Bayesian Mixture Modeling of Gene-Environment and Gene-Gene Interactions: Supplementary Material

Jon Wakefield[1,2], Frank De Vocht[1,3] and Rayjean J. Hung[1,4]

[1]International Agency for Research on Cancer, Lyon, France [2]Departments of Statistics and Biostatistics, University of Washington, Seattle, Washington, USA. [3]Occupational and Environmental Health Research Group, School of Translational Medicine, Faculty of Medical and Human Sciences, The University of Manchester. [4]Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada

Difficulties with the Full and Single Model Strategies

Suppose we wish to test m interactions at level α and let $T_i = 0/1$ represent a random variable representing non-reject/reject for interaction i, $i = 1, ..., m$.

We first consider inference using the full model that contains all SNPs, exposures and interactions. For concreteness suppose that we test via the (square root) Wald statistic $Z_i = \widehat{\delta}_i / \widehat{\text{s.e.}}(\widehat{\delta}_i)$ and reject if $|Z_i| > z_\alpha$ where $z_\alpha = \Phi^{-1}(1-\alpha/2)$, so that we have exploited the asymptotic normal distribution of the MLE: $\widehat{\boldsymbol{\delta}} \sim N(\boldsymbol{\delta}, \boldsymbol{V})$. This model provides unbiased estimation of the relevant interactions, but at the expense of an increased variance, which occurs because of the number of parameters that are simultaneously estimated, as we demonstrate shortly. For each test we have the correct (asymptotic) coverage. If the tests were independent (which corresponds to V being diagonal) then the total number of rejections, $T = \sum_{i=1}^{m} T_i$, is Binomial(m, α). With correlated tests the situation is more complex since although $E[T] = m\alpha$, the variance takes the form

$$\text{var}(T) = m\alpha(1-\alpha) + 2\sum_{i<j} \text{cov}(T_i, T_j) \tag{1}$$

where

$$\text{cov}(T_i, T_j) = \Pr(T_i = T_j = 1) - \alpha^2 \tag{2}$$

and

$$\Pr(T_i = T_j = 1) = \Pr(Z_i > z_\alpha, Z_j > z_\alpha). \tag{3}$$

Therefore the variance of the total number of rejections is inflated when the tests are postively correlated (in the limit all null hypotheses will be either not rejected or rejected). This dependence means that setting an α level to control a meaningful criteria is difficult. For example, the family-wise error rate (FWER), which controls the probability of one or more false discovery, is a popular criteria. The Bonferroni correction takes as threshold $\alpha = \alpha^\star/m$, with $\alpha^\star$ the required FWER, will control FWER but will be highly conservative in the case of dependent tests. Similarly, procedures that control

the FDR (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003) were initially developed in the contex of independent tests, and cannot simply be adjusted to the dependent test situation.

Consider the generic logistic regression model

$$\text{logit } \mu_i = \alpha + \sum_{p=1}^{P} x_{ip}\theta_p$$

for $i = 1, ..., n$ individuals for which we have a Bernoulli outcome. The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by the inverse of the information, which is given by:

$$\begin{bmatrix} \sum_{i=1}^n \mu_i & \sum_{i=1}^n x_{i1}\mu_i & ... & \sum_{i=1}^n x_{iP}\mu_i \\ \sum_{i=1}^n x_{i1}\mu_i & \sum_{i=1}^n x_{i1}^2\mu_i & ... & \sum_{i=1}^n x_{i1}x_{iP}\mu_i \\ ... & ... & ... & ... \\ \sum_{i=1}^n x_{iP}\mu_i & \sum_{i=1}^n x_{i1}x_{iP}\mu_i & ... & \sum_{i=1}^n x_{iP}^2\mu_i \end{bmatrix}^{-1}$$

The off-diagonal elements of the lower $p \times p$ portion of this matrix behave as weighted covariance terms. Hence if we have high correlation between the elements x_{ij} and x_{ik}, $j, k = 1, ..., P$, $j \neq k$, then we will obtain a highly unstable (and close to singular) information matrix and, under inversion, we will obtain a variance-covariance matrix that contains many large elements, and correlations close to unity. Hence the full model approach can be highly unstable, and give large standard errors, if we have correlated exposures.

In the single model approach we only have a single interaction so that correlated exposures is not an issue in terms of the fitting. However, there will be bias in the estimation of the quantity we would like to estimate. Further, the number of rejections, T, is still not binomial since the test outcomes are not independent when the exposures are correlated. Equations (1)–(3) still hold but determining the level of the dependence requires a little more detail. Specifically, consider the two logistic models for a generic individual:

$$\begin{aligned} \text{logit } \mu &= \alpha + \beta_1 x + \gamma_1 z_j + \delta_j x \times z_j \\ \text{logit } \mu &= \alpha + \beta_1 x + \gamma_1 z_k + \delta_k x \times z_k \end{aligned}$$

for SNP genotypes x, and exposures z_j and z_k. Fitting these models yield Wald statistics

$$\begin{aligned} Z_j &= \frac{\hat{\delta}_j}{\widehat{\text{s.e.}}(\hat{\delta}_j)} \\ Z_k &= \frac{\hat{\delta}_k}{\widehat{\text{s.e.}}(\hat{\delta}_k)} \end{aligned}$$

and both the numerator and denominator of these statistics will be positively correlated if z_j and z_k are positively correlated. Hence again the distribution of the number of

rejections depends on the level of dependence. Since we are not adjusting for additional exposures there will be bias in using the single model when the exposures are correlated. This bias arises because the regression coefficient for the interaction is picking up not only the effect of the exposure included in the model, but all other exposure-SNP effects that are correlated with the included exposure.

For the mixture model there will be some bias since the prior is centered at zero, but the variance is vastly reduced due to the reduced number of parameters in the model (since many of the parameters will be effectively set to zero).

Simulation Results

In the first situation all main effects and interactions are null and we see wide interval estimates under the full model for the interactions; the single model intervals are slightly narrower, but there are spurious false positives under both the full (3) and single (2) model approaches. The mixture estimates are wider when spurious results occur in the other approaches, but all intervals correctly contain 0.

The second simulation is also null with two correlated exposures (in all cases the correlation was approximately 0.75). The full model again gives the widest intervals, and two false positives. The single and Bayesian models do not produce false positives for this simulation. The single model interval widths are between the full model and mixture estimates.

In the third simulation there are two non-null SNPs, 2 null exposures and 40 null interactions. The SNP effects are weak here (the relative risk corresponding to a single copy of the minor allele is 1.2); the mixture model misses the second SNP, while the full model interval includes the true value and zero. The main exposure effects are very wide under the full model and there are 2 false positives. The single and mixture models perform better, with all intervals containing zero, and the mixture models have very narrow intervals.

In simulation 4 there is one non-null SNP and a single non-null exposure. The full and single models erroneously flag interaction 12 as significant. The mixture model misses the non-null SNP, while the full model does not detect the non-null exposure.

In simulation 5 the exposure effects (which correspond to correlated exposures) are poorly estimated under the full model. Three interactions are incorrectly flagged as significant under the full model, while for the mixture and single models all intervals correctly contain zero.

In simulation 6 we have 20 SNPs with one non-null and a single exposure which is non-null, with one interaction between the non-null SNP and exposure. Here all of the methods perform satisfactorily, though the interval estimates are shorter under the mixture model, which is reflected in the greatly reduced MSEs, when compared to the single and full models.

In simulation 7 we have 2 out of 20 non-null SNPs, 2 (uncorrelated) exposures with 1 non-null, and 1 non-null interaction (from 40). The true interaction was identified by all three methods, with the shortest interval for the mixture model. Interval estimates for the interactions from the full and single models gave, respectively, two and one type II errors.

Simulation 8 had the same set up as simulation 7, but with correlated exposures. Here,

