



Stockholms
universitet

A new method for combining data from heterogeneous sources

Pär Villner

Masteruppsats 2022:7
Matematisk statistik
Juni 2022

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

A new method for combining data from heterogeneous sources

Pär Villner*

June 2022

Abstract

In this thesis, we introduce a new method for summarizing data from heterogeneous sources. Assuming that there is a true data-generating model for a given phenomenon, we construct a statistical model that aims to include the true data-generating model as a special case. Inference on the parameters in the postulated model is made by performing simulations that replicate the data-generating process behind observed data. If the results from the simulations are sufficiently close to the observed data, the model is deemed plausible. The plausible model can then be used to calculate quantities of interest, such as risk measures.

The new method allows us to make inferences based on data from widely different sources. Examples of these could be summary statistics or raw data; results from research studies with heterogeneous experimental designs and study populations; general scientific facts or results of lab experiments.

In this thesis, we limit the applications to data where traditionally one would use meta-analysis. This is a set of methods that summarize the results of several studies by calculating the weighted average of reported intervention effects. We argue that meta-analysis faces disadvantages that the new method can avoid.

The performance of the new method is explored in two extensive simulation studies, and we also apply the method to data that was previously used in a traditional meta-analysis.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: par.villner@ki.se. Supervisor: Matteo Bottai.

Acknowledgments

I am deeply grateful to my supervisor Matteo Bottai, both for supervising me and for conceiving the method that is the topic of this thesis. During the last 20 months, I have received more advice and encouragement from Matteo than I could have hoped for. I am also grateful to the Biostatistics Core Facility at Karolinska Institutet for providing me with an office space and letting me participate in their meetings and workshops.

I would like to extend thanks to David Miller and his colleagues at the Environmental Protection Agency in Washington D.C. who have been collaborators in the development of the new method. Their feedback has been extremely valuable, as it has given me an insight into how the method can be applied in realistic scenarios.

Lastly, I would like to thank my older brother Tomas Villner, without whom I doubt I would have studied mathematics or taken an interest in biostatistics. For as long as I can remember, Tomas has been my best friend and my greatest intellectual mentor. Tomas received his education as a physician at Karolinska Institutet in Solna. Although he only took a single introductory course in statistics, Tomas was fascinated by probability theory and took every chance to discuss it. I remember a conversation in which Tomas made the distinction between the Bayesian and frequentist definition of probability, although he had never heard of these concepts. In September last year, Tomas passed away after several years of illness. I dedicate this thesis to Tomas and also his children, Ines and Lo.

Contents

1	Introduction	6
1.1	The aim of the thesis	6
1.2	An outline of the thesis	7
2	An overview of meta-analysis	8
2.1	The omnibus test	8
2.1.1	Limitations of the omnibus test	9
2.2	Standard methods in meta-analysis	9
2.3	Calculating a weighted average	10
2.3.1	The Fixed Effect model	10
2.3.2	The Random Effects model	11
2.4	Meta-regression, network analysis and Bayes	13
2.5	Limitations of Fixed Effect and Random Effects models	13
3	An intuitive explanation of the HS method	15
3.1	The Mother Nature’s Model and the Postulated Model	15
3.2	The inferential procedure	16
3.3	Calculating relevant quantities	16
3.4	An example of inference with the HS method	17
3.5	“Meta-analysis” with the HS method	18
3.6	The limitations of standard meta-analysis revisited	19
3.7	Topics for future research	19
4	Inference with the HS method	21
4.1	Sampling-based inference	21
4.2	An algorithm for finding a plausible θ	23
4.3	Density approximation techniques	23
4.3.1	Kernel density estimators	24
4.3.2	Copula estimators	26
4.4	Correcting for multiple studies	29
4.5	Likelihood-based inference	31
4.6	Similar methods	32
4.6.1	Diggle & Gratton’s method	32
4.6.2	Approximate Bayesian Computation	33
4.6.3	Sensitivity analysis and imputation	34
5	Computational aspects of the inference	35
5.1	Implementation of the kernel density estimator	35
5.2	Implementation of the copula estimator	35
5.3	Correction for multiple studies	35

6	Simulation Study I	37
6.1	The data-generating models	37
6.1.1	Mother Nature's Model 1	37
6.1.2	Mother Nature's Model 2	38
6.2	The simulation algorithm	38
6.3	Computational details of the inferential procedure	39
6.4	Simulations results	40
6.5	Difficulties in the simulations	42
6.5.1	Postulated Model 1	42
6.5.2	Postulated Model 2	43
7	Simulation Study II	45
7.1	The Mother Nature's Model 3	45
7.2	The simulation algorithm	46
7.3	Computational details of the inferential procedure	47
7.4	Simulations results	47
8	An analysis on the association between Paraquat exposure and Parkinson's Disease	48
8.1	The meta-analysis by Ntzani et. al.	48
8.2	Plan for performing an analysis	49
8.3	Step 1. Construct a Postulated Model	50
8.4	Step 2. Recreate the summary statistics and find a plausible θ	51
8.4.1	Computational details of the inferential procedure	52
8.4.2	The effect of adding new statistics and studies	53
8.5	Step 3: Calculating a measure of association	53
8.6	Adding a study of PD prevalence	55
8.7	Comparison with the Ntzani et. al. study	56
8.8	Some suggested improvements	57
9	Conclusions	59
10	Graphs	61
10.1	Postulated Model 1, $J=1$, $n=10$	62
10.2	Postulated Model 1, $J=1$, $n=100$	63
10.3	Postulated Model 1, $J=1$, $n=1000$	64
10.4	Postulated Model 1, $J=5$, $n=10$	65
10.5	Postulated Model 1, $J=5$, $n=100$	66
10.6	Postulated Model 1, $J=5$, $n=1000$	67
10.7	Postulated Model 1, $J=20$, $n=10$	68
10.8	Postulated Model 1, $J=20$, $n=100$	69
10.9	Postulated Model 1, $J=20$, $n=1000$	70
10.10	Postulated Model 1, $J=100$, $n=10$	71
10.11	Postulated Model 1, $J=100$, $n=100$	72

