Online Appendix for Chapter 3 of thesis

Alexandra C Gillett, Ammar Al-Chalabi and Cathryn M Lewis

Contents

1 Risk estimation using the LTM

Note, as in Chapter 3 of the main thesis ★ denotes novel work.

1.1 ★ Novel variation of the approximate risk estimation method

The approximate risk estimation methods, via the liability threshold model (LTM), using the Pearson-Aitken (PA) formula select on the disease status of the relative first. That is, they find the *approximate* distribution of either:

$$\text{(a)} \quad \begin{bmatrix} L_I | L_R > T \\ M_I | L_R > T \end{bmatrix} \text{ when } Y_R = 1, \text{ or,}$$

$$\text{(b)} \quad \begin{bmatrix} L_I | L_R \leq T \\ M_I | L_R \leq T \end{bmatrix} \text{ when } Y_R = 0.$$

They then use this *approximate* distribution and select on $\{M_I = m_I\}$ to obtain an *approximate* distribution for either: (a) $L_I|\{M_I = m_I, L_R > T\}$, or, (b) $L_I|\{M_I = m_I, L_R \leq T\}$.

However, if we instead select on $M_I = m_I$ first, we can use standard statistical theory for multivariate normal distributions, as in the exact method, to gain the joint distribution of:

$$\begin{bmatrix} L_I|M_I = m_I \\ L_R|M_I = m_I \end{bmatrix}$$

which, provided that L_I, L_R and M_I are trivariate normal, is not an approximation. We can then use the PA formula to gain the *approximate* distribution of either: (a) $L_I|\{M_I = m_I, L_R > T\}$, or, (b) $L_I|\{M_I = m_I, L_R \leq T\}$.

The error in risk estimation for PA approximation methods is *cumulative*. We therefore hypothesise that by removing one layer of approximation, by selecting on the measured genetic component first, we will improve the accuracy of risk estimates. The reduction in error may be trivial in the case where we are only conditioning on 2 variables ($\{M_I = m_I\}$ and $\{Y_R = y\}$; $y = 0, 1$), however, as this number increases the improvement in accuracy would also increase.

Additionally, the updated methodology presented in Section 3.3.2 of the main thesis, obtains risk estimates using the law of total probability, and therefore *sums* over multiple probabilities. If the PA approximation is used to estimate probabilities in this case, the improvement in a single probability in the summation may be trivial, but the improvement in the sum of all of these probabilities may be substantial.

We derive this updated PA approximate method, which we shall denote by $PA2$. We start by applying standard multivariate normal statistical theory to obtain the following joint distribution of $L_I|\{M_I = m_I\}$ and $L_R|\{M_I = m_I\}$:

$$\begin{bmatrix} L_I|M_I = m_I \\ L_R|M_I = m_I \end{bmatrix} \sim N(\mu_{m_I}, \Sigma_{m_I})$$

where:

$$\mu_{m_I} = \begin{bmatrix} m_I \\ rm_I \end{bmatrix} \quad \text{and} \quad \Sigma_{m_I} = \begin{bmatrix} 1 - V_M & r(h_L^2 - V_M) \\ r(h_L^2 - V_M) & 1 - r^2 V_M \end{bmatrix}$$

as in the exact method (see Section 3.3.1, 'Review of the exact method', in the main thesis for details). We then apply the PA formula, selecting on the disease status of relative $\{R\}$, to gain an approximate distribution of conditional liability to disease for individual $\{I\}$, and so estimate: (a) $p(Y_I = 1|M_I = m_I, Y_R = 1)$, and, (b) $p(Y_I = 1|M_I = m_I, Y_R = 0)$.

(a) Relative {R} is affected

When $Y_R = 1$, we want to apply the PA formula to the relevant distribution given in main thesis (Chapter 3) to find the approximate distribution of $L_I|\{M_I = m_I, L_R > T\}$. We define:

$$X' = L_R|\{M_I = m_I\} \sim N(rm_I, 1 - r^2 V_M)$$
$$Y' = L_I|\{M_I = m_I\} \sim N(m_I, 1 - V_M)$$

To obtain the distribution of $Y'|X' > T$, we first find the updated mean ($\mu^*_{X'}$) and variance ($V^*_{X'}$) for $X'|X' > T$. Since X' is normally distributed, $X'|X' > T$ is a truncated normal distribution. Using

equations presented in the Chapter 3 of the main thesis for the mean and variance of an upper tailed, truncated normal distribution, we obtain:

$$\mu^*_{X'} = rm_I + \sqrt{1 - r^2 V_M}\lambda(\alpha)$$

and

$$V^*_{X'} = (1 - r^2 V_M)\Big(1 - \lambda(\alpha)\big(\lambda(\alpha) - \alpha\big)\Big)$$

where:

- $\alpha = \frac{T - rm_I}{\sqrt{1 - r^2 V_M}}$,
- $\lambda(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)}$,
- $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)$; the probability density function of the standard normal distribution,
- $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$; the CDF of the standard normal distribution, and,
- T is the disease threshold.

Using these updated distribution parameters in the PA formula we get the following approximate distribution for $L_I|\{M_I = m_I, L_R > T\}$:

$$L_I|\{M_I = m_I, L_R > T\} \sim N(\mu_{1,PA2}, \sigma^2_{1,PA2}) \tag{1}$$

where:

$$\begin{aligned}\mu_{1,PA2} &= \mu^*_Y \\ &= m_I + \frac{r(h^2_L - V_M)}{\sqrt{1 - r^2 V_M}}\lambda(\alpha)\end{aligned}$$

and

$$\begin{aligned}\sigma^2_{1,PA2} &= V^*_Y \\ &= 1 - V_M - \frac{r^2(h^2_L - V_M)^2}{1 - r^2 V_M}\lambda(\alpha)\big(\lambda(\alpha) - \alpha\big)\end{aligned}$$

and α and $\lambda(\alpha)$ are as defined above.

Using this approximate distribution we can estimate risk as:

$$\begin{aligned}p(Y_I = 1|M_I = m_I, Y_R = 1) &= p(L_I > T|M_I = m_I, L_R > T) \\ &= p(\{L_I|M_I = m_I, L_R > T\} > T) \\ &= 1 - \Phi\Big(\frac{T - \mu_{1,PA2}}{\sigma_{1,PA2}}\Big)\end{aligned} \tag{2}$$

(b) Relative {R} is unaffected

When $Y_R = 0$, we want to apply the PA formula to the relevant distribution given in Chapter 3 of the main thesis to find the approximate distribution of $L_I|\{M_I = m_I, L_R \leq T\}$. Let us define X' and Y' as above.

To obtain the distribution of $Y'|X' \leq T$, we need the updated mean ($\mu^*_{X'}$) and variance ($V^*_{X'}$) for $X'|X' \leq T$. $X'|X' \leq T$ is a truncated normal distribution. Using the equations presented in Chapter 3 of the main thesis for the mean and variance of an lower tailed, truncated normal distribution, we obtain:

$$\mu^*_{X'} = rm_I - \sqrt{1 - r^2 V_M}\Upsilon(\alpha)$$

and

$$V^*_{X'} = (1 - r^2 V_M)\Big(1 - \Upsilon(\alpha)\big(\Upsilon(\alpha) + \alpha\big)\Big)$$

where:

- $\alpha = \frac{T - rm_I}{\sqrt{1 - r^2 V_M}}$,
- $\Upsilon(\alpha) = \frac{\phi(\alpha)}{\Phi(\alpha)}$,
- $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)$; the probability density function of the standard normal distribution,
- $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$; the CDF of the standard normal distribution, and,
- T is the disease threshold.

Using $\mu^*_{X'}$ and $V^*_{X'}$ in the PA formula we get the following approximate distribution for $L_I|\{M_I = m_I, L_R \leq T\}$:

$$L_I|\{M_I = m_I, L_R \leq T\} \sim N(\mu_{0,PA2}, \sigma^2_{0,PA2}) \tag{3}$$

where:

$$\begin{aligned}\mu_{0,PA2} &= \mu^*_Y \\ &= m_I - \frac{r(h^2_L - V_M)}{\sqrt{1 - r^2 V_M}}\Upsilon(\alpha)\end{aligned}$$

and

$$\begin{aligned}\sigma^2_{0,PA2} &= V^*_Y \\ &= 1 - V_M - \frac{r^2(h^2_L - V_M)^2}{1 - r^2 V_M}\Upsilon(\alpha)\big(\Upsilon(\alpha) + \alpha\big)\end{aligned}$$

and α and $\Upsilon(\alpha)$ are as defined above.

Using this approximate distribution we can estimate risk as:

$$\begin{aligned}p(Y_I = 1|M_I = m_I, Y_R = 0) &= p(L_I > T|M_I = m_I, L_R \leq T) \\ &= p(\{L_I|M_I = m_I, L_R \leq T\} > T) \\ &= 1 - \Phi\Big(\frac{T - \mu_{0,PA2}}{\sigma_{0,PA2}}\Big)\end{aligned} \tag{4}$$

The distributions obtained using the new $PA2$ method; defined in Equations (1) and (3), differ from those of the original PA approximation methods. Does the updated PA approach improve the accuracy of risk estimates? We shall now explore this using simulation.

1.2 Simulation study

We now explore the comparative performance of $PA2$ and the original PA approximation method, from here-on-in denoted by $PA1$, in 2 simulation scenarios. We assume that risk estimates from the exact method are the closest to the true risk, and therefore PA approximations closer to the exact method are deemed superior. Each simulation scenario will now be described and results provided.

Simulation group 1:

Here we compare the risk equations derived using $PA2$ (described above) and the original PA methods. We call the original PA methods $PA1$ here, and they are outlined in Chapter 3 of the main thesis.

This is done for a variety of relative types and disease model inputs, described next.

Simulation set up 1:

We compare $PA1$ (the existing method of So et al., 2011) and $PA2$ (the updated method) estimates for $p(Y_I = 1|M_I = m_I, Y_R = y)$ with the following varying inputs:

- $y = \{0, 1\}$,
- $R = \{1^{st}\text{degree relative}, 2^{nd}\text{degree relative}, 3^{rd}\text{degree relative}\}$; also denoted $R = \{1, 2, 3\}$,
- $K = \{0.001, 0.01, 0.05, 0.1, 0.2\}$,
- $h^2_L = \{0.1, 0.2, ..., 0.9, 1\}$,
- $\rho = \{0.1, 0.25, 0.5, 0.75, 0.9\}$ where $V_M = \rho h^2_L$, and,
- 1000 m_I values, spanning an appropriate range*.

Note: $V_M = Var[M]$.

$*$: m_I's are generated in the following manner:

1. 998 points are generated using the distribution for M_I: $M_I \sim N(0, V_M = \rho h^2_L)$.
2. 2 additional values are added to these 998 points. They are found by generating 10,000 values using the distribution $M_I \sim N(0, V_M = \rho h^2_L)$ and then taking the minimum and maximum. This is to ensure that we capture values at the extremes of the M_I distribution.

Simulation results 1:

First we note that across all simulations within this group, both methods produce similar and accurate estimates; the updated method, $PA2$, performs equally as well as the existing method, $PA1$. In fact, for all input combinations, 100% of results from $PA2$ and 99.98% of results from $PA1$ are the same as the exact method when rounded to 2 decimal places. The tiny proportion of cases where $PA1$ differ from the exact method when rounded to 2 decimal places all (marginally) over-estimate risk, and, occur when:

- relative, R, is a 1^{st} degree relative,

- this 1^{st} degree relative is unaffected ($Y_R = 0$),
- the narrow sense heritability $h^2_L \geq 0.5$, and,
- M_I is a value in the extreme upper tail of the distribution. How extreme depends upon the model parameters: $\{K, h^2_L, V_M\}$.

Prior to simulation we hypothesised that, although differences between $PA1$ and $PA2$ might be negligible in this simulation group, $PA2$ would provide estimates that are consistently closer to the exact method than $PA1$. This is due to the reduced number of approximate distributions used to generate the approximation. This was not the case across all simulation input combinations. However, the largest deviations from the exact method occur for the $PA1$ method.

Despite the very small differences in risk estimates we shall look at any trends where $PA2$ may perform better than $PA1$, and vice versa. Firstly, when $R = 1$ and $Y_R = 1$:

- Median risk difference is approximately 0, and the interquartile range narrow, for both $PA1$ and $PA2$ when narrow-sense heritability $h^2_L \leq 0.3$.
- For a given K, the accuracy of both methods decrease for high h^2_L and low ρ. This is shown by:
  - $PA1$ and $PA2$ either under or overestimate risk (relative to the exact method) when a disease is highly heritable and the proportion of h^2_L accounted for by M is low.
  - There is increased variability in risk difference for both methods relative to the exact method.

  These problems are more noticeable for $PA1$. That is shown by:
  - When both $PA1$ and $PA2$ *overestimate* disease risk, the median risk difference from $PA2$ tends to be closer to 0.
  - $PA1$ has more 'extreme' risk differences than $PA2$. This is because $PA1$ tends to over/ underestimate the risk of disease for individuals with high M.
- Additionally, as ρ increases, and so $V_M = \rho h^2_L$ increases, the accuracy of $PA2$ is greater than $PA1$, as the variability of risk difference for $PA1$ is larger.
- For both $PA1$ and $PA2$, deviations in median risk difference from 0 are larger for increasing K.

These trends can be seen for all combinations of R and Y_R; $R = 1, 2, 3$ and $Y_R = 0, 1$. The magnitude of the risk difference for $PA1$ and $PA2$ is greatest for 1^{st} degree relatives ($R = 1$) simulations. The improvement in accuracy of $PA2$ over $PA1$ is most notable when $R = 1$ and $Y_R = 0$. For your reference, Figures (1) - (30), Section 1.2.1 below, provide stratified box-plots of the risk difference (y-axis) against narrow-sense heritability (x-axis) by PA method (rows) and $\rho = \frac{V_M}{h^2_L}$ (columns) for all combinations of $\{R = i, D_R = d_R, K = j\}$ ($\{R = 1, 2, 3; D_R = 0, 1; K = 0.001, 0.01, 0.05, 0.1, 0.2\}$).

Conclusion: $PA2$, and the existing, $PA1$, Pearson-Aiken approximate methods perform equally well when estimating $p(Y_I = 1|Y_R = y, M_I = m_I)$ over a wide range of M_I values. $PA2$ is preferred: 1. when the relative is unaffected, 2. for high values of M_I, and, 3. when ρ is high (as $V_M \to h^2_L$). These are perhaps scenarios when M_I is more informative about the disease status of individual $\{I\}$ than the family history of disease, and so conditioning on M_I first is at its most advantageous.

Simulation group 2:

In this simulation group we consider risk estimation when we have disease status available for a single relative, plus measurements for sets of disease associated variables. For example, these measurements could be:

- polygenic risk scores from co-morbid or possibly correlated diseases (such as schizophrenia, bipolar and major depressive disorder), and/ or,
- clinically available measurements such as, for cardiovascular disease, variables that may influence disease risk include systolic and diastolic blood pressure, BMI, hours of exercise per week, and, HDL and LDL cholesterol levels.

We simulate a variety of measurement types. In particular, we vary:

- the number of measured variables (5 or 10),
- the level of correlation between measured variables, and,
- the amount of information each variable contributes to the variability in disease on the liability scale, and whether this information is attributable to genetic or environmental causes of disease.

Additionally, we assume that $H_L^2 = h_L^2$, and we vary narrow-sense heritability and the lifetime prevalence of disease. For family history information we assume that there is information about 1 affected 1^{st} degree relative across all simulations. We now provide a detailed description of the simulation set up.

Simulation set up 2:

We assume that disease is explained by a liability threshold model, with no dominance ($H_L^2 = h_L^2$):

$$L = A + E \quad \sim N(0, 1)$$

and that both A and E each can be broken down into 10 independent variables:

$$A = \sum_{i=1}^{10} A_i$$

and:

$$E = \sum_{i=1}^{10} E_i$$

where $A_i \sim N(0, \frac{1}{10}h_L^2)$ and $E_i \sim N(0, \frac{1}{10}(1 - h_L^2))$. We then simulate a variety of variables that each measure some, but not all, of the causal components, A_i and E_i; $i = 1, 2, ..., 10$, plus noise.

In these simulations we vary:

- narrow-sense heritability; $h_L^2 = \{0.2, 0.5, 0.8\}$,
- lifetime disease prevalence; $K = \{0.01, 0.05, 0.15\}$,
- the number of measured variables; $p = 5, 10$, and,
- the variance of error for each variable; $\sigma^2_{error} = 1, 0.2$.

When the number of measured variables is 5 ($p = 5$), we have run 4 simulation types, each with the measured variables defined in a different way. We denote each of these simulation types by 'type 5.i', where $i = 1, 2, 3, 4$.

Type 5.1 defines the measured variables M_1 to M_5 as in Table (1). Each variable measures 4 out of 10 polygenic components and 4 out of 10 environmental components of disease, and therefore are all equally correlated with liability to disease. This correlation is 0.3381. They are also equally correlated with each other with common correlation of 0.07143.

Causal variables	Measured variables				
	M_1	M_2	M_3	M_4	M_5
A_1	1	1	0	0	0
A_2	1	0	1	0	0
A_3	1	0	0	1	0
A_4	1	0	0	0	1
A_5	0	1	1	0	0
A_6	0	1	0	1	0
A_7	0	1	0	0	1
A_8	0	0	1	1	0
A_9	0	0	1	0	1
A_{10}	0	0	0	1	1
E_1	1	1	0	0	0
E_2	1	0	1	0	0
E_3	1	0	0	1	0
E_4	1	0	0	0	1
E_5	0	1	1	0	0
E_6	0	1	0	1	0
E_7	0	1	0	0	1
E_8	0	0	1	1	0
E_9	0	0	1	0	1
E_{10}	0	0	0	1	1
$error$[1]	1	1	1	1	1

[1] 2 simulation scenarios for $error$: a. $error = Z \sim N(0, 1)$, and, b. $error \sim N(0, 0.2)$

Table example: $M_1 = A_1 + A_2 + A_3 + A_4 + E_1 + E_2 + E_3 + E_4 + error$

Table 1: Creating the measured variables for Simulation Group 2; type 5.1. 5 measured variables (M_1 - M_5). Each entry $\{i, j\}$, represents the weighted contribution of causal variable i to measured variable j

Type 5.2 defines the measured variables M_1 to M_5 as in Table (2). Each M_i; $i = 1, 2, ..., 5$, variable measures 2 polygenic causal components and 2 environmental causal components. Therefore all are equally correlated with liability to disease, with correlation 0.1826. All measured variables are uncorrelated.

Type 5.3 defines the measured variables M_1 to M_5 as in Table (3). Like in type 5.2 all measured variables are uncorrelated. However, now:

- M_1 and M_2 measure genetic risk factors only; each variable measures 4 out of the 10 polygenic causal components,
- M_3 measures both genetic and environmental risk factors; 2 polygenic components and 2 environmental components are captured, and,

Causal variables	Measured variables				
	M_1	M_2	M_3	M_4	M_5
A_1	1	0	0	0	0
A_2	1	0	0	0	0
A_3	0	1	0	0	0
A_4	0	1	0	0	0
A_5	0	0	1	0	0
A_6	0	0	1	0	0
A_7	0	0	0	1	0
A_8	0	0	0	1	0
A_9	0	0	0	0	1
A_{10}	0	0	0	0	1
E_1	1	0	0	0	0
E_2	1	0	0	0	0
E_3	0	1	0	0	0
E_4	0	1	0	0	0
E_5	0	0	1	0	0
E_6	0	0	1	0	0
E_7	0	0	0	1	0
E_8	0	0	0	1	0
E_9	0	0	0	0	1
E_{10}	0	0	0	0	1
$error$[1]	1	1	1	1	1

[1] 2 simulation scenarios for $error$: a. $error = Z \sim N(0,1)$, and, b. $error \sim N(0, 0.2)$

Table example: $M_1 = A_1 + A_2 + E_1 + E_2 + error$

Table 2: Creating the measured variables for Simulation Group 2; type 5.2. 5 measured variables (M_1 - M_5). Each entry $\{i, j\}$, represents the weighted contribution of causal variable i to measured variable j

- M_4 and M_5 measure environmental risk factors only; each variable measures 4 out of the 10 environmental causal components.

Finally, type 5.4 defines the measured variables M_1 to M_5 as in Table (4). All measured variables have the same amount of correlation with liability to disease (0.5963). They share the same, high level of correlation between themselves, with a common correlation of 0.75.

When the number of measured variables is 10 ($p = 10$), we have run 2 simulation types, again, each type defines the measured variables differently.

Firstly, 10 uncorrelated measured variables are created as in Table (5). This is called simulation type 10.1. Each variable measures 1 causal polygenic and 1 causal environmental component, and is correlated with L by 0.0909.

Secondly, another set of uncorrelated measured variables are created, shown in Table (6) and referred to as type 10.2. Here:

- variables M_1 to M_5 measure (independent) genetic risk factors only; each variable measures 2 causal polygenic components, and,
- variables M_6 to M_{10} measure (independent) genetic risk factors only; each variable measures 2 causal environmental components.

Causal variables	Measured variables				
	M_1	M_2	M_3	M_4	M_5
A_1	1	0	0	0	0
A_2	1	0	0	0	0
A_3	1	0	0	0	0
A_4	1	0	0	0	0
A_5	0	1	0	0	0
A_6	0	1	0	0	0
A_7	0	1	0	0	0
A_8	0	1	0	0	0
A_9	0	0	1	0	0
A_{10}	0	0	1	0	0
E_1	0	0	0	0	1
E_2	0	0	0	0	1
E_3	0	0	0	0	1
E_4	0	0	0	0	1
E_5	0	0	0	1	0
E_6	0	0	0	1	0
E_7	0	0	0	1	0
E_8	0	0	0	1	0
E_9	0	0	1	0	0
E_{10}	0	0	1	0	0
$error$[1]	1	1	1	1	1

[1] 2 simulation scenarios for $error$: a. $error = Z \sim N(0,1)$, and, b. $error \sim N(0, 0.2)$

Table example: $M_1 = A_1 + A_2 + A_3 + A_4 + error$

Table 3: Creating the measured variables for Simulation Group 2; type 5.3. 5 measured variables (M_1 - M_5). Each entry $\{i, j\}$, represents the weighted contribution of causal variable i to measured variable j

As noted before, we assume that we know of 1 affected 1^{st} degree relative across all simulations.

For $10,000$ samples we generate A_i and E_i; $i = 1, 2, ..., 10$. Using these we generate the measured variables for each of the simulation types listed above. We then estimate risk using the measured variables and family history information via the exact method, $PA1$ (the original method, selecting on family history first) and $PA2$ (the updated version, selecting on the measured variables first). We compare the risk difference between each PA method and the exact method in the results below.

Causal variables	Measured variables				
	M_1	M_2	M_3	M_4	M_5
A_1	1	1	1	1	0
A_2	1	1	1	1	0
A_3	1	1	1	0	1
A_4	1	1	1	0	1
A_5	1	1	0	1	1
A_6	1	1	0	1	1
A_7	1	0	1	1	1
A_8	1	0	1	1	1
A_9	0	1	1	1	1
A_{10}	0	1	1	1	1
E_1	1	1	1	1	0
E_2	1	1	1	1	0
E_3	1	1	1	0	1
E_4	1	1	1	0	1
E_5	1	1	0	1	1
E_6	1	1	0	1	1
E_7	1	0	1	1	1
E_8	1	0	1	1	1
E_9	0	1	1	1	1
E_{10}	0	1	1	1	1
$error$[1]	1	1	1	1	1

[1] 2 simulation scenarios for $error$: a. $error = Z \sim N(0,1)$, and, b. $error \sim N(0, 0.2)$

Table example: $M_1 = A_1 + A_2 + A_3 + A_4 + A_5 + A_6 + A_7 + A_8 + E_1 + E_2 + E_3 + E_4 + E_5 + E_6 + E_7 + E_8 + error$

Table 4: Creating the measured variables for Simulation Group 2; type 5.4. 5 measured variables (M_1 - M_5). Each entry $\{i, j\}$, represents the weighted contribution of causal variable i to measured variable j

Causal variables	Measured variables									
	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
A_1	1	0	0	0	0	0	0	0	0	0
A_2	0	1	0	0	0	0	0	0	0	0
A_3	0	0	1	0	0	0	0	0	0	0
A_4	0	0	0	1	0	0	0	0	0	0
A_5	0	0	0	0	1	0	0	0	0	0
A_6	0	0	0	0	0	1	0	0	0	0
A_7	0	0	0	0	0	0	1	0	0	0
A_8	0	0	0	0	0	0	0	1	0	0
A_9	0	0	0	0	0	0	0	0	1	0
A_{10}	0	0	0	0	0	0	0	0	0	1
E_1	1	0	0	0	0	0	0	0	0	0
E_2	0	1	0	0	0	0	0	0	0	0
E_3	0	0	1	0	0	0	0	0	0	0
E_4	0	0	0	1	0	0	0	0	0	0
E_5	0	0	0	0	1	0	0	0	0	0
E_6	0	0	0	0	0	1	0	0	0	0
E_7	0	0	0	0	0	0	1	0	0	0
E_8	0	0	0	0	0	0	0	1	0	0
E_9	0	0	0	0	0	0	0	0	1	0
E_{10}	0	0	0	0	0	0	0	0	0	1
$error$[1]	1	1	1	1	1	1	1	1	1	1

[1] 2 simulation scenarios for $error$: a. $error = Z \sim N(0,1)$, and, b. $error \sim N(0, 0.2)$

Table example: $M_1 = A_1 + E_1 + error$

Table 5: Creating the measured variables for Simulation Group 2; type 10.1. 10 measured variables (M_1 - M_{10}). Each entry $\{i, j\}$, represents the weighted contribution of causal variable i to measured variable j

Causal variables	Measured variables									
	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
A_1	1	0	0	0	0	0	0	0	0	0
A_2	1	0	0	0	0	0	0	0	0	0
A_3	0	1	0	0	0	0	0	0	0	0
A_4	0	1	0	0	0	0	0	0	0	0
A_5	0	0	1	0	0	0	0	0	0	0
A_6	0	0	1	0	0	0	0	0	0	0
A_7	0	0	0	1	0	0	0	0	0	0
A_8	0	0	0	1	0	0	0	0	0	0
A_9	0	0	0	0	1	0	0	0	0	0
A_{10}	0	0	0	0	1	0	0	0	0	0
E_1	0	0	0	0	0	1	0	0	0	0
E_2	0	0	0	0	0	1	0	0	0	0
E_3	0	0	0	0	0	0	1	0	0	0
E_4	0	0	0	0	0	0	1	0	0	0
E_5	0	0	0	0	0	0	0	1	0	0
E_6	0	0	0	0	0	0	0	1	0	0
E_7	0	0	0	0	0	0	0	0	1	0
E_8	0	0	0	0	0	0	0	0	1	0
E_9	0	0	0	0	0	0	0	0	0	1
E_{10}	0	0	0	0	0	0	0	0	0	1
$error$[1]	1	1	1	1	1	1	1	1	1	1

[1] 2 simulation scenarios for $error$: a. $error = Z \sim N(0, 1)$, and, b. $error \sim N(0, 0.2)$

Table example: $M_1 = A_1 + A_2 + error$

Table 6: Creating the measured variables for Simulation Group 2; type 10.2. 10 measured variables (M_1 - M_{10}). Each entry $\{i, j\}$, represents the weighted contribution of causal variable i to measured variable j

Simulation results 2:

The PA approximate methods perform equally well across the majority of simulated $\underline{M}$ values. The largest difference in risk for each PA method relative to the exact method was:

- 0.0006 for $PA2$; occurring in the 10 variable simulation, type 2, with $\{h^2_L = 0.80, K = 0.15, \sigma^2_{error} = 1\}$, and,
- -0.0026 for $PA1$; occurring in the 10 variable simulation, type 2, with $\{h^2_L = 0.80, K = 0.01, \sigma^2_{error} = 0.2\}$.

These can be seen in Figures (31) and (42), in Section 1.2.1 below.

For both PA approximate methods, we rounded the estimated risk to 2 decimal places (dp), and calculated the risk difference for this rounded risk relative to the rounded exact method risk. That is:

$$rd_{PAi,2dp} = round(risk_{PAi}, dp = 2) - round(risk_{exact}, dp = 2)$$

where $i = 1, 2$. We then calculated the % of times that this rounded risk difference differed from 0.00, for all simulation scenarios. These are displayed in Tables (7) to (10). Using these and Figures (31) - (42) we see that across all of the simulations defined above, for both $PA1$ and $PA2$, there was a *decrease* in accuracy, relative to the exact method, with:

- increasing noise in measured variables,
- increasing heritability of the disease; with $H^2_L = h^2_L$ in all simulations, and,
- increasing lifetime disease prevalence, K.

In particular, as h^2_L and K increase, both methods tend to have an increased median risk difference (> 0), and an increased interquartile range. This reduction in accuracy tends to be greater for $PA1$. Additionally, the risk difference plots for the $PA1$ method often show a tail of outliers, with $PA1$ risk estimates comparatively far from the exact method.

Conclusion: As in Simulation Group 1, $PA2$ and $PA1$ perform equally well, compared to the exact method, for the majority of measured variable- simulation scenarios run here. Even when 10 variables are measured and incorporated, the performance of $PA1$ and $PA2$ are very similar. However, the 10 variable scenarios considered here contain variables where a large proportion of the variance is noise. Performance differences may be greater with reduced noise, and measured variables capturing important non-genetic risk information. In general $PA2$ has a median risk difference closer to 0, with lower variability and fewer outliers.

Simulations here are illustrative, and by no means exhaustive. $PA1$ and $PA2$ may differ in performance in other settings. However, the trends across both simulation groups show that, here at least, both PA methods are good approximations for risk calculated using the exact method. $PA2$ is at least as good as $PA1$, and may be preferred when:

- an individual has an extremely high polygenic risk score (or measured variable),
- disease heritability is estimated to be high, and,
- the disease is common.

For this reason, the PA approximate method used in subsequent work in this chapter will be $PA2$.

	{Noise:Total} variance ratio[1]					Correlation											% risk difference (2 d.p.) $!= 0$[2]					
						M_i	L_I					L_R					$K = 0.01$		$K = 0.05$		$K = 0.15$	
Type	M_1	M_2	M_3	M_4	M_5	M_j	$M_{I,1}$	$M_{I,2}$	$M_{I,3}$	$M_{I,4}$	$M_{I,5}$	$M_{I,1}$	$M_{I,2}$	$M_{I,3}$	$M_{I,4}$	$M_{I,5}$	$PA2$	$PA1$	$PA2$	$PA1$	$PA2$	$PA1$
$h_L^2 = 0.2$																						
1	0.71	0.71	0.71	0.71	0.71	0.07	0.34	0.34	0.34	0.34	0.34	0.03	0.03	0.03	0.03	0.03	0.01	0.02	0.00	0.00	0.03	0.04
2	0.83	0.83	0.83	0.83	0.83	0.00	0.18	0.18	0.18	0.18	0.18	0.02	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.00	0.03	0.03
3	0.93	0.93	0.83	0.76	0.76	0.00	0.08	0.08	0.18	0.28	0.28	0.04	0.04	0.02	0.00	0.00	0.02	0.01	0.01	0.02	0.03	0.11
4	0.56	0.56	0.56	0.56	0.56	0.33	0.60	0.60	0.60	0.60	0.60	0.06	0.06	0.06	0.06	0.06	0.00	0.00	0.00	0.00	0.00	0.00
$h_L^2 = 0.5$																						
1	0.71	0.71	0.71	0.71	0.71	0.07	0.34	0.34	0.34	0.34	0.34	0.08	0.08	0.08	0.08	0.08	0.03	0.10	0.11	0.21	0.22	0.67
2	0.83	0.83	0.83	0.83	0.83	0.00	0.18	0.18	0.18	0.18	0.18	0.05	0.05	0.05	0.05	0.05	0.19	0.16	0.25	0.13	0.49	0.81
3	0.83	0.83	0.83	0.83	0.83	0.00	0.18	0.18	0.18	0.18	0.18	0.09	0.09	0.05	0.00	0.00	0.07	0.11	0.19	0.39	0.53	0.89
4	0.56	0.56	0.56	0.56	0.56	0.33	0.60	0.60	0.60	0.60	0.60	0.15	0.15	0.15	0.15	0.15	0.01	0.07	0.02	0.03	0.06	0.24
$h_L^2 = 0.8$																						
1	0.71	0.71	0.71	0.71	0.71	0.07	0.34	0.34	0.34	0.34	0.34	0.14	0.14	0.14	0.14	0.14	0.37	0.26	0.70	1.23	1.16	3.83
2	0.83	0.83	0.83	0.83	0.83	0.00	0.18	0.18	0.18	0.18	0.18	0.07	0.07	0.07	0.07	0.07	0.53	0.48	1.07	1.29	2.73	4.55
3	0.76	0.76	0.83	0.93	0.93	0.00	0.28	0.28	0.18	0.08	0.08	0.14	0.14	0.07	0.00	0.00	0.60	0.50	0.80	1.47	2.33	4.84
4	0.56	0.56	0.56	0.56	0.56	0.33	0.60	0.60	0.60	0.60	0.60	0.24	0.24	0.24	0.24	0.24	0.04	0.13	0.14	0.49	0.27	2.27

[1] $\frac{\sigma^2_{error}}{\sigma^2_{M_i}}$; $i = 1, 2, ..., 5$

[2] % of risk differences (between PA method $\{i\}$ and the exact method; $i = 1, 2$) which are not $= 0.00$, *after* rounding to 2 decimal places
$PA1$ is the original PA approximate method, presented by So et al. [2011]
$PA2$ is the updated PA approximate method

Table 7: Summary for 5 variable simulations, type 1-4, $\sigma^2_{error} = 1$

	{Noise:Total} variance ratio[1]					Correlation											% risk difference (2 d.p.) $!= 0$[2]					
						M_i	L_I					L_R					$K = 0.01$		$K = 0.05$		$K = 0.15$	
Type	M_1	M_2	M_3	M_4	M_5	M_j	$M_{I,1}$	$M_{I,2}$	$M_{I,3}$	$M_{I,4}$	$M_{I,5}$	$M_{I,1}$	$M_{I,2}$	$M_{I,3}$	$M_{I,4}$	$M_{I,5}$	$PA2$	$PA1$	$PA2$	$PA1$	$PA2$	$PA1$
$h_L^2 = 0.2$																						
1	0.33	0.33	0.33	0.33	0.33	0.17	0.52	0.52	0.52	0.52	0.52	0.05	0.05	0.05	0.05	0.05	0.00	0.00	0.00	0.01	0.01	0.01
2	0.50	0.50	0.50	0.50	0.50	0.00	0.32	0.32	0.32	0.32	0.32	0.03	0.03	0.03	0.03	0.03	0.00	0.02	0.00	0.01	0.03	0.02
3	0.71	0.71	0.50	0.38	0.38	0.00	0.15	0.15	0.32	0.44	0.44	0.08	0.08	0.03	0.00	0.00	0.00	0.00	0.00	0.06	0.02	0.16
4	0.20	0.20	0.20	0.20	0.20	0.60	0.80	0.80	0.80	0.80	0.80	0.08	0.08	0.08	0.08	0.08	0.00	0.00	0.00	0.00	0.00	0.01
$h_L^2 = 0.5$																						
1	0.33	0.33	0.33	0.33	0.33	0.17	0.52	0.52	0.52	0.52	0.52	0.13	0.13	0.13	0.13	0.13	0.00	0.05	0.00	0.05	0.04	0.30
2	0.50	0.50	0.50	0.50	0.50	0.00	0.32	0.32	0.32	0.32	0.32	0.08	0.08	0.08	0.08	0.08	0.07	0.11	0.09	0.13	0.17	0.64
3	0.50	0.50	0.50	0.50	0.50	0.00	0.32	0.32	0.32	0.32	0.32	0.16	0.16	0.08	0.00	0.00	0.04	0.07	0.14	0.49	0.16	1.03
4	0.20	0.20	0.20	0.20	0.20	0.60	0.80	0.80	0.80	0.80	0.80	0.20	0.20	0.20	0.20	0.20	0.00	0.00	0.01	0.01	0.00	0.08
$h_L^2 = 0.8$																						
1	0.33	0.33	0.33	0.33	0.33	0.17	0.52	0.52	0.52	0.52	0.52	0.21	0.21	0.21	0.21	0.21	0.03	0.08	0.06	0.38	0.16	2.01
2	0.50	0.50	0.50	0.50	0.50	0.00	0.32	0.32	0.32	0.32	0.32	0.13	0.13	0.13	0.13	0.13	0.21	0.22	0.41	1.01	1.06	3.77
3	0.38	0.38	0.50	0.71	0.71	0.00	0.44	0.44	0.32	0.15	0.15	0.22	0.22	0.13	0.00	0.00	0.12	0.34	0.38	1.35	0.69	3.74
4	0.20	0.20	0.20	0.20	0.20	0.60	0.80	0.80	0.80	0.80	0.80	0.32	0.32	0.32	0.32	0.32	0.00	0.02	0.01	0.12	0.00	0.55

[1] $\frac{\sigma^2_{error}}{\sigma^2_{M_i}}$; $i = 1, 2, ..., 5$

[2] % of risk differences (between PA method $\{i\}$ and the exact method; $i = 1, 2$) which are not $= 0.00$, *after* rounding to 2 decimal places
$PA1$ is the original PA approximate method, presented by So et al. [2011]
$PA2$ is the updated PA approximate method