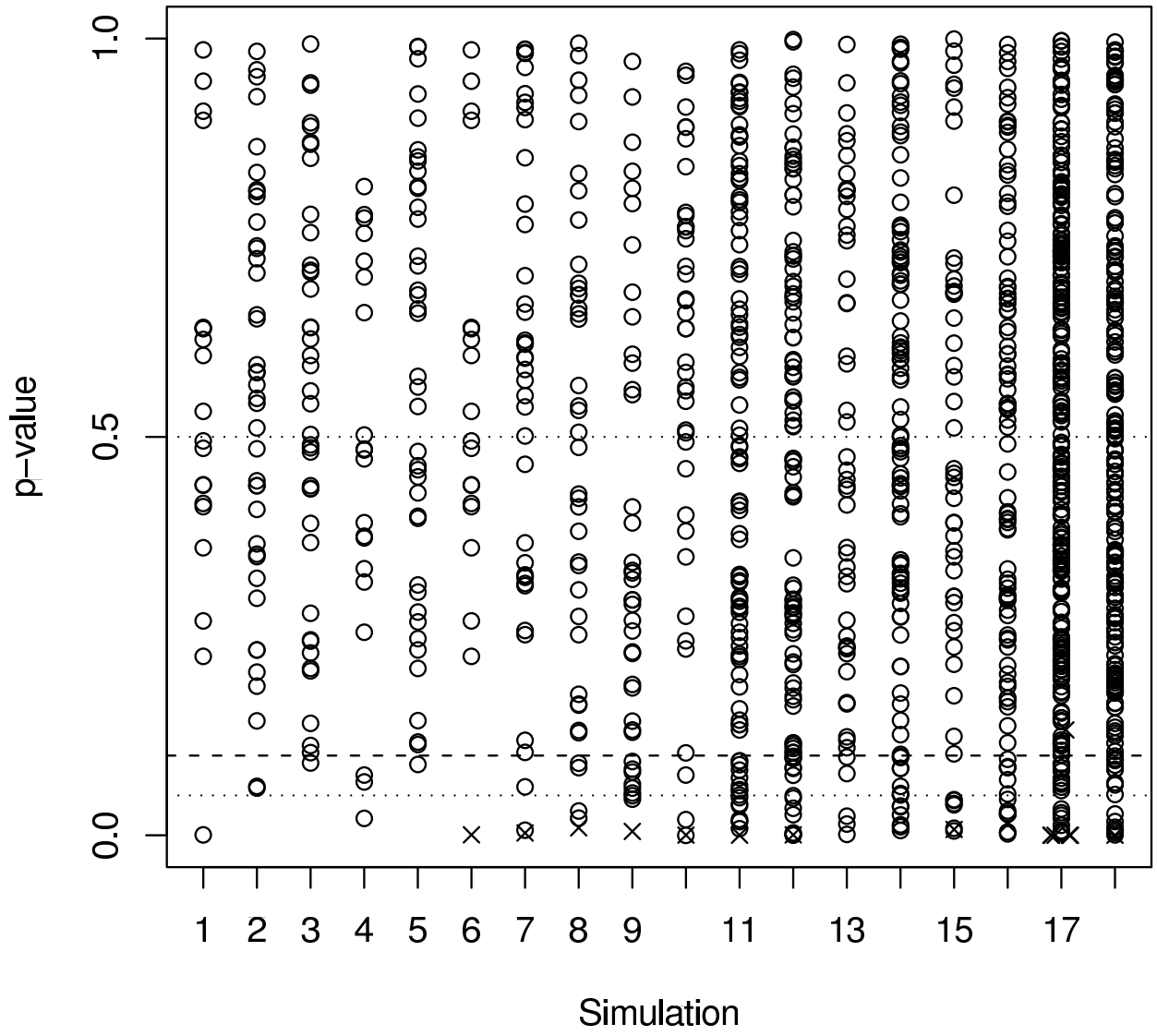

Figure 1: p-values from single model analyses for interactions for all simulations. Reference line at $p = 0.5$ and thresholds of 0.10 and 0.05 also drawn. Circles denote null effects, and crosses non-null effects.

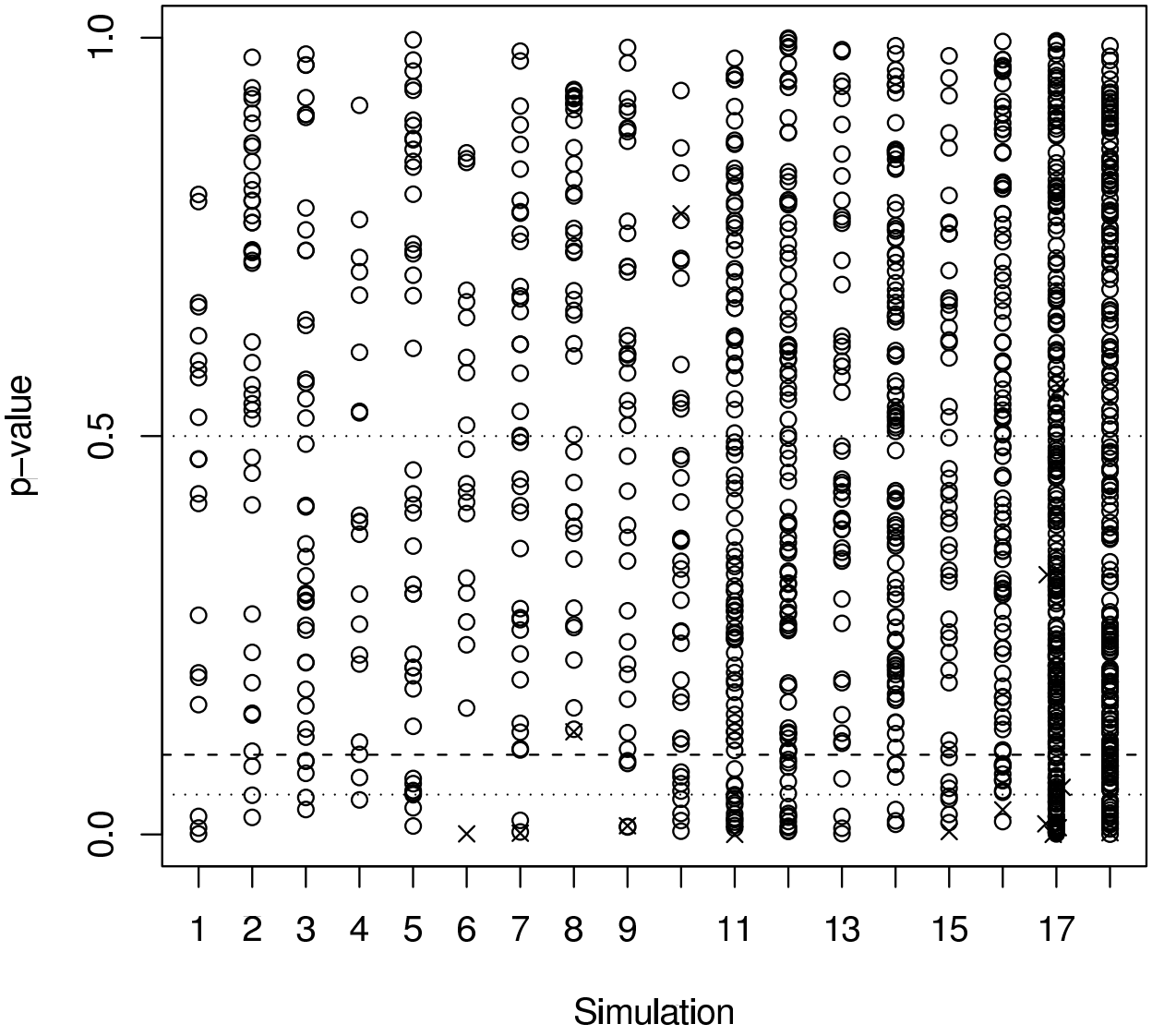

Figure 2: p-values from full model analyses for interactions for all simulations. Reference line at $p = 0.5$ and thresholds of 0.10 and 0.05 also drawn. Circles denote null effects, and crosses non-null effects.

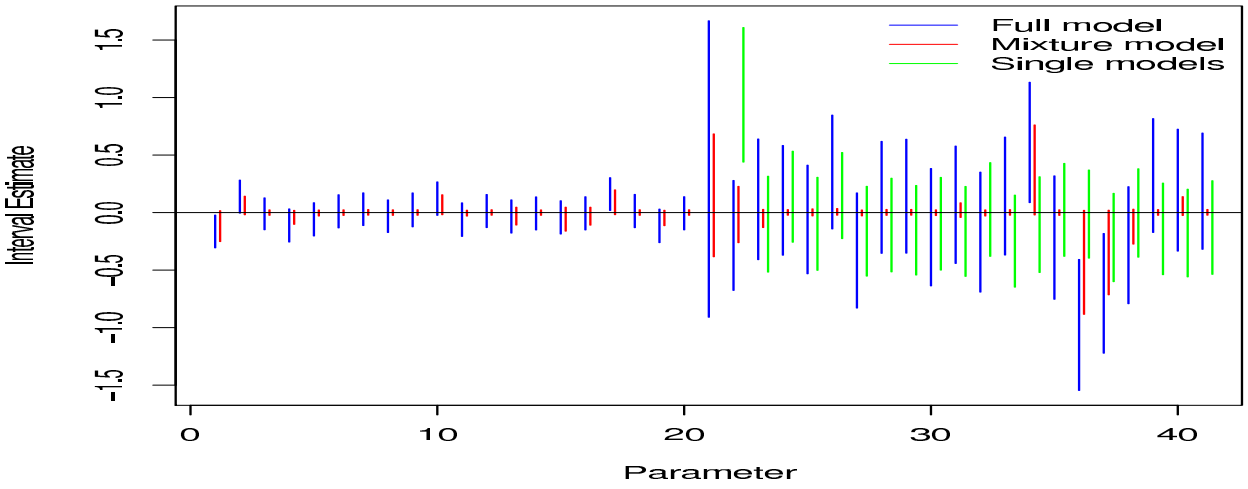

Figure 3: Simulation 1. 20/0 null/non-null SNPs, 1/0 null/non-null exposures, 20/0 null/non-null interactions.

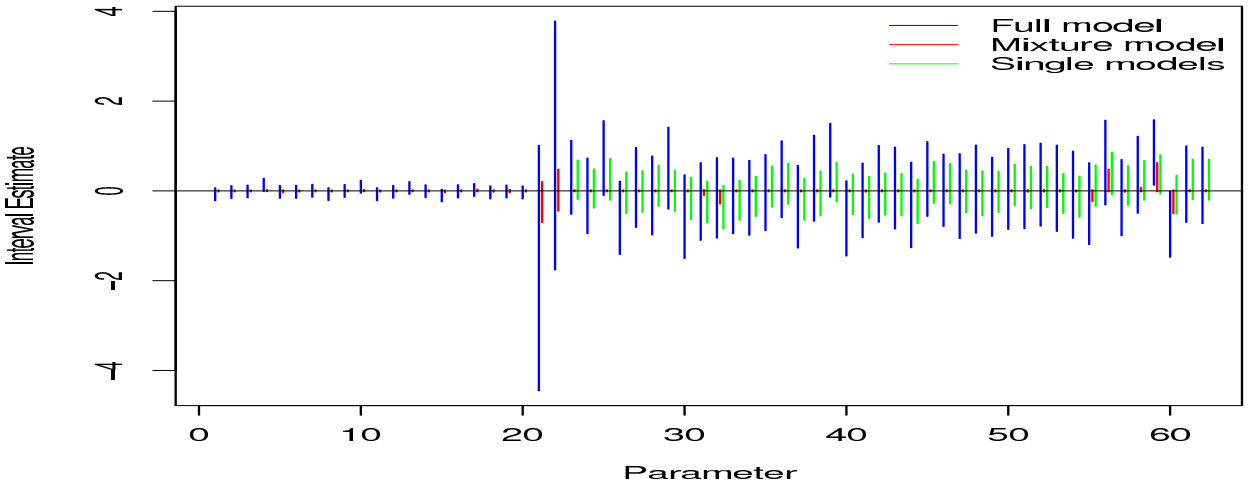

Figure 4: Simulation 2. 20/0* null/non-null SNPs, 2/0 null/non-null exposures, 40/0 null/non-null interactions.