10.12	Postulated Model 1, $J=100$, $n=1000$	73
10.13	Postulated Model 1, $J=5$, $n=10$, $c_4 = c_6 = 3$	74
10.14	Postulated Model 1, $J=5$, $n=100$, $c_4 = c_6 = 3$	75
10.15	Postulated Model 1, $J=5$, $n=1000$, $c_4 = c_6 = 3$	76
10.16	Postulated Model 2, $J=1$, $n=10$	77
10.17	Postulated Model 2, $J=1$, $n=100$	78
10.18	Postulated Model 2, $J=1$, $n=1000$	79
10.19	Postulated Model 2, $J=5$, $n=10$	80
10.20	Postulated Model 2, $J=5$, $n=100$	81
10.21	Postulated Model 2, $J=5$, $n=1000$	82
10.22	Postulated Model 2, $J=20$, $n=10$	83
10.23	Postulated Model 2, $J=20$, $n=100$	84
10.24	Postulated Model 2, $J=20$, $n=1000$	85
10.25	Postulated Model 2, $J=100$, $n=10$	86
10.26	Postulated Model 2, $J=100$, $n=100$	87
10.27	Postulated Model 2, $J=100$, $n=1000$	88
10.28	Postulated Model 3	89
10.29	Postulated Model 3	90

11 Sources

91

1 Introduction

Statistical inference is based on data collected in scientific studies. For instance, we may assess the effect of a medical treatment based on a randomized trial with participants sampled from a given population. If it turns out that a similar experiment was performed on another group of participants from the same population, one can pool the groups of participants together, and perform an analysis on the larger dataset.

There may be other studies that investigate the same research question as the first study, but that do so in slightly different ways, e.g. because the participants were sampled from a different population or because the treatment effect was assessed using a different measure. If we want to use traditional statistical methods in such cases, we have to choose between two alternatives:

1. Only use a subset of results that came from similar experimental situations. This can lead to low statistical power.
2. Use a larger subset of results despite the differences. This could result in biased, or even nonsensical, results.

If several studies investigate the same research question, they should all contain valuable information. There must be a way of using all the available information.

In this thesis, we describe a new statistical method with the potential of combining data from heterogeneous data sources. The method is built upon a simple and intuitive idea: that for any phenomenon, there is a statistical model which generates the data. We call this the Mother Nature's Model. Gaining full knowledge of the Mother Nature's Model may be impossible, but based on scientific expertise, we may design a statistical model which is a good approximation of the Mother Nature's model. We call this our Postulated Model. To investigate if our Postulated Model is plausible, we perform simulations where the experimental procedure behind the observed data is recreated. If the Postulated Model can be used to generate data sufficiently close to the observed data, the Postulated Model is deemed plausible. We can then use the plausible Postulated Model to calculate quantities of interests, such as risks. Because the interest is in the model that generated the data, a very broad range of data sources can be used.

We call the new method the Heterogeneous Sources method (the HS method).

The HS method can be used in a variety of contexts. In this thesis, however, we limit the applications to meta-analysis. This is a group of statistical methods that aim at summarizing the results of many research studies, and the HS method has the potential to enrich this field.

1.1 The aim of the thesis

The HS method is still under development. As explained in sections 3 and 4, a solid theoretical foundation is yet lacking. With this thesis, we aim to describe the basic ideas underlying the HS method; provide details for applying the method; and report on the results of two simulation studies, as well as the result of an analysis of real data. We also give possible suggestions for further developments of the HS method.

1.2 An outline of the thesis

The remainder of the thesis is structured as follows:

- In section 2, we introduce meta-analysis. We look at both the history of the field and the methods that are currently being used, as well as point out limitations of these methods.
- In section 3, we describe the HS method and explain how it can be applied to meta-analysis.
- In section 4, we discuss inference with the HS method, which we later use in the simulation studies.
- In section 5, we outline some computational details regarding the application of the HS method.
- In sections 6 and 7, we present the results of two simulation studies that show potentials and limitations of the HS method.
- In section 8, we show how an analysis of real data can be performed with the HS method.
- In section 9, we give a short summary and make some suggestions for future research.

2 An overview of meta-analysis

Meta-analysis is a set of methods that aim to use existing research to improve statistical power. While such methods have existed for a long time, meta-analysis has increased dramatically in popularity during the last decades.

In this section, we first describe early methods of meta-analysis and then we describe current methods in closer detail. We end the section by considering some limitations that the current methods of meta-analysis suffer from.

2.1 The omnibus test

This section is based on Hedges (1992).

Scientists have always taken the results of previous studies into account when conducting their research. To this day, it often takes the form of a “scientific review”, where the researcher discusses results from studies that they find relevant. Such an analysis does not, in general, result in a single conclusion, and what studies are deemed relevant is up to the researcher.