Table 8: Summary for 5 variable simulations, type 1-4, $\sigma^2_{error} = 0.2$

Type	{Noise:Total} variance ratio[1]		Correlation					% risk difference (2 d.p.) $!= 0$[2]					
			M_i	L_I		L_R		$K = 0.01$		$K = 0.05$		$K = 0.15$	
	M_1 - M_5	M_6 - M_7	M_j	M_1 - M_5	M_6 - M_7	M_1 - M_5	M_6 - M_7	$PA2$	$PA1$	$PA2$	$PA1$	$PA2$	$PA1$
$h_L^2 = 0.2$													
1	0.91	0.91	0.00	0.10	0.10	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
2	0.96	0.86	0.00	0.04	0.15	0.02	0.00	0.03	0.03	0.05	0.07	0.01	0.04
$h_L^2 = 0.5$													
1	0.91	0.91	0.00	0.10	0.10	0.02	0.02	0.08	0.09	0.24	0.11	0.51	0.61
2	0.91	0.91	0.00	0.10	0.10	0.05	0.00	0.16	0.17	0.16	0.29	0.58	0.88
$h_L^2 = 0.8$													
1	0.91	0.91	0.00	0.10	0.10	0.04	0.04	0.75	0.60	1.11	1.23	3.55	4.70
2	0.86	0.96	0.00	0.15	0.04	0.07	0.00	0.56	0.43	0.85	1.57	2.74	4.64

[1] $\frac{\sigma^2_{error}}{\sigma^2_{M_i}}$; $i = 1, 2, ..., 10$

[2] % of risk differences (between PA method $\{i\}$ and the exact method; $i = 1, 2$) which are not $= 0.00$, *after* rounding to 2 decimal places
$PA1$ is the original PA approximate method, presented by So et al. [2011]
$PA2$ is the updated PA approximate method

Table 9: Summary for 10 variable simulations, type 1-4, $\sigma^2_{error} = 1$

Type	{Noise:Total} variance ratio[1]		Correlation					% risk difference (2 d.p.) $!= 0$[2]					
			M_i	L_I		L_R		$K = 0.01$		$K = 0.05$		$K = 0.15$	
	M_1 - M_5	M_6 - M_7	M_j	M_1 - M_5	M_6 - M_7	M_1 - M_5	M_6 - M_7	$PA2$	$PA1$	$PA2$	$PA1$	$PA2$	$PA1$
$h_L^2 = 0.2$													
1	0.67	0.67	0.00	0.18	0.18	0.02	0.02	0.00	0.03	0.00	0.00	0.00	0.02
2	0.83	0.56	0.00	0.08	0.27	0.04	0.00	0.01	0.03	0.01	0.01	0.04	0.15
$h_L^2 = 0.5$													
1	0.67	0.67	0.00	0.18	0.18	0.05	0.05	0.09	0.24	0.21	0.17	0.28	0.73
2	0.67	0.67	0.00	0.18	0.18	0.09	0.00	0.08	0.17	0.09	0.40	0.31	1.40
$h_L^2 = 0.8$													
1	0.67	0.67	0.00	0.18	0.18	0.07	0.07	0.44	0.32	0.73	1.10	1.70	4.18
2	0.56	0.83	0.00	0.27	0.08	0.13	0.00	0.32	0.42	0.58	1.56	1.09	4.38

[1] $\frac{\sigma^2_{error}}{\sigma^2_{M_i}}$; $i = 1, 2, ..., 10$

[2] % of risk differences (between PA method $\{i\}$ and the exact method; $i = 1, 2$) which are not $= 0.00$, *after* rounding to 2 decimal places
$PA1$ is the original PA approximate method, presented by So et al. [2011]
$PA2$ is the updated PA approximate method

Table 10: Summary for 10 variable simulations, type 1-4, $\sigma^2_{error} = 0.2$

1.2.1 Pearson-Aitken approximate methods simulation: Additional graphs

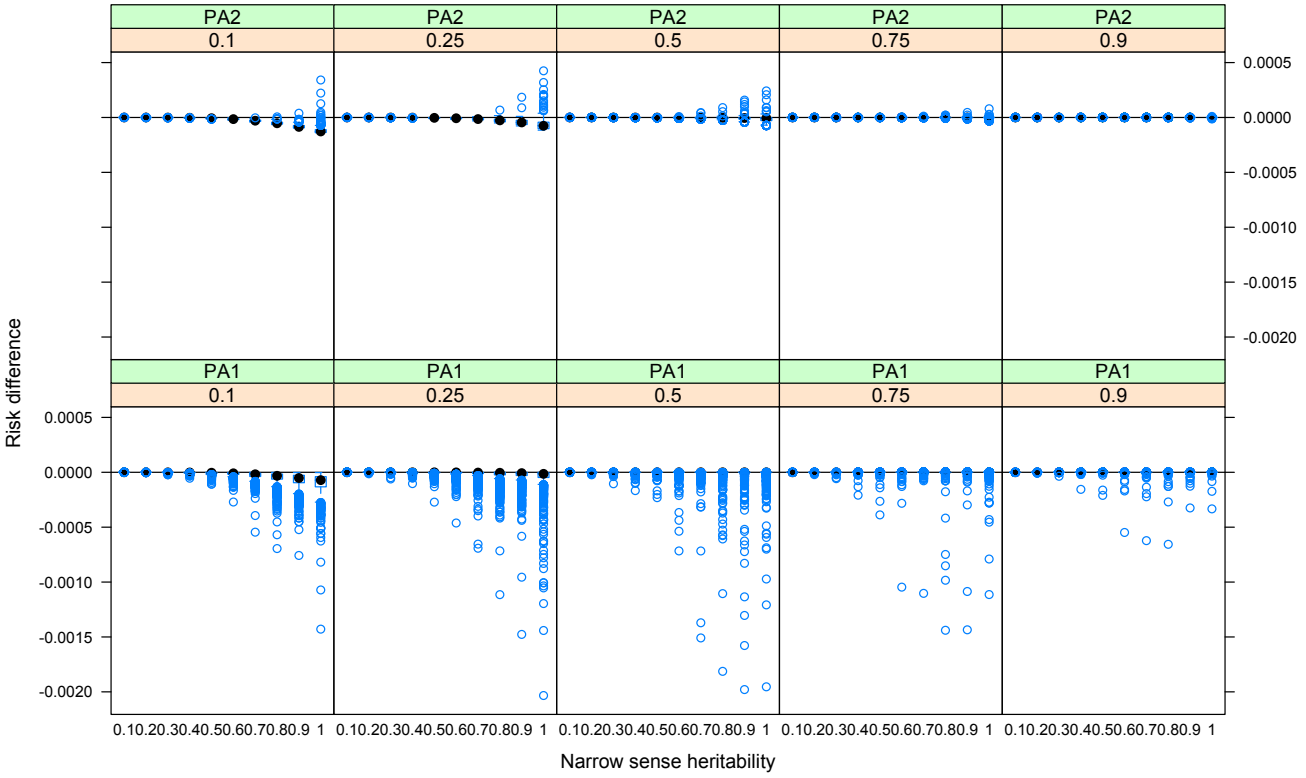

Figure 1: $\{R = 1, Y_R = 1, K = 0.001\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

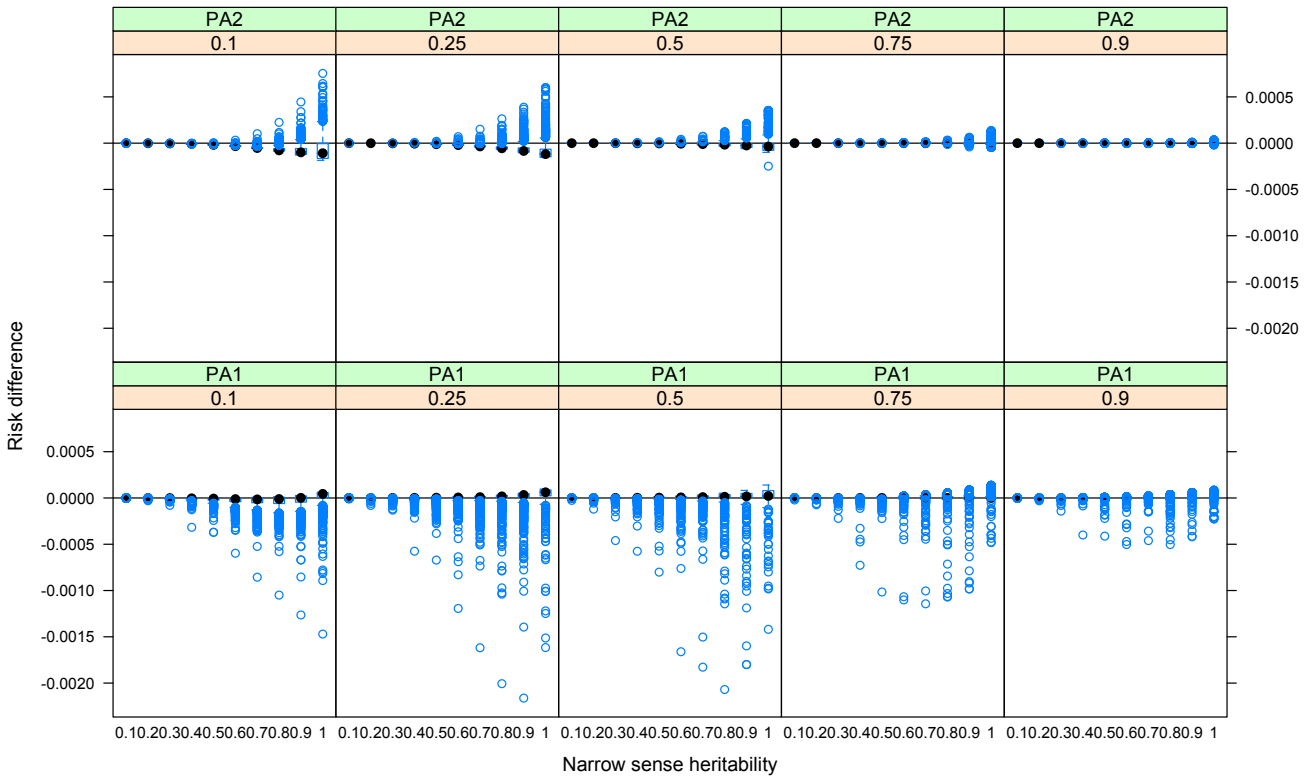

Figure 2: $\{R = 1, Y_R = 1, K = 0.01\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

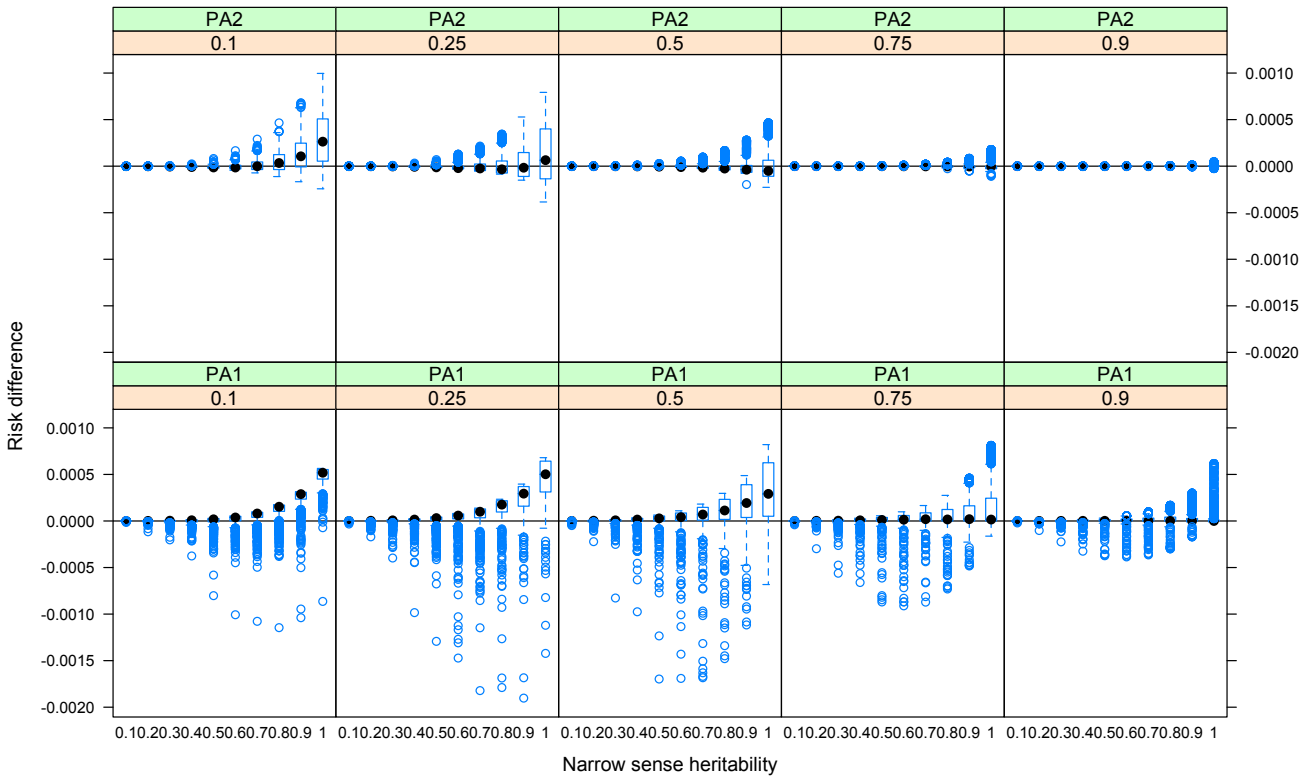

Figure 3: $\{R = 1, Y_R = 1, K = 0.05\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

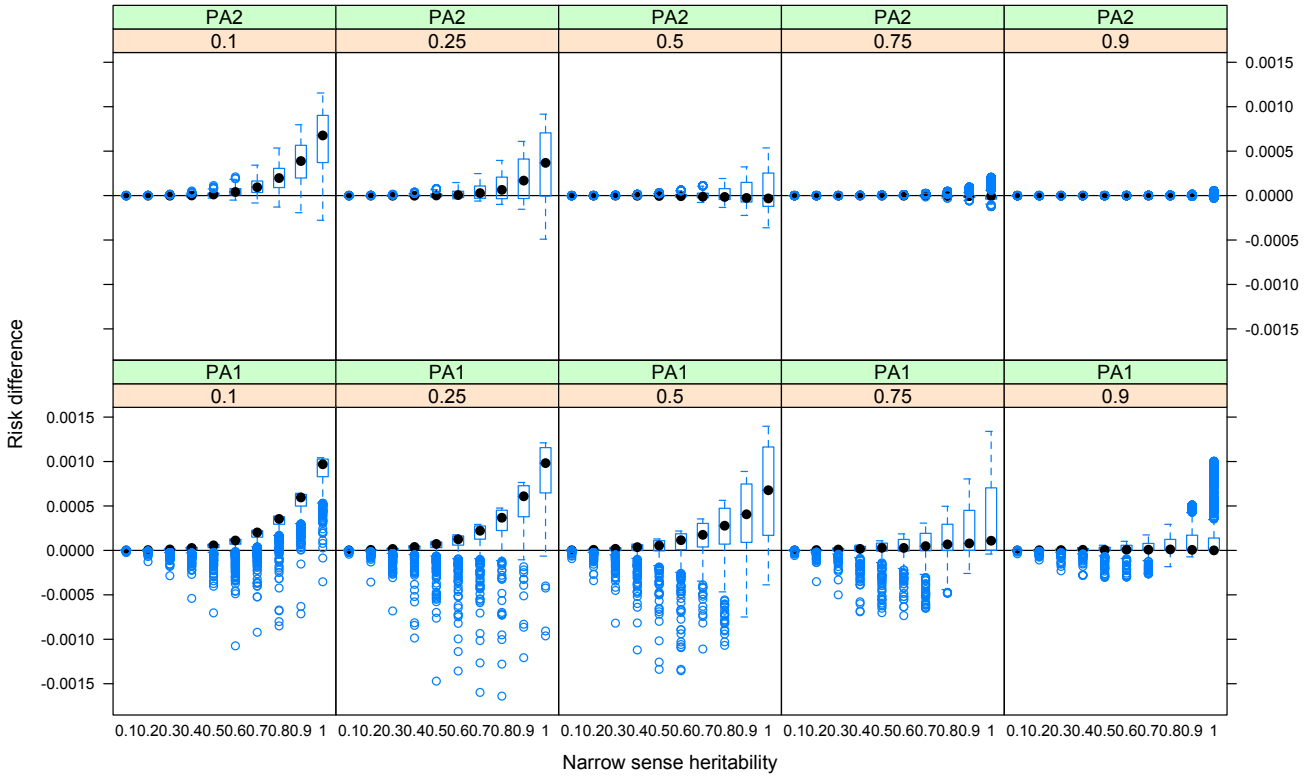

Figure 4: $\{R = 1, Y_R = 1, K = 0.1\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

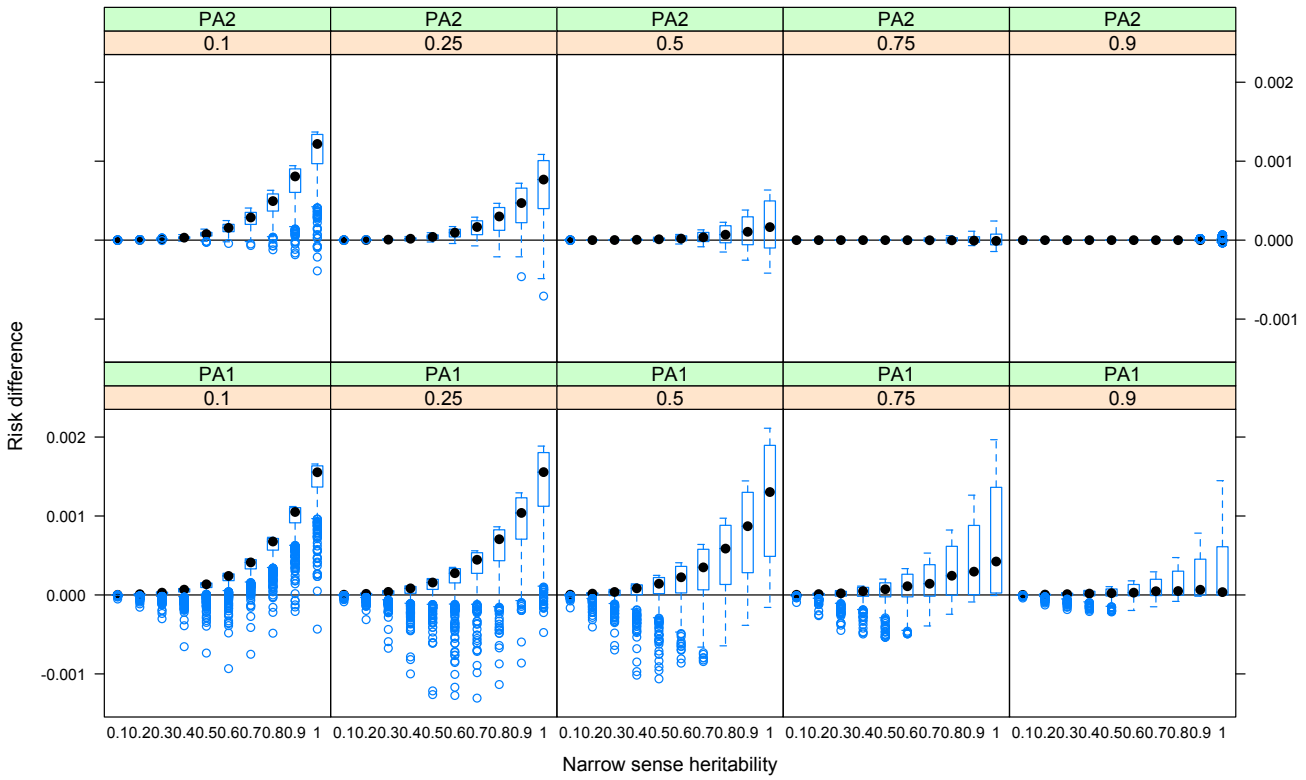

Figure 5: $\{R = 1, Y_R = 1, K = 0.2\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

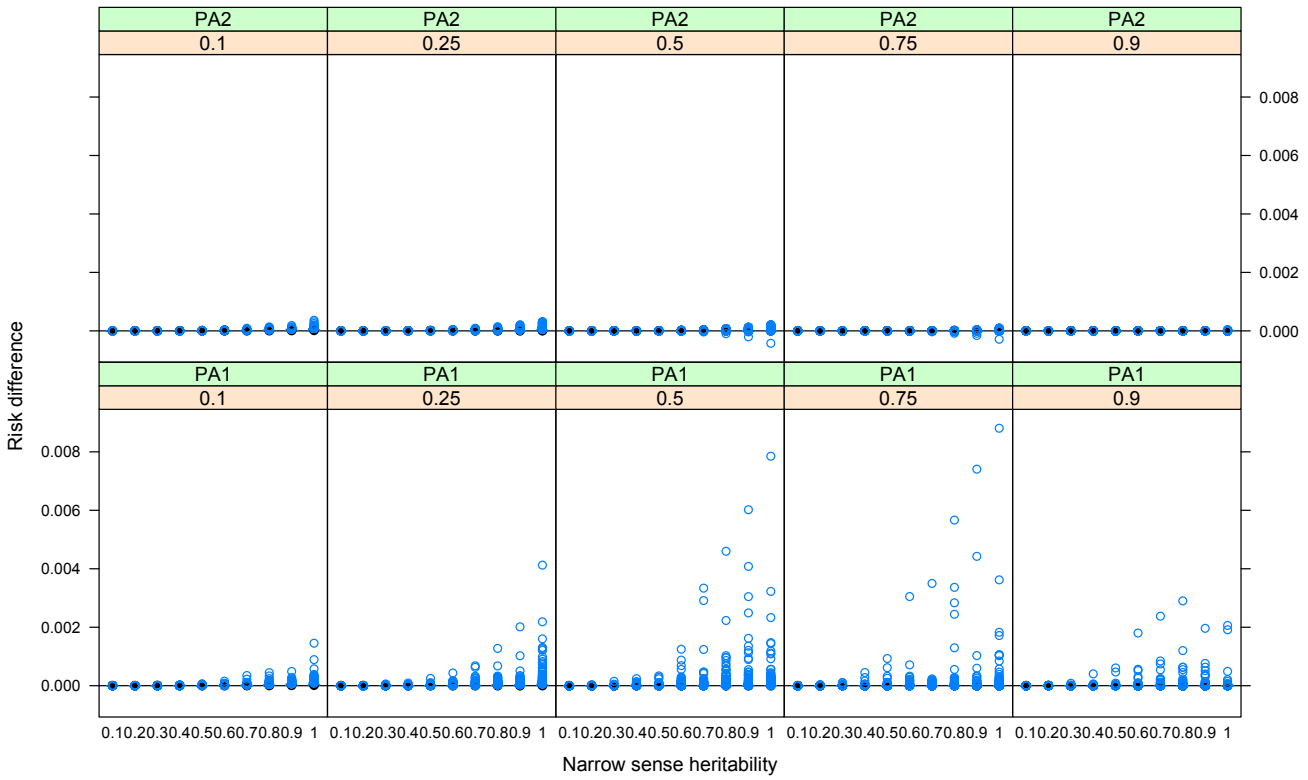

Figure 6: $\{R = 1, Y_R = 0, K = 0.001\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

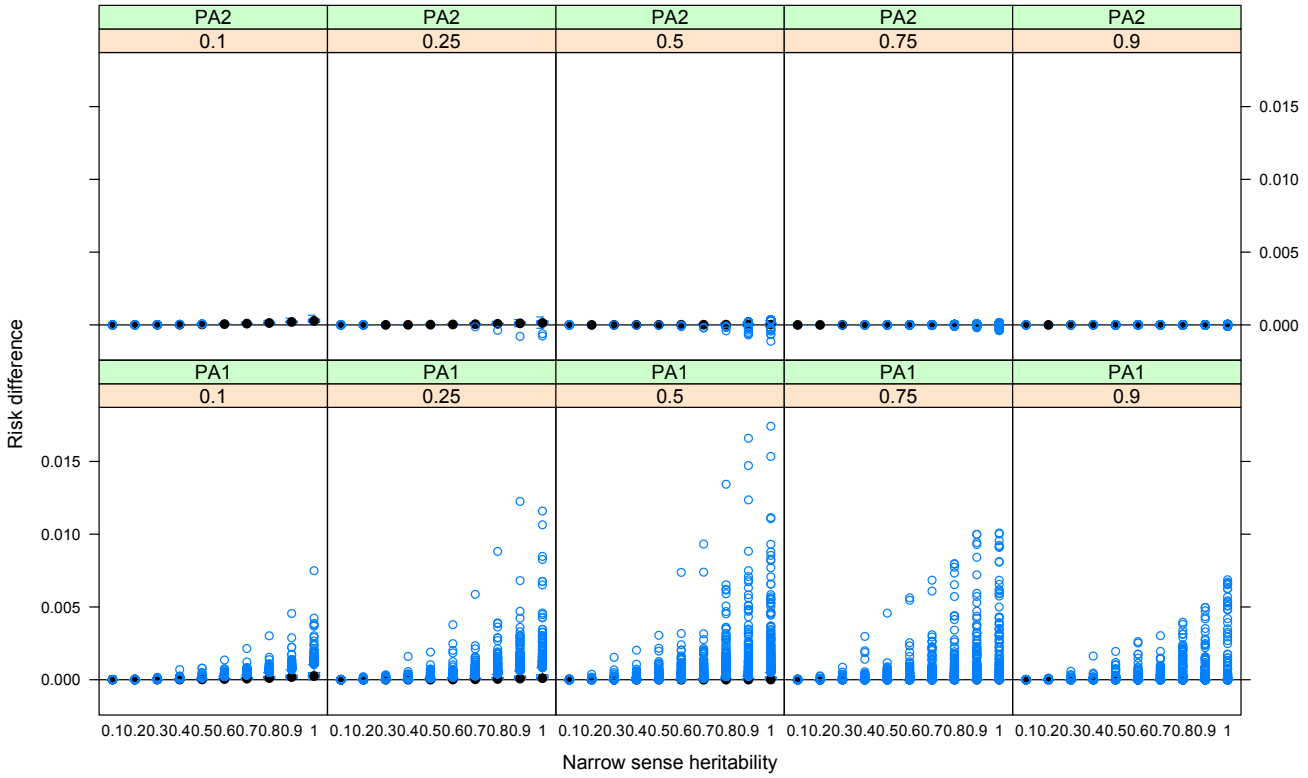

Figure 7: $\{R = 1, Y_R = 0, K = 0.01\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

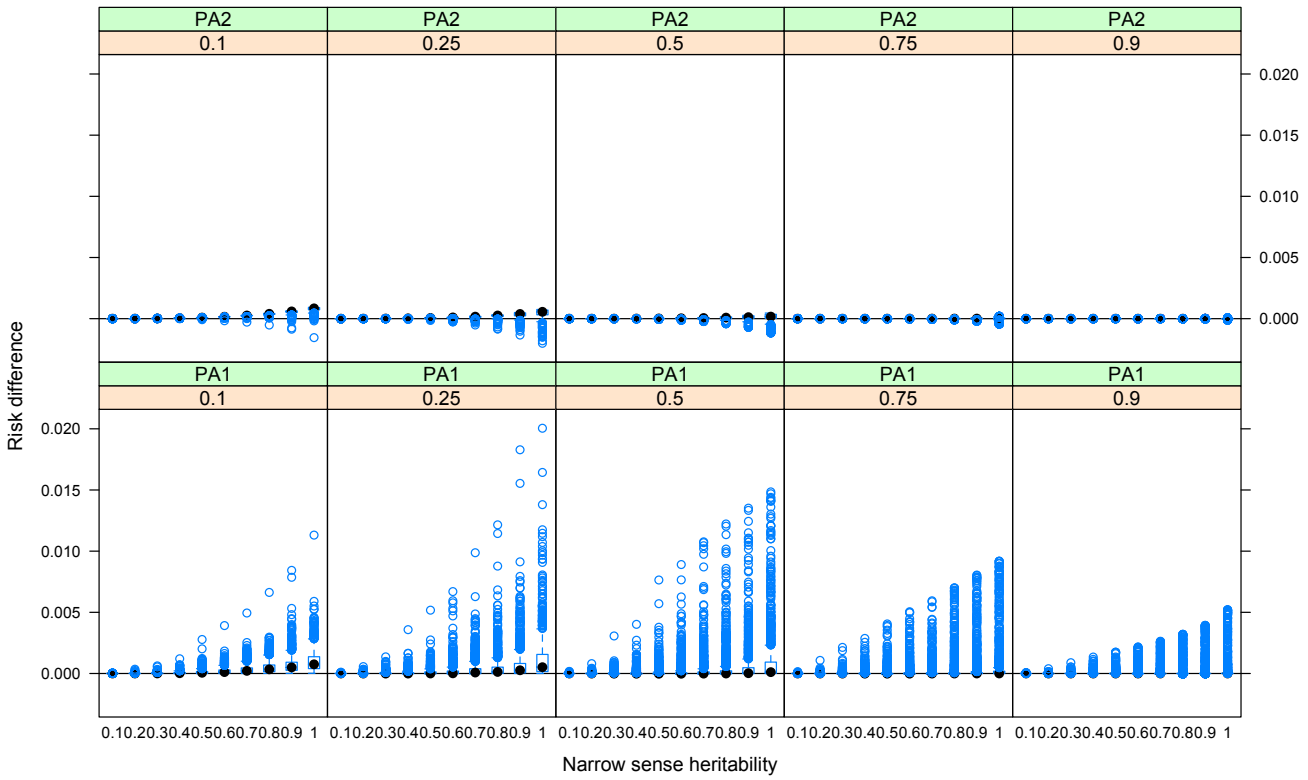

Figure 8: $\{R = 1, Y_R = 0, K = 0.05\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

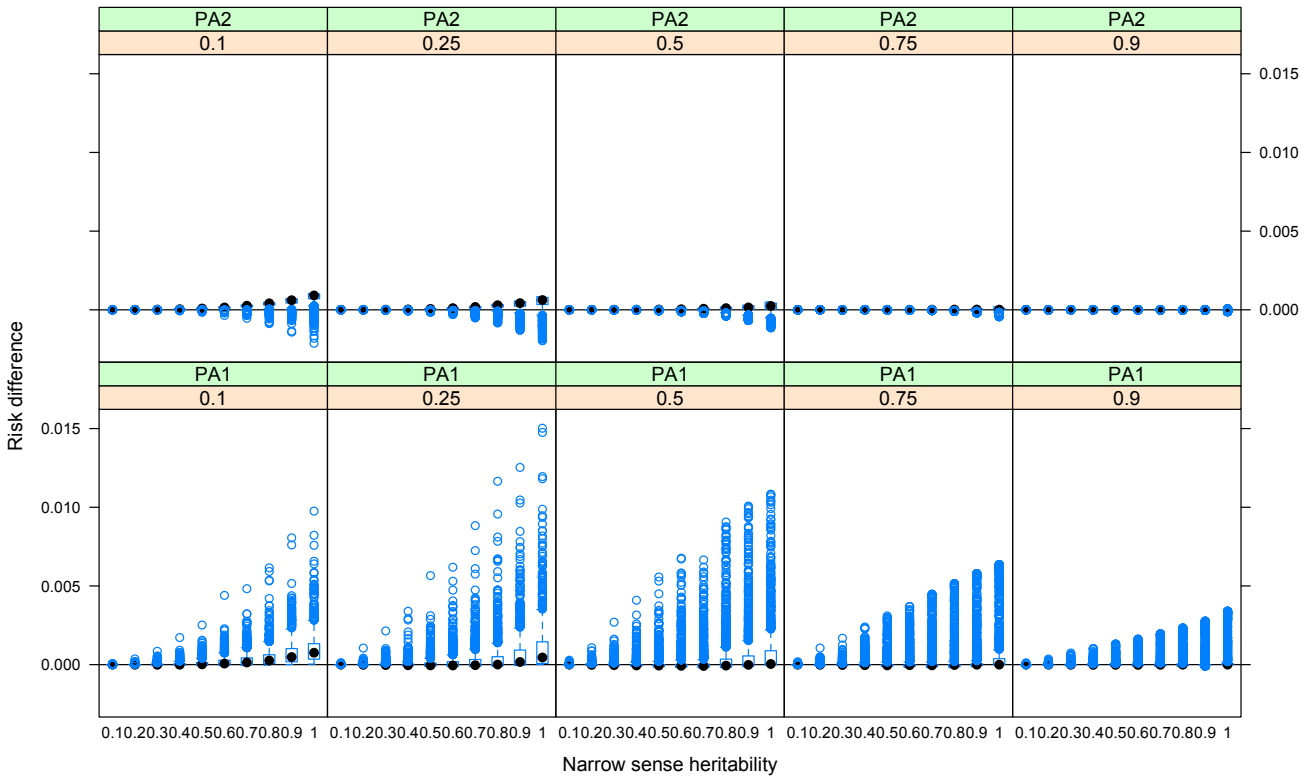

Figure 9: $\{R = 1, Y_R = 0, K = 0.1\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

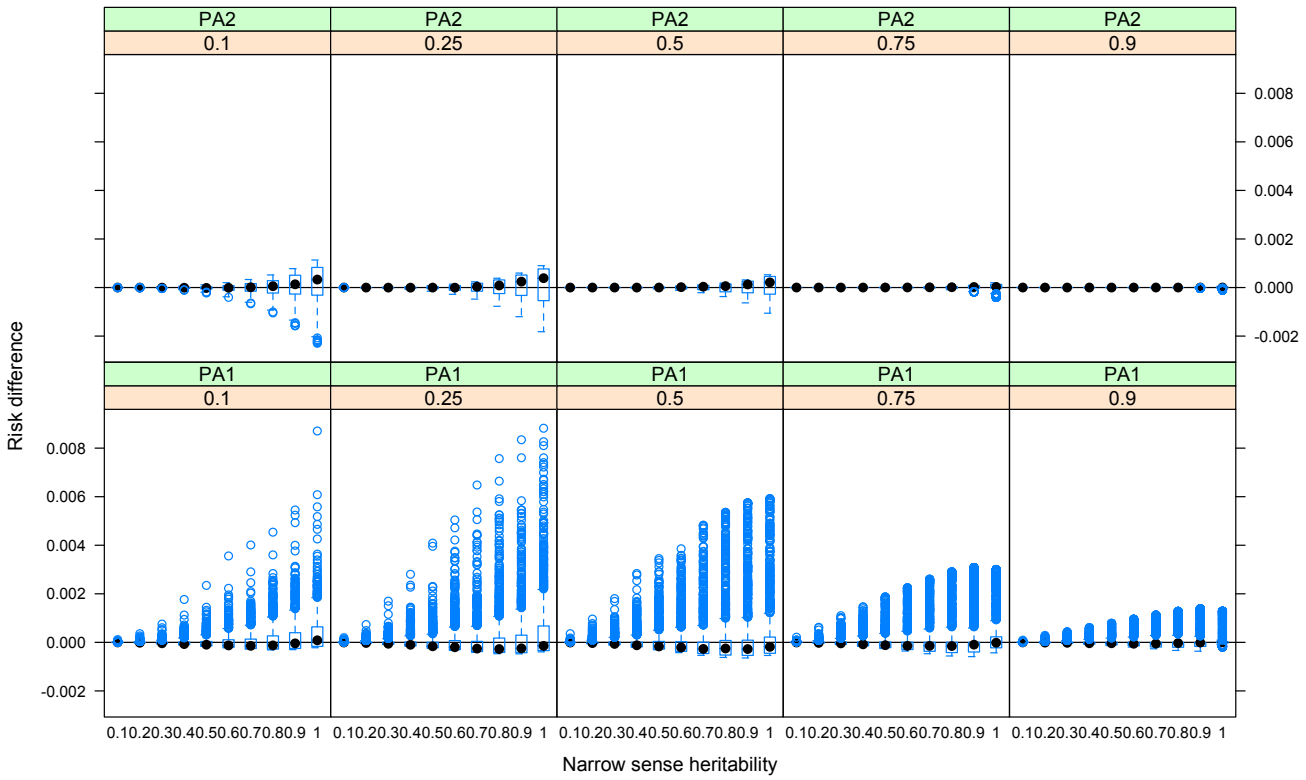

Figure 10: $\{R = 1, Y_R = 0, K = 0.2\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

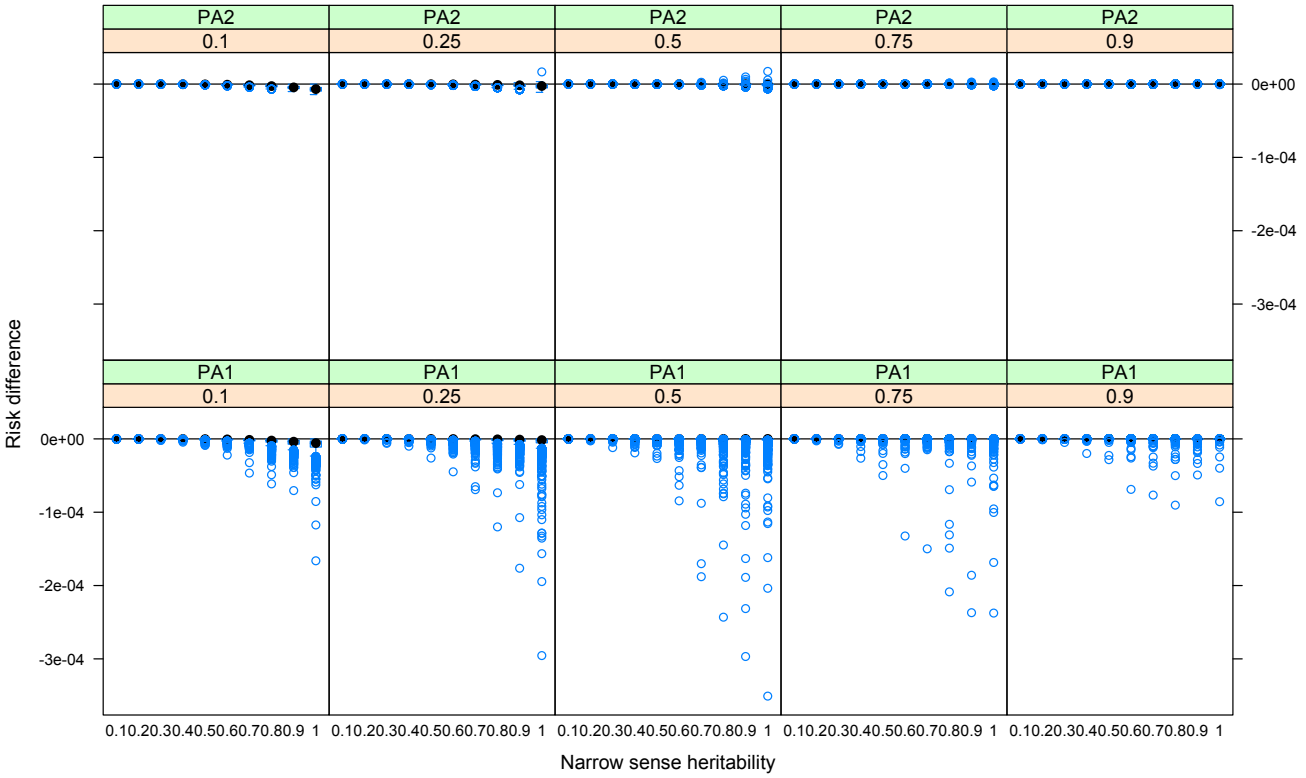

Figure 11: $\{R = 2, Y_R = 1, K = 0.001\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