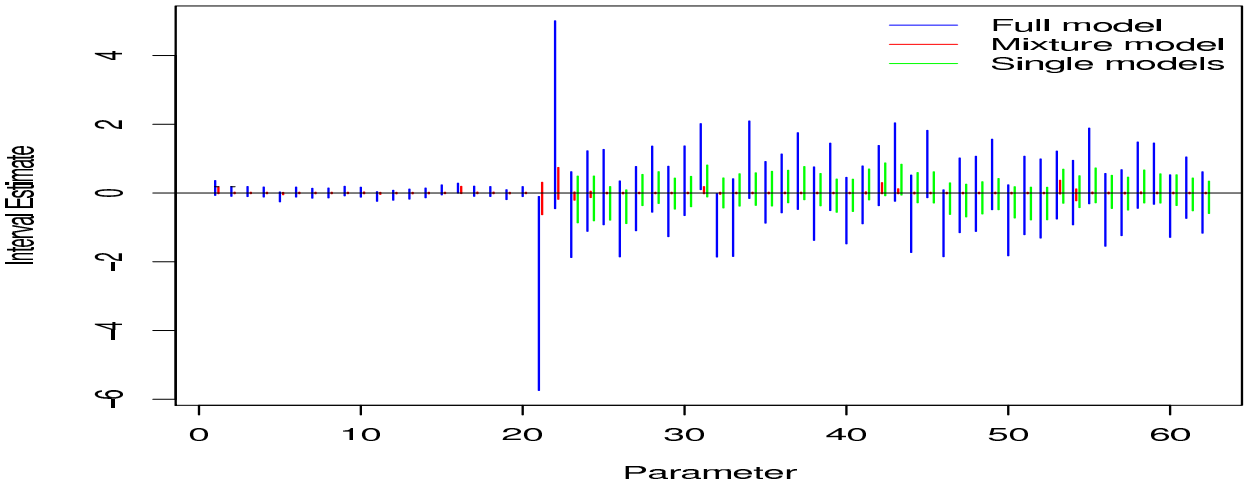

Figure 5: Simulation 3. 20/2 null/non-null SNPs, 2/0* null/non-null exposures, 40/0 null/non-null interactions.

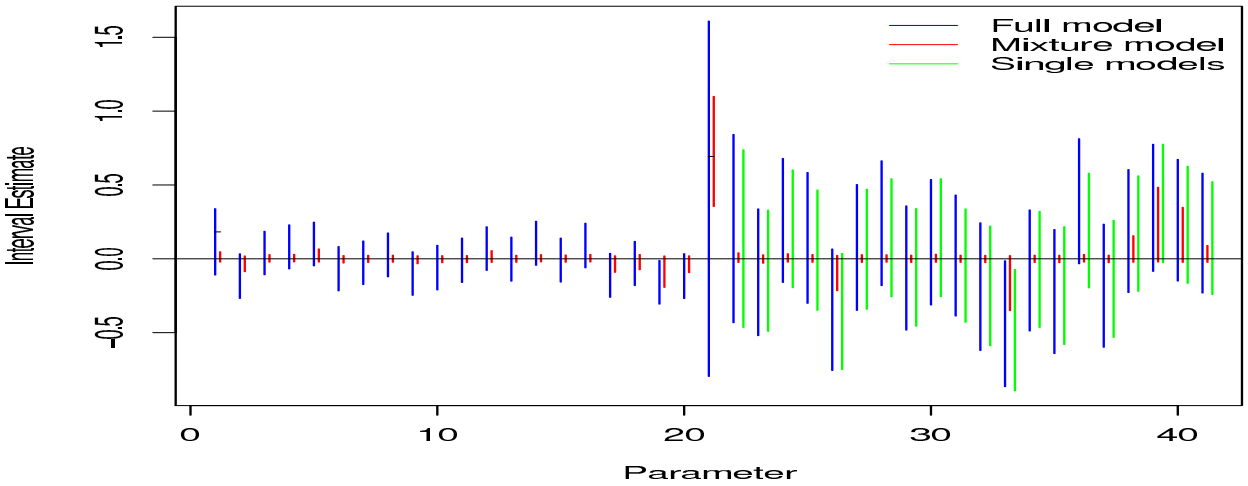

Figure 6: Simulation 4. 20/1 null/non-null SNPs, 1/1 null/non-null exposures, 20/0 null/non-null interactions.

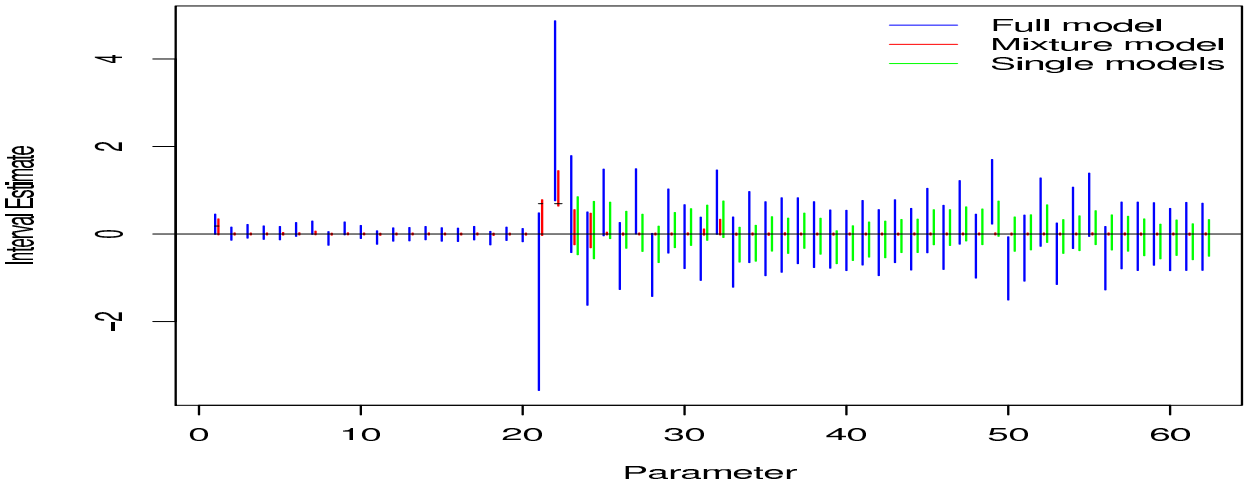

Figure 7: Simulation 5. 20/1 null/non-null SNPs, 2/2* null/non-null exposures, 40/0 null/non-null interactions.

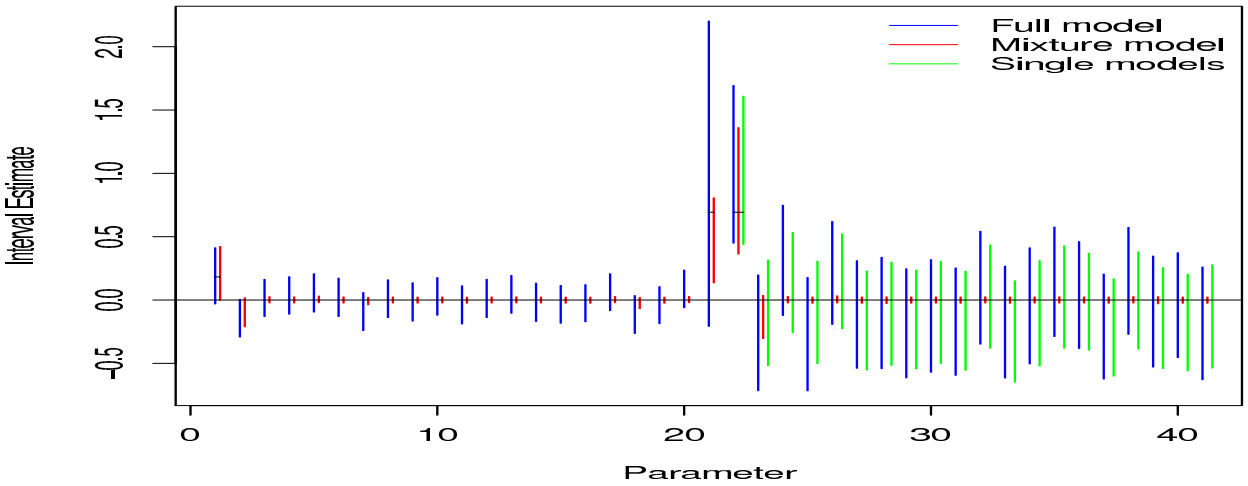

Figure 8: Simulation 6. 20/1 null/non-null SNPs, 1/1 null/non-null exposures, 20/1 null/non-null interactions.