Systematic summaries of previous research began appearing in 18th century astronomy. Adrien-Marie Legendre along with other scientists, combined data from a series of observations to make precise estimates of general patterns in planetary motion. This was the foundation of the least-squares method for solving a system of linear equations.

At the start of the 20th century, meta-analysis, as we know it today, started to develop. Methods were developed to summarize the hypothesis tests from several, independent research studies. For n studies, each estimating an effect parameter $\lambda_i, i = 1, \dots, n$, there are n separate null hypotheses to be tested:

$$H_{0,i} : \lambda_i = 0, i = 1, \dots, n.$$

where large values of the test statistic leads to a rejection of the null hypothesis.

The so-called omnibus test was developed to investigate if the true effect is 0 in all studies. That is, the null hypothesis of the omnibus test is

$$H_0 : \lambda_1 = \dots = \lambda_n = 0.$$

Several ways of testing this hypothesis were suggested. All of them use that the one-tailed p-value of the i th study is

$$p_i = P(T_i > t_i | \lambda_i = 0)$$

where t_i is the obtained test statistic of the i th study. All T_i are assumed to be continuous and independent of each other. Under $H_{0,i}$,

$$p_i \sim U(0, 1), \tag{1}$$

by the probability integral transform. This property is used in the methods described below.

By the minimum-p method, we reject H_0 when

$$\min(p_i) < 1 - (1 - \alpha)^{1/k}$$

for a desired significance level α .

The inverse χ^2 method exploits that under H_0 and independence, we have that $-2 \sum_i \log(p_i) \sim \chi_{df=2n}^2$. This leads to a test where H_0 is rejected when

$$-2 \sum_i \log(p_i) > C$$

for a C corresponding to a suitable quantile of the χ^2 distribution with $2n$ degrees of freedom.

A third method is the inverse normal method. Set $\Phi(z_i) = p_i$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Define Z such that

$$Z = \frac{z_1 + \dots + z_n}{\sqrt{n}} = \frac{\Phi^{-1}(p_1) + \dots + \Phi^{-1}(p_n)}{\sqrt{n}}.$$

It can be shown that $Z \sim N(0, 1)$, so a hypothesis test with Z as a pivot statistic can be performed.

2.1.1 Limitations of the omnibus test

In many fields, including biostatistics, we are interested in the size and direction of the intervention effect λ_i . The omnibus test is in general not well-suited for drawing conclusions on these matters. Rejection of H_0 only tells us that one $H_{0,i}$ is false. This is a rather weak claim, especially if there are many λ_i 's.

2.2 Standard methods in meta-analysis

Meta-analysis as we know it today is centered around the idea of estimating the size of an intervention effect, rather than simply testing the hypothesis that the effect is different from zero. The phrase “meta-analysis” was coined by Gene Glass (1977). In the following decades, meta-analysis became increasingly popular, particularly in the social sciences. Lately, applications in natural science and biostatistics have become more common. A meta-analysis is usually performed using the following steps:

1. Decide on the intervention effect of interest, e.g. the effect of a new diet on BMI; the change in survival for patients who undergo a cancer treatment; or the possible increase in the risk of Parkinson's disease for farmers exposed to a pesticide.
2. Decide a statistic with which to measure the intervention effect, e.g. odds ratio, risk ratio or hazard rate.
3. Set methodological standards that all research studies should live up to. The methodological standards can specify major issues, such as what experimental

design can be used, or they can specify minor details, such as the number of participants included in the studies or the age ranges of the participants.

4. Collect research articles. This is done by searching databases containing scientific journals, and potentially also by collecting unpublished papers.
5. After removing the articles that do not live up to the methodological standards that were set, a weighted average of the reported intervention effects is calculated. The weighted average of the intervention effect usually serves as the main result of a meta-analysis.
6. Additionally, an analysis of the variation of reported intervention effects can be performed. For instance, one may investigate how features of the participants or the study design affect the intervention effect.

2.3 Calculating a weighted average

The weighted average of the intervention effect is the main result of almost all meta-analyses. There are two dominant frameworks for generating the weights: Fixed Effect and Random Effects models. These two frameworks make different assumptions about the collected research studies and how they are related to each other.

2.3.1 The Fixed Effect model

This section is based on Borenstein, et. al. (2007, p 6-12).

Within the Fixed Effect framework, we assume that all studies are estimating the same underlying intervention effect. If different studies report different results, it is merely due to random sampling error. The reported intervention effect of the i th study can be represented as

$$t_i = \mu^f + \epsilon_i$$

where μ^f is the true effect common to all studies and ϵ_i is the sampling error of the i th study, $\text{var}(\epsilon_i) = v_i^2$.

When using a Fixed Effect model, we are implicitly assuming that all studies used the same methodology and sampled participants from the same population. These assumptions lead to a natural process for calculating the weighted average:

$$\bar{t}_w = \frac{\sum_i^n t_i w_i}{\sum_i^n w_i}, \quad (2)$$

where $w_i = \frac{1}{\hat{v}_i^2}$ and \hat{v}_i^2 is the observed variance of the i th study. The size of \hat{v}_i^2 largely depends on the sample size of the i th study. Therefore, using the Fixed Effect model to calculate an average intervention effect is very similar to pooling together results from different studies and taking the average.