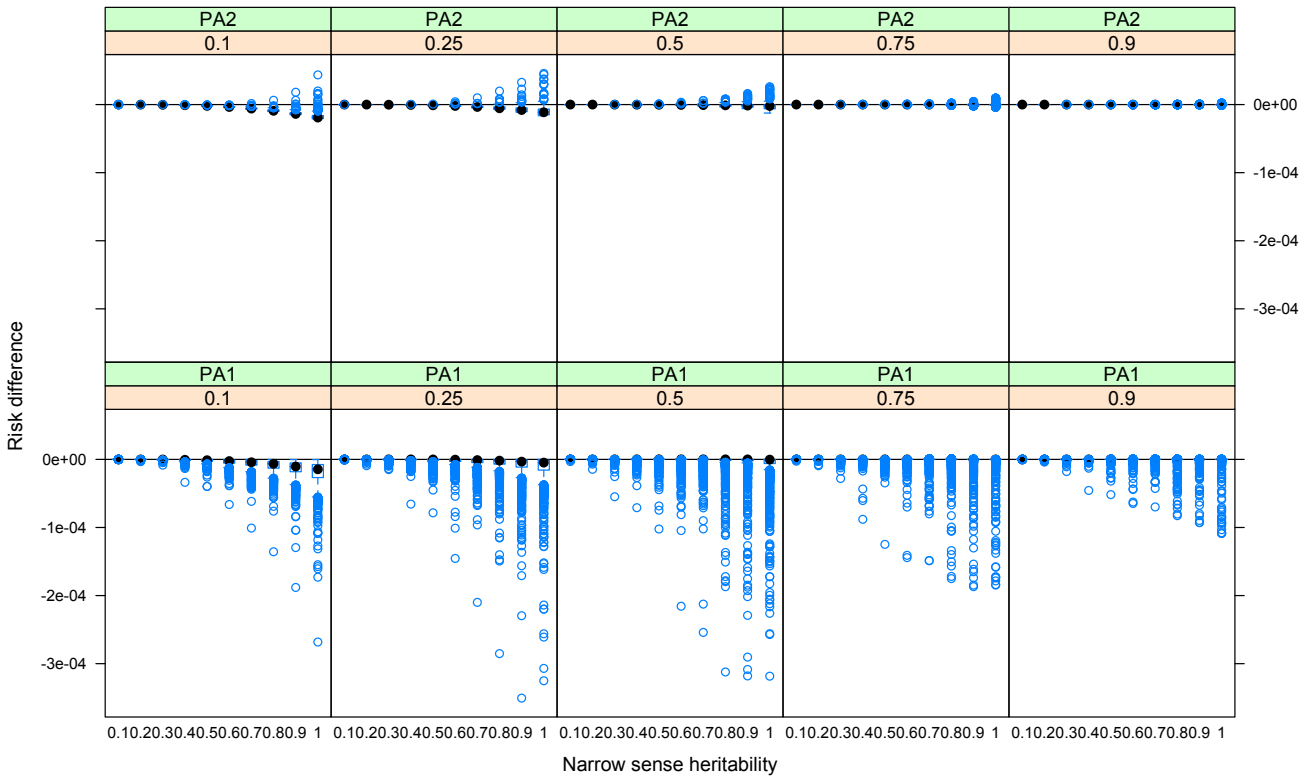

Figure 12: $\{R = 2, Y_R = 1, K = 0.01\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

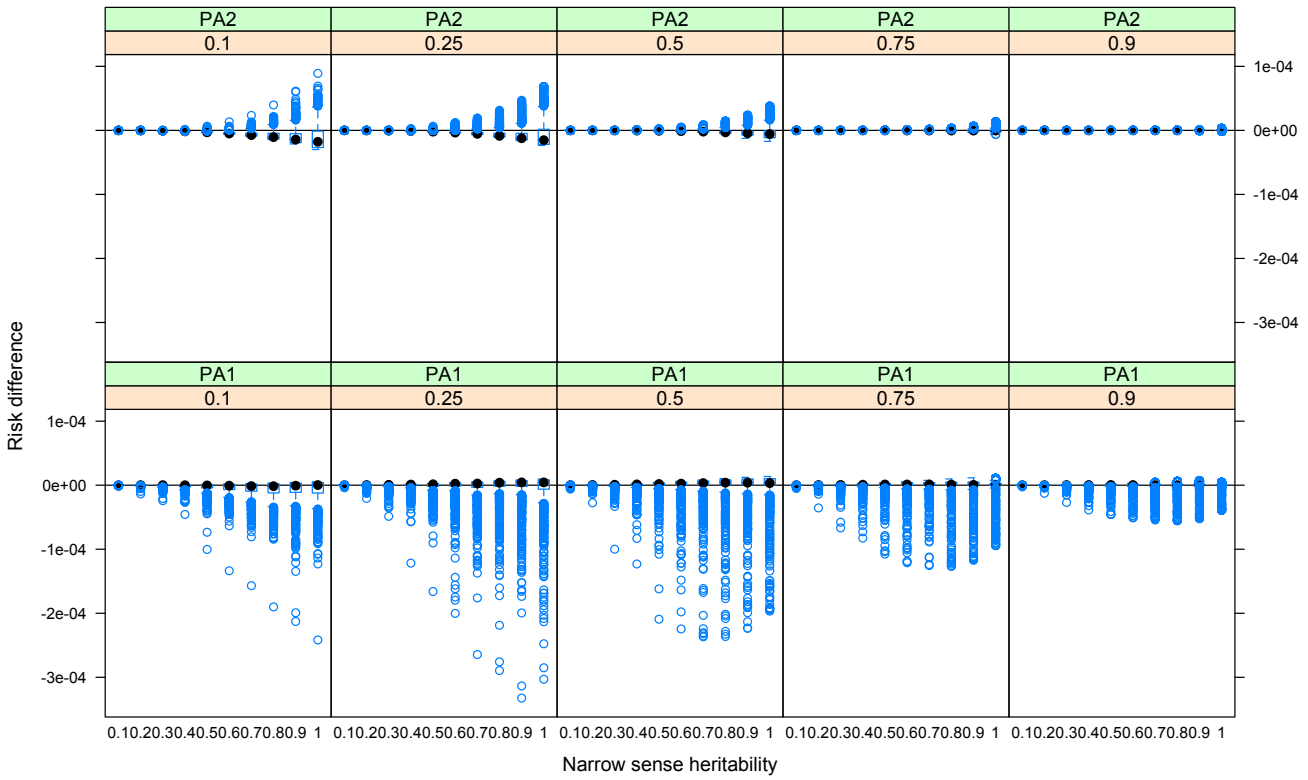

Figure 13: $\{R = 2, Y_R = 1, K = 0.05\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

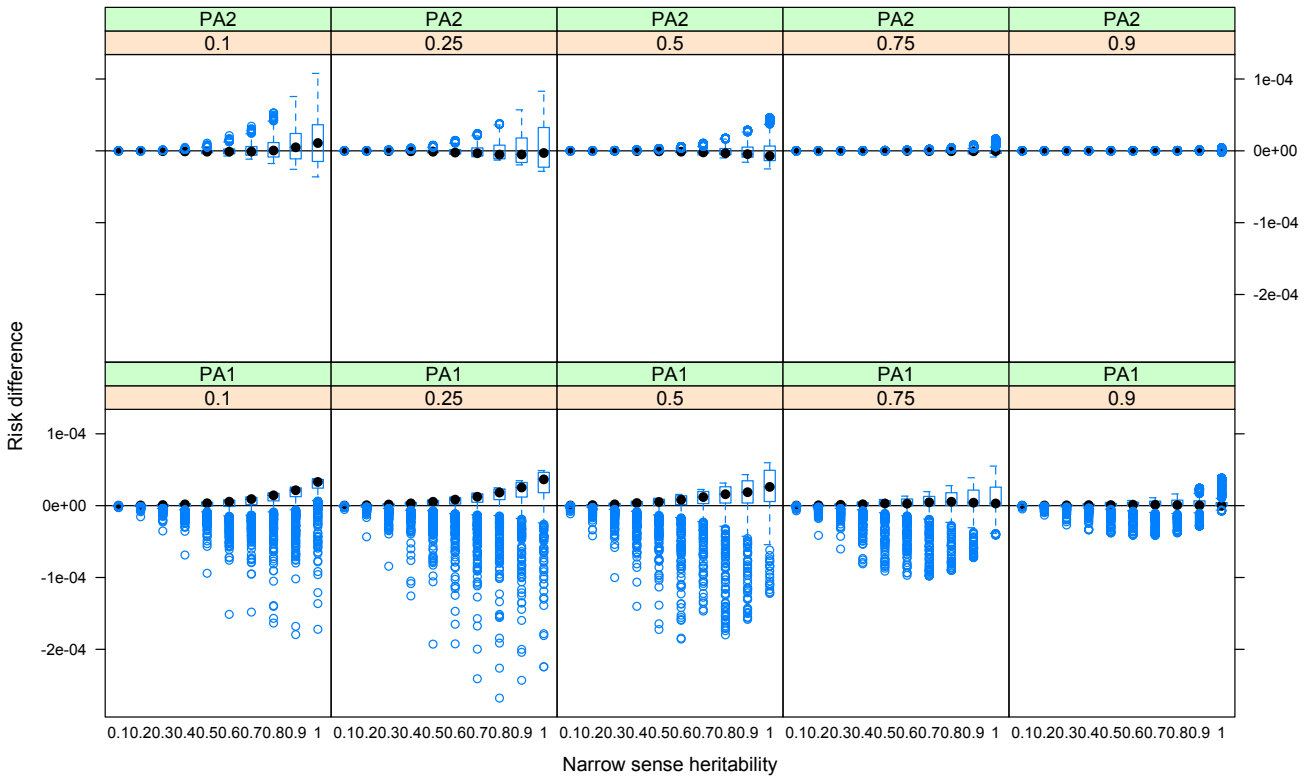

Figure 14: $\{R = 2, Y_R = 1, K = 0.1\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

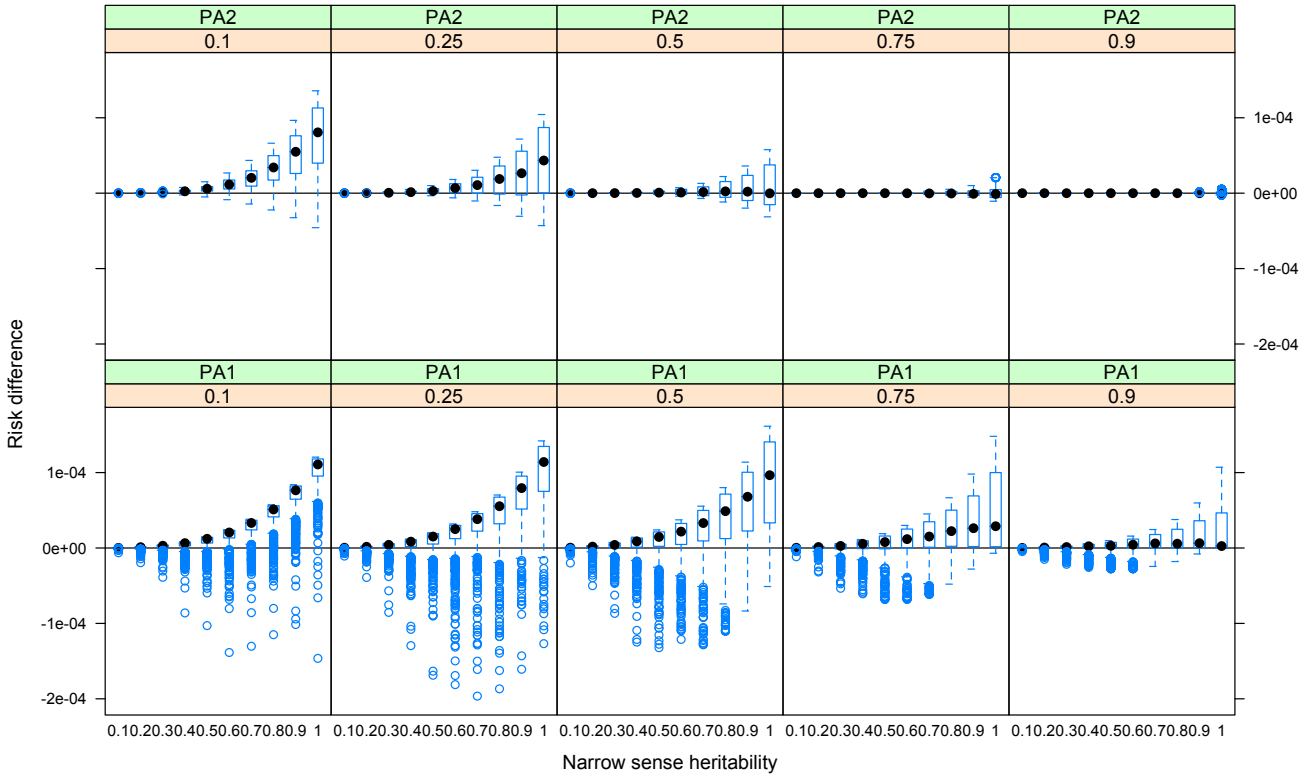

Figure 15: $\{R = 2, Y_R = 1, K = 0.2\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

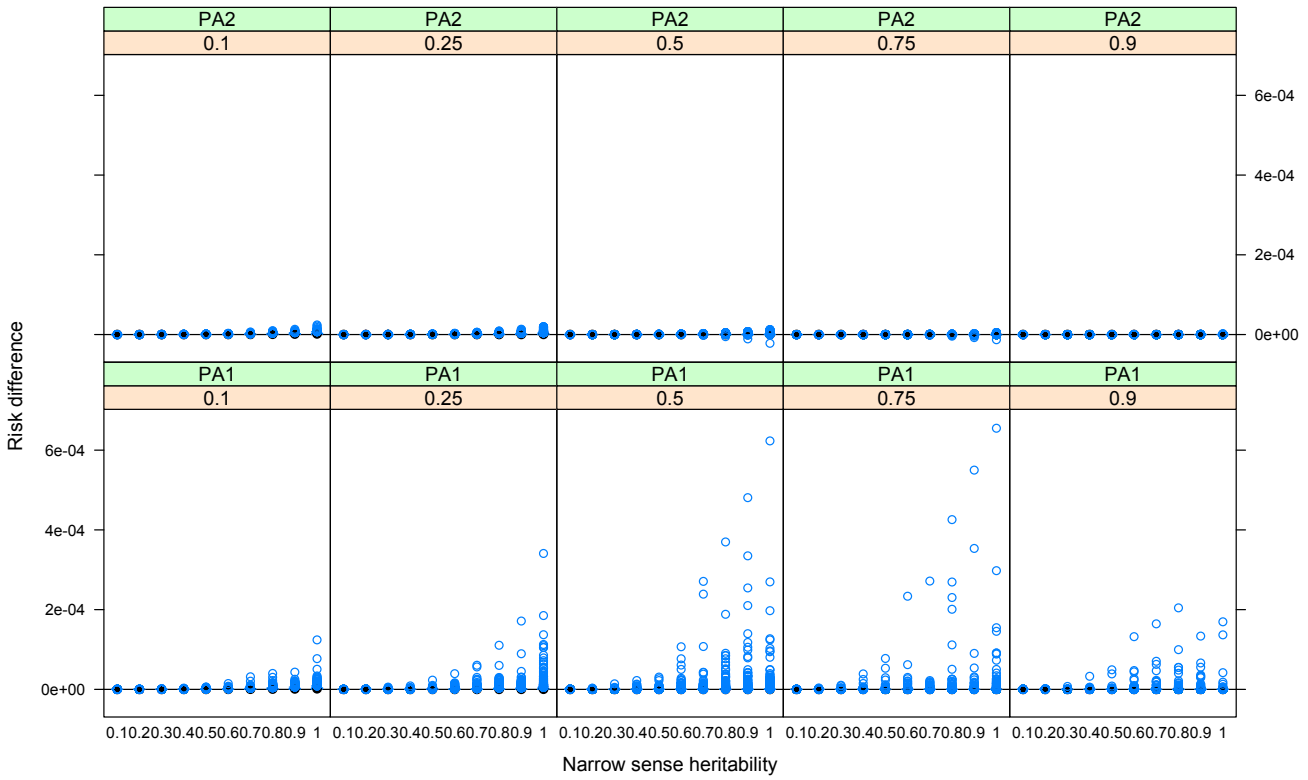

Figure 16: $\{R = 2, Y_R = 0, K = 0.001\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

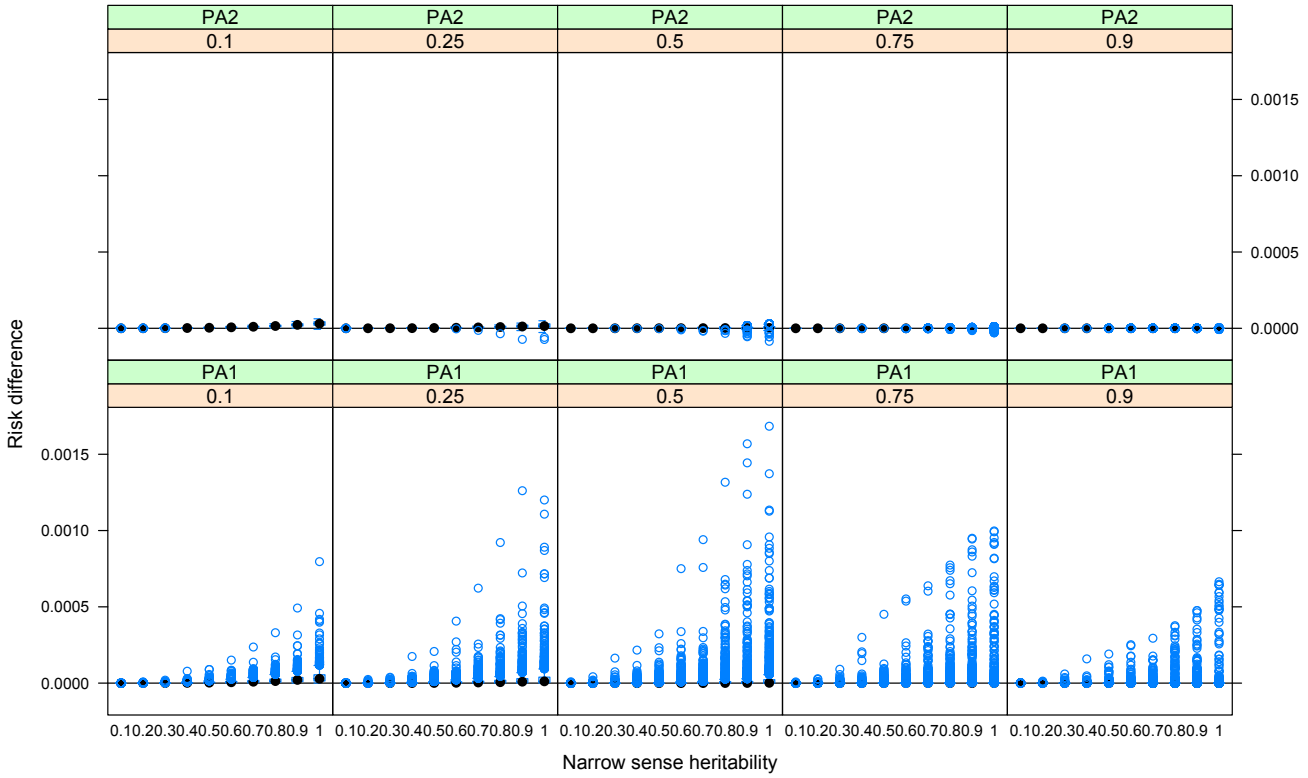

Figure 17: $\{R = 2, Y_R = 0, K = 0.01\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

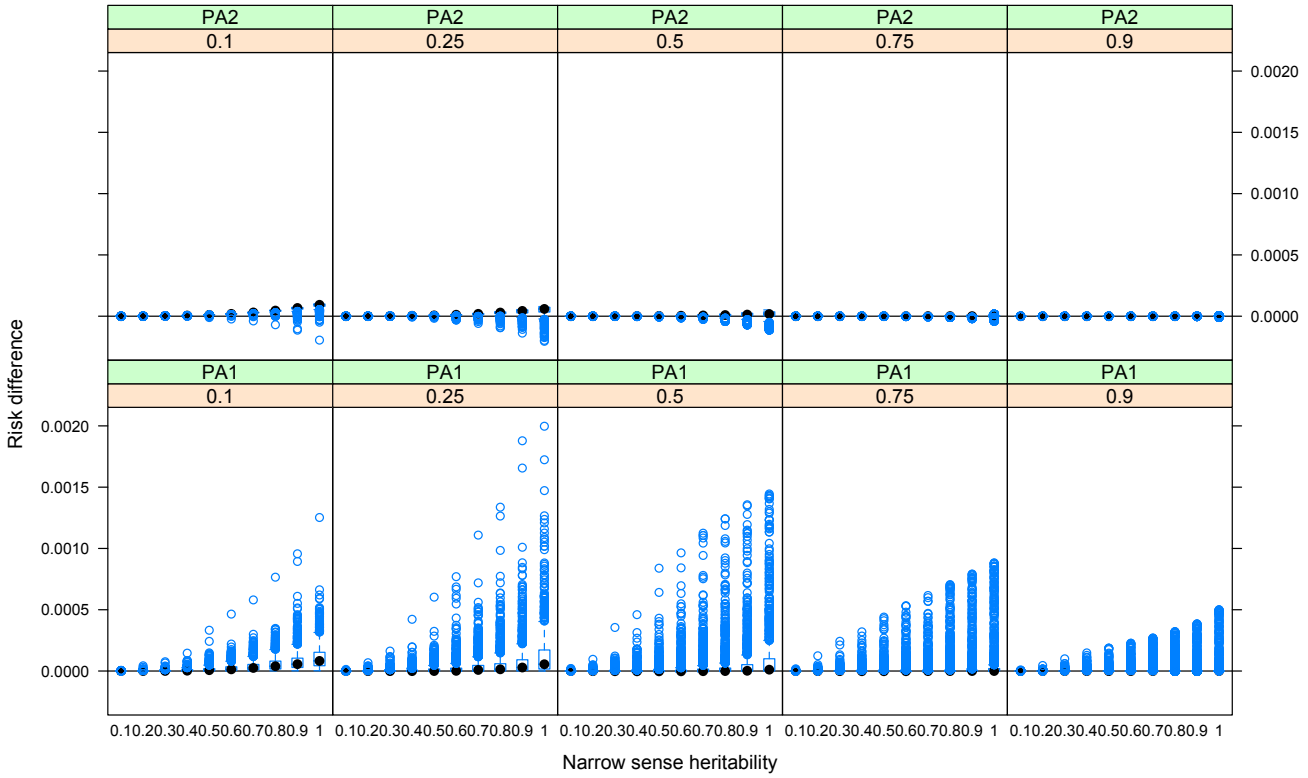

Figure 18: $\{R = 2, Y_R = 0, K = 0.05\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

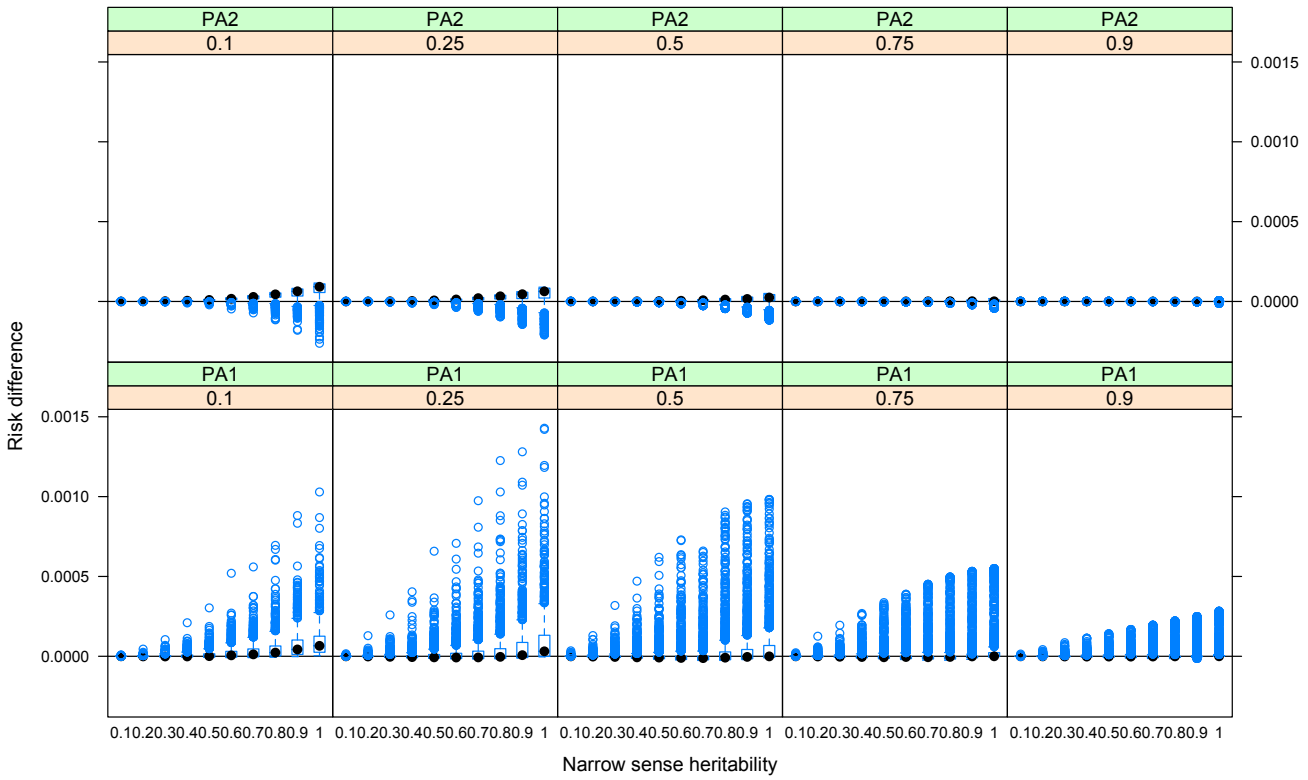

Figure 19: $\{R = 2, Y_R = 0, K = 0.1\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

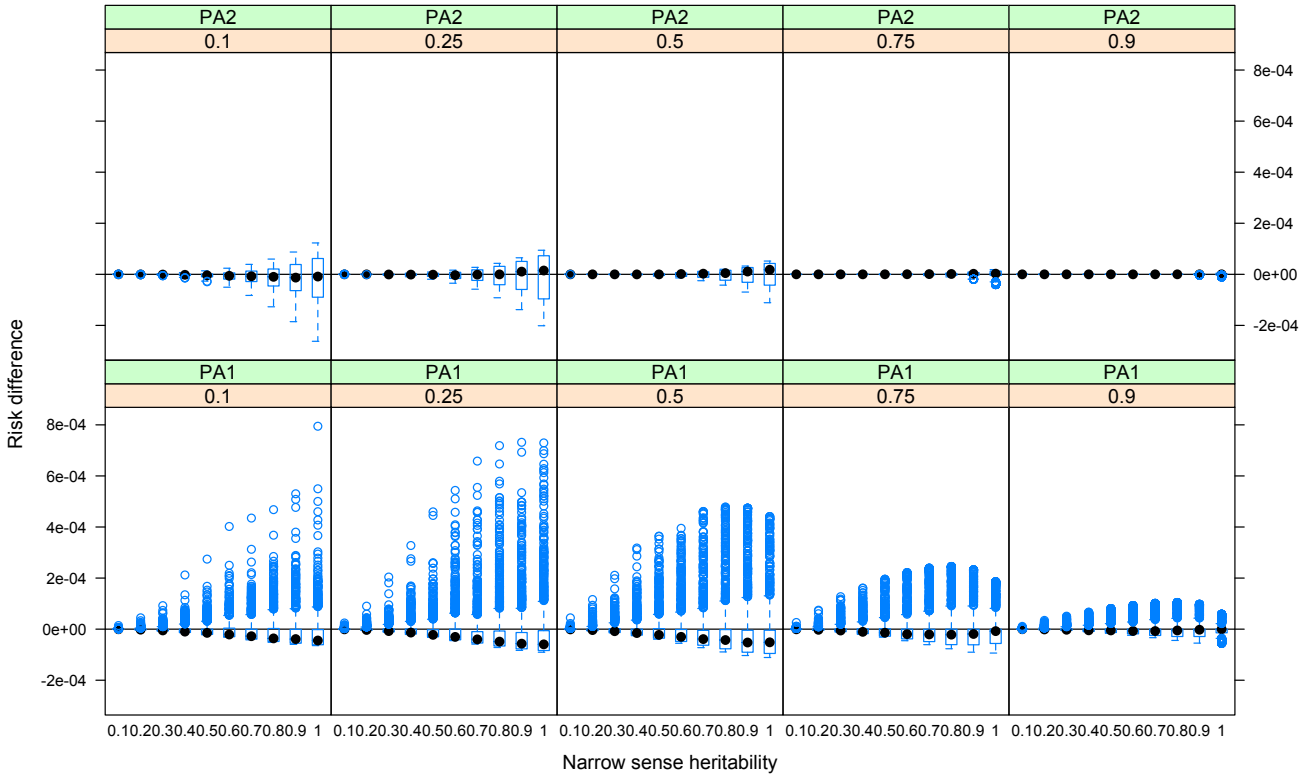

Figure 20: $\{R = 2, Y_R = 0, K = 0.2\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

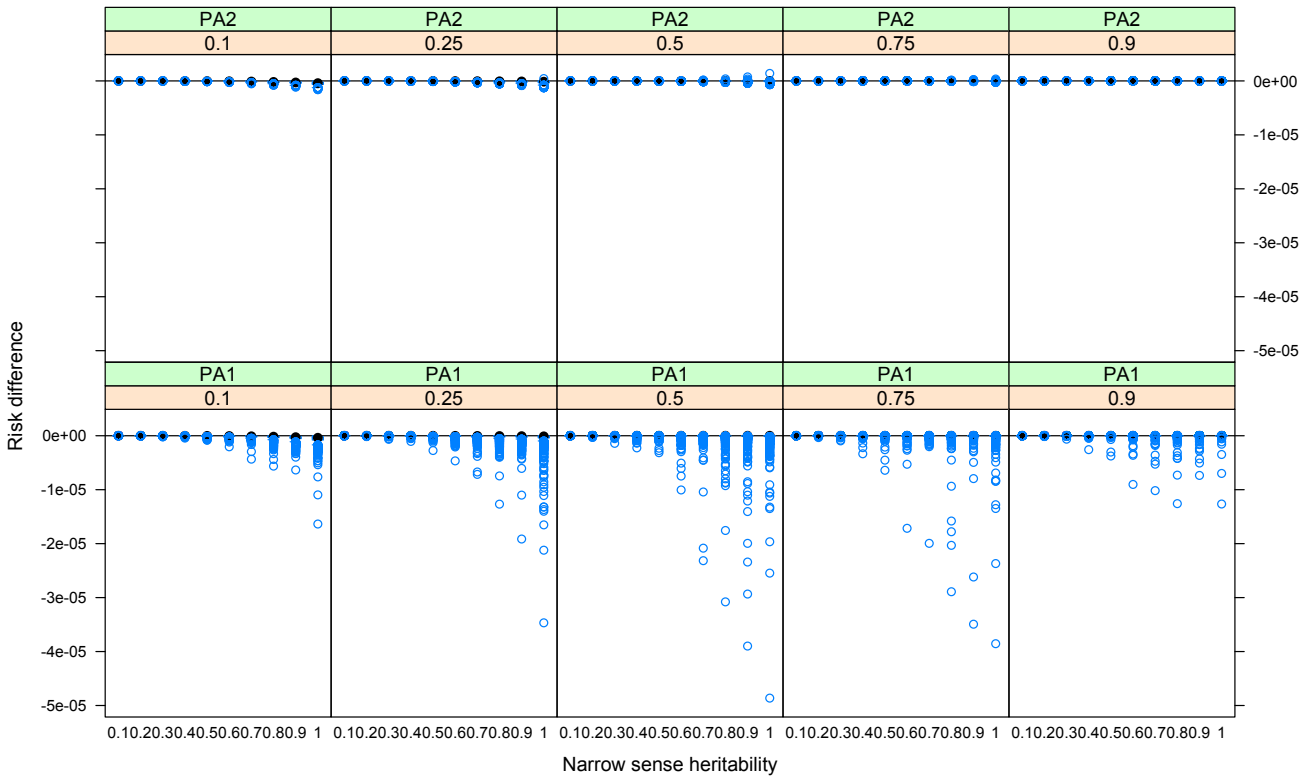

Figure 21: $\{R = 3, Y_R = 1, K = 0.001\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

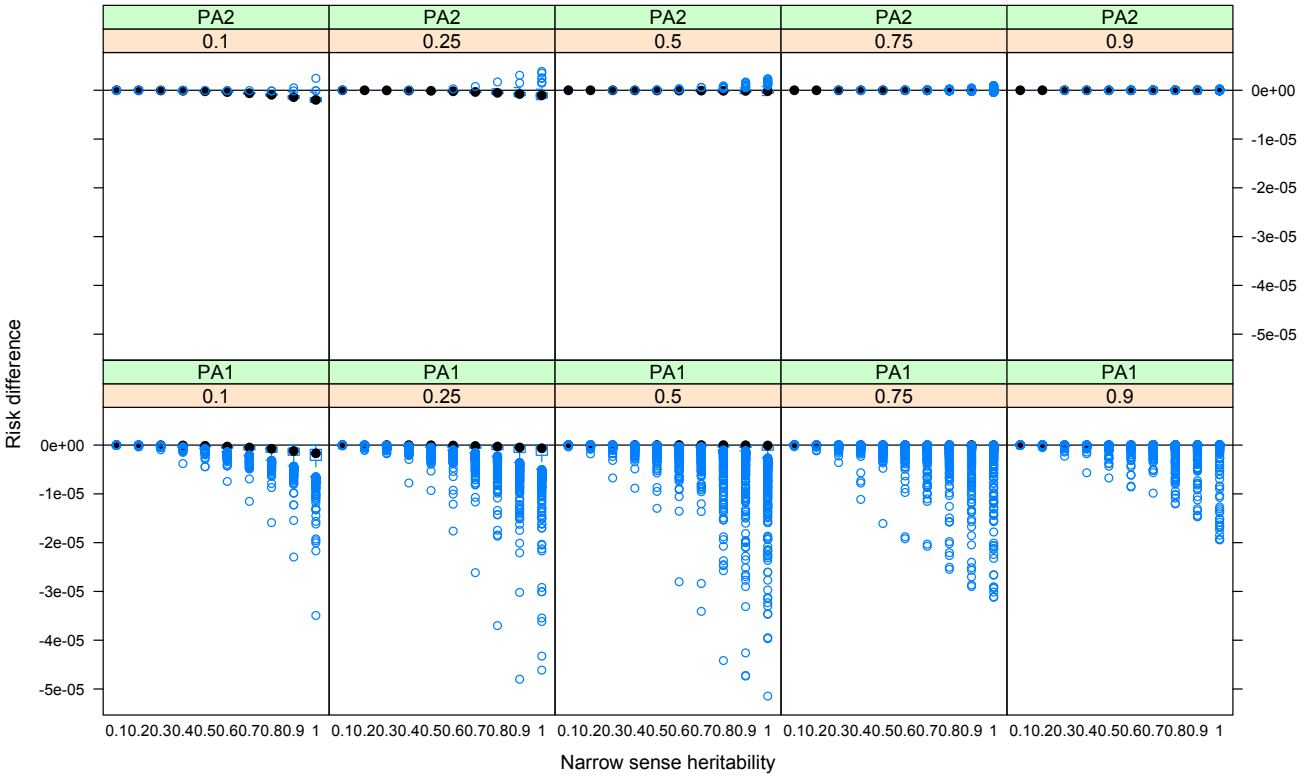

Figure 22: $\{R = 3, Y_R = 1, K = 0.01\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