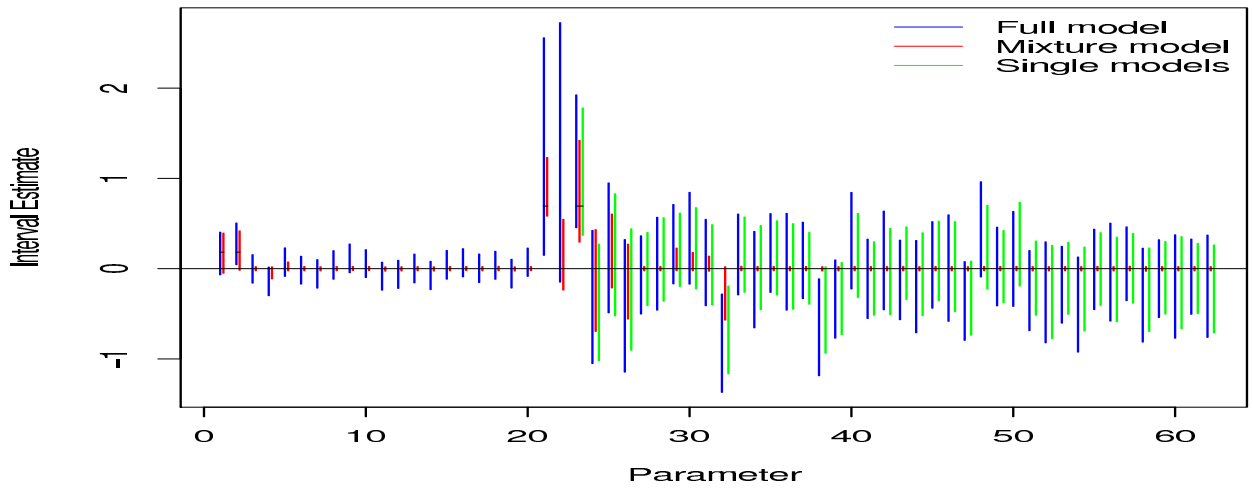

Figure 9: Simulation 7. 20/2 null/non-null SNPs, 2/1 null/non-null exposures, 40/1 null/non-null interactions.

the interval estimates for both the full and the mixture models contain the true interaction value but also zero, whereas the single model has an interval that (correctly) does not contain zero. However, the single models also incorrectly flag the interaction between SNP 1 and exposure 2 as significant — this association is induced because of the correlation between the exposures, single models do not control for such confounding. An additional false positive interaction is also present for the single model. A 90% interval for the non-null interaction excludes the value 0.

In simulation 9 we have 2 non-null correlated exposures and a single non-null interaction. All models flag this interaction as significant (via a 95% interval). The full and single models each have 2 type II errors for interactions.

In simulation 10, as in simulation 9, we have two non-null SNPs, and two non-null correlated exposures with a single interaction but we double the case and control sample sizes. This gives a halving of the MSE for the single and full models. For the mixture model the MSE is quite similar, which reflects that most of the parameters are shrunk to the narrow prior range, regardless of sample size. The exposure intervals are very wide under the full model, while the mixture model intervals are shorter but centered at the wrong locations. Each of the methods falsely picks out a second interaction between the second exposure and SNP 1. The single model has a total of 2 false positive interactions and the full model 4, while the mixture model has 1 (the interaction that contains the correct SNP but the other exposure — the second most significant result is the correct non-null interaction).

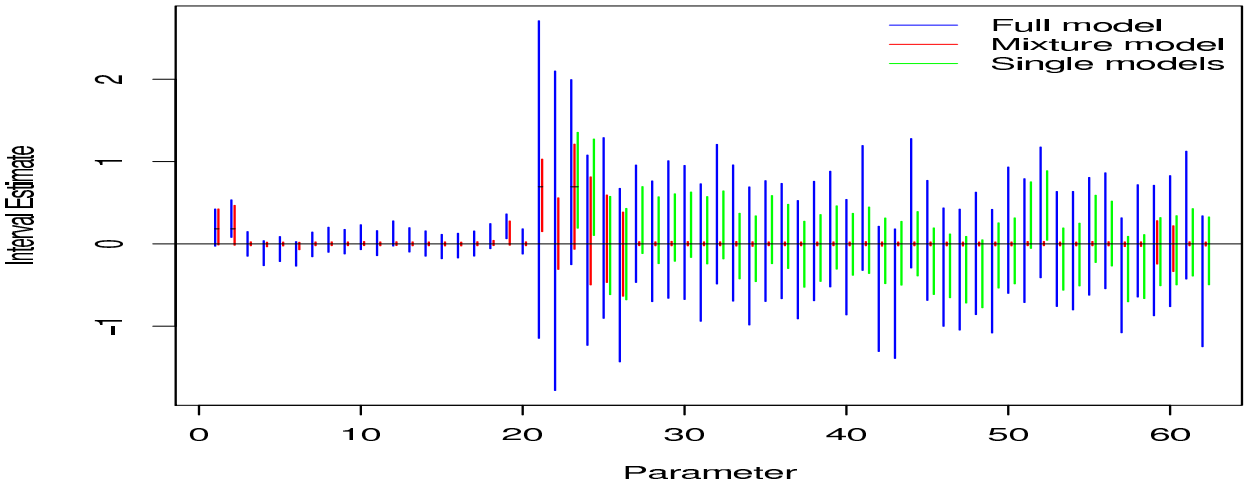

Figure 10: Simulation 8. 20/2 null/non-null SNPs, 2/1* null/non-null exposures, 40/1 null/non-null interactions.

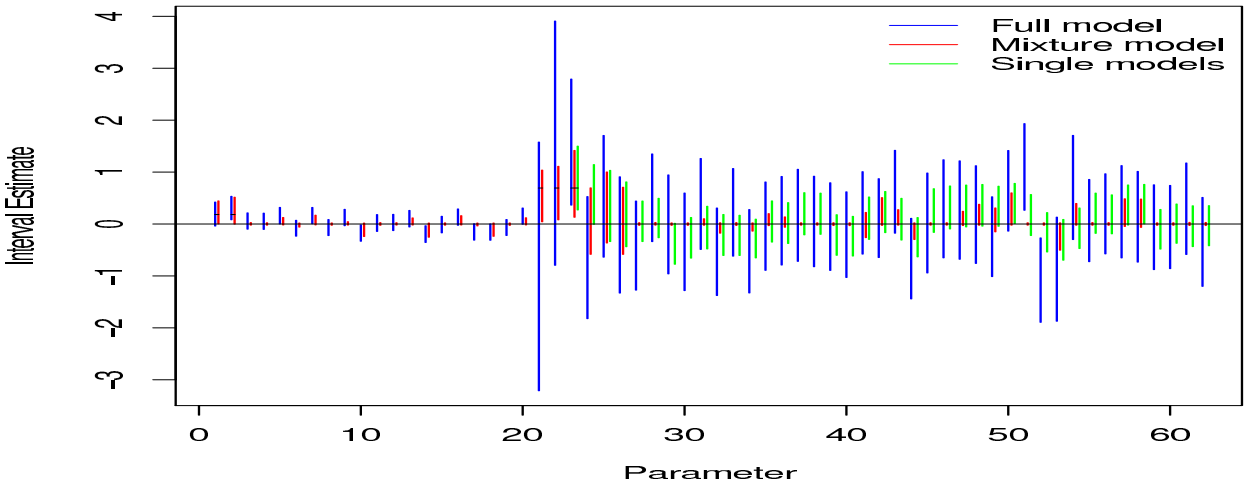

Figure 11: Simulation 9. 20/2 null/non-null SNPs, 2/2* null/non-null exposures, 40/1 null/non-null interactions.

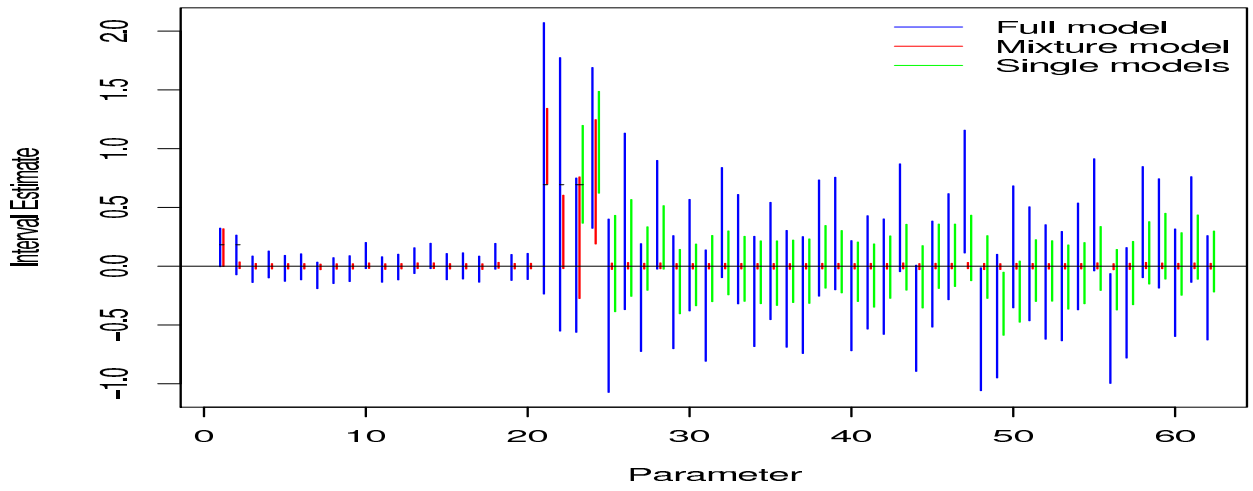

Figure 12: Simulation 10. 20/2 null/non-null SNPs, 2/2* null/non-null exposures, 40/1 null/non-null interactions.