A natural estimate of the true intervention effect μ^f is the weighted average of the intervention effect:

$$\hat{\mu}^f = \bar{t}_w.$$

2.3.2 The Random Effects model

With a Random Effects model, it is assumed that different studies are estimating different true intervention effects. This is because the studies differ in ways that are relevant for the intervention effect, e.g. that they use different experimental designs and sample participants from different populations.

Despite these differences, it makes sense to calculate a weighted average of intervention effects. This is because we are assuming that there is a distribution of intervention effects. We consider the true intervention effects of the studies included in a meta-analysis to be a random sample from this distribution of intervention effects. The weighted average of intervention effects is an estimate of the average in this distribution.

The reported intervention effect of the i th study is defined as

$$t_i = \theta_i + e_i$$

where e_i is the sampling error with variance $\text{var}(e_i) = v_i^2$.

θ_i is the true intervention effect of the i th study and $\theta_i = \mu^r + \delta_i$. Here, μ^r is the mean of the distribution of intervention effects, and δ_i represents how the true intervention effect of the i th study deviates from μ^r . $\text{Var}(\delta) = \Delta^2$ is a measure of the variation of true intervention effects in the distribution.

Thus, the Random Effects model assumes that two types of sampling are taking place. First, we are taking a random sample from the distribution of true intervention effects θ . The expected true intervention effect is μ^r and the variance is Δ^2 . Then, traditional sampling error occurs due to the selection process of study participants.

There exist several methods for estimating the parameters in the Random Effects model. As Bodnar (2016) notes, the differences are small with a large sample size, and are not relevant for this thesis. Therefore, we will only describe the estimation method proposed by DerSimonian & Laird (1986) in an influential paper. Their estimator of μ^r is

$$\hat{\mu}^r = \frac{\sum_i w_i^r t_i}{\sum_i w_i^r}.$$

Here $w_i^r = (\hat{v}_i^2 + \hat{\Delta}^2)^{-1}$ is the inverse of the sample variance estimate and an estimate of the between-study variance Δ^2 . The latter can be found by considering

$$Q = \sum_i^n w_i (t_i - \bar{t}_w)^2, \quad (3)$$

for \bar{t}_w and w_i as in (2). Q is a measure of the variation among the reported intervention effects. Under the assumption that there is no between-study variance, meaning that $\Delta^2 = 0$, Q is asymptotically χ^2 distributed with $n - 1$ degrees of freedom. This implies that Q has the degrees of freedom $n - 1$ as its expected value.

We can use this to estimate Δ^2 by

$$\hat{\Delta}^2 = \frac{\max(0, (Q - (n - 1)))}{C}$$

for

$$C = \sum_i^n w_i - \frac{\sum_i^n w_i^2}{\sum_i^n w_i}.$$

This estimator is derived from

$$E(Q) = \Delta^2 \left(\sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i} \right) + n - 1.$$

$\hat{\Delta}^2$ can now be inserted into w_i^r and used with the estimator $\hat{\mu}^r$.

Measures of heterogeneity

This section is based on Higgins & Thompson (2002a).

Meta-analyses often provide a measure of the heterogeneity of the reported study results. In particular, it is common to assess whether all studies included in the meta-analysis can be assumed to share the same true intervention effect. These assessments often guide whether a Fixed Effect or Random Effects model is used.

We have seen that if all studies share the same intervention effect, then $Q \sim \chi_{df=n-1}^2$. If $Q > C$ for a C corresponding to a suitable quantile in the χ_{df}^2 distribution, we reject the hypothesis that all studies have the same intervention effect, because $\Delta^2 > 0$.

The Q statistic is only asymptotically χ^2 distributed, however. With few studies, the Q test is underpowered. There are alternatives, with greater statistical power. One popular such measure is the I^2 statistic, which gives a measure of the proportion of between-study variation that is due to differences in the true intervention effect rather than sample variance:

$$I^2 = \frac{\hat{\Delta}^2}{\hat{\Delta}^2 + \hat{v}^2},$$

with $\hat{v}^2 = E(\hat{v}_i^2)$.

Using the Q or I^2 statistics to decide between a Fixed Effect and Random Effects model may seem reasonable since both of these tests are easy to perform. However, the Random Effects and the Fixed Effect models use different assumptions regarding how the summary statistics were generated. It is not obvious that the Q or I^2 statistics are good indicators of which model we should choose. It seems wiser to look at the actual methodology behind the studies, as this is what should guide the decision of what model framework we use.

2.4 Meta-regression, network analysis and Bayes

Calculating the average intervention effect with a Fixed Effect or Random Effects model is likely the most common meta-analytical method, but there are other tools as well.

Meta-regression is a method to further explore potential sources of heterogeneity of study results. As the name suggests, meta-regression is a type of regression analysis used to estimate the effect that a particular study feature has on the expected intervention effect. For instance, if participants in different studies were of different ages, we may estimate the effect age has on the intervention effect. Unfortunately, meta-regression is seldom applicable because many studies are required. (Thompson 2002)

Network analysis is a way of indirectly comparing the effects of several related interventions, in situations where they have not been compared in a single study. The idea is that if interventions A and B were compared in study 1, and intervention B and C were compared in study 2, then assuming transitivity, we can compare A and C via B. (Hu, et.al. 2020)

Lastly, although it has historically been common to perform meta-analysis within a frequentist framework, it is increasingly popular to use Bayesian methods. As pointed out in Bodnar (2016), there are several benefits to applying a Bayesian perspective. For example, uncertainties regarding the between-study variance Δ^2 can be modeled, and subject-matter knowledge (for instance grounded in previous meta-analyses) can be used in formulating the prior distributions of the μ and Δ^2 parameters.