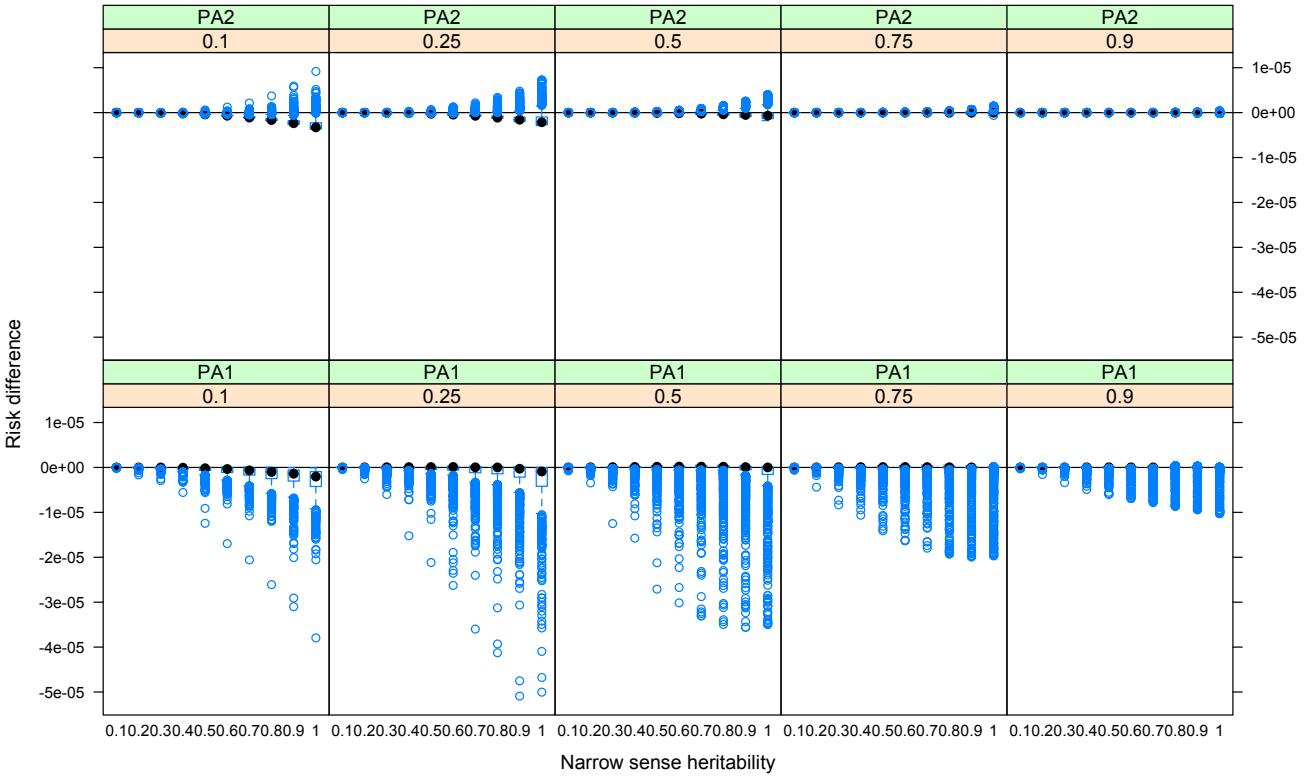

Figure 23: $\{R = 3, Y_R = 1, K = 0.05\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

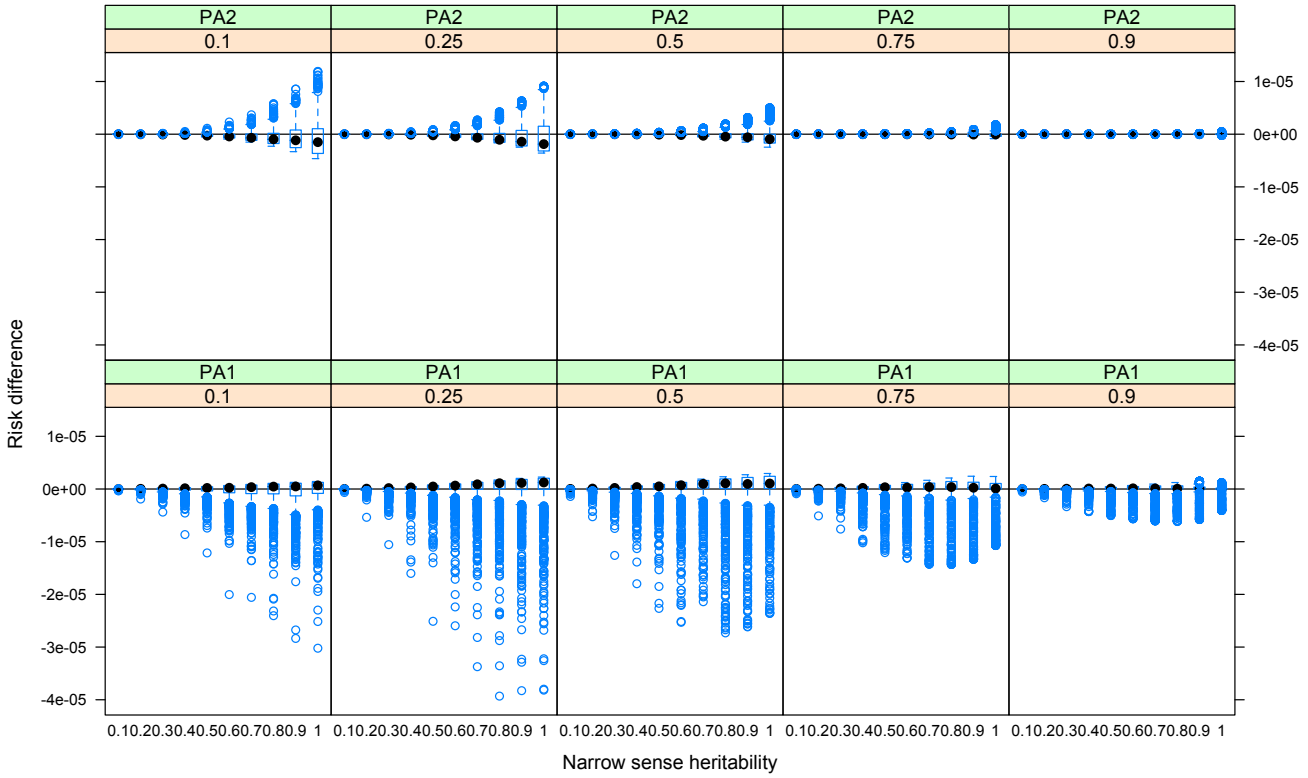

Figure 24: $\{R = 3, Y_R = 1, K = 0.1\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

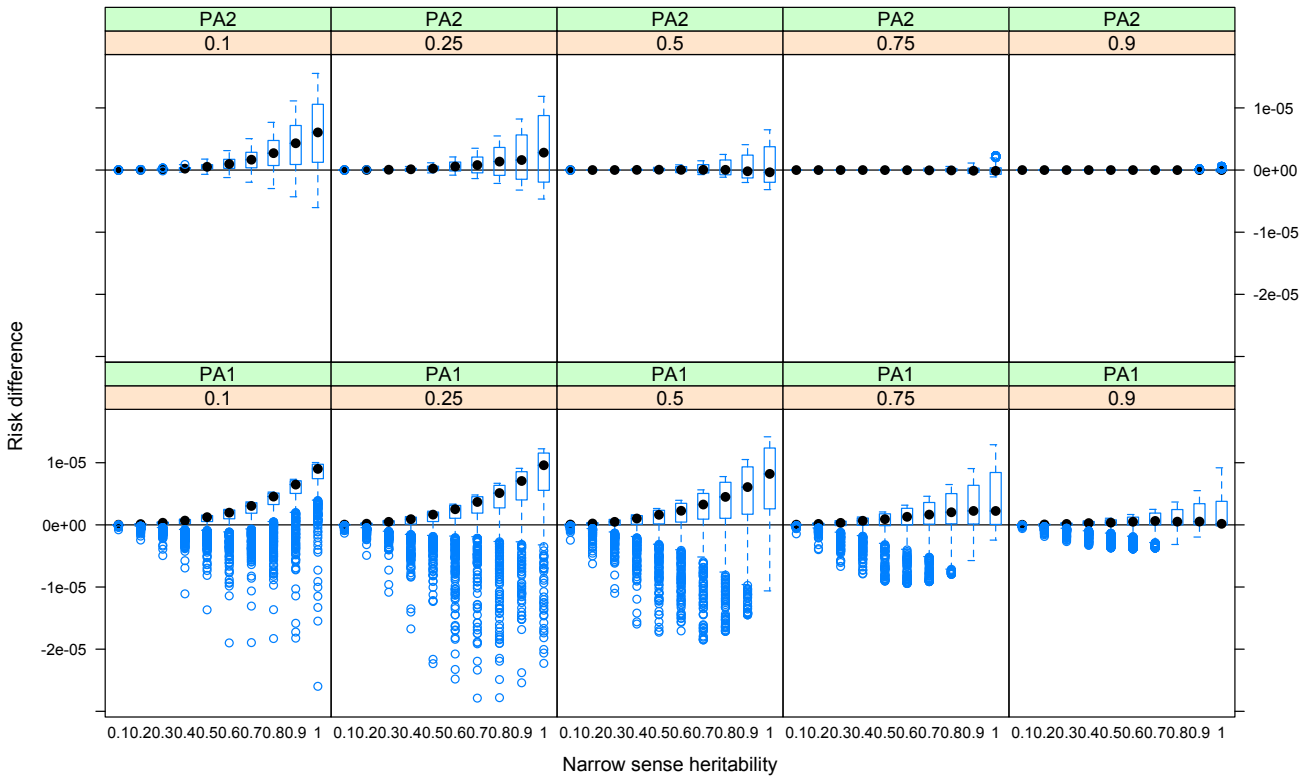

Figure 25: $\{R = 3, Y_R = 1, K = 0.2\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

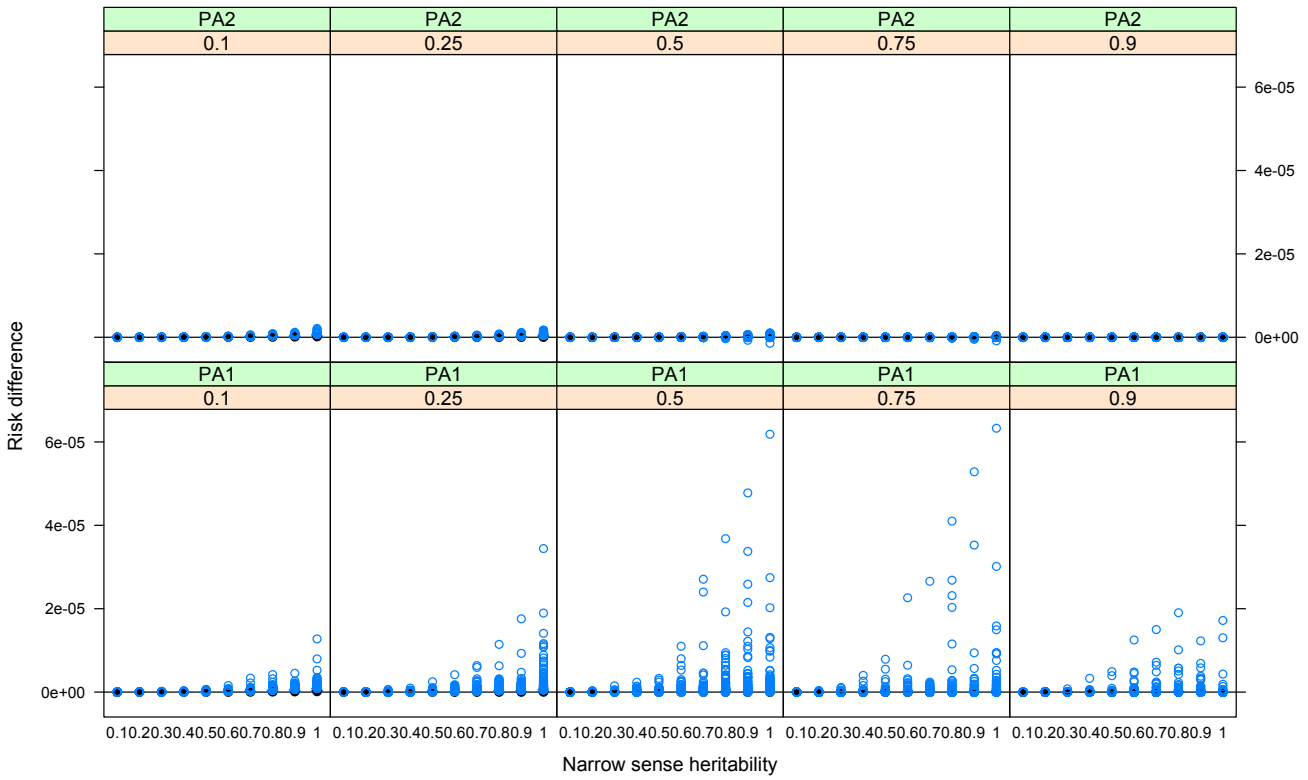

Figure 26: $\{R = 3, Y_R = 0, K = 0.001\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

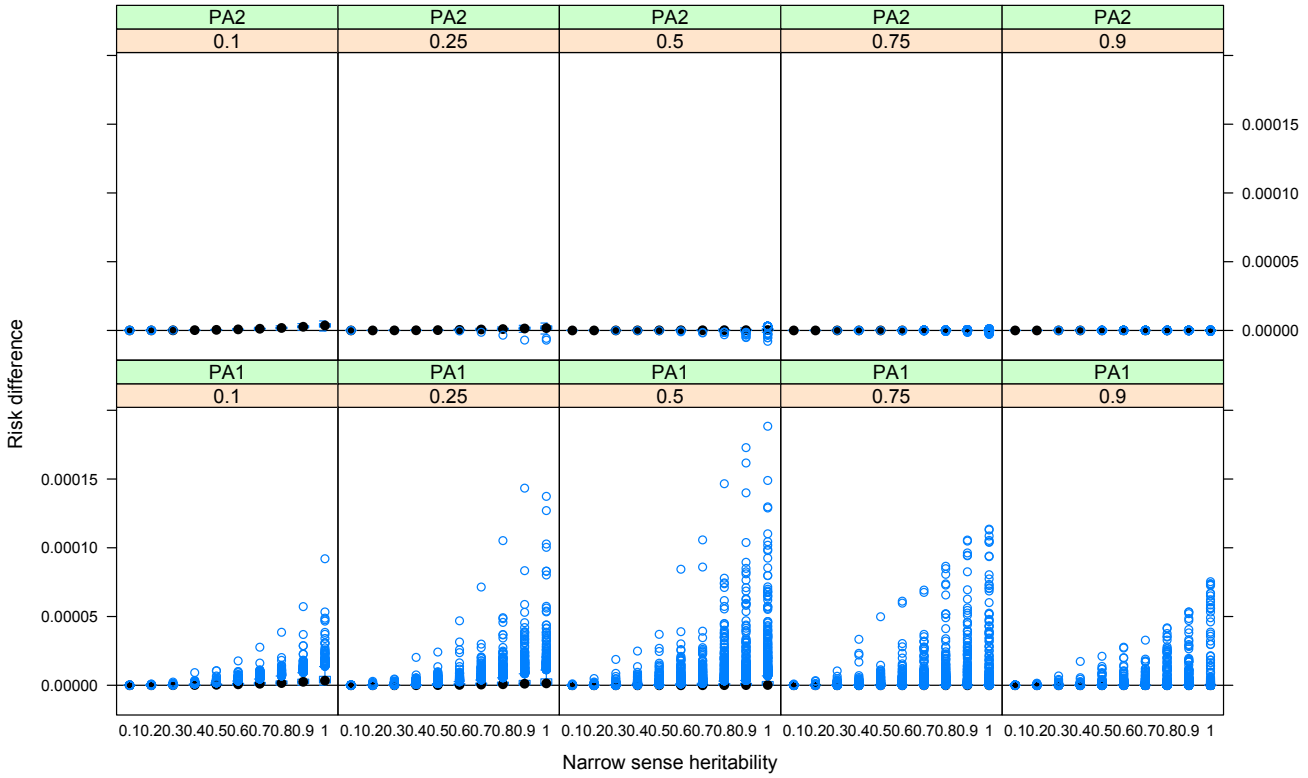

Figure 27: $\{R = 3, Y_R = 0, K = 0.01\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

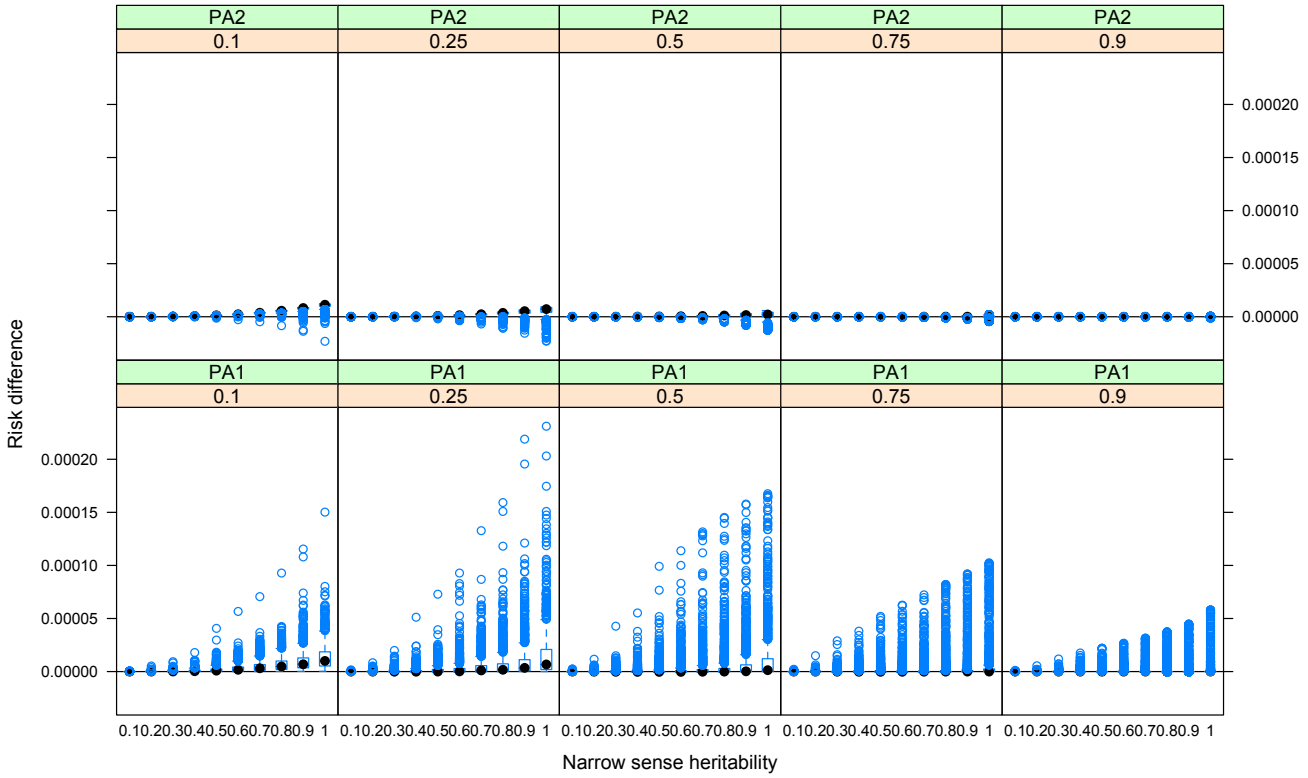

Figure 28: $\{R = 3, Y_R = 0, K = 0.05\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h^2_L}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

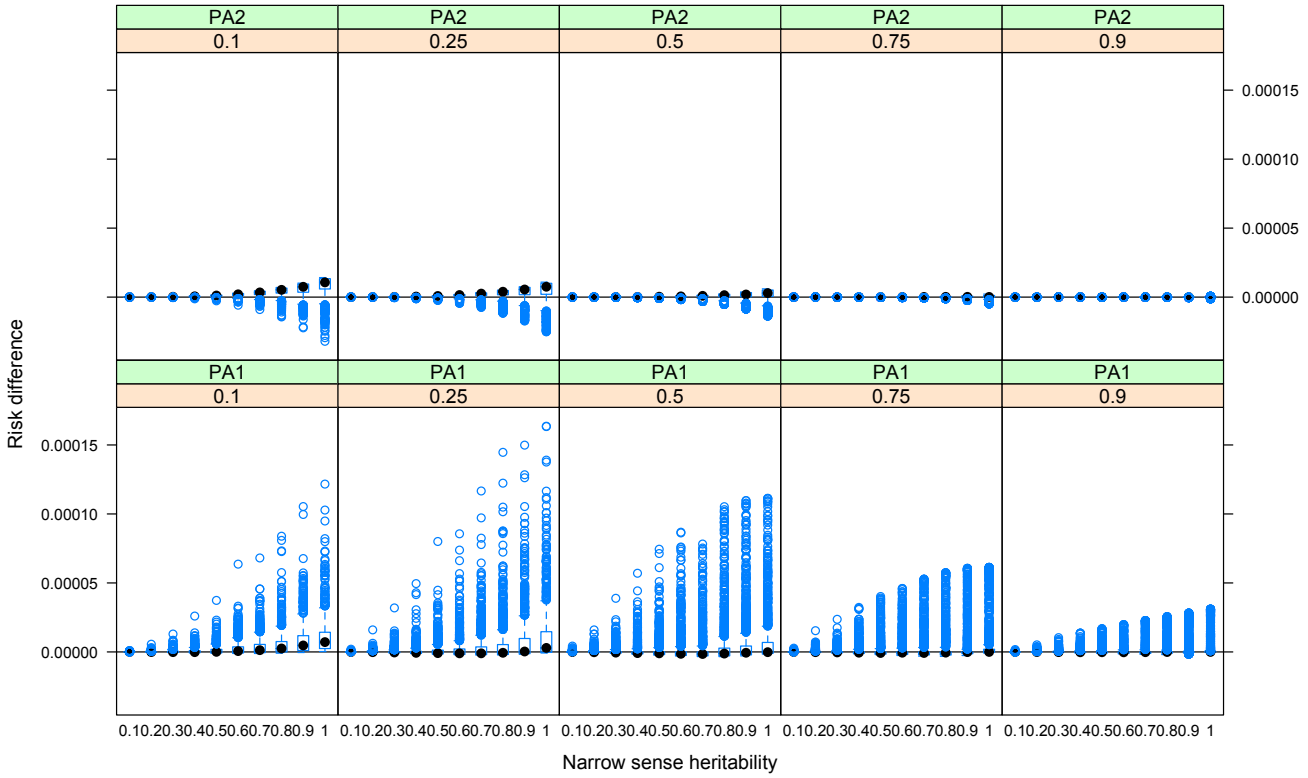

Figure 29: $\{R = 3, Y_R = 0, K = 0.1\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

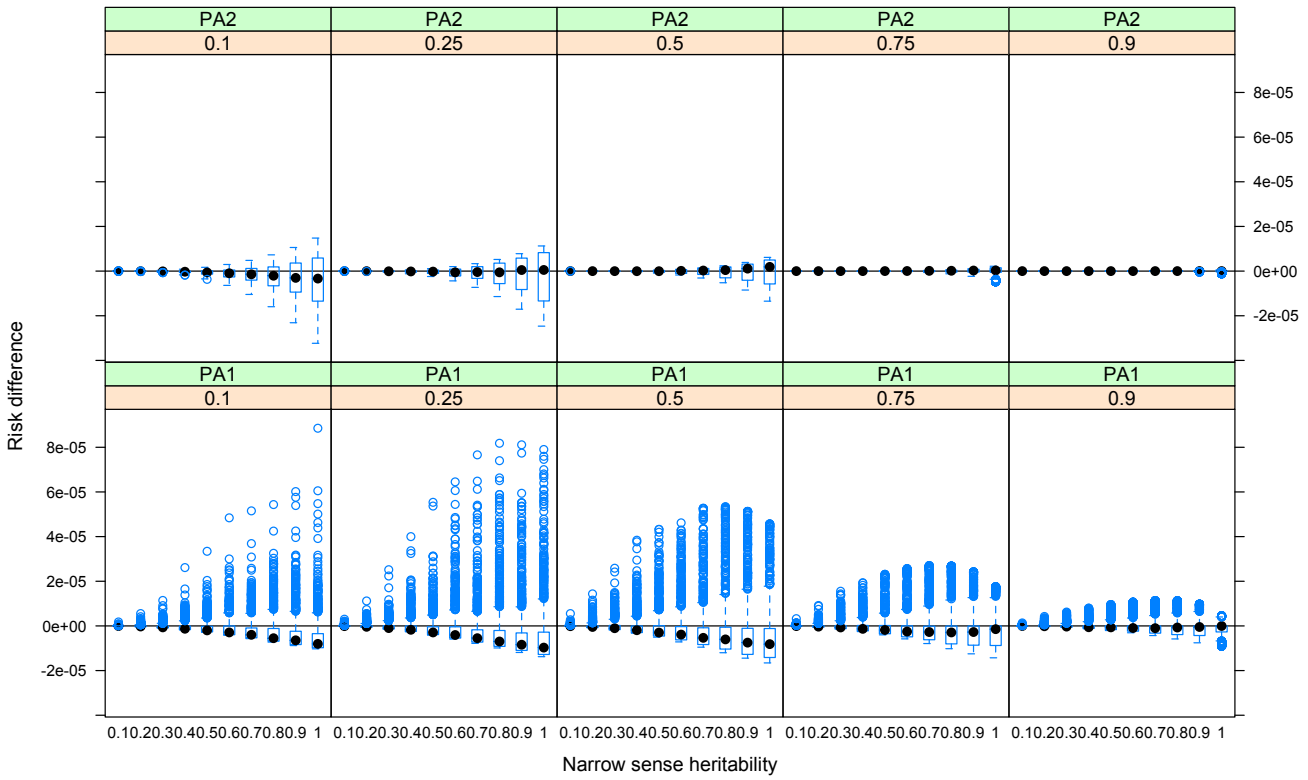

Figure 30: $\{R = 3, Y_R = 0, K = 0.2\}$ PA simulation 1 results. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. $\rho = \frac{V_M}{h_L^2}$; column 1 = $\{\rho = 0.1\}$, column 2 = $\{\rho = 0.25\}$, column 3 = $\{\rho = 0.5\}$, column 4 = $\{\rho = 0.75\}$, and, column 5 = $\{\rho = 0.9\}$.

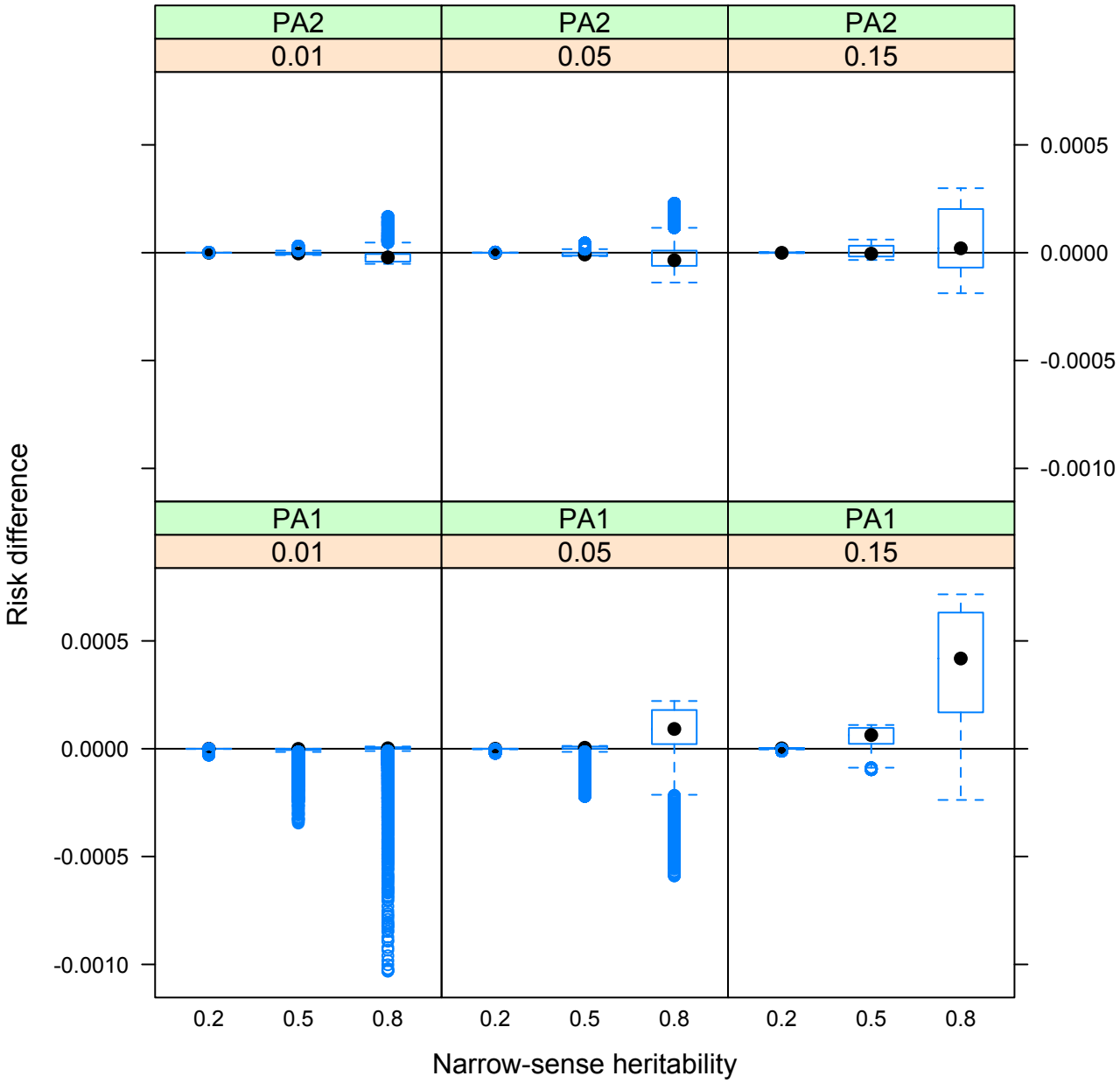

Figure 31: PA simulation 2 results; 5 variables, type 1 with $\sigma^2_{error} = 1$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

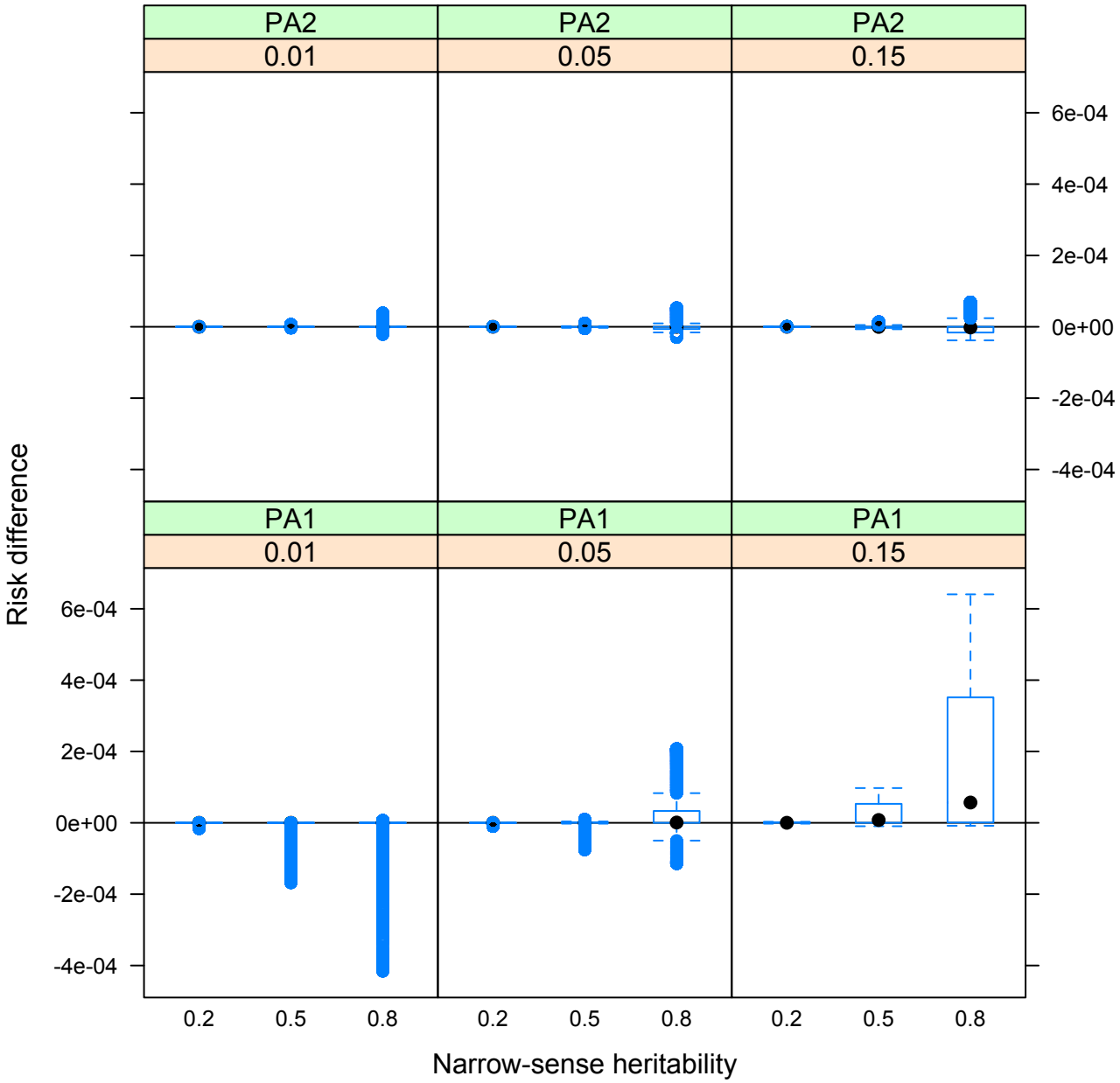

Figure 32: PA simulation 2 results; 5 variables, type 1 with $\sigma^2_{error} = 0.2$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

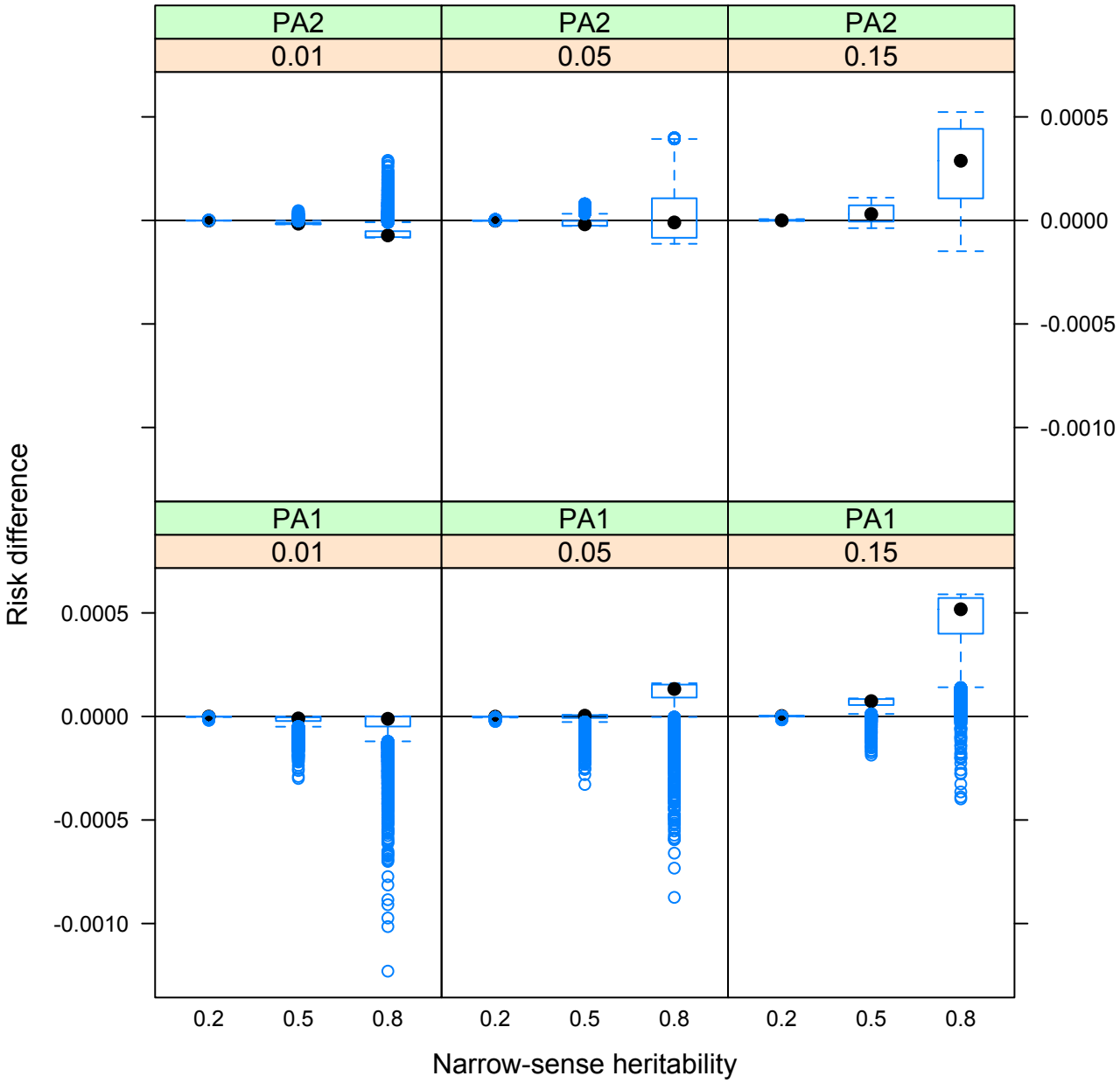

Figure 33: PA simulation 2 results; 5 variables, type 2 with $\sigma^2_{error} = 1$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