In simulation 11 there is 1 non-null SNP from 20, and 1 non-null exposure from 5 uncorrelated exposures. All models pick out the correct non-null interaction, but again there are a number of false positives in the single (7%) and (particularly) full models (11%), due to the increased number of exposures. An association with SNP 8 is falsely flagged by the full model and under the mixture model we see much uncertainty in the posterior for the interactions associated with this SNP, these are because $\Pr(T_8 = 1|\boldsymbol{y}) = 0.27$, and so the wider mixture prior is considered 27% of the time.

In simulation 12 we have the same set-up as 11, but with correlated exposures. This is reflected in increased MSEs for the full and mixture models. The conclusions are similar to simulation 11. For the single model interactions between SNP 1 and all other exposures are flagged as significant. The mixture model just fails to call the true interaction significant at the 95% level, but does so at the 90% level.

In simulation 13 we increase the number of SNPs to 50, with a single null exposure. The full and single model approaches each have 3 type I errors for interactions, while the mixture model again has 0 false discoveries. The interval estimate associated with the single exposure is exceedingly wide under the full model, reflecting the increased number of parameters to estimate.

In simulation 14 we take the set-up of simulation 13 and increase the number of exposures to 2. The interval estimates for the full model are increased, while for the single models they remain relatively unchanged (since the number of parameters in each model does not increase in the single model framework). There are 6 and 3 falsely discovered interactions

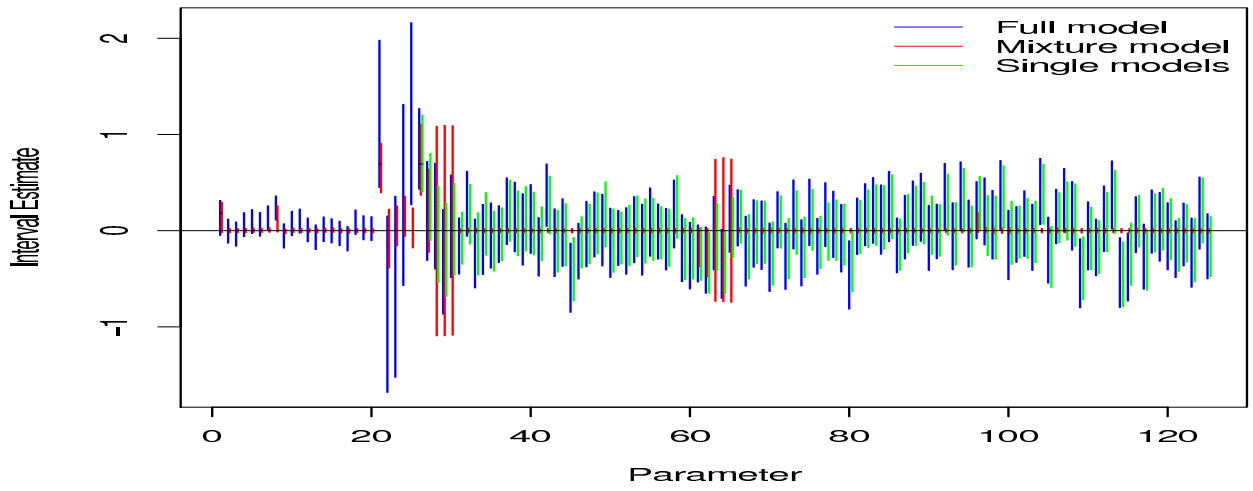

Figure 13: Simulation 11. 20/1 null/non-null SNPs, 5/1 null/non-null exposures, 100/1 null/non-null interactions.

under the single and full models, respectively.

In simulation 15 we gave 50 SNPs with 5 non-null and a single exposure. The type I error rates are high for the single (12%) and full (8%) models, but 0% for the mixture model.

In simulation 16 we have 2 non-null SNPs out of 50 and two correlated exposures. The single model performs badly here with 7 (9%) false discoveries. The intervals are widest under the full model which is reflected in the largest MSE of the three methods. Under the mixture model the single non-null interaction is the most significant, and a 90% interval does not contain zero (though a 95% interval does).

In simulation 17 we have 50 SNPs and 5 correlated exposures, 3 of which are non-null, and the full model preforms very badly with 32 out of 244 truly null interactionss being erroneously flagged as significant. The mixture model has no false discoveries and the single models approach has 9 false discoveries. None of the methods perform particularly well at detecting the true non-null interactions with the single model approach performing best at the 95% level, though it seriously overestimates each of the interactions (since it is not just the effect of the single exposure that is being estimated, but the combined effect of the 5 correlated exposures). Under the mixture model the true non-null interactions all have at least 80% posterior probability of being non-null.

In simulation 18 we increase the number of SNPs to 100 with 2 non-null, and have two, correlated, non-null exposures. There is 1 single non-null interaction (out of 200).

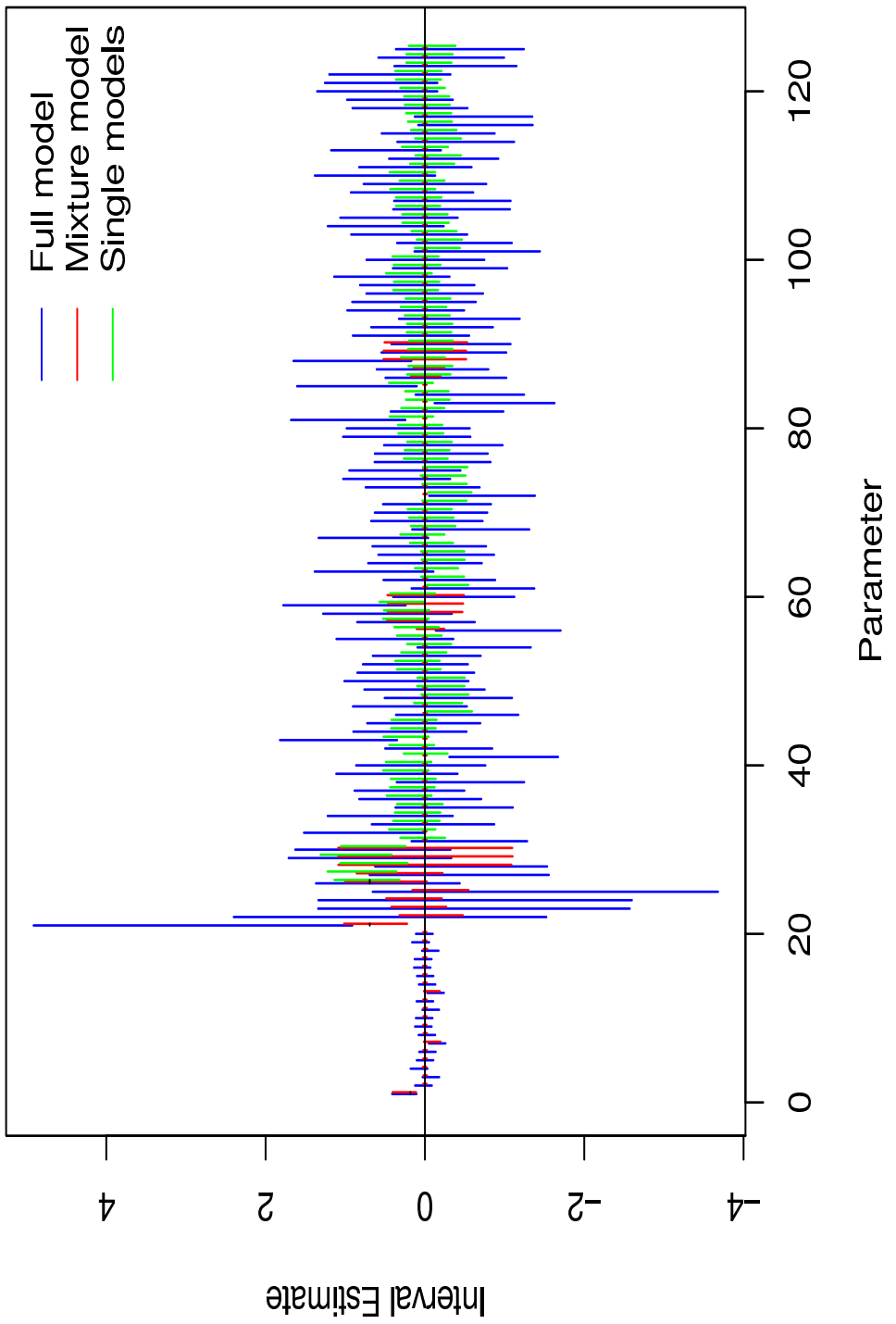

Figure 14: Simulation 12. 20/1 null/non-null SNPs, 5/1* null/non-null exposures, 100/1 null/non-null interactions.