2.5 Limitations of Fixed Effect and Random Effects models

Several criticisms have been directed against meta-analysis. The most common critique is that by combining results of studies that are different from each other, a comparison is not meaningful. A related critique is that we risk averaging over poorly designed studies, so that the conclusion is similarly poor. Meta-analysis is also prone to publication bias, since it is predominantly published studies that are included in an analysis. (Borenstein, et. al. 2009, chapter 43)

In this thesis, we focus our attention on two difficulties that are related to each other:

1. If we set strict methodological standards on the studies included in our meta-analysis, then the studies share similarities in terms of experimental design and study populations, meaning that the weighted average that we calculate is easy to interpret. However, the stricter our standards are, the fewer studies can be included in the analysis.
2. If we set loose methodological standards, we can include more studies in our analysis, but it is less clear what the weighted average is an average of.

If we try to avoid the first problem, we get more of the other problem and vice versa.

The Fixed Effect model avoids the second problem, by assuming that all studies investigate the same question in the same way. However, in biostatistics, it is very seldom the case that several studies are sufficiently similar. Therefore, the Fixed Effect model is usually not applicable in biostatistics.

The Random Effects model should be able to avoid the first problem, since it allows for differences between studies. The consequence of having loose methodological standards, however, is that the impact of differences between studies is included in the random effects parameter. The interpretation of this parameter is often unclear. Some questions that might come up include: what does it mean that there is a population of intervention effects? Does it consist of the true intervention effects of all possible ways of conducting a study? Is there any reason to believe that our sample from this population is random?

A meta-analysis could suffer from both problems at once. According to Davey et. al. (2011), the number of studies included in a Random Effects meta-analysis in the Cochrane meta-analysis database is fewer than five in 75% of the cases. As Bodnar (2016) points out, this also means that the estimator of the between-study variance Δ^2 becomes unstable.

An example of the selection procedure of studies in a Random Effects meta-analysis can be seen in the flow-chart by Meyer-Baron, et. al. (2014) which has been reproduced in Figure 1. We see that out of 1165 potentially relevant articles, only 22 are used. Many studies that investigated the same research question were excluded because they used a slightly different study design from what the authors of the meta-analysis had set out for.

If many studies are investigating related research questions there is valuable information in all of them. It is important to find a way of extracting this information. Calculating a weighted average may just not be the best approach.

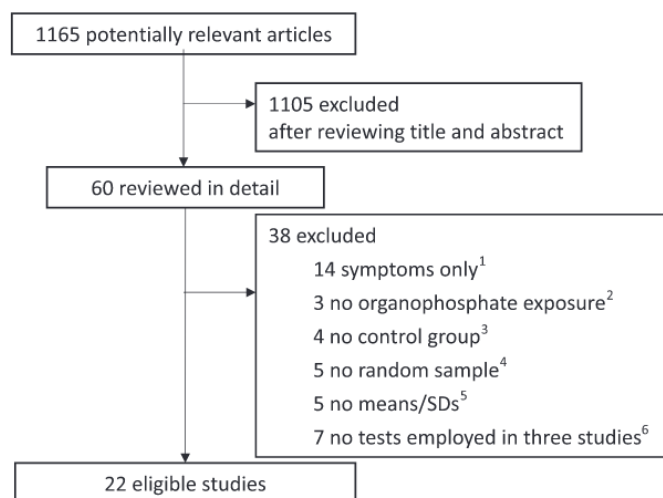


Figure 1: A flowchart from Meyer-Baron (2014) showing the process of including studies in a Random Effects meta-analysis.

3 An intuitive explanation of the HS method

The topic of this thesis is a new statistical method conceived by Matteo Bottai at Karolinska Institutet. We refer to it as the HS method, since it is a method that lets us synthesize data from heterogeneous sources. While the method could potentially be used in a broader context, we mainly discuss it here from the perspective of meta-analysis in biostatistics.

In this section, we describe the central idea behind the the method and briefly explain how it can be used to analyze data. We also argue that the HS method has the potential to avoid the limitations of standard meta-analysis.

No proofs are given and mathematical notation is limited to the strictly necessary. The aim is to convey the intuition underlying the HS method.

3.1 The Mother Nature's Model and the Postulated Model

For a specific set of variables of interest, we consider the true data-generating model, marginalized over the non-relevant variables. We call this the Mother Nature's Model.

Definition 3.1 (The Mother Nature's Model). Let $Y \in \mathbb{R}^q$ be a vector with variables of scientific interest and $X^M \in \mathbb{R}^p$ be a vector of explanatory variables. Set $W^M = (X^M, Y)$ and parameter vector $\theta^M \in \Theta_M \subseteq \mathbb{R}^d$ where Θ_M is a finite subset of values in a well-defined d -dimensional real hyperspace. The Mother Nature's Model of W^M is a parametric joint probability function $f_M(w^M|\theta^M) : \mathbb{R}^{p+q} \otimes \mathbb{R}^d \mapsto \mathbb{R}_+$.

Definition 3.2 (Non-relevant variables). For $Z \in \mathbb{R}^v$ such that $Z \notin W^M$, we say that Z is non-relevant with regard to W^M if $f_D(w^M) \approx f_D(w^M|z)$, where f_D is a true data-generating model.