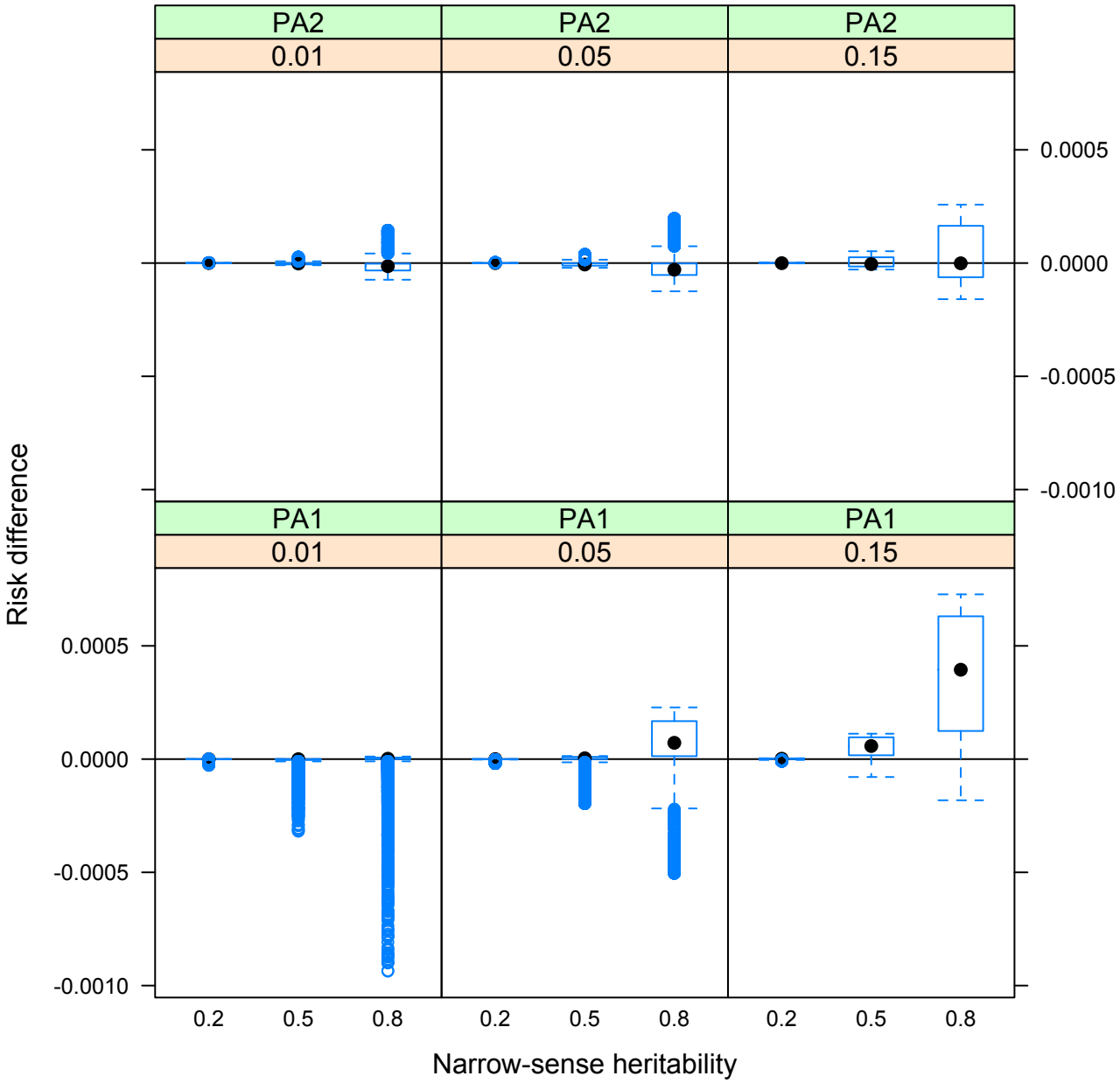

Figure 34: PA simulation 2 results; 5 variables, type 2 with $\sigma^2_{error} = 0.2$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

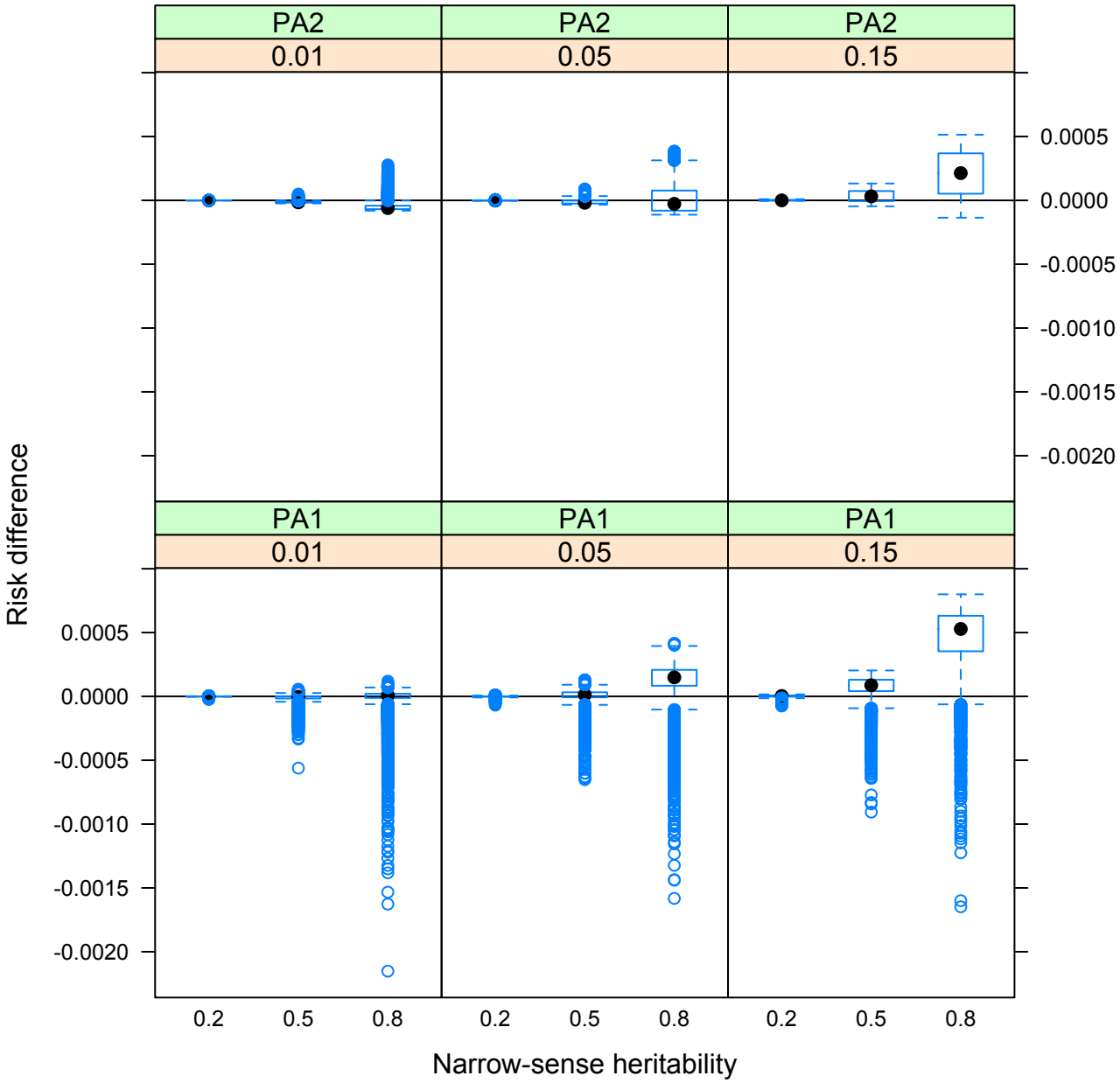

Figure 35: PA simulation 2 results; 5 variables, type 3 with $\sigma^2_{error} = 1$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

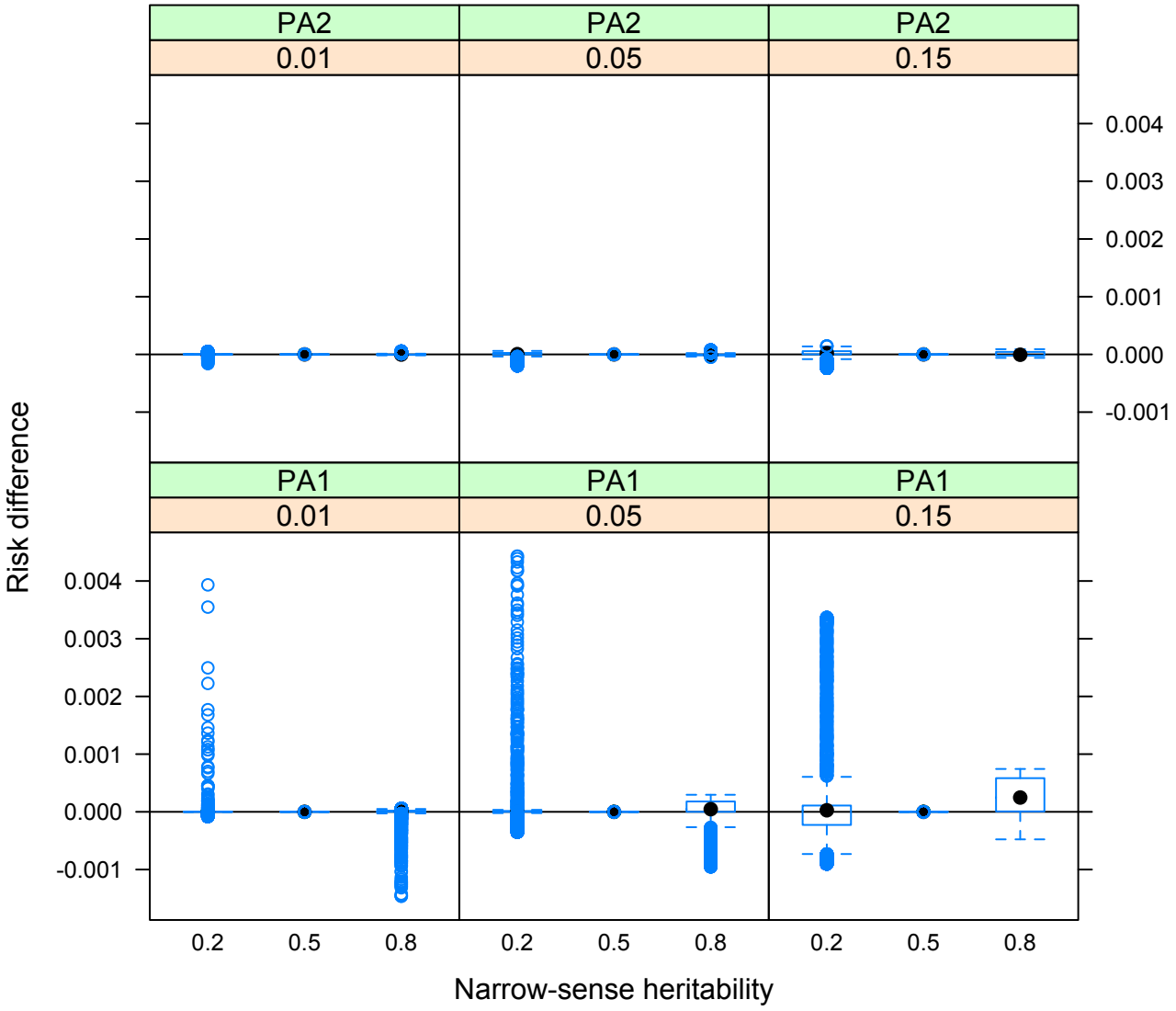

Figure 36: PA simulation 2 results; 5 variables, type 3 with $\sigma^2_{error} = 0.2$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

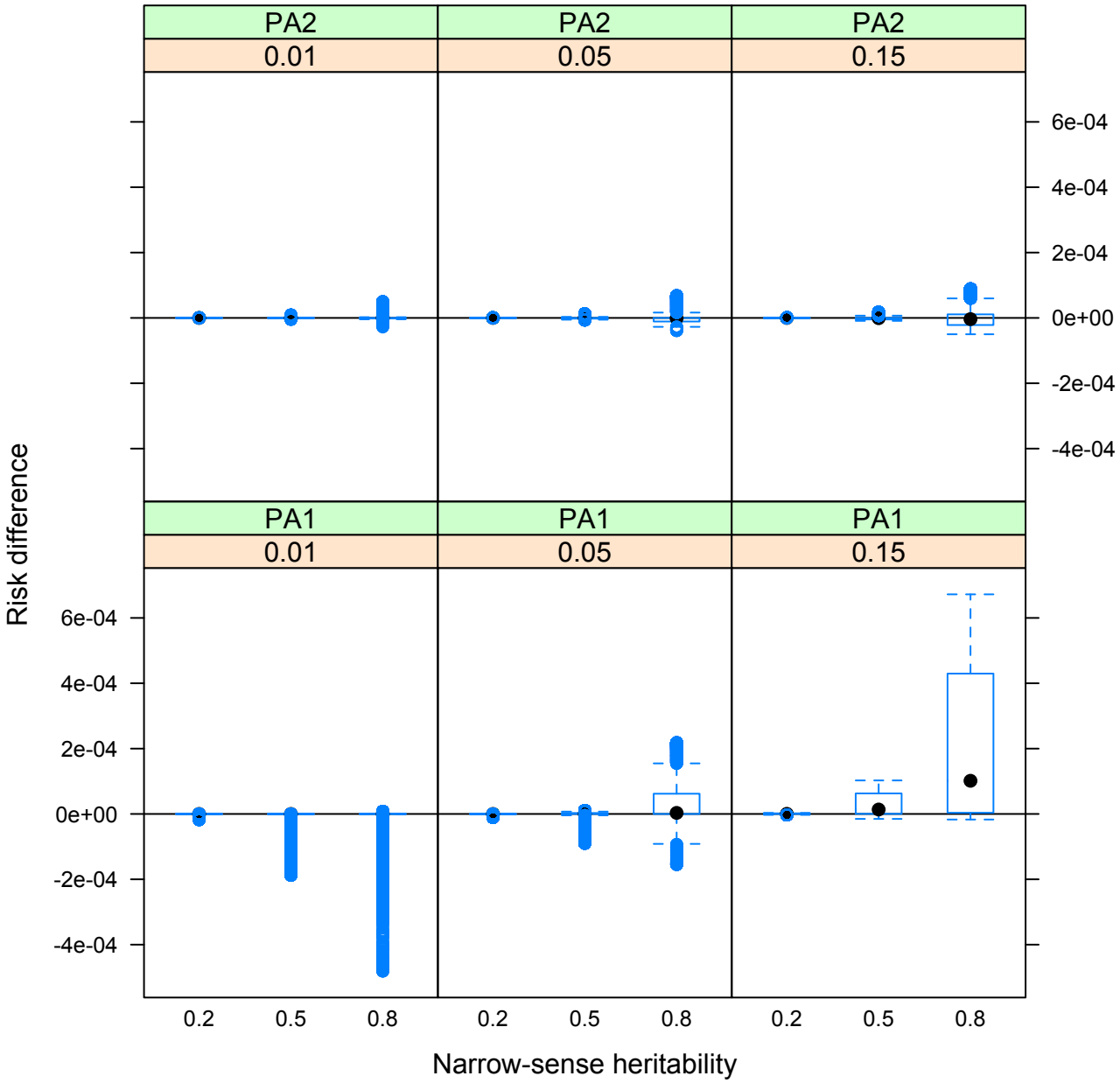

Figure 37: PA simulation 2 results; 5 variables, type 4 with $\sigma^2_{error} = 1$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

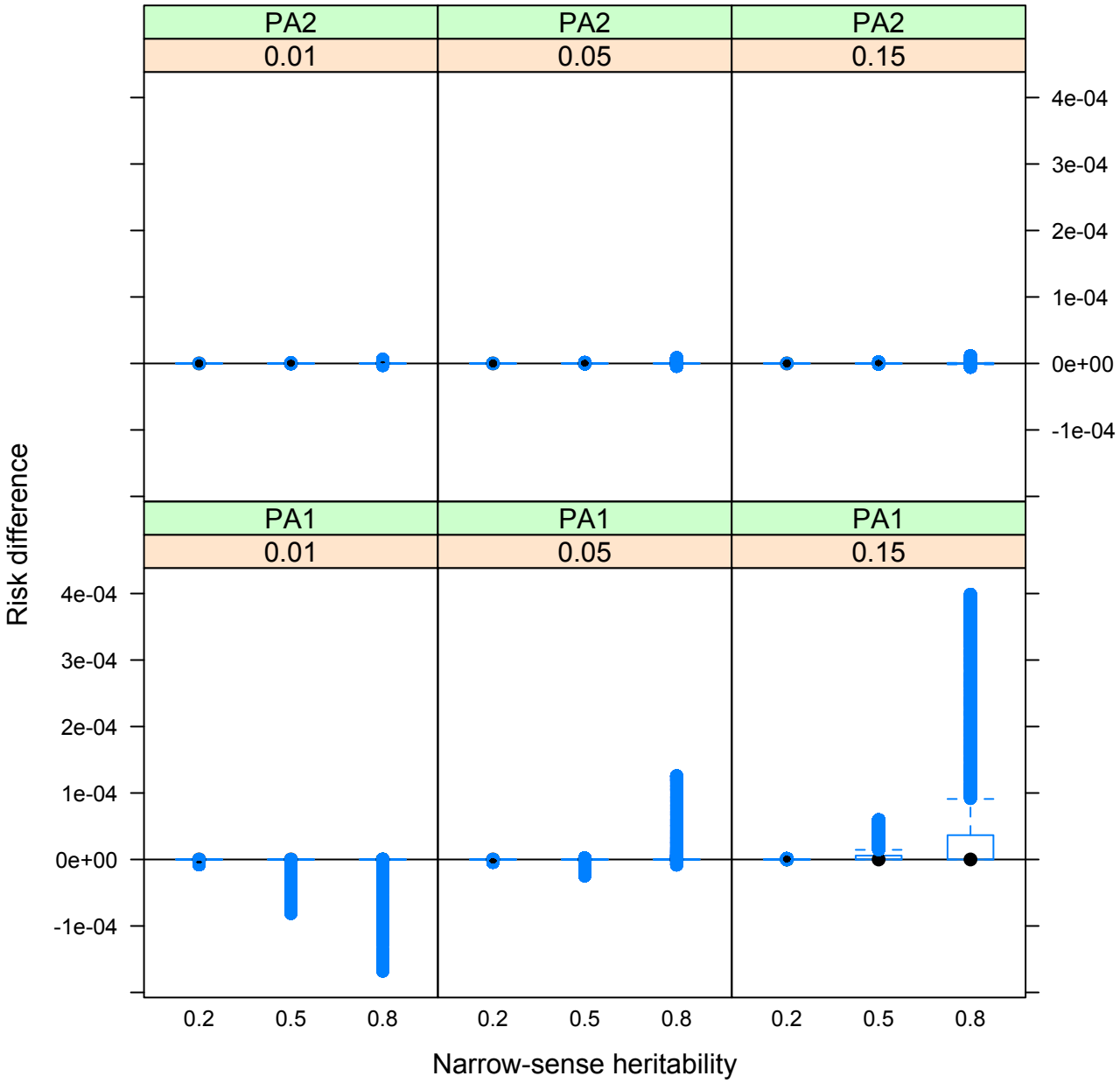

Figure 38: PA simulation 2 results; 5 variables, type 4 with $\sigma^2_{error} = 0.2$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

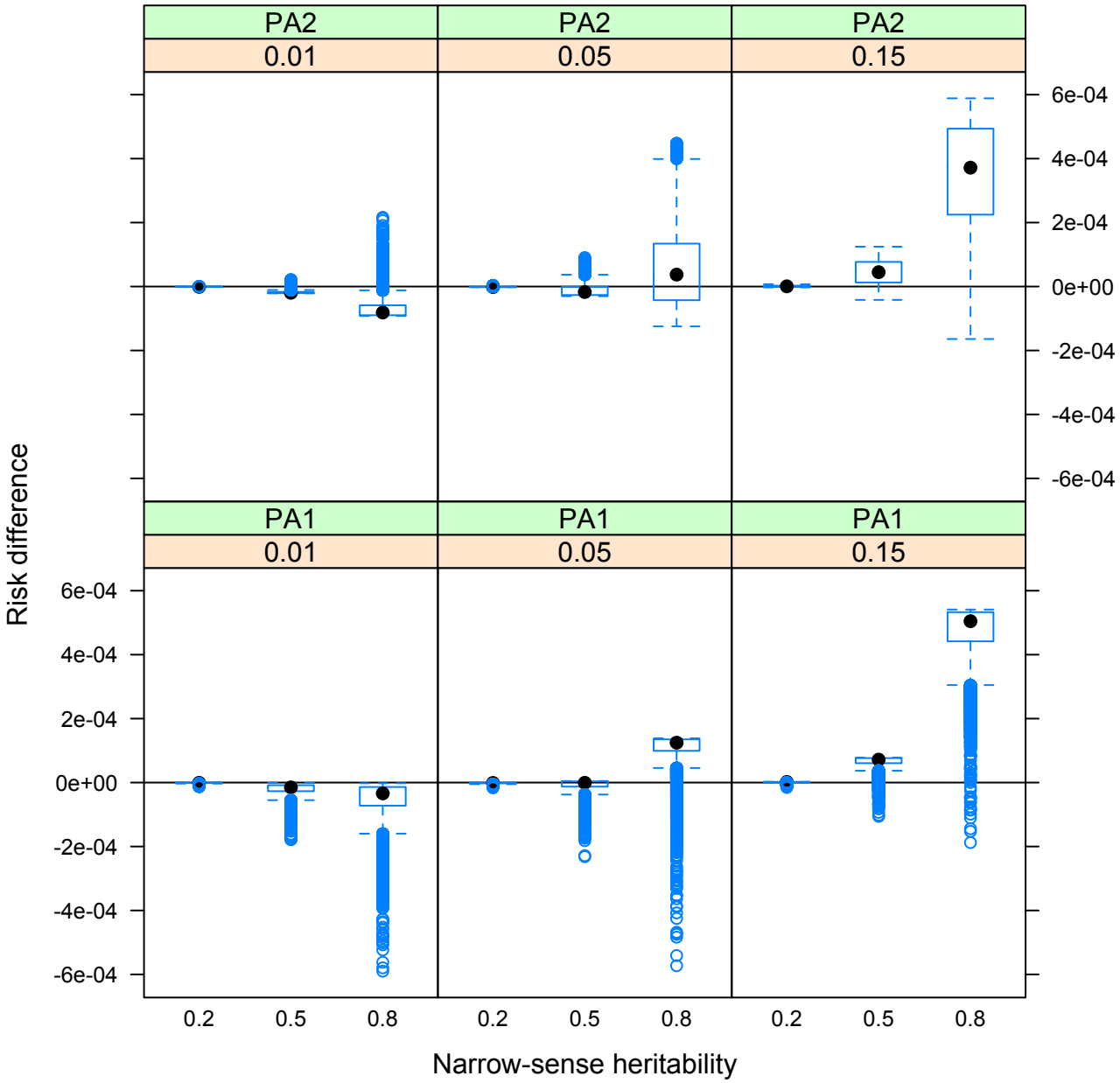

Figure 39: PA simulation 2 results; 10 variables, type 1 with $\sigma^2_{error} = 1$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

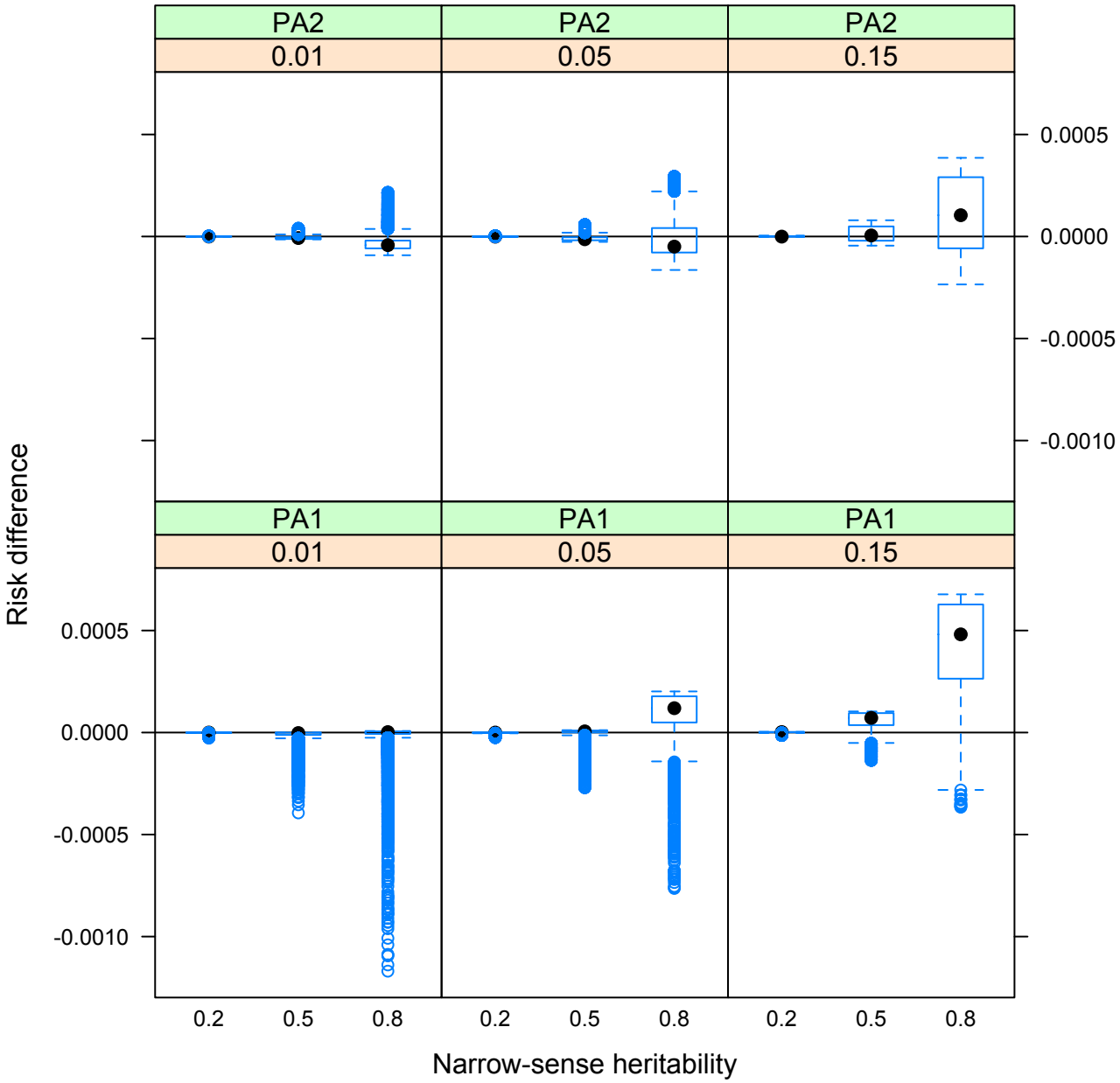

Figure 40: PA simulation 2 results; 10 variables, type 1 with $\sigma^2_{error} = 0.2$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

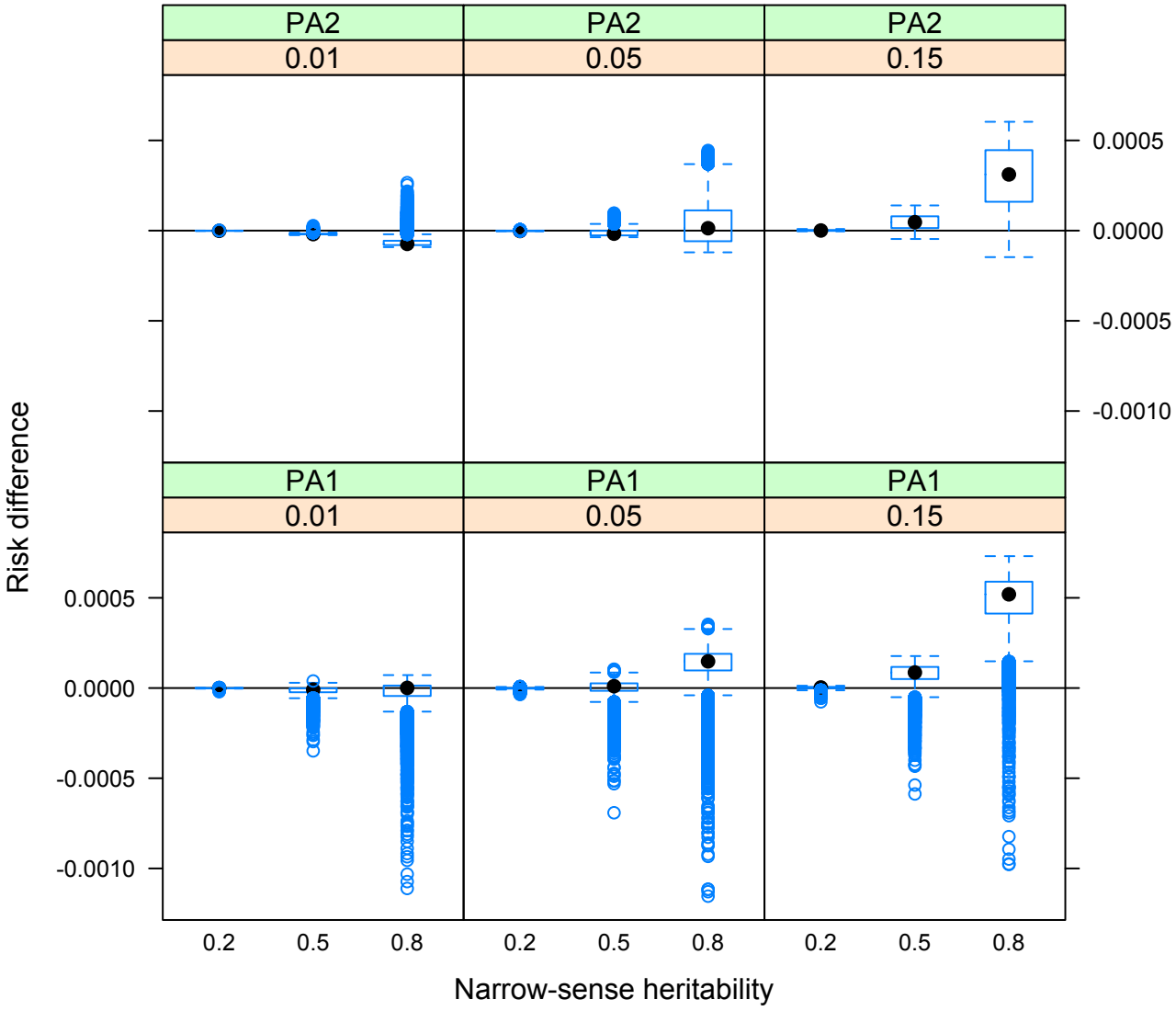

Figure 41: PA simulation 2 results; 10 variables, type 2 with $\sigma^2_{error} = 1$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

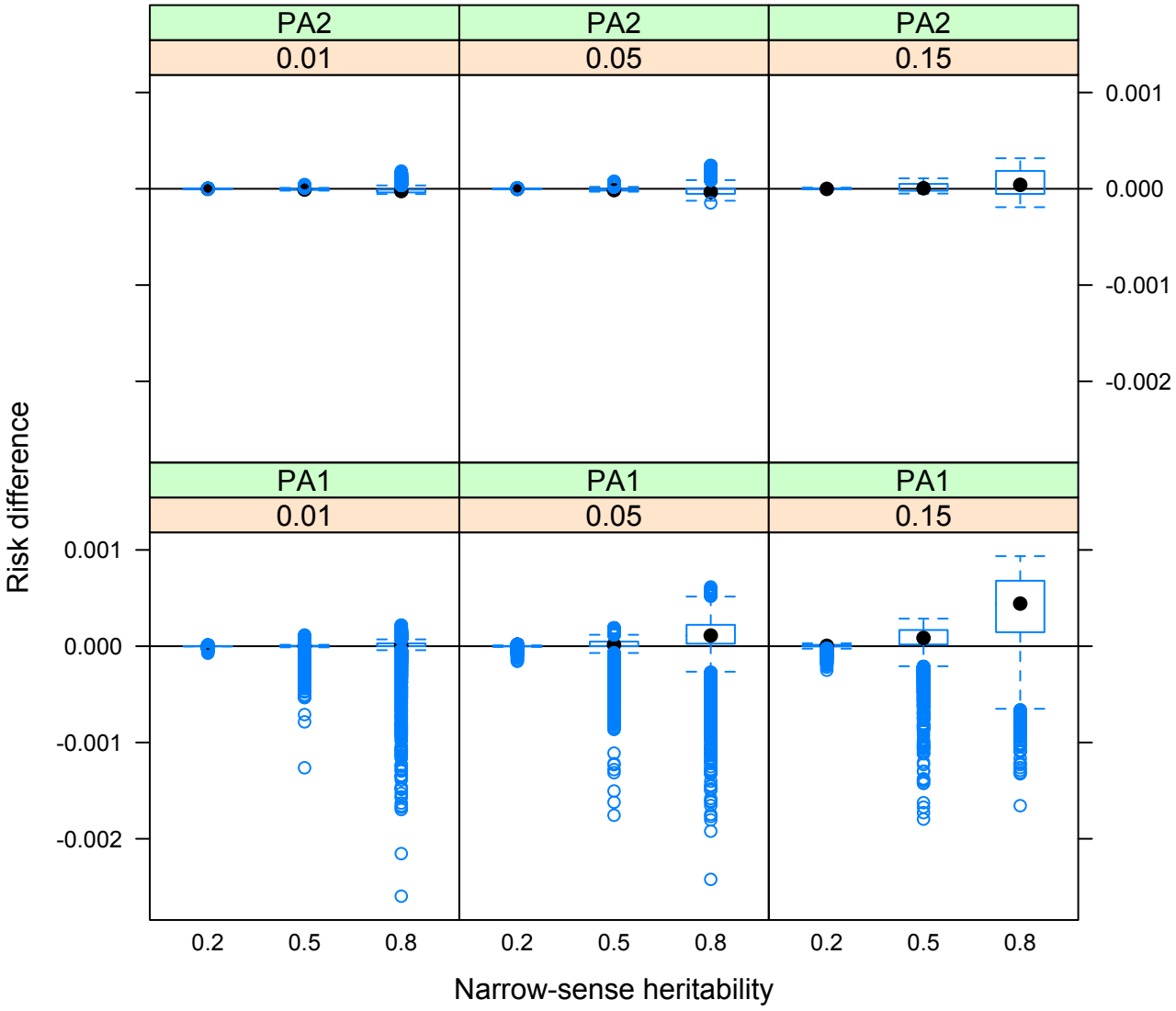

Figure 42: PA simulation 2 results; 10 variables, type 2 with $\sigma^2_{error} = 0.2$. Box-plot of risk difference (PA method risk minus exact method risk) against narrow-sense heritability by: 1. PA method; top row = $PA2$ and bottom row = $PA1$, and, 2. K; column 1 = $\{K = 0.01\}$, column 2 = $\{K = 0.05\}$, and, column 3 = $\{K = 0.15\}$

2 Risk estimation via the LTMM

2.1 ★ LTMM with 1 major locus and S relatives

In Section 3.3.2 of the main thesis we provide formulae to estimate the risk of disease for an individual $\{I\}$, given:

- a measured polygenic variable for individual $\{I\}$; M_I,
- a known major locus variable for individual I; $G_{1,I}$, and,
- the disease status of a relative $\{R\}$; Y_R.

We now extend the family history scenario to S relatives, compared to the 1 relative previously considered. That is we wish to estimate:

$$p(Y_I = 1 | G_{1,I} = g_{1,I}, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}) \tag{5}$$

where:

- Y_I is the disease status for individual $\{I\}$,
- $\underline{R} = [R_1, ..., R_S]^T$ is a vector denoting the S relatives of individual $\{I\}$,
- $\underline{Y}_{\underline{R}} = [Y_{R_1}, ..., Y_{R_S}]^T$ is the vector containing the disease status variable for the S relatives of individual $\{I\}$,
- $\underline{y}_{\underline{R}} = [y_{R_1}, ..., y_{R_S}]^T$ is the vector containing the *observed* disease status for the S relatives of individual $\{I\}$,
- $G_{1,I}$ is the major locus variable for individual $\{I\}$ which counts the number of risk alleles that $\{I\}$ carries at the major locus,
- $g_{1,I}$ is the observed number of risk alleles carried by individual $\{I\}$ at the major locus,
- M_I is the measurable, additive genetic component variable for individual $\{I\}$, and,
- m_I is the observed, measurable, additive genetic component for individual $\{I\}$.

To gain an estimate for Equation (5) we first need to find:

$$p(Y_I = 1 | G_{1,I} = g_{1,I}, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}}) \tag{6}$$

where:

- $\underline{G}_{1,\underline{R}}$ is the vector containing the major locus variable for the S relatives of individual $\{I\}$,
- $\underline{g}_{1,\underline{R}}$ is the vector containing the observed risk allele counts at the major locus for the S relatives of individual $\{I\}$, and,
- all else as defined above.

The results for Equation (6) are then used to calculate Equation (5). Equation (6) is discussed in Scenario (A) and Equation (5) is discussed in Scenario (B). For both scenarios 2 methods are presented: 1. an exact method, and 2. a Pearson-Aitken approximate method.

We now present Scenario (A), starting with the exact method.

