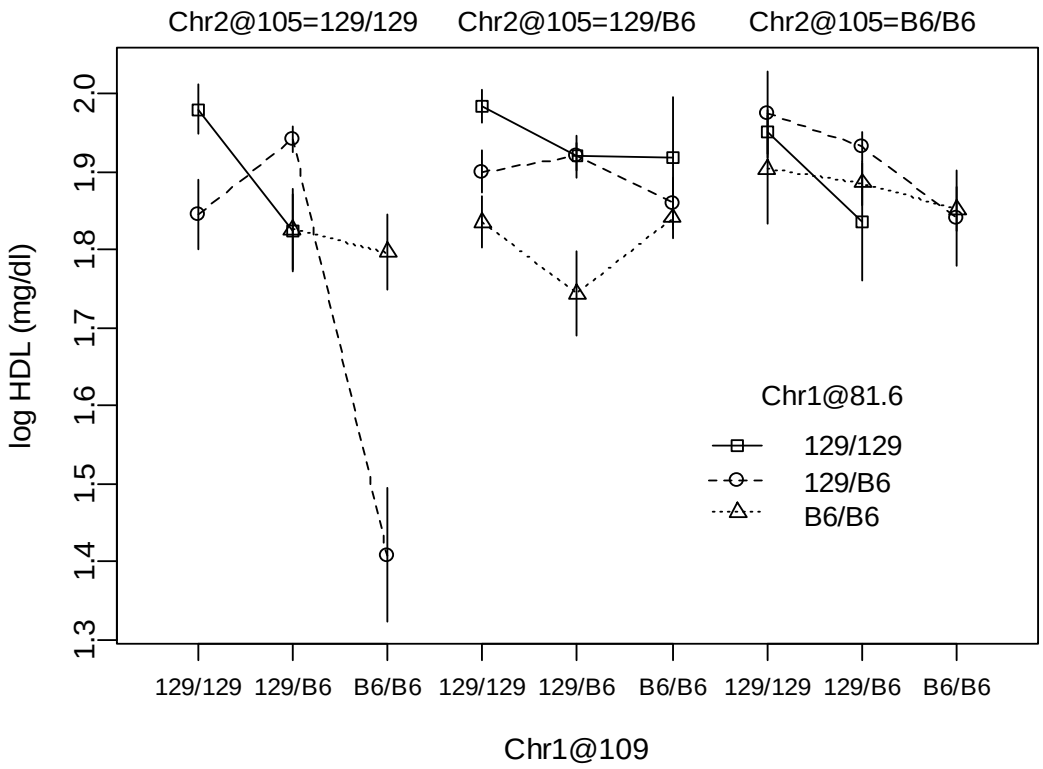
- The average number of falsely identified variables was 1.8 out of 993 using SPV algorithm.
- Revealed all the 7 causal variables:
  1. The SPV algorithm: 87 of 100 data sets
  2. The partition retention method: 1 of 5 data sets

Real Example: QTL study of high density lipoprotein cholesterol (Ishimori et al, 2004)

- One of the objectives was to identify loci controlling the plasma HDL levels.
- C57BL/6J (B6) and 129S1/SvImJ (129) mice were mated to produce the (B6 129) F_1 progeny, which were intercrossed to produce 294 female F_2 progeny.
- Female B6 mice have low plasma HDL levels and are susceptible to atherosclerosis
- The plasma HDL concentrations and genotypes of 113 markers of the 294 F_2 progeny are available from the QTL archive at the website of Jackson Laboratory

Location	(marker name)	df	SS	%Var	p-value	Ishimori, et al.
Chr1@101.2	(D1MIT406)	2	0.156	2.656	0.0017	
Chr12@22	(D12MIT172)	2	0.172	2.929	0.0009	Chr12@20
Chr9@26	(D9MIT129)	2	0.164	2.792	0.0013	Chr9@24
Chr8@43	(D8MIT248)	2	0.132	2.240	0.0046	Chr8@44
Chr1@81.6	(D1MIT159)	18	0.748	12.731	0.0000	Chr1@80
Chr1@109	(D1MIT210)	18	1.031	17.543	0.0000	Chr1@104
Chr2@105	(D2MIT148)	18	0.821	13.960	0.0000	Chr2@90
Chr1@81.6:Chr1@109		12	0.622	10.586	0.0000	Chr1@80:Chr1@104
Chr1@81.6:Chr2@105		12	0.450	7.649	0.0004	
Chr1@109:Chr2@105		12	0.678	11.539	0.0000	Chr1@104:Chr2@90
Chr1@81.6:Chr1@109:Chr2@105		8	0.224	3.814	0.0191	
Total		291	5.879	47.676		

One category of five mice with log HDL levels, 1.62, 1.53, 1.48, 1.2 and 1.2 which fall below the 5th percentile of the 292 observations.



Since low HDL levels relate to the occurrence of atherosclerosis, the combination of these three markers may be a genetic predictor for the risk.

Conclusion

- The main idea is to stepwise pare down the total variation of responses so that the remaining influential sets of factors have increased chances of being identified
- It outperformed four available methods in the simulation studies
- Computation time is always a concern for methods involving exhaustive screening with permutations for determining thresholds.

Conclusion

- The main objective of the SPV algorithm is to identify multiple-factor interactions, rather than to measure their effects.
- Once the influential terms are detected in the first phase using the SPV algorithm, it should be relatively easy to find an appropriate model for a much smaller number of terms.
- For binary outcome or ordinal response variables, the SPV algorithm may not be an appropriate choice.

Thank you!