Remark 3.1. We can view the Mother Nature's Model as a theoretical universal data-generating model f_D marginalized over the non-relevant variables Z , $f_M(w^M|\theta^M) = \int_Z f_D(w^M, z)dz$. Albeit an interesting philosophical question, we do not go deeper into what the true data-generating model f_D is. Whether we consider it an entity existing in the Platonic sense, or merely a useful fiction, should matter little for this thesis.

The Mother Nature's Model of any phenomenon is unknown to us, but based on scientific expertise and prior research, we can construct a model that we postulate for Mother Nature's.

Definition 3.3 (The Postulated Model). Define Y as in Definition 3.1 and let $X \in \mathbb{R}^k, k \geq p$ be a vector of explanatory variables. Set $W = (Y, X)$ and let $\theta \in \Theta \subseteq \mathbb{R}^e$ where Θ is a finite subset of values in a well-defined e -dimensional real hyperspace. The Postulated Model $f_P(w|\theta) : \mathbb{R}^{q+k} \otimes \mathbb{R}^e \mapsto \mathbb{R}_+$ is a parametric joint probability function that aims at containing the Mother Nature's Model as a special case, in the sense that the parametric family of f_P contains the parametric family of f_M .

Remark 3.2. We assume that there is a value of $\theta \in \Theta$ such that the Postulated Model is approximately equal to the Mother Nature's Model over the relevant variables, that is

$$f_M(w|\theta^M) \approx f_P(w|\theta).$$

for some $\theta \in \Theta$.

3.2 The inferential procedure

The role of scientific experts is vital in applying the HS method. Without scientific expertise, it is extremely difficult to identify a Postulated Model belonging to a parametric family that includes the parametric family of the Mother Nature's Model, as required by Definition 3.3. In the remainder of the thesis, we do not discuss the role of scientific experts in detail.

However, scientific expertise alone is likely not sufficient to find a f_P and θ such that, by Remark 3.2, $f_P \approx f_M$. For that, we need to make inference based on data. If the model that we postulate for Mother Nature is a good approximation, it should be able to produce data that is similar to the real data. This assumption leads us to the following inferential procedure:

We begin by gathering data related to the Postulated Model. We then use our Postulated Model to perform simulations where we recreate the experimental procedure used to generate the data. Any kind of data is useful, as long as it is related to the Postulated Model and it is possible to replicate the experimental procedure. It can be data from cohort studies, randomized trials, case-control studies or studies with any other experimental design. The data itself may come in different forms. Raw data with variable values for each study participant can be used, as well as summary statistics, such as mean values and odds ratios. Participants in the studies may come from different populations with different characteristics. The data can also come in the form of facts about a whole population, such as the true proportion of males and females in Germany, or the number of deaths in Finland in a given year. The only limit to the type of data that can be used is that it must relate to the variables in the Postulated Model and we must know enough about how the data was generated so that we can replicate the procedure.

The reason why the HS method can handle such heterogeneity is that all data was generated by the same Mother Nature's Model. If the Postulated Model is plausible, it should be able to generate similar data. Differences in terms of experimental design, study population and reported statistics are accounted for in the simulation process, where we sample participants from the same populations, perform the same experiments and calculate the same statistics as those in the experiments.

If the data generated with the Postulated Model is not close to the observed data, then the Postulated Model is deemed implausible. We make changes and perform new simulations. Once the observed data and the simulated data are sufficiently close, we say that the Postulated Model is plausible.

3.3 Calculating relevant quantities

A plausible Postulated Model can be interesting in itself, but in most cases, we imagine that the main interest lies in calculating a conditional probability, $f(y|x^*)$, where $X^* \subseteq$

X , or a function thereof, such as an odds ratio, a risk difference or a hazard rate. For example, we may be interested in how the risk of contracting cancer is different between people exposed to and those not exposed to a pesticide, and how this difference depends on the age or sex of a person.

A plausible Postulated Model can be used to approximate the quantity of interest, using $f_P(y|x^*)$ with $X^* \subseteq X$.

This makes a plausible Postulated Model an extremely powerful tool. We can use it to calculate any quantity of interest. This is in stark contrast to statistical methods in general and standard meta-analysis in particular, where the result of an analysis is the quantity of interest expressed with a particular measure.

3.4 An example of inference with the HS method

As an example of how the HS method could be used, assume that we have the following Postulated Model:

MODEL 1

$$Z \sim Be(g[\theta_1])$$

$$X \sim Be(g[\theta_2 + \theta_3 Z])$$

$$Y \sim Be(g[\theta_4 + \theta_5 X + \theta_6 Z + \theta_7 XZ])$$

with $g(a) = (1 + \exp(-a))^{-1}$, $\{X, Y, Z\} \in \{0, 1\}$ and $\theta_i \in \mathbb{R}, i \in \{1, \dots, 7\}$.

Further assume that we are interested in the association between X and Y , and that we have gathered the following data from three different studies:

- Observations of the X and Y variables for m individuals: $(x_i, y_i), i = 1, \dots, m$
- The proportion out of o participants for which $X = 1$.
- The coefficients of a logistic regression model of Y given Z , based on k participants.

All this data can be used to make inference, since we can perform simulations with the Postulated Model.