Scenario (A)

(A.1) Exact method

Assuming that $\underline{Y}_{\underline{R}} \perp G_{1,I} | \{\underline{G}_{1,\underline{R}}, M_I\}$, then we start by re-writing Equation (6) as:

$$\begin{aligned} & p(Y_I = 1 | G_{1,I} = g_{1,I}, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}}) \\ & = \frac{p(Y_I = 1, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}} | G_{1,I} = g_{1,I}, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}}, M_I = m_I)}{p(\underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}} | \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}}, M_I = m_I)} \end{aligned}$$

To calculate this risk using multivariate integration we need to define the following joint, conditional distribution:

$$\begin{bmatrix} L_I | G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I \\ L_{R_1} | G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I \\ \vdots \\ L_{R_S} | G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I \end{bmatrix} = \begin{bmatrix} L_I | G_{1,I} = g_{1,I}, M_I = m_I \\ L_{R_1} | G_{1,R_1} = g_{1,R_1}, M_I = m_I \\ \vdots \\ L_{R_S} | G_{1,R_S} = g_{1,R_S}, M_I = m_I \end{bmatrix}$$

where we assume that:

- $L_I \perp \underline{G}_{1,\underline{R}} | \{G_{1,I}, M_I\}$,
- $L_{R_s} \perp G_{1,I} | \{G_{1,R_s}, M_I\}$; $s = 1, 2, ..., S$, and,
- $L_{R_s} \perp G_{1,R_t} | \{G_{1,R_s}, M_I\}$; $s, t = 1, 2, ..., S$ but $s \neq t$.

(A.1.0) The required joint distribution

We start with the following multivariate normal distribution:

$$\begin{bmatrix} L_I | G_{1,I} = g_{1,I} \\ L_{R_1} | G_{1,R_1} = g_{1,R_1} \\ \vdots \\ L_{R_S} | G_{1,R_S} = g_{1,R_S} \\ M_I \end{bmatrix} \sim N_{S+2}(\mu*, \Sigma*) \tag{7}$$

where:

$$\circ \quad \mu* = \begin{bmatrix} h(g_{1,I}) \\ h(g_{1,R_1}) \\ \vdots \\ h(g_{1,R_S}) \\ 0 \end{bmatrix}$$

$$\circ \quad \Sigma^* = \begin{bmatrix} \Sigma^*_I & \Sigma^*_{I,\underline{R}} & \Sigma^*_{I,M_I} \\ \Sigma^*_{\underline{R},I} & \Sigma^*_{\underline{R}} & \Sigma^*_{\underline{R},M_I} \\ \Sigma^*_{M_I,I} & \Sigma^*_{M_I,\underline{R}} & \Sigma^*_{M_I} \end{bmatrix}$$

and:

- $h(g_1) = \beta_1 I_{g_1=1} + \beta_2 I_{g_1=2}$, with:
  - $L|\{G_1 = g_1\} \sim N(h(g_1), 1 - V_{h(G_1)})$,
  - $V_{h(G_1)} = Var[h(G_1)] = Var[\beta_1 I_{G_1=1} + \beta_2 I_{G_1=2}]$,
- $\Sigma^*_I = 1 - V_{h(G_1)}$,
- $\Sigma^*_{I,\underline{R}} = \Sigma^{*T}_{\underline{R},I}$ is an 1xS matrix where the i^{th} column is:

$$\Sigma^*_{I,\underline{R}}(1,i) = r_{I,R_i}\big(h^2_L - V_{A_{h(G_1)}}\big) + \theta_{I,R_i}\big(H^2_L - h^2_L - V_{D_{h(G_1)}}\big)$$

  for $i = 1, ..., S$, and where:
  - $H^2_L = Var[G] = Var[A + D] = V_A + V_D$ is the broad-sense heritability,
  - $h^2_L = Var[A] = V_A$ is the narrow-sense heritability,
  - $V_{A_{h(G_1)}} = \big(a_1 + d_1(1 - 2f_1)\big)^2 2f_1(1 - f_1)$ is the part of the additive genetic variation (narrow-sense heritability) determined by G_1 where:
    - $G_1 \sim Binom(2, f_1)$,
    - $a_1 = \frac{\beta_2}{2}$,
    - $d_1 = \beta_1 - \frac{\beta_2}{2}$, and
    - $E[L|G_1 = g_1] = \beta_1 I_{g_1=1} + \beta_2 I_{g_1=2}$,
  - $V_{D_{h(G_1)}} = \big(d_1 2f_1(1 - f_1)\big)^2$ is the part of the quasi-dominant variation, V_D, determined by G_1,
  - r_{I,R_i} is the coefficient of relatedness between individual $\{I\}$ and relative $\{R_i\}$; $i = 1, 2, ..., S$,
  - θ_{I,R_i} is the coefficient of coancestry between individual $\{I\}$ and relative $\{R_i\}$; $i = 1, 2, ..., S$.

  For more details on the variance components above please see Section 3.3.2 of the main thesis. For r and θ values between 'common' relative pairs please see Table B.1, Appendix B of the main thesis.
- $\Sigma^*_{I,M_I} = \Sigma^{*T}_{M_I,I} = V_M$, where:
  - $M \sim N(0, V_M)$
- $\Sigma^*_{\underline{R}}$ is an SxS matrix where the i^{th} row, j^{th} column is:

$$\Sigma^*_{\underline{R}}(i,j) = \begin{cases} 1 - V_{h(G_1)} & \text{when } i = j, \\ r_{R_i,R_j}\big(h^2_L - V_{A_{h(G_1)}}\big) + \theta_{R_i,R_j}\big(H^2_L - h^2_L - V_{D_{h(G_1)}}\big) & \text{when } i \neq j; \end{cases}$$

  for $i, j = 1, 2, ..., S$, and where:
  - r_{R_i,R_j} is the coefficient of relatedness between $\{R_i\}$ and $\{R_j\}$; $i, j = 1, 2, ...S$,
  - θ_{R_i,R_j} is the coefficient of coancestry between relatives $\{R_i\}$ and $\{R_j\}$; $i, j = 1, 2, ..., S$.
- $\Sigma^*_{\underline{R},M_I} = \Sigma^{*T}_{M_I,\underline{R}}$ is a 1xS matrix where the i^{th} column is:

$$\Sigma^*_{\underline{R},M_I}(1,i) = r_{I,R_i} V_M$$

  for $i = 1, 2, ...S$, and,

- $\Sigma^*_{M_I} = V_M$.

Then, applying standard statistical theory to the distribution in Equation (7), we gain the required joint distribution:

$$\begin{bmatrix} L_I|G_{1,I} = g_{1,I}, M_I = m_I \\ L_{R_1}|G_{1,R_1} = g_{1,R_1}, M_I = m_I \\ \vdots \\ L_{R_S}|G_{1,R_S} = g_{1,R_S}, M_I = m_I \end{bmatrix} \sim N_{S+1}(\mu, \Sigma) \ , \tag{8}$$

where:

$$\circ \quad \mu = \begin{bmatrix} h(g_{1,I}) + m_I \\ h(g_{1,R_1}) + r_{I,R_1} m_I \\ \vdots \\ h(g_{1,R_S}) + r_{I,R_S} m_I \end{bmatrix},$$

$$\circ \quad \Sigma = \begin{bmatrix} \Sigma_I & \Sigma_{I,\underline{R}} \\ \Sigma_{\underline{R},I} & \Sigma_{\underline{R}} \end{bmatrix} \ ;$$

and:

- $\Sigma_I = 1 - V_{h(G_1)} - V_M$,
- $\Sigma_{I,\underline{R}} = \Sigma^T_{\underline{R},I}$ is a 1xS matrix such that the i^{th} column is:

$$\Sigma_{I,\underline{R}}(1,i) = r_{I,R_i}\big(h^2_L - V_{A_{h(G_1)}} - V_M\big) + \theta_{I,R_i}\big(H^2_L - h^2_L - V_{D_{h(G_1)}}\big)$$

  for $i = 1, 2, ...S$, and,

- $\Sigma_{\underline{R}}$ is an SxS matrix where the i^{th} row, j^{th} column entry is:

$$\Sigma_{\underline{R}}(i,j) = \begin{cases} 1 - V_{h(G_1)} - r^2_{I,R_i} V_M & \text{when } i = j, \\ r_{R_i,R_j}\big(h^2_L - V_{A_{h(G_1)}}\big) + \theta_{R_i,R_j}\big(H^2_L - h^2_L - V_{D_{h(G_1)}}\big) - r_{I,R_i} r_{I,R_j} V_M & \text{when } i \neq j; \end{cases}$$

  for $i, j = 1, 2, ..., S$.

The distribution in Equation (8) is then used in multivariate integration to calculate the risk in Equation (6). As an example, let us assume that we know the disease status and major loci profiles of S relatives to individual I, where relatives $\{R_1, R_2, ..., R_\xi\}$ are *affected* and relatives $\{R_{\xi+1}, R_{\xi+2}, ..., R_S\}$ are *unaffected*. Then:

$$\begin{aligned} &p(Y_I = 1|G_{1,I} = g_{1,I}, M_I = m_I, Y_{R_1} = 0, ..., Y_{R_\xi} = 0, Y_{R_{\xi+1}} = 1, ..., Y_{R_S} = 1, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}}) \\ &= \frac{\underline{\int}^T_{-\infty} \underline{\int}^\infty_T \int^\infty_T f_{L_I, \underline{L}_{\underline{R}}|G_{1,I}, \underline{G}_{1,\underline{R}}, M_I}(x, y_1, \ldots, y_S|g_{1,I}, \underline{g}_{1,\underline{R}}, m_I) dx dy_1 \ldots dy_S}{\underline{\int}^T_{-\infty} \underline{\int}^\infty_T f_{\underline{L}_{\underline{R}}|\underline{G}_{1,\underline{R}}, M_I}(y_1, \ldots, y_S|\underline{g}_{1,\underline{R}}, m_I) dy_1 \ldots dy_S} \end{aligned}$$

where:

- $\underline{\int}^\infty_T = \int^\infty_T \ldots \int^\infty_T$ contains the integrals for the ξ affected relatives, and,
- $\underline{\int}^T_{-\infty} = \int^T_{-\infty} \ldots \int^T_{-\infty}$ contains the integrals for the remaining $S - \xi$ unaffected relatives.

(A.2) Pearson-Aitken approximate method

Using the distribution of:

$$\begin{bmatrix} L_I|G_{1,I}=g_{1,I},G_{1,R_1}=g_{1,R_1},...,G_{1,R_S}=g_{1,R_S},M_I=m_I \\ L_{R_1}|G_{1,I}=g_{1,I},G_{1,R_1}=g_{1,R_1},...,G_{1,R_S}=g_{1,R_S},M_I=m_I \\ \cdots \\ L_{R_S}|G_{1,I}=g_{1,I},G_{1,R_1}=g_{1,R_1},...,G_{1,R_S}=g_{1,R_S},M_I=m_I \end{bmatrix} = \begin{bmatrix} L_I|G_{1,I}=g_{1,I},M_I=m_I \\ L_{R_1}|G_{1,R_1}=g_{1,R_1},M_I=m_I \\ \cdots \\ L_{R_S}|G_{1,R_S}=g_{1,R_S},M_I=m_I \end{bmatrix}$$

given in Equation (8), we iteratively apply the PA approximation to gain the required estimate.

Scenario (B)

Recall that in Scenario (B) we wish to calculate:

$$p(Y_I=1|G_{1,I}=g_{1,I},M_I=m_I,\underline{Y}_{\underline{R}}=\underline{y}_{\underline{R}})$$

We shall now describe solutions via: (B.1) the exact method, and, (B.2) a Pearson-Aitken approximate method.

(B.1) Exact method

Assuming that:

- $\underline{Y}_{\underline{R}} \perp G_{1,I}|\{M_I,\underline{G}_{1,\underline{R}}\}$, and,
- $\underline{G}_{\underline{R}} \perp M_I|G_{1,I}$,

we can write:

$$\begin{aligned} &p(Y_I=1|G_{1,I}=g_{1,I},M_I=m_I,\underline{Y}_{\underline{R}}=\underline{y}_{\underline{R}}) \\ &= \frac{p(Y_I=1,\underline{Y}_{\underline{R}}=\underline{y}_{\underline{R}}|G_{1,I}=g_{1,I},M_I=m_I)}{p(\underline{Y}_{\underline{R}}=\underline{y}_{\underline{R}}|G_{1,I}=g_{1,I},M_I=m_I)} \\ &= \frac{\sum_{\underline{g}_{1,\underline{R}}} p(Y_I=1,\underline{Y}_{\underline{R}}=\underline{y}_{\underline{R}}|G_{1,I}=g_{1,I},M_I=m_I,\underline{G}_{1,\underline{R}}=\underline{g}_{1,\underline{R}})p(G_{1,I}=g_{1,I},\underline{G}_{1,\underline{R}}=\underline{g}_{1,\underline{R}})}{\sum_{\underline{g}_{1,\underline{R}}} p(\underline{Y}_{\underline{R}}=\underline{y}_{\underline{R}}|M_I=m_I,\underline{G}_{1,\underline{R}}=\underline{g}_{1,\underline{R}})p(G_{1,I}=g_{1,I},\underline{G}_{1,\underline{R}}=\underline{g}_{1,\underline{R}})} \end{aligned}$$

where $\sum_{\underline{g}_{1,\underline{R}}} = \sum_{g_{1,R_1}=0}^{2}\sum_{g_{1,R_2}=0}^{2}\cdots\sum_{g_{1,R_S}=0}^{2}$.

$p(G_{1,I}=g_{1,I},\underline{G}_{1,\underline{R}}=\underline{g}_{1,\underline{R}})$ will depend upon the relationships between $\{I,R_1,R_2,\ldots,R_S\}$.

The joint distribution in Equation (8) can then be used to estimate this risk.

Currently, bespoke code would need to be written for every new $\{I,R_1,R_2,\ldots,R_S\}$ scenario. This is because $p(G_{1,I}=g_{1,I},\underline{G}_{1,\underline{R}}=\underline{g}_{1,\underline{R}})$ will vary with every new scenario, and there would be too many possibilities to program.

(B.2) Pearson-Aitken approximate method

We present the PA method here, but we cannot show a final risk equation because this will differ depending on the disease status of the S relatives, and the relationships between the family members.

We again use the law of total probability, and this time write:

$$
\begin{aligned}
&p(Y_I = 1|G_{1,I} = g_{1,I}, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}) \\
&= \sum_{\underline{g}_{1,\underline{R}}} p(Y_I = 1|G_{1,I} = g_{1,I}, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}}) \\
&\quad p(\underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}}|G_{1,I} = g_{1,I}, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}) \\
&= \frac{\sum_{\underline{g}_{1,\underline{R}}} \omega_{\underline{g}_{1,\underline{R}}} \theta_{R_S,\underline{g}_{1,\underline{R}}} (\prod_{s=1}^{S} \theta_{R_s,\underline{g}_{1,\underline{R}}}) p(G_{1,I} = g_{1,I}, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}})}{\sum_{\underline{g}_{1,\underline{R}}} \theta_{R_S,\underline{g}_{1,\underline{R}}} (\prod_{s=1}^{S} \theta_{R_s,\underline{g}_{1,\underline{R}}}) p(G_{1,I} = g_{1,I}, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}})}
\end{aligned} \tag{9}
$$

where:

- $\omega_{\underline{g}_{1,\underline{R}}} = p(Y_I = 1|G_{1,I} = g_{1,I}, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}})$
- $\theta_{R_s,\underline{g}_{1,\underline{R}}} = p(Y_{R_s} = y_{R_s}|Y_{R_{s+1}} = y_{R_{s+1}}, ..., Y_{R_S} = y_{R_S}, G_{1,I} = g_{1,I}, M_I = m_I, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}})$, for $s = 1, 2, ..., S-1$, and,
- $\theta_{R_S,\underline{g}_{1,\underline{R}}} = p(Y_{R_S} = y_{R_S}|G_{1,I} = g_{1,I}, M_I = m_I, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}})$

We then start with the joint conditional distribution of:

$$
\begin{bmatrix}
L_I|G_{1,I} = g_{1,I}, M_I = m_I \\
L_{R_1}|G_{1,R_1} = g_{1,R_1}, M_I = m_I \\
\vdots \\
L_{R_S}|G_{1,R_S} = g_{1,R_S}, M_I = m_I
\end{bmatrix} =
\begin{bmatrix}
L_I|G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I \\
L_{R_1}|G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I \\
\vdots \\
L_{R_S}|G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I
\end{bmatrix}
$$

provided in Equation (8).

$\theta_{R_S,\underline{g}_{1,\underline{R}}} = p(Y_{R_S} = y_{R_S}|G_{1,I} = g_{1,I}, M_I = m_I, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}})$ can be calculated using the marginal distribution of $L_{R_S}|G_{1,R_S} = g_{1,R_S}, M_I = m_I$ in Equation (8).

Then we iteratively apply the PA approximation S times in total. At each iteration we store the resulting distribution. The first $S-1$ iterations provide the joint distributions of:

$$
\begin{bmatrix}
L_I|G_{1,I} = g_{1,I}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S} \\
L_{R_1}|G_{1,R_1} = g_{1,R_1}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S} \\
\vdots \\
L_{R_{s-1}}|G_{1,R_S} = g_{1,R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S}
\end{bmatrix} =
$$

$$
\begin{bmatrix}
L_I|G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S} \\
L_{R_1}|G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S} \\
\vdots \\
L_{R_{s-1}}|G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S}
\end{bmatrix}
$$

for $s = S, S-1, ..., 2$.

From each of these joint conditional distributions we can extract the marginal distribution of $L_{R_{s-1}}|G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S}$. Therefore we can calculate $\theta_{R_{s-1},\underline{g}_{1,\underline{R}}} = p(Y_{R_{s-1}} = y_{R_{s-1}}|Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S}, G_{1,I} = g_{1,I}, M_I = m_I, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}})$, for $s = S, S-1, ..., 2$.

At the final (S^{th}) PA approximation application, we gain the conditional distribution of $L_I|G_{1,I} = g_{1,I}, G_{1,R_1} = g_{1,R_1}, ..., G_{1,R_S} = g_{1,R_S}, M_I = m_I, Y_{R_1} = y_{R_1}, ..., Y_{R_S} = y_{R_S}$. From this $\omega_{\underline{g}_{1,\underline{R}}} = p(Y_I = 1|G_{1,I} = g_{1,I}, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}})$ can be found.

All of these probabilities are entered into Equation (9) and, along with $p(G_{1,I} = g_{1,I}, \underline{G}_{1,\underline{R}} = \underline{g}_{1,\underline{R}})$, can be used to approximate the required risk.

2.2 ★ LTMM with Q major loci and 1 relative

Section 3.3.2 of the main thesis covers the estimation of disease risk for an individual, $\{I\}$, conditional on:

- a measured polygenic variable for individual I; M_I,
- a known major locus variable for individual I; $G_{1,I}$, and,
- the disease status of a relative R; Y_R.

Here we extend this to when there are Q known major loci. We assume:

- that all major loci are bi-allelic (this is for simplicity as results are easily extended to the multi-allelic case), and so $G_q \sim Bi(2, f_q)$,
- there are no interactions *between* major loci (on the liability scale),

We start by describing the disease model used here. We then show when this model is identifiable, and so usable. We then presents risk calculation results for Scenario (A), a formula for the risk of disease conditional on:

- $M_I = m_I$,
- $\underline{G}_I = \underline{g}_I$,
- $Y_R = y$, and,
- $\underline{G}_R = \underline{g}_R$,

for $y = 0, 1$, and where $\underline{G} = [G_1, G_2, ..., G_Q]^T$, denotes the Q major loci random variables. Therefore $\underline{G}_I$ contains the major loci random variables for individual I and $\underline{G}_R$ contains the major loci random variables for individual R. $\underline{g}_I$ contains the *observed* major loci values for individual I and $\underline{g}_R$ contains the *observed* major loci values for individual R.

We start with this probability calculation because the results from Scenario (A) are then used in Scenario (B), where we present results for calculating the disease risk of an individual I conditional on:

- $M_I = m_I$,
- $\underline{G}_I = \underline{g}_I$, and,
- $Y_R = y$.

The disease model

Let Y denote disease status, such that $Y = 1$ for affected and $Y = 0$ for unaffected. We say:

$$Y \sim Bi(1, K)$$

where $K = p(Y = 1)$ is the disease prevalence.

We assume that disease status is underlined by a latent variable, L, which is called the liability to disease, where:

$$\begin{aligned} K &= p(Y = 1) \\ &= p(L > T) \end{aligned}$$

where T is the disease threshold. We define:

$$L = h(\underline{G}) + M + U$$

where:

- $\underline{G} = [G_1, G_2, ..., G_Q]^T$ is the vector containing the Q major loci random variables, where $G_q \sim Bi(2, f_q)$; $q = 1, 2, ..., Q$,
- h is the function defining the relationship between the Q major loci ($\underline{G}$) and liability to disease (L),
- M is a measured polygenic component, and,
- U is a variable capturing the residual, unmeasured disease variables.

We define this model, and its parameters, such that $Var[L] = 1$. We assume that:

- $h(\underline{G}) = \sum_{q=1}^{Q} \sum_{j=0}^{2} \beta_{qj} I_{G_q=j}$; h is a linear function, and so there are no interactions *between* major loci on the *liability scale*.
- $M \sim N(0, V_M)$; V_M is the variability in L attributable to M.
- $U \sim N(0, V_U)$l V_U is the variability in L attributable to U.
- U can be decomposed such that $U = U_G + E$, where:
  - $U_G \sim N(0, V_{U_G})$ is the unmeasured genetic component, and,
  - $E \sim N(0, V_E)$ is the environmental component (assumed to be unmeasured here).

As in the single major locus case, L is not normally distributed, but is a *mixture* of normal distributions. In the single major locus case L is a mixture of 3 normals. Here L is a mixture of 3^Q normals, where Q is the number of major loci. These Q normal distributions are:

$$\begin{aligned} L|\{\underline{G} = \underline{g}\} &= \sum_{q=1}^{Q} \sum_{j=0}^{2} \beta_{qj} I_{g_q=j} + M + U \\ &\sim N(\sum_{q=1}^{Q} \sum_{j=0}^{2} \beta_{qj} I_{g_q=j}, 1 - V_{h(\underline{G})}) \end{aligned} \tag{10}$$

where:

- $1 - V_{h(\underline{G})} = V_M + V_U$, and,
- $\beta_{q0} = 0$, and so $G_q = 0$ is the reference category, for $q = 1, 2, ..., Q$.

To use this model we need to know how to calculate model parameters $\{T\}$ and $\{\beta_{qj} : q = 1, 2, ..., Q; j = 1, 2\}$.

Defining model parameters

Calculating T:

We have defined $[\beta_{10}, \beta_{20}, ..., \beta_{Q0}]^T = [0, 0, ..., 0]^T$. Therefore:

$$L|\{G_1 = 0, ..., G_Q = 0\} = M + U \quad \sim N(0, 1 - V_{h(\underline{G})})$$

and,

$$Y\ \{G_1 = 0, ..., G_Q = 0\} \sim Bi(1, K_{\underline{0}})$$

where:

$$\begin{aligned}
K_{\underline{0}} &= K(G_1 = 0, ..., G_Q = 0) \\
&= p(Y = 1 | G_1 = 0, ..., G_Q = 0) \\
&= p(L | \{G_1 = 0, ..., G_Q = 0\} > T) \\
&= p(L > T | G_1 = 0, ..., G_Q = 0) \\
&= p(Z > \frac{T}{\sqrt{1 - V_{h(\underline{G})}}}) \\
&= 1 - p(Z \leq \frac{T}{\sqrt{1 - V_{h(\underline{G})}}}) \\
&= 1 - \Phi\Big(\frac{T}{\sqrt{1 - V_{h(\underline{G})}}}\Big)
\end{aligned}$$

and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{1}{2}z^2)dz$.

Rearranging the above gives:

$$T = \sqrt{1 - V_{h(\underline{G})}}\,\Phi^{-1}\big(1 - K_{\underline{0}}\big)$$

Therefore, to find T we need $K_{\underline{0}}$ and $V_{h(\underline{G})}$.

$K_{\underline{0}}$ will need to be calculated from data. Thus, for our purposes, where we assume that we have no raw data, $K_{\underline{0}}$ will need to be available in the literature. If the major loci are all rare then it can be assumed that $K_{\underline{0}} \approx K$.

$V_{h(\underline{G})}$ is the variance of $h(\underline{G})$. That is:

$$\begin{aligned}
V_{h(\underline{G})} &= var[h(\underline{G})] \\
&= var[\sum_{q=1}^{Q}\sum_{j=1}^{2} \beta_{qj} I_{G_q=j}] \\
&= E[\big(\sum_{q=1}^{Q}\sum_{j=1}^{2} \beta_{qj} I_{G_q=j}\big)^2] - \big(E[\sum_{q=1}^{Q}\sum_{j=1}^{2} \beta_{qj} I_{G_q=j}]\big)^2 \\
&= \sum_{q=1}^{Q}\sum_{j=1}^{2}\sum_{l=1}^{Q}\sum_{m=1}^{2} \beta_{qj}\beta_{lm} p(G_q = j, G_l = m) - \beta_{qj}\beta_{lm} p(G_q = j) p(G_l = m) \qquad (11)
\end{aligned}$$

Results from the liability/ probit model in Equation (10), and therefore the βs, may be available. If so these βs can be used, with the joint distribution of G_q and G_l; $q, l = 1, 2, ..., Q$, to calculate $V_{h(\underline{G})}$. If the βs are not available in the literature, then the following section describes how to calculate them, and $V_{h(\underline{G})}$.

Calculating β_{qj} for $q = 1, 2, ..., Q$ and $j = 1, 2$:

If not we shall show what is needed for them to be calculated. Recall that:

$$L|\{\underline{G}=\underline{g}\}=\sum_{q=1}^{Q}\sum_{j=0}^{2}\beta_{qj}I_{g_q=j}+M+U \quad \sim N(\sum_{q=1}^{Q}\sum_{j=0}^{2}\beta_{qj}I_{g_q=j},1-V_{h(\underline{G})})$$

and,

$$Y|\{\underline{G}=\underline{g}\}\sim Bi(1,K(g_1,...,g_Q))$$

where:

$$\begin{aligned}K(g_1,...,g_Q)&=p(Y=1|G_1=g_1,...,G_Q=g_Q)\\&=p(L>T|G_1=g_1,...,G_Q=g_Q)\\&=p(Z>\frac{T-\sum_{q=1}^{Q}\sum_{j=0}^{2}\beta_{qj}I_{g_q=j}}{\sqrt{1-V_{h(\underline{G})}}})\\&=1-\Phi\Big(\frac{T-\sum_{q=1}^{Q}\sum_{j=0}^{2}\beta_{qj}I_{g_q=j}}{\sqrt{1-V_{h(\underline{G})}}}\Big)\end{aligned}$$

Rearranging the above gives:

$$\begin{aligned}\sum_{q=1}^{Q}\sum_{j=0}^{2}\beta_{qj}I_{g_q=j}&=T-\sqrt{1-V_{h(\underline{G})}}\Phi^{-1}\big(1-K(g_1,...,g_Q)\big)\\&=\sqrt{1-V_{h(\underline{G})}}\Big(\Phi^{-1}\big(1-K_{\underline{0}}\big)-\Phi^{-1}\big(1-K(g_1,...,g_Q)\big)\Big)\end{aligned}$$

By definition, $\beta_{q0}=0$ for $q=1,2,...,Q$. Therefore the above is equivalent to:

$$\begin{aligned}\sum_{q=1}^{Q}\sum_{j=1}^{2}\beta_{qj}I_{g_q=j}&=T-\sqrt{1-V_{h(\underline{G})}}\Phi^{-1}\big(1-K(g_1,...,g_Q)\big)\\&=\sqrt{1-V_{h(\underline{G})}}\Big(\Phi^{-1}\big(1-K_{\underline{0}}\big)-\Phi^{-1}\big(1-K(g_1,...,g_Q)\big)\Big)\end{aligned}$$

The simplest way to calculate model coefficients $\{\beta_{q1},\beta_{q2}\}$ for a major locus q is to use the penetrance function for this major locus given all other major loci are 0.

For example, to find $\{\beta_{11},\beta_{12}\}$ we require $V_{h(\underline{G})}$ and:

$$\{K(0,0,...,0),K(1,0,...,0),K(2,0,...,0)\}=\{K_{\underline{0}},K(1,0,...,0),K(2,0,...,0)\}$$

That is:

$$\begin{aligned}\{&p(Y=1|G_1=0,G_2=0,...,G_Q=0),\\&p(Y=1|G_1=1,G_2=0,...,G_Q=0),\\&p(Y=1|G_1=2,G_2=0,...,G_Q=0)\}\end{aligned}$$

If we have this variance and partial penetrance function then:

$$\beta_{11}=\sqrt{1-V_{h(\underline{G})}}\Big(\Phi^{-1}\big(1-K_{\underline{0}}\big)-\Phi^{-1}\big(1-K(1,0,...,0)\big)\Big)$$

and

$$\beta 12=\sqrt{1-V_{h(\underline{G})}}\Big(\Phi^{-1}\big(1-K_{\underline{0}}\big)-\Phi^{-1}\big(1-K(2,0,...,0)\big)\Big)$$

The partial penetrance function will need to be estimated from data.

However, recall from Equation (11), that $V_{h(\underline{G})}$ is a function of β_{qj}; $q = 1, 2, ..., Q$ and $j = 1, 2$. Given this, can we make this model identifiable? Yes, and here is how...

If we assume that the Q major loci are independent then:

$$V_{h(\underline{G})} = \sum_{q=1}^{Q} V_{h(G_q)}$$

where:

$$\begin{aligned} V_{h(G_q)} &= var[\sum_{j=1}^{2} \beta_{qj} I_{G_q=j}] \\ &= \beta_{q1}^2 p(G_q = 1)(1 - p(G_q = 1)) - 2\beta_{q1}\beta_{q2} p(G_q = 1)p(G_q = 2) + \beta_{q2}^2 p(G_q = 2)(1 - p(G_q = 2)) \end{aligned}$$

Let us denote that the probability of disease given the risk allele count for the i^{th} major locus is 1, and the risk allele count is 0 for all other major loci by $K_{g_i=1,\underline{0}}$. Similarly, let us denote that the probability of disease given the risk allele count for the i^{th} major locus is 2, and the risk allele count is 0 for all other major loci by $K_{g_i=2,\underline{0}}$. Then we can write:

$$\beta_{q1} = \sqrt{1 - V_{h(\underline{G})}}\Big(\Phi^{-1}\big(1 - K_{\underline{0}}\big) - \Phi^{-1}\big(1 - K_{g_i=1,\underline{0}}\big)\Big)$$

and,

$$\beta_{q2} = \sqrt{1 - V_{h(\underline{G})}}\Big(\Phi^{-1}\big(1 - K_{\underline{0}}\big) - \Phi^{-1}\big(1 - K_{g_i=2,\underline{0}}\big)\Big)$$

Inserting these β formulae into the equation for $V_{h(G_q)}$ gives:

$$V_{h(G_q)} = (1 - V_{h(\underline{G})})C_q$$

where:

$$\begin{aligned} C_q =& p(G_q = 1)(1 - p(G_q = 1))\Big(\Phi^{-1}\big(1 - K_{\underline{0}}\big) - \Phi^{-1}\big(1 - K_{g_q=1,\underline{0}}\big)\Big)^2 \\ &- 2p(G_q = 1)p(G_q = 2)\Big(\Phi^{-1}\big(1 - K_{\underline{0}}\big) - \Phi^{-1}\big(1 - K_{g_q=1,\underline{0}}\big)\Big)\Big(\Phi^{-1}\big(1 - K_{\underline{0}}\big) - \Phi^{-1}\big(1 - K_{g_q=2,\underline{0}}\big)\Big) \\ &+ p(G_q = 2)(1 - p(G_q = 2))\Big(\Phi^{-1}\big(1 - K_{\underline{0}}\big) - \Phi^{-1}\big(1 - K_{g_q=2,\underline{0}}\big)\Big)^2 \end{aligned}$$

C_q can be calculated if the partial penetrance function, $\{K_{\underline{0}}, K_{g_q=1,\underline{0}}, K_{g_q=2,\underline{0}}\}$, and f_q are known.

Then:

$$\begin{aligned} V_{h(\underline{G})} &= \sum_{q=1}^{Q} V_{h(G_q)} \\ &= (1 - V_{h(\underline{G})}) \sum_{q=1}^{Q} C_q \\ &= \frac{\sum_{q=1}^{Q} C_q}{1 + \sum_{q=1}^{Q} C_q} \end{aligned}$$

$V_{h(\underline{G})}$ can be calculated if we know the partial penetrance function, $\{K_{\underline{0}}, K_{g_q=1,\underline{0}}, K_{g_q=2,\underline{0}}\}$, and f_q, for $q = 1, 2, ..., Q$. And therefore T and β_{qj} can be calculated.

We have shown that the disease model in Equation (10) is usable in our circumstance when:

- the Q major loci are independent,
- there are no between major loci interactions on the liability scale,
- $\underline{G}$, M and U are independent and do no interact on the liability scale,
- M and U are normally distributed,
- f_q, where $G_q \sim Bi(2, f_q)$; $q = 1, 2, ..., Q$, are known, and,
- $\{K_{\underline{0}}, K_{g_q=1,\underline{0}}, K_{g_q=2,\underline{0}}\}$; $q = 1, 2, ..., Q$, are known.

Given all of these are true then this liability threshold framework can be used to estimate risk. Our particular focus will be:

- Scenario (A) - the risk of disease for an individual, I, conditional on:
  - $M_I = m_I$,
  - $\underline{G}_I = \underline{g}_I$,
  - $Y_R = y$; $y = 0, 1$, and,
  - $\underline{G}_R = \underline{g}_R$.
- Scenario (B) - the risk of disease for an individual, I, conditional on:
  - $M_I = m_I$,
  - $\underline{G}_I = \underline{g}_I$, and,
  - $Y_R = y$; $y = 0, 1$.

An exact and PA approximate method will be presented for both scenarios.

Let us start with Scenario (A).

Scenario (A)

In this scenario we wish to calculate:

$$p(Y_I = 1 | M_I = m_I, \underline{G}_I = \underline{g}_I, Y_R = y, \underline{G}_R = \underline{g}_R) \tag{12}$$

for $y = 0, 1$ and where:

- Y_I is the disease status variable for individual $\{I\}$,
- Y_R is the disease status variable for relative $\{R\}$,
- $\underline{G}_I = [G_{1,I}, G_{2,I}, ..., G_{Q,I}]^T$ is a vector containing the risk allele count random variables for the Q known major loci of individual $\{I\}$,
- $\underline{g}_I = [g_{1,I}, g_{2,I}, ..., g_{Q,I}]^T$ is a vector containing the observed risk allele counts for the Q known major loci of individual $\{I\}$,
- $\underline{G}_R = [G_{1,R}, G_{2,R}, ..., G_{Q,R}]^T$ is a vector containing the risk allele count random variables for the Q known major loci of relative $\{R\}$,

- $\underline{g}_R = [g_{1,R}, g_{2,R}, ..., g_{Q,R}]^T$ is a vector containing the observed risk allele counts for the Q known major loci of relative $\{R\}$,
- M_I is the random variable for the measurable, *additive* genetic component of individual $\{I\}$, and,
- m_I is the observed measurable, *additive* genetic component of individual $\{I\}$.

We shall now consider 2 methods for calculating risk: (A.1) the exact method, and, (A.2) a PA approximate method. We shall now describe the exact method.

(A.1) Exact method:

Let us start by re-writing the probability we wish to calculate:

$$\begin{aligned} & p(Y_I = 1 | M_I = m_I, \underline{G}_I = \underline{g}_I, Y_R = y, \underline{G}_R = \underline{g}_R) \\ &= \frac{p(Y_I = 1, Y_R = y | M_I = m_I, \underline{G}_I = \underline{g}_I, \underline{G}_R = \underline{g}_R)}{p(Y_R = y | M_I = m_I, \underline{G}_I = \underline{g}_I, \underline{G}_R = \underline{g}_R)} \\ &= \frac{p(Y_I = 1, Y_R = y | M_I = m_I, \underline{G}_I = \underline{g}_I, \underline{G}_R = \underline{g}_R)}{p(Y_R = y | M_I = m_I, \underline{G}_R = \underline{g}_R)} \end{aligned}$$

for $y = 0, 1$, and assuming that $Y_R \perp \underline{G}_I | \{M_I, \underline{G}_R\}$.

To calculate this risk via multivariate integration (the exact method) we need to define the following conditional joint distribution:

$$\begin{bmatrix} L_I | \underline{G}_I = \underline{g}_I, \underline{G}_R = \underline{g}_R, M_I = m_I \\ L_R | \underline{G}_I = \underline{g}_I, \underline{G}_R = \underline{g}_R, M_I = m_I \end{bmatrix} = \begin{bmatrix} L_I | \underline{G}_I = \underline{g}_I, M_I = m_I \\ L_R | \underline{G}_R = \underline{g}_R, M_I = m_I \end{bmatrix}$$

We shall now derive this distribution.

(A.1.0) The joint conditional distribution required for calculating risk

We start with the trivariate normal distribution:

$$\begin{bmatrix} L_I | \underline{G}_I = \underline{g}_I \\ L_R | \underline{G}_R = \underline{g}_R \\ M_I \end{bmatrix} \sim N(\mu, \Sigma)$$

where:

- $\mu = \begin{bmatrix} \sum_{q=1}^{Q} \sum_{j=1}^{2} \beta_{qj} I_{g_{q,I}=j} \\ \sum_{q=1}^{Q} \sum_{j=1}^{2} \beta_{qj} I_{g_{q,R}=j} \\ 0 \end{bmatrix}$
- $\Sigma = \begin{bmatrix} 1 - V_{h(\underline{G})} & C_{1,2} & C_{1,3} \\ C_{2,1} & 1 - V_{h(\underline{G})} & C_{2,3} \\ C_{3,1} & C_{3,2} & V_M \end{bmatrix}$

and:

- $C_{1,2} = C_{2,1} = Cov[\{L_I | \underline{G}_I = \underline{g}_I\}, \{L_R | \underline{G}_R = \underline{g}_R\}]$,
- $C_{1,3} = C_{3,1} = Cov[\{L_I | \underline{G}_I = \underline{g}_I\}, M_I]$, and,
- $C_{2,3} = C_{3,2} = Cov[\{L_R | \underline{G}_R = \underline{g}_R\}, M_I]$.