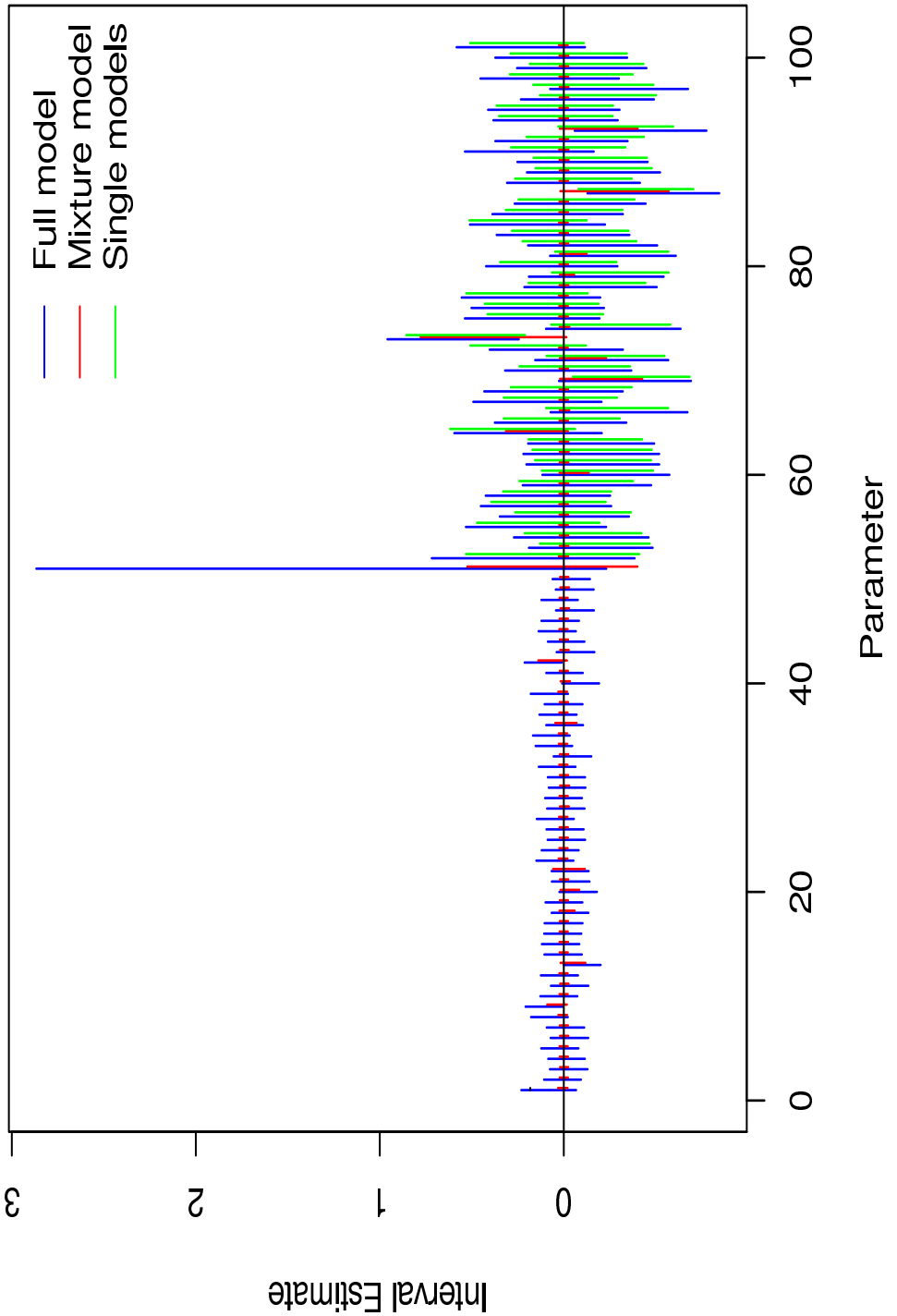

Figure 15: Simulation 13. 50/1 null/non-null SNPs, 1/0 null/non-null exposures, 50/1 null/non-null interactions.

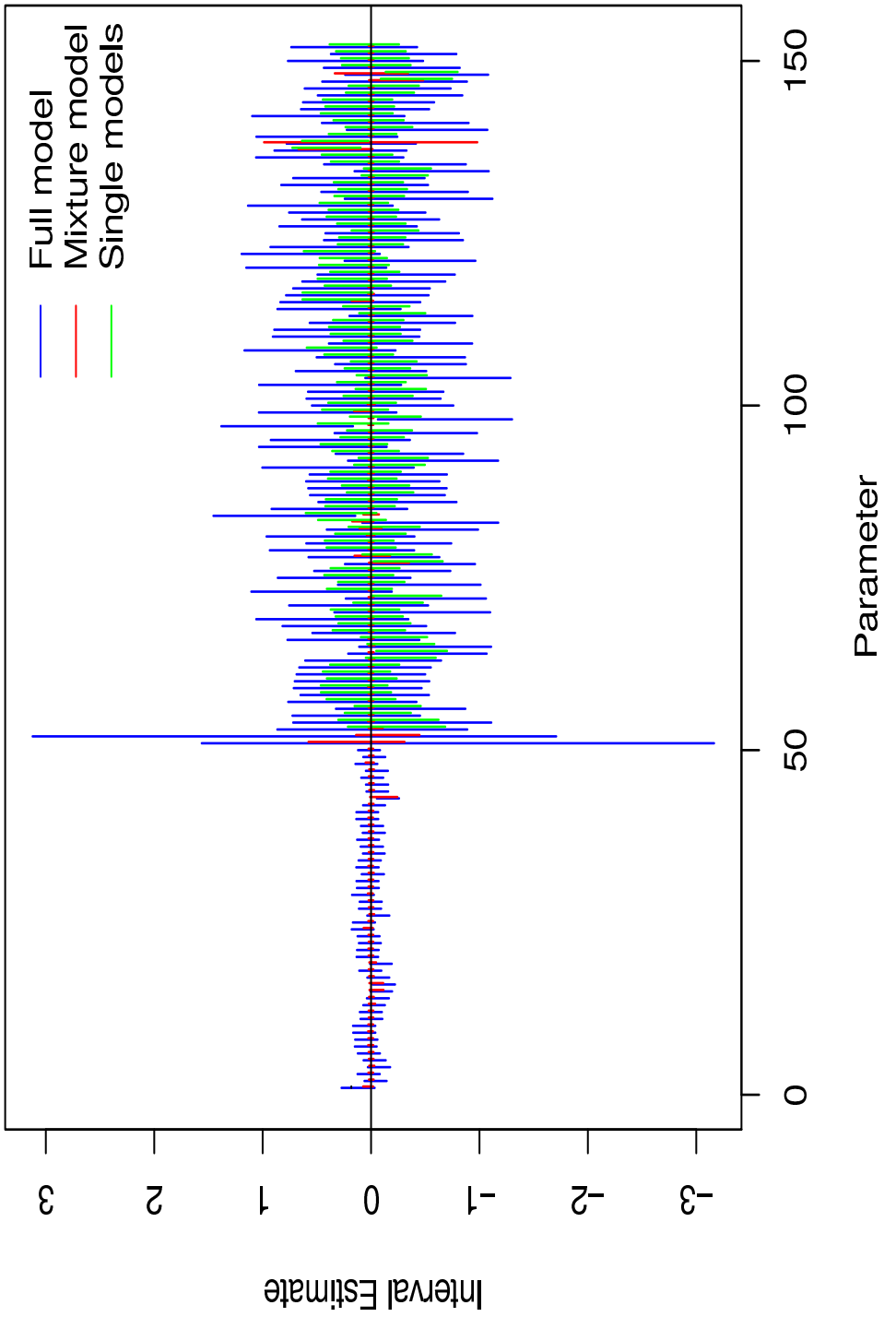

Figure 16: Simulation 14. 50/1 null/non-null SNPs, 2/0* null/non-null exposures, 100/0 null/non-null interactions.