For the first data set, we could generate X and Y values for m individuals and compare the simulated values and the real values with a suitable distance function. For the second data set, we could simulate o individuals and calculate the proportion that has $X = 1$, and compare this proportion with the proportion reported in the research study. For the third data set, we could generate k individuals and perform logistic regression of Y given Z . The coefficient estimates from our simulation can then be compared with the reported coefficient estimates.

Plausible values on the θ coefficients in our Postulated Model are found by searching for θ s that can reproduce results similar to the results from the research studies. Once we have a plausible Postulated Model, we can use it to calculate any probability or

association measure related to the phenomenon we are interested in. For instance, we can use Model 1 with plausible θ values to express the association between X and Y in terms of an odds ratio, risk rate, risk difference or any other measure we can think of, using the marginal probability distribution.

3.5 “Meta-analysis” with the HS method

As already stated, a Postulated Model can be used to recreate any type of data and functions of data, including summary statistics. This is useful as meta-analyses are typically based on summary statistics. The equivalent of a meta-analysis with the HS method can be performed in the following way:

1. Based on scientific knowledge, design a Postulated Model of the phenomenon that we are interested in. This means that even if we are interested in the association of e.g. a exposure to a pesticide and Parkinson’s disease, we design a Postulated Model of Parkinson’s disease, with more variables than just pesticide exposure, to the extent we believe them to be important.
2. Collect data related to the Postulated Model. The data can be related to any variables in the Postulated Model, not only the variables whose association we are ultimately interested in. The experiments that generated the data may follow different designs, and participants may be sampled from different populations.
3. Perform simulations that recreate the collected data. This includes sampling participants from the same population as in the studies; simulating exposure to the same treatment; and calculating the same summary statistics. When performing these simulations, we may also model the biases we suspect to be present in the available studies, e.g. non-response bias, self-report bias or publication bias.
4. If the results from the simulations are not sufficiently close to the results reported in the studies, the Postulated Model is implausible. In that situation, we must make changes to our model and try again.
5. Once we can generate results that are sufficiently close to the reported results, we say that we have a plausible Postulated Model that we can use to calculate the association measures of interest.

This procedure is quite different from a standard meta-analysis. Not only can we include a much wider range of data in our analysis – it is required if we are to make precise inference for the parameters in our Postulated Model. Fortunately, much useful information that we often bypass can be used from research papers. For instance, if a study reports on the association between pesticide exposure and cancer, a standard meta-analysis only looks at the measure of that particular association. With the HS method, we can use the distribution of age, sex, education level and many other variables. Journal papers are usually full of these facts, and using the HS method we can exploit them.

3.6 The limitations of standard meta-analysis revisited

In section 2, we pointed out two limitations of traditional meta-analysis, namely that we have to choose between imposing strict methodological standards – meaning that the weighted average still makes sense but there are few studies to include. Or imposing looser methodological standards – meaning that we can include more studies but the weighted average is less meaningful.

As we have seen, the HS method has the potential to handle both of these problems. We can include a wide range of studies in our analysis in a way that makes sense. Differences in experimental design and sampling populations are considered in the simulations. With regards to the flowchart in Figure 1, using the HS method, we would be able to include all of the 60 studies that were reviewed in detail, rather than having to ignore 38 of them.

The HS method also has the potential to handle the other limitations that were briefly mentioned, namely publication bias and problems caused by poor studies. The publication bias can be modeled in the simulation procedure, as can other biases such as non-response bias. Poor study design is not a problem. As long as we can follow the steps that the original researchers took, the results of the studies are useful with the HS method. This point is elaborated on in Simulation Study II, where we show that a flawed study can be included in an analysis.

3.7 Topics for future research

Making inference on the Postulated Model is different from traditional statistical inference. Below we list some of the questions that have to be further explored.

The size of the Postulated Model

According to Definition 3.3, the Postulated Model can be larger than the Mother Nature’s Model, as long as the Mother Nature’s Model is included in the Postulated Model. How large can the Postulated Model be? When employing traditional statistical methods, we prefer parsimonious models because they (a) are easy to interpret and because (b) unnecessary parameters may absorb noise from the sampling procedure. Regarding (a), if our interest is in the Postulated Model in itself, then the interpretability issue is relevant for us as well. But as we pointed out in Section 3.3, we are often interested in a conditional probability, and a comprehensive Postulated Model does not affect our appreciation of that probability. Regarding (b), we see in section 4 that the sampling procedure is taken into account when estimating the parameter values.

Still, large models are difficult to work with, particularly if we have to make inference for many parameters, so there is no reason to use extravagant Postulated Models. Striking the balance between having a Postulated Model that includes the Mother Nature’s Model as a special case, and having a Postulated Model that is easy to work with is certainly not a trivial concern.

Sampling and asymptotics

Similar to the Random Effects model, it is not obvious how data included in an analysis with the HS method should be considered to have been sampled. Is it a sample from all research studies that could have been performed? Or is it rather not a sample in the traditional sense? Should we treat it as facts about the world?

Similarly, if we want to consider the asymptotic results of our analysis, then does the asymptotic relate to the number of studies or the number of participants included in studies?

Finding the answers to these questions may be difficult, and it may have a big impact on the theoretical foundation of the HS method. For instance, it may determine if we place the method in a frequentist or Bayesian framework.

Sufficient data

A Postulated Model is plausible when it is consistent with observed data. What data are we referring to? A natural starting point is to include data related to the variables that are most pertinent to our research question. This gives information about the parameters related to the observed data – but as we see in Simulation Study I in section 6, it also gives information about parameters related to variables in the model that are not in the observed data.