Let us start with $\mathbf{C_{1,3} = C_{3,1}}$:

$$\begin{aligned} C_{1,3} = C_{3,1} &= Cov[\{L_I|\underline{G}_I = \underline{g}_I\}, M_I] \\ &= E[M_I(M_I + U_I)] \\ &= E[M_I^2] \\ &= V_M \end{aligned}$$

Followed by $\mathbf{C_{2,3} = C_{3,2}}$:

$$\begin{aligned} C_{2,3} = C_{3,2} &= Cov[\{L_R|\underline{G}_R = \underline{g}_R\}, M_I] \\ &= E[M_I(M_R + U_R)] \\ &= E[M_I M_R] \\ &= rV_M \end{aligned}$$

Finally, we shall define $\mathbf{C_{1,2} = C_{2,1}}$. Deriving this covariance is more involved than the others above.

$$\begin{aligned} C_{1,2} = C_{2,1} &= Cov[\{L_I|\underline{G}_I = \underline{g}_I\}, \{L_R|\underline{G}_R = \underline{g}_R\}] \\ &= E[(M_I + U_I)(M_R + U_R)] \\ &= E[M_I M_R] + E[U_I U_R] \\ &= Cov[M_I, M_R] + Cov[U_I, U_R] \\ &= rV_M + Cov[U_I, U_R] \end{aligned}$$

Therefore to derive $C_{1,2} = C_{2,1}$ we need to know $Cov[U_I, U_R]$.

What is $\mathbf{Cov[U_I, U_R]}$?

U is the unmeasured risk component in Equation (10). It can be split into an unmeasured genetic (U_G) and an unmeasured environmental component (E):

$$U = U_G + E \quad \sim N(0, V_U)$$

where:

- $U_G \sim N(0, V_{U_G})$, and,
- $E \sim N(0, V_E)$; $V_E = 1 - H_L^2$.

Focusing on U_G, this is the total genetic component, G, minus what has been measured. Here that is the major loci $\underline{G}$ and the additive component M. That is:

$$U_G = G - h(\underline{G}) - M$$

The total genetic component G is typically decomposed into an *additive* component, A, and an *independent* (quasi) dominant component, D. Therefore:

$$\begin{aligned} U &= U_G + E \\ &= G - h(\underline{G}) - M + E \\ &= A + D - h(\underline{G}) - M + E \end{aligned}$$

M is by definition additive. $h(\underline{G})$ contains both additive and dominant information. We write:

$$U = A_{res} + D_{res} + E$$

where:

- A_{res} is the *residual* additive component; $A_{res} = A - M - A_{h(\underline{G})}$, defined such that $E[A_{res}] = 0$,
- D_{res} is the *residual* quasi-dominant component; $D_{res} = D - D_{h(\underline{G})}$, defined such that $E[D_{res}] = 0$, and,
- $h(\underline{G})$ can be decomposed into an additive, $A_{h(\underline{G})}$, and quasi-dominant component, $D_{h(\underline{G})}$; $h(\underline{G}) = A_{h(\underline{G})} + D_{h(\underline{G})}$.

We define:

$$\circ \quad V_A = var[A] = \sum_{i=1}^{N} 2f_i(1-f_i)\big(\beta_{i1}(1-2f_i) + \beta_{i2}f_i\big)^2$$

and

$$\circ \quad V_D = var[D] = \sum_{i=1}^{N} \big(f_i(1-f_i)(2\beta_{i1} - \beta_{i2})\big)^2$$

where N is the *total* number of causal loci, all assumed to be bi-allelic, and L is re-written as:

$$\begin{aligned} L &= h(\underline{G}) + M + U \\ &= h(\underline{G}) + (M' - E[M']) + (U' - E[U']) \\ &= (h(\underline{G}) + M' + U') - E[M' + U'] \\ &= (\sum_{i=1}^{N}\sum_{j=1}^{2} \beta_{ij} I_{G_i=j} + E) - E[M' + U'] \end{aligned}$$

where:

- G_i is a random variable counting the number of risk alleles at the i^{th} causal genetic loci; $G_i \sim Bi(2, f_I)$ and f_i is the risk allele frequency.
- $I_{G_i=j}$ is an indicator variable which equals 1 if $\{G_i = j\}$, and 0 otherwise; $i = 1, 2, ..., N$ and $j = 1, 2$.
- $M = M' - E[M'] \quad \sim N(0, V_M)$ is the measured, additive genetic component, and therefore M' is the non-zero-centred measured, additive genetic component.
- $U = U' - E[U'] \quad \sim N(0, V_U)$ is the unmeasured component in the disease model, and therefore U' is the non-zero-centred unmeasured component.
- Assuming the 1^{st} Q loci are the known major loci then:

$$M + U = \sum i = Q + 1^N \sum_{j=1}^{2} \beta ij I_{G_i=j} + E - E[M' + U']$$

Assuming that there are J loci in the measured additive component M and, for the ease of writing, the total N risk loci are ordered such that:

- loci $\{1, 2, ..., Q\}$ are the Q known major loci,
- loci $\{Q+1, Q=2, ..., Q+J\}$ are the J loci contained in M, and,
- loci $\{Q+J+1, Q+J+2, ..., N\}$ are the remaining unmeasured genetic risk loci;

we can then write:

$$\circ \quad V_A = V_{A_{h(\underline{G})}} + V_M + V_{A_{res}}$$
$$= h_L^2$$

and

$$\circ \quad V_D = V_{D_{h(\underline{G})}} + V_{D_{res}}$$
$$= H_L^2 - h_L^2$$

where:

- $V_{A_{h(\underline{G})}} = \sum_{q=1}^{Q} 2f_q(1-f_q)\big(\beta_{q1}(1-2f_q) + \beta_{q2}f_q\big)^2$
- $V_M = \sum_{j=Q+1}^{Q+J} 2f_j(1-f_j)\big(\beta_{j1}(1-2f_j) + \beta_{j2}f_j\big)^2$
- $V_{A_{res}} = h_L^2 - V_{A_{h(\underline{G})}} - V_M$
- $V_{D_{h(\underline{G})}} = \sum_{q=1}^{Q} \big(f_q(1-f_q)(2\beta_{q1} - \beta_{q2})\big)^2$
- $V_{D_{res}} = H_L^2 - h_L^2 - V_{D_{h(\underline{G})}}$

Therefore:

$$Cov[U_I, U_R] = E[U_I U_R]$$
$$= E[(A_{res,I} + D_{res,I} + E_I)(A_{res,R} + D_{res,R} + E_R)]$$

By definition, $A \perp D$, and so:

- $A_{res,I} \perp D_{res,R}$, and,
- $D_{res,I} \perp A_{res,R}$.

We have assumed that the environmental risk variables captured in E and genetic risk loci are independent; $E \perp A$ and $E \perp D$, and so:

- $E_I \perp A_{res,R}$,
- $E_I \perp D_{res,R}$,
- $E_R \perp A_{res,I}$, and,
- $E_R \perp D_{res,I}$.

We have also assumed that environmental risk variables do not cluster within families, and so $E_I \perp E_R$. Thus:

$$\begin{aligned} Cov[U_I, U_R] &= E[A_{res,I}A_{res,R}] + E[D_{res,I}D_{res,R}] \\ &= Cov[A_{res,I}, A_{res,R}] + Cov[D_{res,I}, D_{res,R}] \\ &= rV_{A_{res}} + \theta V_{D_{res}} \\ &= r(h_L^2 - V_{A_{h(\underline{G})}} - V_M) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}}) \end{aligned}$$

This means that the covariance between $\{L_I|\underline{G}_I = \underline{g}_I\}$ and $\{L_R|\underline{G}_R = \underline{g}_R\}$ is:

$$\begin{aligned} C_{1,2} = C_{2,1} &= Cov[\{L_I|\underline{G}_I = \underline{g}_I\}, \{L_R|\underline{G}_R = \underline{g}_R\}] \\ &= Cov[M_I, M_R] + Cov[U_I, U_R] \\ &= rV_M + r(h_L^2 - V_{A_{h(\underline{G})}} - V_M) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}}) \\ &= r(h_L^2 - V_{A_{h(\underline{G})}}) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}}) \end{aligned}$$

Thus, the joint distribution of $L_I|\{\underline{G}_I = \underline{g}_I\}$, $L_R|\{\underline{G}_R = \underline{g}_R\}$ and M_I, required to calculate our desired risk in Equation (12) is:

$$\begin{bmatrix} L_I|\underline{G}_I = \underline{g}_I \\ L_R|\underline{G}_R = \underline{g}_R \\ M_I \end{bmatrix} \sim N(\mu, \Sigma) \tag{13}$$

where:

$$\mu = \begin{bmatrix} h(\underline{g}_I) \\ h(\underline{g}_R) \\ 0 \end{bmatrix}$$

and,

$$\Sigma = \begin{bmatrix} 1 - V_{h(\underline{G})} & r(h_L^2 - V_{A_{h(\underline{G})}}) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}}) & V_M \\ r(h_L^2 - V_{A_{h(\underline{G})}}) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}}) & 1 - V_{h(\underline{G})} & rV_M \\ V_M & rV_M & V_M \end{bmatrix}$$

Then applying standard statistical theory to Equation (13), we gain the joint distribution of $L_I|\{\underline{G}_I = \underline{g}_I, M_I = m_I\}$ and $L_R|\{\underline{G}_R = \underline{g}_R, M_I = m_I\}$ as:

$$\begin{bmatrix} L_I|\underline{G}_I = \underline{g}_I, M_I = m_I \\ L_R|\underline{G}_R = \underline{g}_R, M_I = m_I \end{bmatrix} \sim N(\mu_{m_I}, \Sigma_{m_I}) \tag{14}$$

where:

$$\mu_{m_I} = \begin{bmatrix} h(\underline{g}_I) + m_I \\ h(\underline{g}_R) + rm_I \end{bmatrix}$$

and,

$$\Sigma_{m_I} = \begin{bmatrix} 1 - V_{h(\underline{G})} - V_M & r(h_L^2 - V_{A_{h(\underline{G})}} - V_M) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}}) \\ r(h_L^2 - V_{A_{h(\underline{G})}} - V_M) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}}) & 1 - V_{h(\underline{G})} - r^2V_M \end{bmatrix}$$

Multivariate integration, using this joint distribution, is then performed to calculate the disease risk for individual $\{I\}$ given $\underline{G}_I = \underline{g}_I$, $M_I = m_I$, $\underline{G}_R = \underline{g}_R$ and: 1. relative $\{R\}$ is affected ($Y_R = 1$), and, 2. $\{R\}$ is unaffected ($Y_R = 0$).

(A.1.1) Exact method when relative {R} is affected

When relative $\{R\}$ is affected, $Y_R = 1$, the formula for risk is:

$$\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = 1, \underline{G}_R = \underline{g}_R) \\
&= \frac{p(Y_I = 1, Y_R = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R)}{p(Y_R = 1|\underline{G}_R = \underline{g}_R, M_I = m_I)} \\
&= \frac{p(L_I > T, L_R > T|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R)}{p(L_R > T|\underline{G}_R = \underline{g}_R, M_I = m_I)} \\
&= \frac{\int_T^\infty \int_T^\infty f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I)dxdy}{1 - \Phi\big(\frac{T - h(\underline{g}_R) - rm_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}}\big)}
\end{aligned}$$

where $f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I)$ is the joint density function for joint distribution in Equation (14).

(A.1.2) Exact method when relative {R} is unaffected

When relative $\{R\}$ is unaffected, $Y_R = 0$, the formula for risk is:

$$\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = 0, \underline{G}_R = \underline{g}_R) \\
&= \frac{p(Y_I = 1, Y_R = 0|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R)}{p(Y_R = 0|\underline{G}_R = \underline{g}_R, M_I = m_I)} \\
&= \frac{p(L_I > T, L_R \leq T|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R)}{p(L_R \leq T|\underline{G}_R = \underline{g}_R, M_I = m_I)} \\
&= \frac{\int_{-\infty}^T \int_T^\infty f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I)dxdy}{\Phi\big(\frac{T - h(\underline{g}_R) - rm_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}}\big)}
\end{aligned}$$

where $f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I)$ is the joint density function for joint distribution in Equation (14).

(A.2) Pearson-Aitken approximation

Next we derive a formula to calculate the risk in Equation (12) using the Pearson-Aitken approximation. To do this we use the joint distribution:

$$\begin{bmatrix} L_I|\underline{G}_I = \underline{g}_I, M_I = m_I \\ L_R|\underline{G}_R = \underline{g}_R, M_I = m_I \end{bmatrix} \quad \sim N(\mu_{m_I}, \Sigma_{m_I})$$

defined in Equation (14), and apply to Pearson-Aitken (PA) approximation such that:

1. $L_R > T$, when relative $\{R\}$ is affected, and,
2. $L_R \leq T$, when relative $\{R\}$ is unaffected.

As we can see from the work for the exact method above, the calculation of risk when there are > 1 major loci is a simple extension of the risk equations in Section 3.3.2 of the main thesis but with, for instance, $V_{A_{h(\underline{G})}}$ instead of $V_{A_{\xi(G_1)}}$. We therefore just present the end equations here.

(A.2.1) PA approximation when relative {R} is affected

By applying the PA formula to the distribution in Equation (14), we obtain an *approximate* distribution for $L_I|\{\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R, L_R > T\}$. This approximate distribution is:

$$L_I|\{\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R, L_R > T\} \sim N(\mu_{1,\underline{g}_R}, \sigma^2_{1,\underline{g}_R}) \tag{15}$$

where:

$$\mu_{1,\underline{g}_R} = h(\underline{g}_I) + m_I + \frac{r(h_L^2 - V_{A_{h(\underline{G})}} - V_M) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}})}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}}\lambda(\alpha)$$

and,

$$\sigma^2_{1,\underline{g}_R} = 1 - V_{h(\underline{G})} - V_M - \frac{\big(r(h_L^2 - V_{A_{h(\underline{G})}} - V_M) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}})\big)^2}{1 - V_{h(\underline{G})} - r^2 V_M}\lambda(\alpha)\big(\lambda(\alpha) - \alpha\big)$$

with $\alpha = \frac{T - h(\underline{g}_R) - r m_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}}$, and, $\lambda(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)}$.

Then the risk of disease for individual $\{I\}$ is approximately:

$$\begin{aligned} &p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R, Y_R = 1) \\ &\quad = p(L_I > T|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R, L_R > T) \\ &\quad = 1 - \Phi\Big(\frac{T - \mu_{1,\underline{g}_R}}{\sigma_{1,\underline{g}_R}}\Big) \end{aligned} \tag{16}$$

where $\mu_{1,\underline{g}_R}$ and $\sigma_{1,\underline{g}_R}$ are defined in Equation (15).

(A.2.2) PA approximation when relative {R} is unaffected

By applying the PA formula to the distribution in Equation (14), we obtain an *approximate* distribution for $L_I|\{\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R, L_R \leq T\}$. This approximate distribution is:

$$L_I|\{\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R, L_R \leq T\} \sim N(\mu_{0,\underline{g}_R}, \sigma^2_{0,\underline{g}_R}) \tag{17}$$

where:

$$\mu_{0,\underline{g}_R} = h(\underline{g}_I) + m_I + \frac{r(h_L^2 - V_{A_{h(\underline{G})}} - V_M) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}})}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}}\Upsilon(\alpha)$$

and,

$$\sigma^2_{0,\underline{g}_R} = 1 - V_{h(\underline{G})} - V_M - \frac{\big(r(h_L^2 - V_{A_{h(\underline{G})}} - V_M) + \theta(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}})\big)^2}{1 - V_{h(\underline{G})} - r^2 V_M}\Upsilon(\alpha)\big(\Upsilon(\alpha) + \alpha\big)$$

with $\alpha = \frac{T - h(\underline{g}_R) - r m_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}}$, and, $\Upsilon(\alpha) = \frac{\phi(\alpha)}{\Phi(\alpha)}$.

The risk of disease for individual $\{I\}$ is then approximately:

$$\begin{aligned} &p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R, Y_R = 0) \\ &\quad = p(L_I > T|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R, L_R \leq T) \\ &\quad = 1 - \Phi\Big(\frac{T - \mu_{0,\underline{g}_R}}{\sigma_{0,\underline{g}_R}}\Big) \end{aligned} \tag{18}$$

where $\mu_{0,\underline{g}_R}$ and $\sigma_{0,\underline{g}_R}$ are defined in Equation (17).

Scenario (B)

In Scenario (B) we wish to calculate:

$$p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y) \tag{19}$$

for $y = 0, 1$; where:

- Y_I is the disease status variable for individual $\{I\}$,
- Y_R is the disease status variable for individual $\{R\}$, a relative of individual $\{I\}$,
- $\underline{G}_I = [G_{1,I}, G_{2,I}, ..., G_{Q,I}]^T$ is the vector containing the Q known major loci variables for individual $\{I\}$, such that; $G_{q,I}$ counts the number of risk alleles at major locus q ($q = 1, 2, ..., Q$),
- $\underline{g}_I = [g_{1,I}, g_{2,I}, ..., g_{Q,I}]^T$ is the vector containing the observed counts of risk alleles for the Q known major loci for individual $\{I\}$,
- M_I is the measurable, additive component for individual $\{I\}$, and,
- m_I is the observed, measurable additive genetic component for individual $\{I\}$.

Again, we shall present an exact method (B.1) and Pearson-Aitken approximate method (B.2).

(B.1) Exact method

We start by using the law of total probability to write:

$$\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y) \\
&= \frac{p(Y_I = 1, Y_R = y|\underline{G}_I = \underline{g}_I, M_I = m_I)}{p(Y_R = y|\underline{G}_I = \underline{g}_I, M_I = m_I)} \\
&= \frac{\sum_{\underline{g}_R} p(Y_I = 1, Y_R = y|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R)p(\underline{G}_R = \underline{g}_R|\underline{G}_I = \underline{g}_I, M_I = m_I)}{\sum_{\underline{g}_R} p(Y_R = y|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R)p(\underline{G}_R = \underline{g}_R|\underline{G}_I = \underline{g}_I, M_I = m_I)}
\end{aligned}$$

where $\sum_{\underline{g}_R} = \sum_{g_{1,R}=0}^{2} \sum_{g_{2,R}=0}^{2} \cdots \sum_{g_{Q,R}=0}^{2}$.

Assuming that:

$$\begin{gathered}
Y_R \perp \underline{G}_I|\{M_I, \underline{G}_R\} \\
\text{and} \\
\underline{G}_R \perp M_I|\underline{G}_I
\end{gathered}$$

then:

$$\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y) \\
&= \frac{\sum_{\underline{g}_R} p(Y_I = 1, Y_R = y|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R)p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I)}{\sum_{\underline{g}_R} p(Y_R = y|M_I = m_I, \underline{G}_R = \underline{g}_R)p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I)}
\end{aligned}$$

Assuming the independence of the Q major loci then:

$$
\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y) \\
&= \frac{\sum_{\underline{g}_R} p(Y_I = 1, Y_R = y|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R) \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I})}{\sum_{\underline{g}_R} p(Y_R = y|M_I = m_I, \underline{G}_R = \underline{g}_R) \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I})}
\end{aligned} \tag{20}
$$

for $y = 0, 1$. The probability $p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I})$ will depend on how $\{I\}$ and $\{R\}$ are related. Please see Appendix B.2.1 of the main thesis for details of this calculation for a variety of relationships.

(B.1.1) When relative {R} is affected

When $Y_R = 1$, the above becomes:

$$
\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = 1) \\
&= \frac{\sum_{\underline{g}_R} p(L_I > T, L_R > T|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R) \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I})}{\sum_{\underline{g}_R} p(L_R > T|M_I = m_I, \underline{G}_R = \underline{g}_R) \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I})} \\
&= \frac{\sum_{\underline{g}_R} \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I}) \int_T^\infty \int_T^\infty f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I) dxdy}{\sum_{\underline{g}_R} \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I}) \int_T^\infty f_{L_R|\underline{G}_R,M_I}(y|\underline{g}_R, m_I) dy}
\end{aligned}
$$

$$
= \frac{\sum_{\underline{g}_R} \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I}) \int_T^\infty \int_T^\infty f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I) dxdy}{\sum_{\underline{g}_R} \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I}) (1 - \Phi\Big(\frac{T - h(\underline{g}_R) - rm_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}}\Big))} \tag{21}
$$

where $f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I)$ is defined using the joint density function of $\{L_I, L_R\}$ given $\{\underline{G}_I = \underline{g}_I, \underline{G}_R = \underline{g}_R, M_I = m_I\}$, given in Scenario (A), Equation (14).

(B.1.2) When relative {R} is unaffected

When $Y_R = 0$, Equation (20) becomes:

$$
\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = 0) \\
&= \frac{\sum_{\underline{g}_R} p(L_I > T, L_R \leq T|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_R = \underline{g}_R) \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I})}{\sum_{\underline{g}_R} p(L_R \leq T|M_I = m_I, \underline{G}_R = \underline{g}_R) \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I})} \\
&= \frac{\sum_{\underline{g}_R} \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I}) \int_{-\infty}^T \int_T^\infty f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I) dxdy}{\sum_{\underline{g}_R} \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I}) \int_{-\infty}^T f_{L_R|\underline{G}_R,M_I}(y|\underline{g}_R, m_I) dy}
\end{aligned}
$$

$$
= \frac{\sum_{\underline{g}_R} \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I}) \int_{-\infty}^T \int_T^\infty f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I) dxdy}{\sum_{\underline{g}_R} \prod_{q=1}^{Q} p(G_{q,R} = g_{q,R}, G_{q,I} = g_{q,I}) \Phi\Big(\frac{T - h(\underline{g}_R) - rm_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}}\Big)} \tag{22}
$$

where $f_{L_I,L_R|\underline{G}_I,\underline{G}_R,M_I}(x, y|\underline{g}_I, \underline{g}_R, m_I)$ is the joint density function of $\{L_I, L_R\}$ given $\{\underline{G}_I = \underline{g}_I, \underline{G}_R = \underline{g}_R, M_I = m_I\}$, given in Scenario (A), Equation (14).

(B.2) Pearson-Aitken approximation

Again, using the law of total probability, but in a slightly different way, we can write:

$$
\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y) \\
&= \sum_{\underline{g}_R} \Big(p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y, \underline{G}_R = \underline{g}_R) \\
&\quad p(\underline{G}_R = \underline{g}_R|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y) \Big) \\
&= \Big[\sum_{\underline{g}_R} \Big(p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y, \underline{G}_R = \underline{g}_R) p(Y_R = y|\underline{G}_R = \underline{g}_R, M_I = m_I) \\
&\quad p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big) \Big] \Big/ \Big[p(Y_R = y|\underline{G}_I = \underline{g}_I, M_I = m_I) p(\underline{G}_I = \underline{g}_I) \Big] \\
&= \Big[\sum_{\underline{g}_R} \Big(p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y, \underline{G}_R = \underline{g}_R) p(Y_R = y|\underline{G}_R = \underline{g}_R, M_I = m_I) \\
&\quad p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big) \Big] \Big/ \Big[\sum_{\underline{g}_R} p(Y_R = y|\underline{G}_R = \underline{g}_R, M_I = m_I) p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big]
\end{aligned} \tag{23}
$$

for $y = 0, 1$; where:

- $p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y, \underline{G}_R = \underline{g}_R)$ can be estimated by the PA approximation outlined in Scenario (A),
- $p(Y_R = y|\underline{G}_R = \underline{g}_R, M_I = m_I)$ can be calculated using the distribution function of:

  $$L_R|\{\underline{G}_R = \underline{g}_R, M_I = m_I\} \sim N(h(\underline{g}_R) + rm_I, 1 - V_{h(\underline{G})} - r^2 V_M)$$

  as derived in Scenario (A) (see Equation (14) for details).

Note that $p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y, \underline{G}_R = \underline{g}_R)$ could also be calculated using the Exact method outlined in Scenario (A). In the Scenario (B) exact method we wrote the probability that we wish to calculate, $p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = y)$, in an equivalent way which explicitly showed the joint distribution of $L_I|\{\underline{G}_I = \underline{g}_I, \underline{G}_R = \underline{g}_R, M_I = m_I\}$ and $L_R|\{\underline{G}_I = \underline{g}_I, \underline{G}_R = \underline{g}_R, M_I = m_I\}$.

(B.2.1) When relative {R} is affected

When $Y_R = 1$:

$$
\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = 1) \\
&= \Big[\sum_{\underline{g}_R} \Big(p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = 1, \underline{G}_R = \underline{g}_R) p(Y_R = 1|\underline{G}_R = \underline{g}_R, M_I = m_I) \\
&\quad p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big) \Big] / \Big[\sum_{\underline{g}_R} p(Y_R = 1|\underline{G}_R = \underline{g}_R, M_I = m_I) p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big] \\
&= \Big[\sum_{\underline{g}_R} \Big(p(L_I > T|\underline{G}_I = \underline{g}_I, M_I = m_I, L_R > T, \underline{G}_R = \underline{g}_R) p(L_R > T|\underline{G}_R = \underline{g}_R, M_I = m_I) \\
&\quad p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big) \Big] / \Big[\sum_{\underline{g}_R} p(L_R > T|\underline{G}_R = \underline{g}_R, M_I = m_I) p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big] \\
&= \Big[\sum_{\underline{g}_R} \Big(p(L_I > T|\underline{G}_I = \underline{g}_I, M_I = m_I, L_R > T, \underline{G}_R = \underline{g}_R) (1 - \Phi\Big(\frac{T - h(\underline{g}_R) - r m_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}} \Big)) \\
&\quad p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big) \Big] / \Big[\sum_{\underline{g}_R} (1 - \Phi\Big(\frac{T - h(\underline{g}_R) - r m_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}} \Big)) p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big]
\end{aligned} \tag{24}
$$

where $p(L_I > T|\underline{G}_I = \underline{g}_I, M_I = m_I, L_R > T, \underline{G}_R = \underline{g}_R)$ can be approximated using Equation (16).

(B.2.2) When relative {R} is unaffected

When $Y_R = 0$:

$$
\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = 0) \\
&= \Big[\sum_{\underline{g}_R} \Big(p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_R = 0, \underline{G}_R = \underline{g}_R) p(Y_R = 0|\underline{G}_R = \underline{g}_R, M_I = m_I) \\
&\quad p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big) \Big] / \Big[\sum_{\underline{g}_R} p(Y_R = 0|\underline{G}_R = \underline{g}_R, M_I = m_I) p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big] \\
&= \Big[\sum_{\underline{g}_R} \Big(p(L_I > T|\underline{G}_I = \underline{g}_I, M_I = m_I, L_R \leq T, \underline{G}_R = \underline{g}_R) p(L_R \leq T|\underline{G}_R = \underline{g}_R, M_I = m_I) \\
&\quad p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big) \Big] / \Big[\sum_{\underline{g}_R} p(L_R \leq T|\underline{G}_R = \underline{g}_R, M_I = m_I) p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big] \\
&= \Big[\sum_{\underline{g}_R} \Big(p(L_I > T|\underline{G}_I = \underline{g}_I, M_I = m_I, L_R \leq T, \underline{G}_R = \underline{g}_R) \Phi\Big(\frac{T - h(\underline{g}_R) - r m_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}} \Big) \\
&\quad p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big) \Big] / \Big[\sum_{\underline{g}_R} \Phi\Big(\frac{T - h(\underline{g}_R) - r m_I}{\sqrt{1 - V_{h(\underline{G})} - r^2 V_M}} \Big) p(\underline{G}_R = \underline{g}_R, \underline{G}_I = \underline{g}_I) \Big]
\end{aligned} \tag{25}
$$

where $p(L_I > T|\underline{G}_I = \underline{g}_I, M_I = m_I, L_R \leq T, \underline{G}_R = \underline{g}_R)$ can be approximated using Equation (18).

2.3 ★ LTMM with Q major loci and S relatives

In Section (2.2) we describe how to calculate the probability that individual $\{I\}$ is affected with a disease of interest given:

- a measured polygenic variable for individual $\{I\}$; M_I,
- Q known major risk loci for individual $\{I\}$; $\underline{G}_I$, and,
- the disease status of a relative of individual $\{I\}$; Y_R.

We now extend this to the scenario where the disease status of S relatives of individual $\{I\}$ are known. That is, we wish to estimate:

$$p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}) \tag{26}$$

where:

- Y_I is the disease status for individual $\{I\}$,
- $\underline{R} = [R_1, ..., R_S]^T$ is a vector denoting the S relatives of individual $\{I\}$,
- $\underline{Y}_{\underline{R}} = [Y_{R_1}, ..., Y_{R_S}]^T$ is the vector containing the disease status variable for the S relatives of individual $\{I\}$,
- $\underline{y}_{\underline{R}} = [y_{R_1}, ..., y_{R_S}]^T$ is the vector containing the observed disease status for the S relatives of individual $\{I\}$,
- $\underline{G}_I = [G_{1,I}, ..., G_{Q,I}]^T$ is the vector containing the random risk allele count variables at the Q major loci for individual $\{I\}$,
- $\underline{g}_I = [g_{1,I}, ..., g_{Q,I}]^T$ is the vector containing the observed risk allele counts at the Q major loci for individual $\{I\}$,
- M_I is the measurable, additive genetic component variable for individual $\{I\}$, and,
- m_I is the observed, measurable, additive genetic component for individual $\{I\}$.

As usual, to gain an estimate of the risk in Equation (26) we first need to find:

$$p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}) \tag{27}$$

where:

- $\underline{G}_{\underline{R}} = [\underline{G}_{R_1}, ..., \underline{G}_{R_S}]^T$ contains the S random vectors; $\underline{G}_{R_s}$ for $s = 1, ..., S$, which themselves contain the Q random major risk loci variables for a relative s,
- $\underline{g}_{\underline{R}} = [\underline{g}_{R_1}, ..., \underline{g}_{R_S}]^T$ contains the S vectors; $\underline{g}_{R_s}$ for $s = 1, ..., S$, which themselves contain the observed major risk loci variables for relative s, and,
- all else as defined above.

We shall discuss how to calculate both of these risks. We start with Equation (27) in Scenario (A), and then use these results to calculate Equation (26) in Scenario (B).

Scenario (A)

(A.1) Exact method

Assuming that $\underline{Y}_{\underline{R}} \perp \underline{G}_I | \{\underline{G}_{\underline{R}}, M_I\}$, then we start by re-writing the Equation (27) as:

$$\begin{aligned} & p(Y_I = 1 | \underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}) \\ &= \frac{p(Y_I = 1, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}} | \underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}, M_I = m_I)}{p(\underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}} | \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}, M_I = m_I)} \end{aligned}$$

To calculate this risk using multivariate integration we need to define the following joint, conditional distribution:

$$\begin{bmatrix} L_I | \underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}, M_I = m_I \\ L_{R_1} | \underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}, M_I = m_I \\ \vdots \\ L_{R_S} | \underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}, M_I = m_I \end{bmatrix} = \begin{bmatrix} L_I | \underline{G}_I = \underline{g}_I, M_I = m_I \\ L_{R_1} | \underline{G}_{R_1} = \underline{g}_{R_1}, M_I = m_I \\ \vdots \\ L_{R_S} | \underline{G}_I = \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I \end{bmatrix}$$

where we assume that:

- $L_I \perp \underline{G}_{\underline{R}} | \{\underline{G}_I, M_I\}$,
- $L_{R_s} \perp \underline{G}_I | \{\underline{G}_{R_s}, M_I\}$; for $s = 1, 2, ..., S$, and,
- $L_{R_s} \perp \underline{G}_{R_t} | \{\underline{G}_{R_s}, M_I\}$; for $s, t = 1, 2, ..., S$ but $s \neq t$.

(A.1.0) The required joint distribution

Using the conditional disease model from Section (2.2), we start with the following multivariate normal distribution:

$$\begin{bmatrix} L_I | \underline{G}_I = \underline{g}_I \\ L_{R_1} | \underline{G}_{R_1} = \underline{g}_{R_1} \\ \vdots \\ L_{R_S} | \underline{G}_I = \underline{G}_{R_S} = \underline{g}_{R_S} \\ M_I \end{bmatrix} \sim N_{S+2}(\mu^*, \Sigma^*) \tag{28}$$

where:

$$\circ\mu^* = \begin{bmatrix} h(\underline{g}_I) \\ h(\underline{g}_{R_1}) \\ \vdots \\ h(\underline{g}_{R_S}) \\ 0 \end{bmatrix}$$

$$\circ\Sigma^* = \begin{bmatrix} \Sigma^*_I & \Sigma^*_{I,\underline{R}} & \Sigma^*_{I,M_I} \\ \Sigma^*_{\underline{R},I} & \Sigma^*_{\underline{R}} & \Sigma^*_{\underline{R},M_I} \\ \Sigma^*_{M_I,I} & \Sigma^*_{M_I,\underline{R}} & \Sigma^*_{M_I} \end{bmatrix}$$

and:

- $h(\underline{g}) = \sum_{i=1}^{Q} \sum_{j=1}^{2} \beta_{ij} I_{g_i = j}$, such that:
  - $L | \{\underline{G} = \underline{g}\} \sim N(h(\underline{g}), 1 - V_{h(\underline{G})})$, and,

- $V_{h(\underline{G})} = Var[h(\underline{G})] = Var[\sum_{i=1}^{Q}\sum_{j=1}^{2}\beta_{ij}I_{G_i=j}]$.

- $\Sigma_I^* = 1 - V_{h(\underline{G})}$,

- $\Sigma_{I,\underline{R}}^* = \Sigma_{\underline{R},I}^{*T}$ is a 1xS matrix where the i^{th} column contains:

$$\Sigma_{I,\underline{R}}^*(1,i) = r_{I,R_i}(h_L^2 - V_{A_{h(\underline{G})}}) + \theta_{I,R_i}(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}})$$

for $i = 1, 2, ..., S$, and where:

  - $H_L^2 = Var[G] = Var[A + D] = V_A + V_D$ is the broad-sense heritability of disease,
  - $h_L^2 = Var[A] = V_A$ is the narrow-sense heritability,
  - $V_{A_{h(\underline{G})}} = \sum_{q=1}^{Q}\big(a_q + d_q(1-2f_q)\big)^2 2f_q(1-f_q)$ is the part of the additive genetic variation (narrow-sense heritability) determined by $\underline{G} = [G_1, ..., G_Q]^T$, where:
    - $G_q \sim Binom(2, f_q)$,
    - $a_q = \frac{\beta_{q2}}{2}$,
    - $d_q = \beta_{q1} - \frac{\beta_{q2}}{2}$, and
    - $E[L|\underline{G} = \underline{g}] = h(\underline{g}) = \sum_{q=1}^{Q}\sum_{l=1}^{2}\beta_{ql}I_{g_q=l} + \sum_{s=1}^{S}\sum_{l=1}^{\nu_s}\beta_{(Q+s)l}I_{e_s=l}$,
  - $V_{D_{h(\underline{G})}} = \sum_{q=1}^{Q}\big(d_q 2f_q(1-f_q)\big)^2$ is the part of the quasi-dominant variation, V_D, determined by $\underline{G}$,
  - r_{I,R_i} is the coefficient of relatedness between individual $\{I\}$ and relative $\{R_i\}$; $i = 1, 2, ..., S$, and,
  - θ_{I,R_i} is the coefficient of coancestry between individual $\{I\}$ and relative $\{R_i\}$; $i = 1, 2, ..., S$.

  For more details on the variance components above please see Section (2.2), Scenario (A), and Section 3.3.2 of the main thesis. For r and θ values between common relative pairs please see Table B.1, Appendix B of the main thesis.

- $\Sigma_{I,M_I}^* = \Sigma_{M_I,I}^{*T} = V_M$ where:
  - $M \sim N(0, V_M)$

- $\Sigma_{\underline{R}}^*$ is an SxS matrix where the i^{th} row, j^{th} column contains:

$$\Sigma_{\underline{R}}^*(i,j) = \begin{cases} 1 - V_{h(\underline{G})} & \text{when } i = j, \\ r_{R_i,R_j}\big(h_L^2 - V_{A_{h(\underline{G})}}\big) + \theta_{R_i,R_j}\big(H_L^2 - h_L^2 - V_{D_{h(\underline{G})}}\big) & \text{when } i \neq j; \end{cases}$$

for $i, j = 1, 2, ..., S$, and where:

  - r_{R_i,R_j} is the coefficient of relatedness between $\{R_i\}$ and $\{R_j\}$; $i, j = 1, 2, ...S$,
  - θ_{R_i,R_j} is the coefficient of coancestry between relatives $\{R_i\}$ and $\{R_j\}$; $i, j = 1, 2, ..., S$.

- $\Sigma_{\underline{R},M_I}^* = \Sigma_{M_I,\underline{R}}^{*T}$ is a 1xS matrix where the i^{th} column is:

$$\Sigma_{\underline{R},M_I}^*(1,i) = r_{I,R_i}V_M$$

for $i = 1, 2, ...S$, and,

- $\Sigma^*_{M_I} = V_M$.