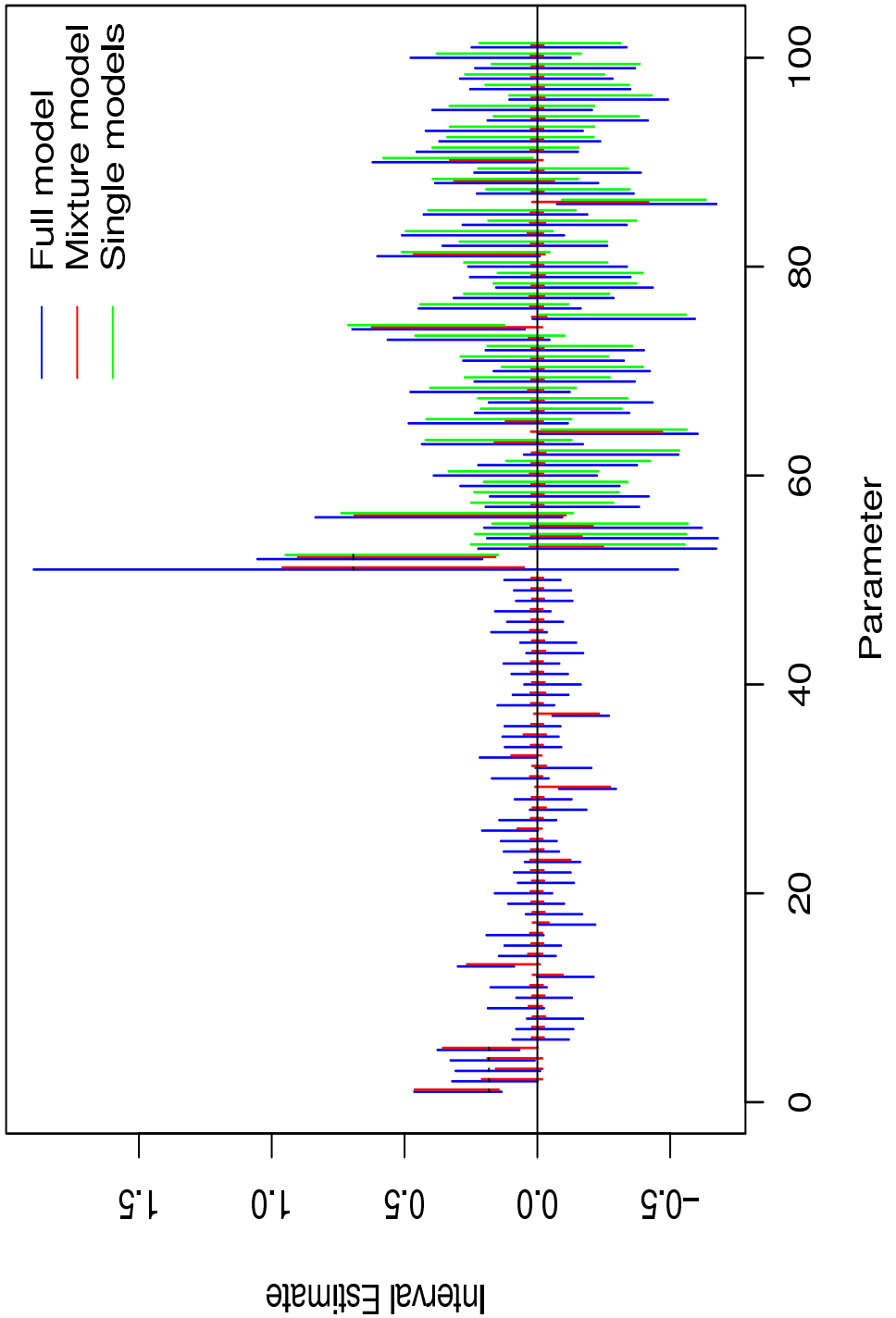

Figure 17: Simulation 15. 50/1 null/non-null SNPs, 2/0* null/non-null exposures, 100/0 null/non-null interactions.

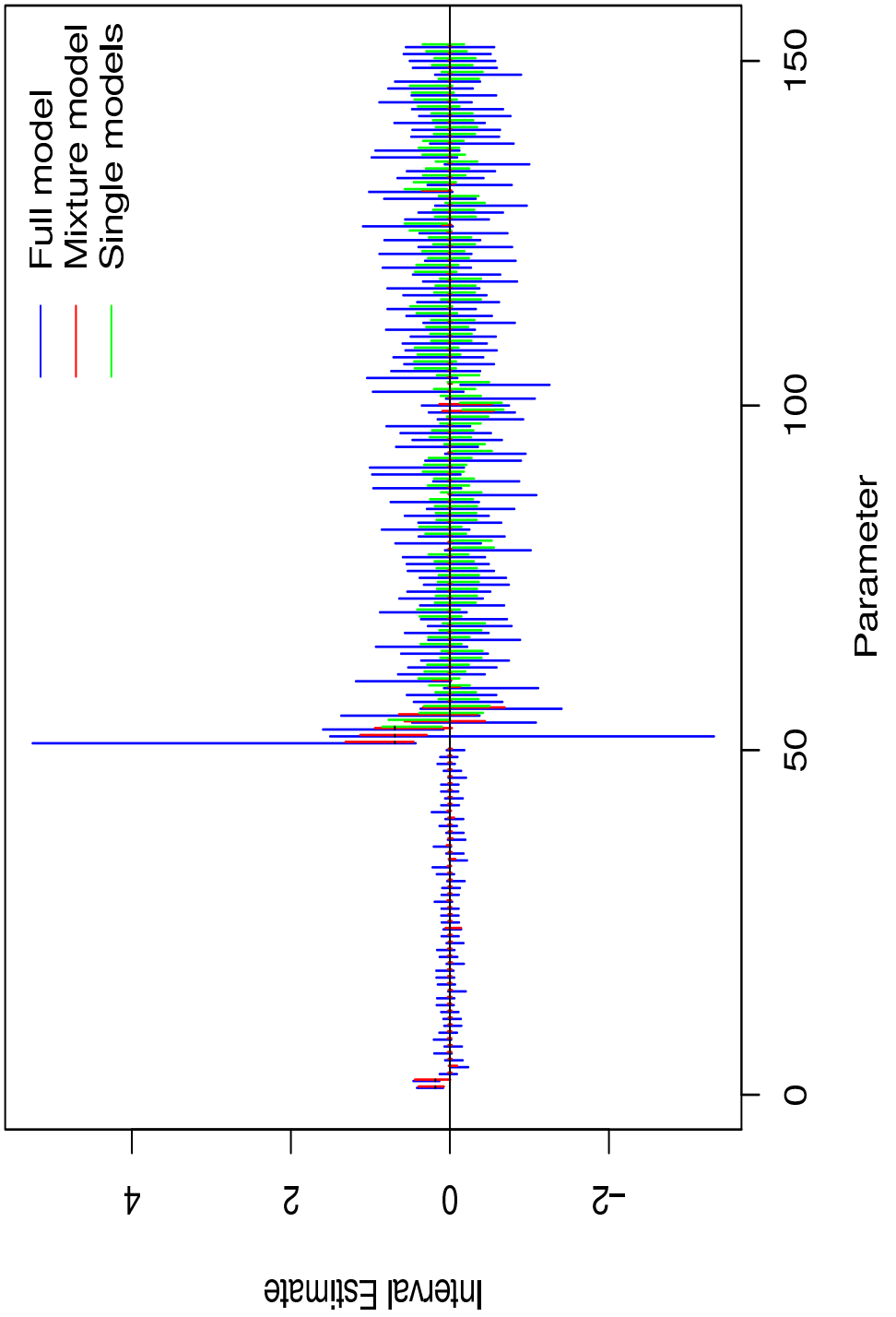

Figure 18: Simulation 16. 50/2 null/non-null SNPs, 2/2* null/non-null exposures, 100/1 null/non-null interactions.

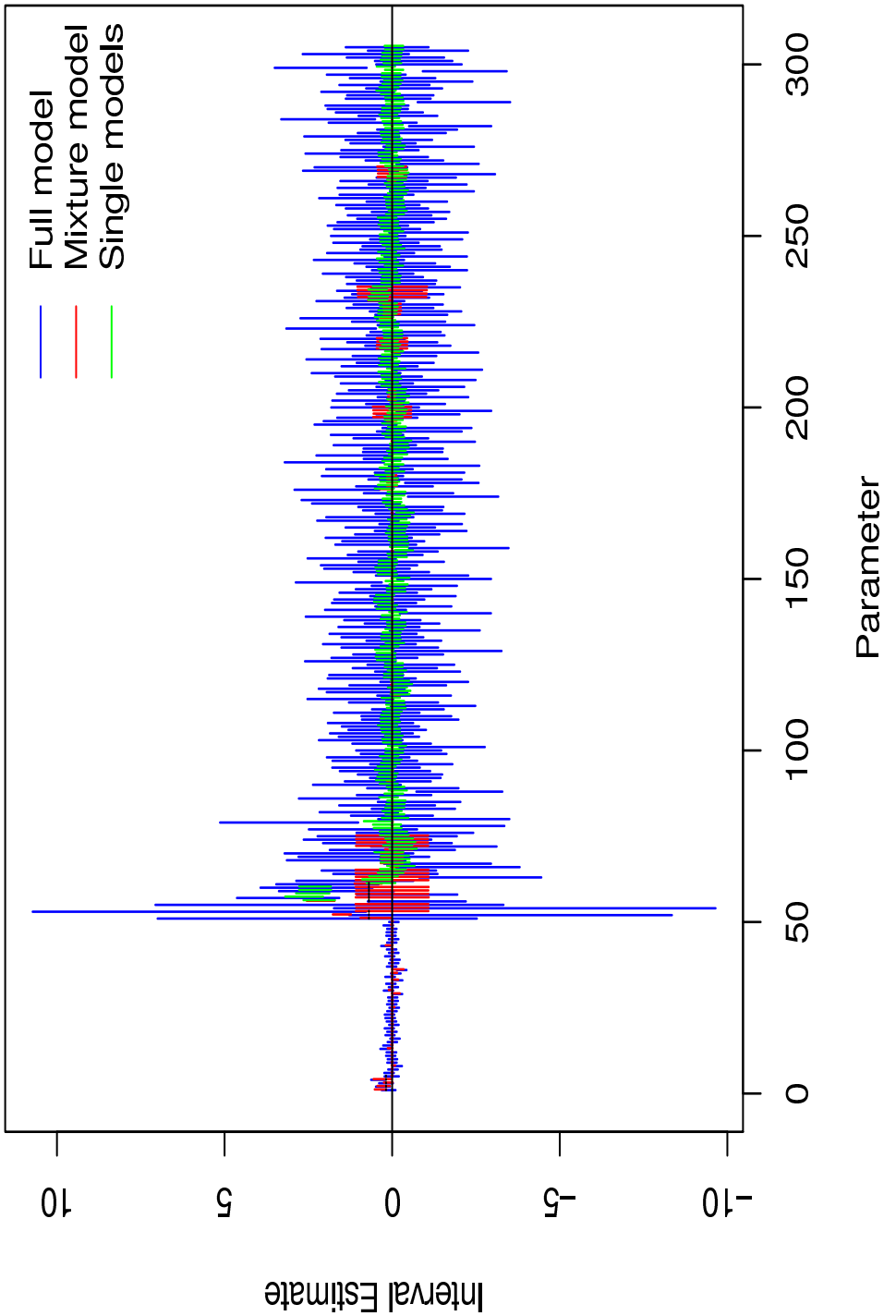

Figure 19: Simulation 17. 50/5 null/non-null SNPs, 5/3* null/non-null exposures, 250/6 null/non-null interactions.