Then, applying standard statistical theory to the distribution in Equation (28), we gain the required joint distribution:

$$\begin{bmatrix} L_I|\underline{G}_I = \underline{g}_I, M_I = m_I \\ L_{R_1}|\underline{G}_{R_1} = \underline{g}_{R_1}, M_I = m_I \\ \vdots \\ L_{R_S}|\underline{G}_I = \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I \end{bmatrix} \sim N_{S+1}(\mu, \Sigma) \tag{29}$$

where:

$$\circ\mu = \begin{bmatrix} h(\underline{g}_I) + m_I \\ h(\underline{g}_{R_1}) + r_{I,R_1} m_I \\ \vdots \\ h(\underline{g}_{R_S}) + r_{I,R_S} m_I \end{bmatrix}$$

$$\circ\Sigma = \begin{bmatrix} \Sigma_I & \Sigma_{I,\underline{R}} \\ \Sigma_{\underline{R},I} & \Sigma_{\underline{R}} \end{bmatrix}$$

and:

- $\Sigma_I = 1 - V_{h(\underline{G})} - V_M$,
- $\Sigma_{I,\underline{R}} = \Sigma^T_{\underline{R},I}$ is a 1xS matrix such that the i^{th} column is:

  $$\Sigma_{I,\underline{R}}(1,i) = r_{I,R_i}\big(h^2_L - V_{A_{h(\underline{G})}} - V_M\big) + \theta_{I,R_i}\big(H^2_L - h^2_L - V_{D_{h(\underline{G})}}\big)$$

  for $i = 1, 2, ..., S$, and,
- $\Sigma_{\underline{R}}$ is an SxS matrix where the i^{th} row and j^{th} column contains:

  $$\Sigma_{\underline{R}}(i,j) = \begin{cases} 1 - V_{h(\underline{G})} - r^2_{I,R_i} V_M & \text{when } i = j, \\ r_{R_i,R_j}\big(h^2_L - V_{A_{h(\underline{G})}}\big) + \theta_{R_i,R_j}\big(H^2_L - h^2_L - V_{D_{h(\underline{G})}}\big) - r_{I,R_i} r_{I,R_j} V_M & \text{when } i \neq j; \end{cases}$$

  for $i, j = 1, 2, ..., S$.

The distribution in Equation (29) is then used in multivariate integration to calculate the risk in Equation (27). As an example, let us assume that we know the disease status and major loci profiles of S relatives to individual $\{I\}$, where relatives $\{R_1, R_2, ..., R_\xi\}$ are affected and relatives $\{R_{\xi+1}, R_{\xi+2}, ..., R_S\}$ are unaffected. Then:

$$\begin{aligned} &p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_{R_1} = 0, ..., Y_{R_\xi} = 0, Y_{R_{\xi+1}} = 1, ..., Y_{R_S} = 1, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}) \\ &= \frac{\underline{\int}^T_{-\infty} \underline{\int}^\infty_T \int^\infty_T f_{L_I, \underline{L}_{\underline{R}}|\underline{G}_I, \underline{G}_{\underline{R}}, M_I}(x, y_1, \ldots, y_S|\underline{g}_I, \underline{g}_{\underline{R}}, m_I) dx dy_1 \ldots dy_S}{\underline{\int}^T_{-\infty} \underline{\int}^\infty_T f_{\underline{L}_{\underline{R}}|\underline{G}_{\underline{R}}, M_I}(y_1, \ldots, y_S|\underline{g}_{1,\underline{R}}, m_I) dy_1 \ldots dy_S} \end{aligned}$$

where:

- $\underline{\int}^\infty_T = \int^\infty_T \ldots \int^\infty_T$ contains the integrals for the ξ affected relatives, and,

- $\underline{\int}_{-\infty}^{T} = \int_{-\infty}^{T} \dots \int_{-\infty}^{T}$ contains the integrals for the remaining $S - \xi$ unaffected relatives.

(A.2) Pearson-Aitken approximate method

Using the distribution of:

$$\begin{bmatrix} L_I|\underline{G}_I = \underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I \\ L_{R_1}|\underline{G}_I = \underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I \\ \dots \\ L_{R_S}|\underline{G}_I = \underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I \end{bmatrix} = \begin{bmatrix} L_I|\underline{G}_I = \underline{g}_I, M_I = m_I \\ L_{R_1}|\underline{G}_{R_1} = \underline{g}_{R_1}, M_I = m_I \\ \dots \\ L_{R_S}|\underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I \end{bmatrix}$$

given in Equation (29), we iteratively apply the PA approximation to gain the required estimate.

Scenario (B)

Recall that in Scenario (B) we wish to calculate:

$$p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}})$$

We shall now describe solutions via: (B.1) the exact method, and, (B.2) a Pearson-Aitken approximate method.

(B.1) Exact method

Assuming that:

- $\underline{Y}_{\underline{R}} \perp \underline{G}_I|\{M_I, \underline{G}_{\underline{R}}\}$, and,
- $\underline{G}_{\underline{R}} \perp M_I|\underline{G}_I$,

we can write:

$$\begin{aligned} &p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}) \\ &= \frac{p(Y_I = 1, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}|\underline{G}_I = \underline{g}_I, M_I = m_I)}{p(\underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}|\underline{G}_I = \underline{g}_I, M_I = m_I)} \\ &= \frac{\sum_{\underline{g}_{\underline{R}}} p(Y_I = 1, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}) p(\underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})}{\sum_{\underline{g}_{\underline{R}}} p(\underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}|M_I = m_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}) p(\underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})} \end{aligned}$$

where $\sum_{\underline{g}_{\underline{R}}} = \sum_{\underline{g}_{R_1}} \sum_{\underline{g}_{R_2}} \dots \sum_{\underline{g}_{R_S}}$, and $\sum_{\underline{g}_{R_s}} = \sum_{g_{1,R_s}=0}^{2} \sum_{g_{2,R_s}=0}^{2} \dots \sum_{g_{Q,R_s}=0}^{2}$; $s = 1, 2, ..., S$.

$p(\underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})$ will depend upon the relationships between $\{I, R_1, R_2, \dots, R_S\}$.

The joint distribution in Equation (29) can then be used to estimate this risk.

Again, as in Section (2.2), bespoke code would need to be written for every new $\{I, R_1, R_2, \dots, R_S\}$ scenario.

(B.2) Pearson-Aitken approximate method

As a formality, and for completeness, we present the PA approximate method. However, it is the authors recommendation that the exact method be used.

We write:

$$
\begin{aligned}
&p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}) \\
&= \sum_{\underline{g}_{\underline{R}}} p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}) \\
&\quad p(\underline{G}_{\underline{R}} = \underline{g}_{\underline{R}}|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}) \\
&= \frac{\sum_{\underline{g}_{\underline{R}}} \omega_{\underline{g}_{\underline{R}}} \theta_{R_S,\underline{g}_{\underline{R}}} (\prod_{s=1}^{S} \theta_{R_s,\underline{g}_{\underline{R}}}) p(\underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})}{\sum_{\underline{g}_{\underline{R}}} \theta_{R_S,\underline{g}_{\underline{R}}} (\prod_{s=1}^{S} \theta_{R_s,\underline{g}_{\underline{R}}}) p(\underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})}
\end{aligned} \tag{30}
$$

where:

- $\omega_{\underline{g}_{\underline{R}}} = p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})$,
- $\theta_{R_s,\underline{g}_{\underline{R}}} = p(Y_{R_s} = y_{R_s}|Y_{R_{s+1}} = y_{R_{s+1}}, ..., Y_{R_S} = y_{R_S}, \underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})$, for $s = 1, 2, ..., S-1$, and,
- $\theta_{R_S,\underline{g}_{\underline{R}}} = p(Y_{R_S} = y_{R_S}|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})$

We then start with the joint conditional distribution of:

$$
\begin{bmatrix}
L_I|\underline{G}_I = \underline{g}_I, M_I = m_I \\
L_{R_1}|\underline{G}_{R_1} = \underline{g}_{R_1}, M_I = m_I \\
\vdots \\
L_{R_S}|\underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I
\end{bmatrix} =
\begin{bmatrix}
L_I|\underline{G}_I = \underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I \\
L_{R_1}|\underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I \\
\vdots \\
L_{R_S}|\underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I
\end{bmatrix}
$$

provided in Equation (29).

$\theta_{R_S,\underline{g}_{\underline{R}}} = p(Y_{R_S} = y_{R_S}|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})$ can be calculated using the marginal distribution of $L_{R_S}|\underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I$ in Equation (29).

Then we iteratively apply the PA approximation S times in total. At each iteration we store the resulting distribution. The first $S-1$ iterations provide the joint distributions of:

$$
\begin{bmatrix}
L_I|\underline{G}_I = \underline{g}_I, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S} \\
L_{R_1}|\underline{G}_{R_1} = \underline{g}_{R_1}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S} \\
\vdots \\
L_{R_{s-1}}|\underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S}
\end{bmatrix} =
$$

$$
\begin{bmatrix}
L_I|\underline{G}_I = \underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S} \\
L_{R_1}|\underline{G}_I = \underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S} \\
\vdots \\
L_{R_{s-1}}|\underline{G}_I = \underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S}
\end{bmatrix}
$$

for $s = S, S-1, ..., 2$.

From each of these joint conditional distributions we can extract the marginal distribution of $L_{R_{s-1}}|\underline{G}_I = \underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I, Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S}$. Therefore we can calculate $\theta_{R_{s-1},\underline{g}_{\underline{R}}} = p(Y_{R_{s-1}} = y_{R_{s-1}}|Y_{R_s} = y_{R_s}, ..., Y_{R_S} = y_{R_S}, \underline{G}_I = \underline{g}_I, M_I = m_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})$, for $s = S, S-1, ..., 2$.

At the final (S^{th}) PA approximation application, we gain the conditional distribution of $L_I|\underline{G}_I = \underline{g}_I, \underline{G}_{R_1} = \underline{g}_{R_1}, ..., \underline{G}_{R_S} = \underline{g}_{R_S}, M_I = m_I, Y_{R_1} = y_{R_1}, ..., Y_{R_S} = y_{R_S}$. From this $\omega_{\underline{g}_{\underline{R}}} = p(Y_I = 1|\underline{G}_I = \underline{g}_I, M_I = m_I, \underline{Y}_{\underline{R}} = \underline{y}_{\underline{R}}, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})$ can be found.

All of these probabilities are entered into Equation (30) and, along with $p(\underline{G}_I = \underline{g}_I, \underline{G}_{\underline{R}} = \underline{g}_{\underline{R}})$, can be used to approximate the required risk.

3 Log linear risk model

3.1 ★ Risk of disease for an individual given a polygenic risk score, environmental risk variables and an unaffected relative

In Section 3.4.2 of the main thesis, we derived a formula to calculate:

$$p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, Y_R = 1)$$

We now wish to derive a formula for:

$$p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, Y_R = 0)$$

Using the notation and independence relations defined in Section 3.4.2 of the main thesis, the disease model definition in that section, and the law of total probability; we write:

$$\begin{aligned}
&p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, Y_R = 0) \\
&= \sum_{\underline{u}_I} p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, \underline{U}_I = \underline{u}_I) \frac{p(Y_R = 0|M_I = m_I, \underline{U}_I = \underline{u}_I)p(\underline{U}_I = \underline{u}_I)}{p(Y_R = 0|M_I = m_I)} \\
&= \sum_{\underline{u}_I} \exp\big(\beta_0\big) \exp\big(\sqrt{V_M} m_I\big) \exp\big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\big) \exp\big(\xi(\underline{u}_I)\big) \frac{p(Y_R = 0|M_I = m_I, \underline{U}_I = \underline{u}_I)p(\underline{U}_I = \underline{u}_I)}{p(Y_R = 0|M_I = m_I)} \\
&= \exp\big(\beta_0\big) \exp\big(\sqrt{V_M} m_I\big) \exp\big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\big) \\
&\quad \sum_{\underline{u}_I} \exp\big(\xi(\underline{u}_I)\big) p(\underline{U}_I = \underline{u}_I) \frac{1 - p(Y_R = 1|M_I = m_I, \underline{U}_I = \underline{u}_I)}{1 - p(Y_R = 1|M_I = m_I)} \\
&= \frac{\exp\big(\beta_0\big) \exp\big(\sqrt{V_M} m_I\big) \exp\big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\big)}{1 - p(Y_R = 1|M_I = m_I)} \Big(\sum_{\underline{u}_I} \exp\big(\xi(\underline{u}_I)\big) p(\underline{U}_I = \underline{u}_I) - \\
&\quad \sum_{\underline{u}_I} \exp\big(\xi(\underline{u}_I)\big) p(\underline{U}_I = \underline{u}_I) p(Y_R = 1|M_I = m_I, \underline{U}_I = \underline{u}_I) \Big)
\end{aligned}$$

Recall from Section 3.4.2 of the main thesis:

$$\begin{aligned}
&p(Y_R = 1|M_I = m_I, \underline{U}_I = \underline{u}_I) \\
&= \exp\big(\beta_0\big) E[\exp\big(\sum_{s=1}^{S} \alpha_{J+s} E_s\big)] \Big(\int \exp\big(\sqrt{V_M} m_R\big) f_{M_R|M_I}(m_R|m_I) dm_R \Big) \\
&\quad \Big(\sum_{\underline{u}_R} \exp\big(\xi(\underline{u}_R)\big) p(\underline{U}_R = \underline{u}_R|\underline{U}_I = \underline{u}_I) \Big)
\end{aligned}$$

and:

$$p(Y_R = 1|M_I = m_I)$$
$$= \exp\big(\beta_0\big)E[\exp\big(\sum_{s=1}^{S}\alpha_{J+s}E_s\big)]E[\exp\big(\xi(\underline{U})\big)]\Big(\int \exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R\Big)$$

Also recall that K can be written as:

$$K = \exp\big(\beta_0\big)E[\exp\big(\sum_{s=1}^{S}\alpha_{J+s}E_s\big)]E[\exp\big(\xi(\underline{U})\big)]E[\exp\big(\sqrt{V_M}M\big)]$$

Using this equation for K we can write:

$$p(Y_R = 1|M_I = m_I, \underline{U}_I = \underline{u}_I)$$
$$= K\Big(\frac{\int \exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R}{E[\exp\big(\sqrt{V_M}M\big)]}\Big)\Big(\frac{\sum_{\underline{u}_R}\exp\big(\xi(\underline{u}_R)\big)p(\underline{U}_R = \underline{u}_R|\underline{U}_I = \underline{u}_I)}{E[\exp\big(\xi(\underline{U})\big)]}\Big)$$

and:

$$p(Y_R = 1|M_I = m_I)$$
$$= K\frac{\int \exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R}{E[\exp\big(\sqrt{V_M}M\big)]}$$

Therefore:

$$p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, Y_R = 0)$$
$$= \frac{\exp\big(\beta_0\big)\exp\big(\sqrt{V_M}m_I\big)\exp\big(\sum_{s=1}^{S}\alpha_{J+s}e_{I,s}\big)}{E[\exp\big(\sqrt{V_M}M\big)] - K\int \exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R}$$
$$\Big(E[\exp\big(\sqrt{V_M}M\big)]\sum_{\underline{u}_I}\exp\big(\xi(\underline{u}_I)\big)p(\underline{U}_I = \underline{u}_I) - K\int \exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R\frac{C_{\underline{U}_I,\underline{U}_R}}{C^2_{\underline{U}}}\Big)$$

where:

$$C_{\underline{U}_I,\underline{U}_R} = \sum_{\underline{u}_I}\sum_{\underline{u}_R}\exp\big(\xi(\underline{u}_I)\big)\exp\big(\xi(\underline{u}_R)\big)p(\underline{U}_I = \underline{u}_I, \underline{U}_R = \underline{u}_R)$$

and:

$$C_{\underline{U}} = \sum_{\underline{u}}\exp\big(\xi(\underline{u})\big)p(\underline{U} = \underline{u}) = E[\exp\big(\xi(\underline{U})\big)]$$

f

Recall:

$$\lambda_{R,\underline{U}} = \frac{C_{\underline{U}_I,\underline{U}_R}}{C^2_{\underline{U}}} = \frac{\lambda_R}{\lambda_{R,M}} = \frac{\lambda_R}{\exp\big(rV_M\big)}$$

Then we write:

$$
\begin{aligned}
&p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, Y_R = 0) \\
&= \frac{\exp\big(\beta_0\big)\exp\big(\sqrt{V_M}m_I\big)\exp\big(\sum_{s=1}^{S}\alpha_{J+s}e_{I,s}\big)}{E[\exp\big(\sqrt{V_M}M\big)] - K\int\exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R} \\
&\quad\Big(E[\exp\big(\sqrt{V_M}M\big)]E[\exp\big(\xi(\underline{U})\big)] - KE[\exp\big(\xi(\underline{U})\big)]\lambda_{R,\underline{U}}\int\exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R\Big) \\
&= \exp\big(\beta_0\big)\exp\big(\sqrt{V_M}m_I\big)\exp\big(\sum_{s=1}^{S}\alpha_{J+s}e_{I,s}\big)E[\exp\big(\xi(\underline{U})\big)] \\
&\quad\frac{\Big(E[\exp\big(\sqrt{V_M}M\big)] - K\lambda_{R,\underline{U}}\int\exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R\Big)}{E[\exp\big(\sqrt{V_M}M\big)] - K\int\exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R} \\
&= K\frac{\exp\big(\sqrt{V_M}m_I\big)}{E[\exp\big(\sqrt{V_M}M\big)]}\frac{\exp\big(\sum_{s=1}^{S}\alpha_{J+s}e_{I,s}\big)}{E[\exp\big(\sum_{s=1}^{S}\alpha_{J+s}E_s\big)]} \\
&\quad\frac{\Big(E[\exp\big(\sqrt{V_M}M\big)] - K\frac{\lambda_R}{\exp\big(rV_M\big)}\int\exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R\Big)}{E[\exp\big(\sqrt{V_M}M\big)] - K\int\exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R}
\end{aligned}
$$

Again, recall that:

$$E[\exp\big(\sqrt{V_M}M\big)] = \exp\big(\frac{1}{2}V_M\big)$$

Then, we just need to calculate:

$$\int\exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R$$

to gain a formula for $p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, Y_R = 0)$.

Recall that the joint distribution for $\{M_R, M_I\}$ is:

$$\begin{bmatrix} M_R \\ M_I \end{bmatrix} \sim N\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\Big)$$

Using standard statistical theory, defined in Section 3.3 of the main thesis we know that:

$$M_R|\{M_I = m_I\} \sim N(rm_I, 1 - r^2)$$

Then:

$$
\begin{aligned}
&\int\exp\big(\sqrt{V_M}m_R\big)f_{M_R|M_I}(m_R|m_I)dm_R \\
&= \frac{1}{\sqrt{2\pi(1-r^2)}}\int\exp\big(\sqrt{V_M}m_R\big)\exp\big(-\frac{1}{2(1-r^2)}(m_R - rm_I)^2\big)dm_R \\
&= \frac{1}{\sqrt{2\pi(1-r^2)}}\int\exp\big(-\frac{1}{2(1-r^2)}(m_R^2 - 2m_R[(1-r^2)\sqrt{V_M} + rm_I] + r^2m_I^2)\big)dm_R \\
&= \frac{1}{\sqrt{2\pi(1-r^2)}}\int\exp\big(-\frac{1}{2(1-r^2)}(\Big(m_R - [(1-r^2)\sqrt{V_M} + rm_I]\Big)^2 \\
&\qquad - (1-r^2)^2V_M - 2rm_I(1-r^2)\sqrt{V_M})\big)dm_R \\
&= \frac{1}{\sqrt{2\pi(1-r^2)}}\exp\big(\frac{(1-r^2)V_M}{2} + r\sqrt{V_M}m_I\big)\int\exp\big(-\Big(\frac{m_R - [(1-r^2)\sqrt{V_M} + rm_I]}{\sqrt{2(1-r^2)}}\Big)^2\big)dm_R
\end{aligned}
$$

Using the following change of variable:

$$x = \frac{m_R - [(1-r^2)\sqrt{V_M} + rm_I]}{\sqrt{2(1-r^2)}}$$

with:

$$\frac{dx}{dm_R} = \frac{1}{\sqrt{2(1-r^2)}}$$

we get:

$$\begin{aligned}
\int \exp\big(\sqrt{V_M} m_R\big) f_{M_R|M_I}(m_R|m_I) dm_R &= \frac{1}{\sqrt{\pi}} \exp\big(\frac{(1-r^2)V_M}{2} + r\sqrt{V_M} m_I\big) \int \exp\big(-x^2\big) dx \\
&= \frac{1}{\sqrt{\pi}} \exp\big(\frac{(1-r^2)V_M}{2} + r\sqrt{V_M} m_I\big)\sqrt{\pi} \\
&= \exp\big(\frac{(1-r^2)V_M}{2} + r\sqrt{V_M} m_I\big) \\
&= \exp\big(\frac{1}{2}V_M\big) \exp\big(-\frac{r^2}{2}V_M\big) \exp\big(r\sqrt{V_M} m_I\big)
\end{aligned}$$

Giving us:

$$\begin{aligned}
&p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, Y_R = 0) \\
&= K \frac{\exp\big(\sqrt{V_M} m_I\big)}{E[\exp\big(\sqrt{V_M} M\big)]} \frac{\exp\big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\big)}{E[\exp\big(\sum_{s=1}^{S} \alpha_{J+s} E_s\big)]} \frac{1 - \lambda_R K \exp\big(-rV_M\big) \exp\big(-\frac{r^2}{2}V_M\big) \exp\big(r\sqrt{V_M} m_I\big)}{1 - K \exp\big(-\frac{r^2}{2}V_M\big) \exp\big(r\sqrt{V_M} m_I\big)}
\end{aligned}$$

If of interest, the relative risk for family history of disease is then:

$$\begin{aligned}
RR_{FH} &= \frac{p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, Y_R = 1)}{p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, Y_R = 0)} \\
&= \frac{\lambda_R - \lambda_R K \exp\big(-\frac{r^2}{2}V_M\big) \exp\big(r\sqrt{V_M} m_I\big)}{\exp\big(rV_M\big) - \lambda_R K \exp\big(-\frac{r^2}{2}V_M\big) \exp\big(r\sqrt{V_M} m_I\big)}
\end{aligned}$$

3.2 ★ Risk of disease for an individual given a polygenic risk score, environmental risk variables and multiple affected relatives

Here we know that N relatives of individual $\{I\}$, denoted by $\{R_1, ..., R_N\}$, are affected. We want to estimate:

$$p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, \underline{Y}_R = \underline{1})$$

where $\underline{Y}_R = [Y_{R_1}, ..., Y_{R_N}]^T$.

Using the total law of probability we write:

$$\begin{aligned}
&p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, \underline{Y}_R = \underline{1}) \\
&= \sum_{\underline{u}_I} p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, \underline{U}_I = \underline{u}_I) p(\underline{U}_I = \underline{u}_I) \frac{p(\underline{Y}_R = \underline{1}|M_I = m_I, \underline{U}_I = \underline{u}_I)}{p(\underline{Y}_R = \underline{1}|M_I = m_I)}
\end{aligned}$$

Using methods similar to Section 3.4.2 of the main thesis, it can be shown that:

$$\begin{aligned}
&p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, \underline{Y}_R = \underline{1}) \\
&= \exp\big(\beta_0\big) \exp\big(\sqrt{V_M} m_I\big) \exp\big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\big) \\
&\quad \frac{\sum_{\underline{u}_I} \sum_{\underline{u}_{\underline{R}}} \exp\big(\xi(\underline{u}_I)\big) \prod_{i=1}^{N} \exp\big(\xi(\underline{u}_{R_i})\big) p(\underline{U}_I = \underline{u}_I, \underline{U}_{\underline{R}} = \underline{u}_{\underline{R}})}{\sum_{\underline{u}_{\underline{R}}} \prod_{i=1}^{N} \exp\big(\xi(\underline{u}_{R_i})\big) p(\underline{U}_{\underline{R}} = \underline{u}_{\underline{R}})}
\end{aligned}$$

Again, similarly to Section 3.4.2 of the main thesis, we can write:

$$\begin{aligned}
\lambda_{\underline{R}} &= \frac{p(Y_I = 1|\underline{Y}_R = \underline{1})}{K} \\
&= \lambda_{\underline{R},M} \lambda_{\underline{R},\underline{U}}
\end{aligned}$$

where:

$$\lambda_{\underline{R},M} = \frac{\int \underline{\int} \exp\big(\sqrt{V_M} m_I\big) \prod_{i=1}^{N} \exp\big(\sqrt{V_M} m_{R_i}\big) f_{M_I,\underline{M}_R}(m_I, \underline{M}_R) dm_I d\underline{m}_R}{\Big(\underline{\int} \prod_{i=1}^{N} \exp\big(\sqrt{V_M} m_{R_i}\big) f_{\underline{M}_R}(\underline{m}_R) d\underline{m}_R\Big)\Big(E[\exp\big(\sqrt{V_M} M\big)]\Big)}$$

and:

$$\lambda_{\underline{R},\underline{U}} = \frac{\sum_{\underline{u}_I} \sum_{\underline{u}_{\underline{R}}} \exp\big(\xi(\underline{u}_I)\big) \prod_{i=1}^{N} \exp\big(\xi(\underline{u}_{R_i})\big) p(\underline{U}_I = \underline{u}_I, \underline{U}_{\underline{R}} = \underline{u}_{\underline{R}})}{\Big(\sum_{\underline{u}_{\underline{R}}} \prod_{i=1}^{N} \exp\big(\xi(\underline{u}_{R_i})\big) p(\underline{U}_{\underline{R}} = \underline{u}_{\underline{R}})\Big)\Big(E[\exp\big(\xi(\underline{U})\big)]\Big)}$$

Therefore:

$$\begin{aligned}
&p(Y_I = 1|M_I = m_I, \underline{E}_I = \underline{e}_I, \underline{Y}_R = \underline{1}) \\
&= \exp\big(\beta_0\big) \exp\big(\sqrt{V_M} m_I\big) \exp\big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\big) \lambda_{\underline{R},\underline{U}} E[\exp\big(\xi(\underline{U})\big)] \\
&= K \frac{\exp\big(\sqrt{V_M} m_I\big)}{E[\exp\big(\sqrt{V_M} M\big)]} \frac{\exp\big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\big)}{E[\exp\big(\sum_{s=1}^{S} \alpha_{J+s} E_s\big)]} \lambda_{\underline{R},\underline{U}}
\end{aligned}$$

To calculate the required risk we therefore need to calculate $\lambda_{\underline{R},\underline{U}} = \frac{\lambda_{\underline{R}}}{\lambda_{\underline{R},M}}$. If an estimate for $\lambda_{\underline{R}}$ is available in the literature, then we just need to calculate $\lambda_{\underline{R},M}$. We cannot derive a generic formula here, because results will depend on the relationships between the relatives and individual $\{I\}$. However, the steps would be to:

1. define the joint, multivariate normal distribution for $[M_I, M_{R_1}, ..., M_{R_N}]^T$ and for $[M_{R_1}, ..., M_{R_N}]^T$,
2. use the PDFs for these distributions in the relevant integrals in $\lambda_{\underline{R},M}$; matrix algebra will need to be used,
3. appropriate, and probably multiple, change of variables will need to be used to calculate integrals, and therefore $\lambda_{\underline{R},M}$.

3.3 ★ Risk of disease for an individual given a polygenic risk score, major risk loci, environmental risk variables and an unaffected relative

In Section 3.4.3 of the main thesis, we showed how to calculate:

$$p(Y_I = 1 | M_I = m_I \underline{G}_I^{'} = \underline{g}_I^{'}, \underline{E}_I = \underline{e}_I, Y_R = 1)$$

That is the probability that an individual $\{I\}$ has the disease of interest, given we observe:

- the polygenic risk score for individual $\{I\}$,
- the Q major risk loci for individual $\{I\}$,
- the S environmental risk variables for individual $\{I\}$, and,
- that a relative to individual $\{I\}$, who is denoted $\{R\}$, is affected with the disease of interest.

We now aim to provide a formula for calculating the above, except that relative $\{R\}$ is unaffected:

$$p(Y_I = 1 | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{E}_I = \underline{e}_I, Y_R = 0)$$

We start by using the law of total probability and writing:

$$\begin{aligned}
&p(Y_I = 1 | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{E}_I = \underline{e}_I, Y_R = 0) \\
&= \sum_{\underline{u}_I} p(Y_I = 1 | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{E}_I = \underline{e}_I, \underline{U}_I = \underline{u}_I) p(\underline{U}_I = \underline{u}_I | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{E}_I = \underline{e}_I, Y_R = 0) \\
&= \sum_{\underline{u}_I} p(Y_I = 1 | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{E}_I = \underline{e}_I, \underline{U}_I = \underline{u}_I) \\
&\quad \frac{p(Y_R = 0 | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{U}_I = \underline{u}_I) p(\underline{U}_I = \underline{u}_I)}{p(Y_R = 0 | M_I = m_I, \underline{G}_I' = \underline{g}_I')} \\
&= \sum_{\underline{u}_I} p(Y_I = 1 | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{E}_I = \underline{e}_I, \underline{U}_I = \underline{u}_I) \\
&\quad \frac{\big(1 - p(Y_R = 1 | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{U}_I = \underline{u}_I)\big) p(\underline{U}_I = \underline{u}_I)}{1 - p(Y_R = 1 | M_I = m_I, \underline{G}_I' = \underline{g}_I')}
\end{aligned}$$

Recall from Section 3.4.3 of the main thesis, that:

$$\begin{aligned}
&p(Y_I = 1 | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{E}_I = \underline{e}_I, \underline{U}_I = \underline{u}_I) \\
&= \exp\big(\beta_0\big) \exp\big(\sqrt{V_M} m_I\big) \exp\Big(\sum_{q=1}^{Q} \sum_{l=1}^{2} \beta_{ql} I_{g_{I,J+q}=l}\Big) \exp\big(\xi(\underline{u}_I)\big) \exp\Big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\Big)
\end{aligned}$$

$$\begin{aligned}
&p(Y_R = 1 | M_I = m_I, \underline{G}_I' = \underline{g}_I', \underline{U}_I = \underline{u}_I) \\
&= \exp\big(\beta_0\big) \Big(\int \exp\big(\sqrt{V_M} m_R\big) f_{M_R | M_I}(m_R | m_I) dm_R \Big) \\
&\quad \Big(\sum_{\underline{g}_R'} \exp\Big(\sum_{q=1}^{Q} \sum_{l=1}^{2} \beta_{ql} I_{g_{R,J+q}=l}\Big) p(\underline{G}_R' = \underline{g}_R' | \underline{G}_I' = \underline{g}_I') \Big) \\
&\quad \Big(\sum_{\underline{u}_R} \exp\big(\xi(\underline{u}_R)\big) p(\underline{U}_R = \underline{u}_R | \underline{U}_I = \underline{u}_I) \Big) \Big(E[\exp\Big(\sum_{s=1}^{S} \alpha_{J+s} E_s\Big)] \Big)
\end{aligned}$$

$$
\begin{aligned}
&p(Y_R = 1|M_I = m_I, \underline{G}'_I = \underline{g}'_I) \\
&= \exp\big(\beta_0\big)\Big(\int \exp\big(\sqrt{V_M} m_R\big) f_{M_R|M_I}(m_R|m_I) dm_R\Big) \\
&\Big(\sum_{\underline{g}'_R} \exp\big(\sum_{q=1}^{Q}\sum_{l=1}^{2} \beta_{ql} I_{g_{R,J+q}=l}\big) p(\underline{G}'_R = \underline{g}'_R|\underline{G}'_I = \underline{g}'_I)\Big) \\
&\Big(E[\exp\big(\xi(\underline{U})\big)]\Big)\Big(E[\exp\big(\sum_{s=1}^{S} \alpha_{J+s} E_s\big)]\Big)
\end{aligned}
$$

$$
\begin{aligned}
K = &\exp\big(\beta_0\big)\Big(E[\exp\big(\sqrt{V_M} M\big)]\Big)\Big(\sum_{\underline{g}'} \exp\big(\sum_{q=1}^{Q}\sum_{l=1}^{2} \beta_{ql} I_{g_{J+q}=l}\big) p(\underline{G}' = \underline{g}')\Big)\Big(E[\exp\big(\xi(\underline{U})\big)]\Big) \\
&\Big(E[\exp\big(\sum_{s=1}^{S} \alpha_{J+s} E_s\big)]\Big)
\end{aligned}
$$

and

$$
\begin{aligned}
\lambda_{R\underline{U}} &= \frac{\sum_{\underline{u}_R}\sum_{\underline{u}_I} \exp\big(\xi(\underline{u}_I)\big) \exp\big(\xi(\underline{u}_R)\big) p(\underline{U}_I = \underline{u}_I, \underline{U}_R = \underline{u}_R)}{E[\exp\big(\xi(\underline{U})\big)]^2} \\
&= \frac{\lambda_R}{\lambda_{RM}\lambda_{R\underline{G}'}}
\end{aligned}
$$

Using these 5 formulae we write:

$$
\begin{aligned}
&p(Y_I = 1|M_I = m_I, \underline{G}'_I = \underline{g}'_I, \underline{E}_I = \underline{e}_I, Y_R = 0) \\
&= K \frac{\exp\big(\sqrt{V_M} m_I\big)}{E[\exp\big(\sqrt{V_M} M\big)]} \frac{\exp\big(\sum_{q=1}^{Q}\sum_{l=1}^{2} \beta_{ql} I_{g_{I,J+q}=l}\big)}{\sum_{\underline{g}'} \exp\big(\sum_{q=1}^{Q}\sum_{l=1}^{2} \beta_{ql} I_{G_{J+q}=l}\big) p(\underline{G}' = \underline{g}')} \frac{\exp\big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\big)}{E[\exp\big(\sum_{s=1}^{S} \alpha_{J+s} E_s\big)]} \\
&\Big[\frac{1 - \lambda_{R\underline{U}} K \frac{\int \exp\big(\sqrt{V_M} m_R\big) f_{M_R|M_I}(m_R|m_I) dm_R}{E[\exp\big(\sqrt{V_M} M\big)]} \frac{\sum_{\underline{g}'_R} \exp\big(\sum_{q=1}^{Q}\sum_{l=1}^{2} \beta_{ql} I_{g_{R,J+q}=l}\big) p(\underline{G}'_R = \underline{g}'_R|\underline{G}'_I = \underline{g}'_I)}{\sum_{\underline{g}'} \exp\big(\sum_{q=1}^{Q}\sum_{l=1}^{2} \beta_{ql} I_{g_{J+q}=l}\big) p(\underline{G}' = \underline{g}')}}{1 - K \frac{\int \exp\big(\sqrt{V_M} m_R\big) f_{M_R|M_I}(m_R|m_I) dm_R}{E[\exp\big(\sqrt{V_M} M\big)]} \frac{\sum_{\underline{g}'_R} \exp\big(\sum_{q=1}^{Q}\sum_{l=1}^{2} \beta_{ql} I_{g_{R,J+q}=l}\big) p(\underline{G}'_R = \underline{g}'_R|\underline{G}'_I = \underline{g}'_I)}{\sum_{\underline{g}'} \exp\big(\sum_{q=1}^{Q}\sum_{l=1}^{2} \beta_{ql} I_{g_{J+q}=l}\big) p(\underline{G}' = \underline{g}')}}\Big]
\end{aligned}
$$

In Section 3.4.2 of the main thesis, it was shown that:

$$
\int \exp\big(\sqrt{V_M} m_R\big) f_{M_R|M_I}(m_R|m_I) dm_R = \exp\big(\frac{1}{2} V_M\big) \exp\big(-\frac{r^2}{2} V_M\big) \exp\big(r\sqrt{V_M} m_I\big)
$$

and

$$
E[\exp\big(\sqrt{V_M} M\big)] = \exp\big(\frac{1}{2} V_M\big)
$$

giving:

$$
\begin{aligned}
&p(Y_I = 1 | M_I = m_I, \underline{G}'_I = \underline{g}'_I, \underline{E}_I = \underline{e}_I, Y_R = 0) \\
&= K \frac{\exp\big(\sqrt{V_M} m_I\big)}{\exp\big(\frac{1}{2} V_M\big)} \frac{\exp\big(\sum_{q=1}^{Q} \sum_{l=1}^{2} \beta_{ql} I_{g_{I,J+q}=l}\big)}{\sum_{\underline{g}'} \exp\big(\sum_{q=1}^{Q} \sum_{l=1}^{2} \beta_{ql} I_{G_{J+q}=l}\big) p(\underline{G}' = \underline{g}')} \frac{\exp\big(\sum_{s=1}^{S} \alpha_{J+s} e_{I,s}\big)}{E[\exp\big(\sum_{s=1}^{S} \alpha_{J+s} E_s\big)]} \\
&\quad \Big[\frac{1 - \lambda_{R\underline{U}} K \exp\big(-\frac{r^2}{2} V_M\big) \exp\big(r\sqrt{V_M} m_I\big) \frac{\sum_{\underline{g}'_R} \exp\big(\sum_{q=1}^{Q} \sum_{l=1}^{2} \beta_{ql} I_{g_{R,J+q}=l}\big) p(\underline{G}'_R = \underline{g}'_R | \underline{G}'_I = \underline{g}'_I)}{\sum_{\underline{g}'} \exp\big(\sum_{q=1}^{Q} \sum_{l=1}^{2} \beta_{ql} I_{g_{J+q}=l}\big) p(\underline{G}' = \underline{g}')}}{1 - K \exp\big(-\frac{r^2}{2} V_M\big) \exp\big(r\sqrt{V_M} m_I\big) \frac{\sum_{\underline{g}'_R} \exp\big(\sum_{q=1}^{Q} \sum_{l=1}^{2} \beta_{ql} I_{g_{R,J+q}=l}\big) p(\underline{G}'_R = \underline{g}'_R | \underline{G}'_I = \underline{g}'_I)}{\sum_{\underline{g}'} \exp\big(\sum_{q=1}^{Q} \sum_{l=1}^{2} \beta_{ql} I_{g_{J+q}=l}\big) p(\underline{G}' = \underline{g}')}} \Big]
\end{aligned}
$$

$$
\sum_{\underline{g}'_R} \exp\big(\sum_{q=1}^{Q} \sum_{l=1}^{2} \beta_{ql} I_{g_{R,J+q}=l}\big) p(\underline{G}'_R = \underline{g}'_R | \underline{G}'_I = \underline{g}'_I)
$$

will need to be derived for each different relationship between $\{I\}$ and $\{R\}$.

References

H.-C. So, J. S. Kwan, S. S. Cherny, and P. C. Sham. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *American Journal of Human Genetics*, 88(5):548–565, 2011.