This scenario therefore represents a situation which is not unrealistic in practice. All 3 methods correctly pick up the true non-null interaction though the full model estimate of this parameter is very wide. The full model performs poorly also in the sense of having a high type I error rate (10% with 20 false positives); the single model approach has 7 type I errors and the mixture model 0.

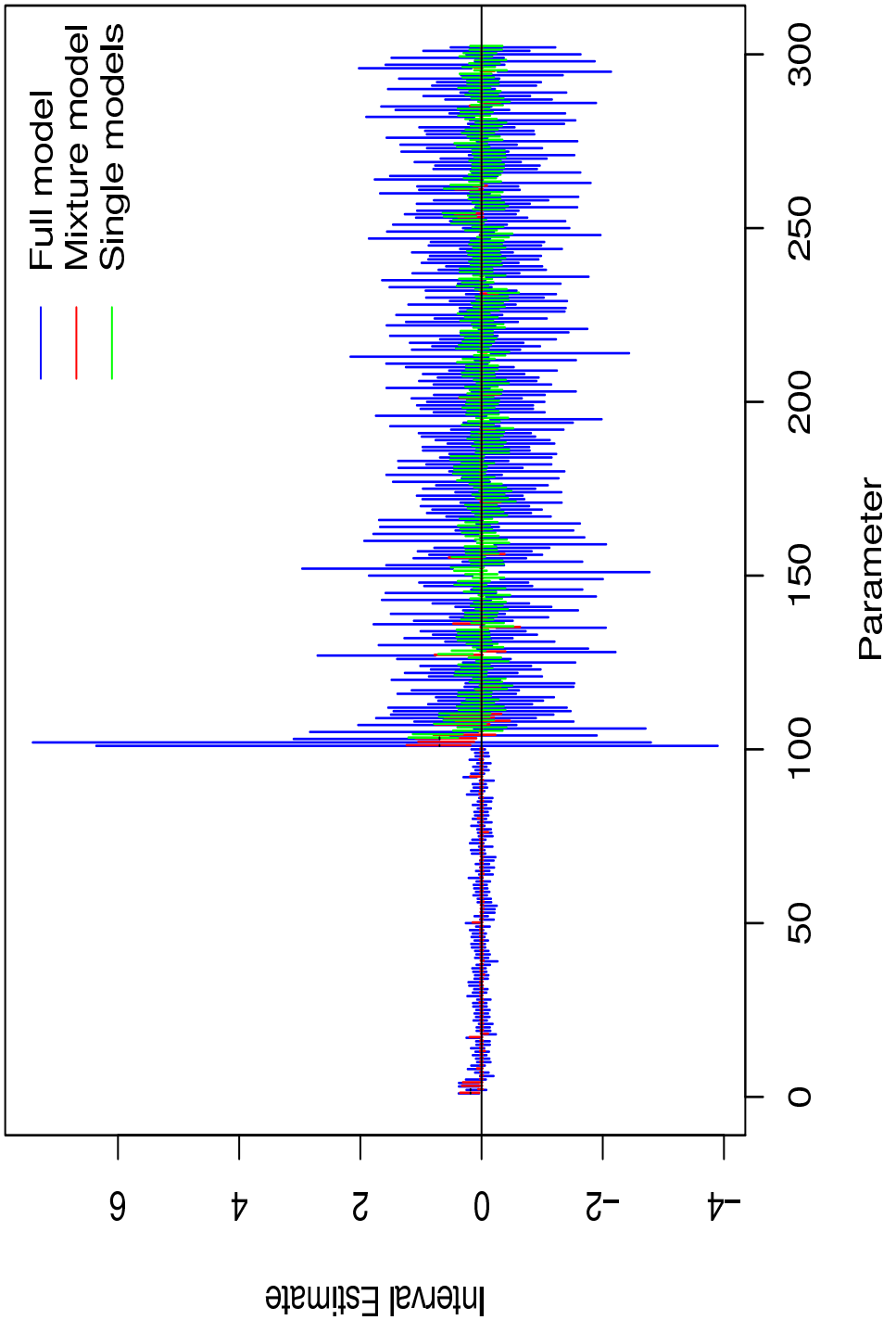

Figure 20: Simulation 18. 100/2 null/non-null SNPs, 2/2* null/non-null exposures, 200/1 null/non-null interactions.

Real Data Example

The 20 SNPs in the real data example are:

CHEK2I157T, loc123688_8034191, ADH3I349V, ADH7_1573496, APEXD148E, ATMA257C, MGMTK178R, ATMF858L, CASP8_D285H, CYP1A1I462V, COX2_8473, GSTM1, GSTT1, P53R72P, MDM2A354G, OGG1S326C, XRCC1_T29C, p21S31R, CCND1G870A, NAT2C282T

WinBUGS code

Below is the code for 1000 cases and 1000 controls, 20 SNPs and 2 exposures.

```
model
{
for (i in 1 : 2000) {
        y[i]~dbern(p[i])
# Logistic linear predictor
        logit(p[i]) <- alpha + SNPmain1[i] + SNPmain2[i] + EXPmain[i] + INTER11[i] +
                                              INTER12[i]  + INTER21[i] + INTER22[i]
        SNPmain1[i] <- beta[1]*X1[i] + beta[2]*X2[i] + beta[3]*X3[i] + beta[4]*X4[i] +
                       beta[5]*X5[i] + beta[6]*X6[i]+ beta[7]*X7[i] + beta[8]*X8[i] +
                       beta[9]*X9[i] +beta[10]*X10[i]
        SNPmain2[i] <- beta[11]*X11[i] + beta[12]*X12[i] + beta[13]*X13[i] +
                       beta[14]*X14[i] + beta[15]*X15[i] + beta[16]*X16[i] +
                       beta[17]*X17[i] + beta[18]*X18[i] + beta[19]*X19[i] +
                       beta[20]*X20[i]
        EXPmain[i]<- gamma[1]*X21[i] + gamma[2]*X22[i]
        INTER11[i] <- delta[1,1]*X1[i]*X21[i] + delta[2,1]*X2[i]*X21[i] +
                      delta[3,1]*X3[i]*X21[i] + delta[4,1]*X4 [i]*X21[i] +
                      delta[5,1]*X5[i]*X21[i] + delta[6,1]*X6[i]*X21[i] +
                      delta[7,1]*X7[i]*X21[i] + delta[8,1]*X8[i]*X21[i] +
                      delta[9,1]*X9[i]*X21[i] + delta[10,1]*X10[i]*X21[i]
        INTER12[i] <- delta[11,1]*X11[i]*X21[i] + delta[12,1]*X12[i]*X21[i] +
                      delta[13,1]*X13[i]*X21[i] + delta[14,1]*X14[i]*X21[i] +
                      delta[15,1]*X15[i]*X21[i] + delta[16,1]*X16[i]*X21[i] +
                      delta[17,1]*X17[i]*X21[i] + delta[18,1]*X18[i]*X21[i] +
                      delta[19,1]*X19[i]*X21[i] + delta[20,1]*X20[i]*X21[i]
        INTER21[i] <- delta[1,2]*X1[i]*X22[i] + delta[2,2]*X2[i]*X22[i] +
                      delta[3,2]*X3[i]*X22[i] + delta[4,2]*X4 [i]*X22[i] +
                      delta[5,2]*X5[i]*X22[i] + delta[6,2]*X6[i]*X22[i] +
                      delta[7,2]*X7[i]*X22[i] + delta[8,2]*X8[i]*X22[i] +
                      delta[9,2]*X9[i]*X22[i] + delta[10,2]*X10[i]*X22[i]
        INTER22[i] <- delta[11,2]*X11[i]*X22[i] + delta[12,2]*X12[i]*X22[i] +
                      delta[13,2]*X13[i]*X22[i] + delta[14,2]*X14[i]*X22[i] +
                      delta[15,2]*X15[i]*X22[i] + delta[16,2]*X16[i]*X22[i] +
                      delta[17,2]*X17[i]*X22[i] + delta[18,2]*X18[i]*X22[i] +
                      delta[19,2]*X19[i]*X22[i] + delta[20,2]*X20[i]*X22[i]
}
# Improper flat prior on the intercept
```

```
alpha~dflat()
# Priors for the SNP main effects
for (j in 1:20){
        prec[j]<- 1/(T[j]*0.310+(1-T[j])*0.000145)
        beta[j]~dnorm(0,prec[j])
        T[j]~dbern(0.2)
}
# Priors for the exposures
for (k in 1:2){
        prec2[k]<-1/(T2[k]*0.310+(1-T2[k])*0.000145)
        gamma[k]~dnorm(0,prec2[k])
        T2[k] <- 1
}
# Priors for the interactions
for (j in 1:20){
        for (k in 1:2){
                prec3[j,k]<-1/(T3[j,k]*0.310 + (1-T3[j,k])*0.000145)
                delta[j,k]~dnorm(0,prec3[j,k])
                T3[j,k] <- T[j]*T2[k]
        }
}
}

#initial starting point 1
list(T=c(1,0,0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),alpha=0,
beta=c(0.18,0.18,0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
gamma=c(0.69,0.69),delta=structure(.Data=c(0,0,0,0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0,0,0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0),.Dim=c(20,2)))

#init starting point 2
list(T=c(0,0,0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),alpha=0,
beta=c(0,0,0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
gamma=c(0,0),delta=structure(.Data=c(0,0,0,0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0,0,0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0),.Dim=c(20,2)))
```

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57*, 289–300.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Science 100*, 9440–9445.