However, including observed data on all the variables in the Postulated Model should yield more precise estimates of all parameters. Therefore, it is not a good strategy to only include studies in an analysis that are similar to each other. Rather, we want to use studies that are investigating different aspects of our Postulated Model: studies that report on different variables for different subpopulations. This point is elaborated on in sections 6, 7 and 8.

At the same time, recreating data related to many variables is time-consuming and for high-dimensional data, the curse of dimensionality may make it difficult to compare simulated and observed data.

For these reasons, we have to find a way to determine what data is sufficient in order to make decent inference on our Postulated Model, given that we are interested in calculating a particular conditional probability, as described in section 3.3.

Generalizing over populations

Different studies may have sampled participants from different strata of the total population. Ideally, we would have data related to all important variables for all relevant strata of the population, but this is unlikely to be the case. It is more likely that some studies report on some variables for one population stratum, and other studies are report on different variables for other strata. It is not obvious how to combine these results into inference on a single Postulated Model.

4 Inference with the HS method

Inference on the Postulated Model is performed by simulating data with a Postulated Model and determining whether it is sufficiently close to the observed data. We can imagine many different measures of the distance between the observed data and simulated data, e.g. Euclidean distance. In this section and the remainder of this thesis, we focus on a particular way of measuring the distance between observed and simulated data, based on the sampling distribution of the simulated data. It is a natural starting point in a meta-analytical context, and it is the inferential method that we use in the simulation studies in sections 6, 7 and 8. After a description and justification of the algorithm that is used, we outline two methods for density approximation that the sampling-based method relies on.

We then briefly discuss another approach to inference, namely likelihood-based inference. This method is still being developed and several theoretical considerations have to be dealt with before an application of this method can be fully understood. We also briefly discuss some other methods that show similarities to the HS method.

As already pointed out, the HS method is described in the context of meta-analysis in this thesis. Among other things, this has the consequence that we only consider inference based on summary statistics in this section, although any kind of observed data could be used with the HS method.

4.1 Sampling-based inference

Our primary interest is in the Postulated Model $f_P(w|\theta)$, as described in Definition 3.2, with $\theta \in \Theta \subseteq \mathbb{R}^e$. In order to make inference on θ , we have access to summary statistics from research studies. For the j th study, $w_j \in \mathbb{R}^{n_j \times (q+k)}$ is a matrix of data on all variables in the Postulated Model for all participants in the study, with n_j being the number of participants of the j th study and $q+k$ the number of variables in the Postulated Model. Note that w_j is the true data matrix and the researchers of the j th study most likely only recorded information on some of the variables. The summary statistic of the j th study is $s_j(w_j) : \mathbb{R}^{n_j \times (q+k)} \mapsto \mathbb{R}^{d_j}$, where d_j can differ between studies.

Given θ , the sampling distribution of $s_j(\cdot)$ is

$$s_j(W) \sim f_j(s_j(w)|f_P(w|\theta)),$$

with the unknown function $f_j : \mathbb{R}^{d_j} \otimes \mathbb{R}^e \mapsto \mathbb{R}_+$. The characteristics of f_j depend not only on θ , but also on what summary statistics s_j are used, how participants were sampled and other details of the experimental design. This means that the f_j 's may differ in many ways; what they have in common is $f_P(w|\theta)$.

f_j can be approximated by using simulated data from the Postulated Model with a density approximation method, such as a kernel density estimator or a copula density estimator. The approximation is denoted

$$\hat{f}_j(s_j(w)|f_P(w|\theta)). \tag{4}$$

We want to make an inference on θ based on how unlikely the reported summary statistics are according to the \hat{f}_j s. To this end, we introduce some concepts.

Definition 4.1 (Significance $\pi_j(\theta)$). Significance with regard to a parameter vector θ and summary statistic $s_j(w_j)$ is

$$\pi_j(\theta) = Pr[\hat{f}_j(s_j(w)|f_P(w|\theta)) < \hat{f}_j(s_j(w_j)|f_P(w_j|\theta))] = \int_{\mathcal{W}} \hat{f}_j(s_j(w)|f_P(w|\theta)) ds_j(w),$$

where $\mathcal{W} = \{w \in \mathbb{R}^{(q+k) \times n_j} : \hat{f}_j(s_j(w)|f_P(w|\theta)) < \hat{f}_j(s_j(w_j)|f_P(w_j|\theta))\}$. Since π_j is a probability, $\pi_j(\theta) \in [0, 1]$. Furthermore

$$\pi(\theta) = \cap_{j=1}^J \pi_j(\theta),$$

where $\pi(\theta) \in [0, 1]$.

Remark 4.1. By Definition 4.1, significance is defined in terms of the sampling distribution \hat{f}_j . Since the samples are from simulations, the approximation can be as close to the true f_j as desired, since the number of simulations can be increased.

Definition 4.2 (Plausibility). We say that θ is plausible at the $1 - \alpha$ level if

$$\pi(\theta) > \alpha,$$

for $\alpha \in (0, 1)$.

Deciding whether a θ is plausible can be viewed as a hypothesis test, where the null-hypothesis H_0 is that $f_P(\cdot|\theta)$ generated the observed data, and H_0 is only rejected if $\pi(\theta) \leq \alpha$. \hat{f}_j approximates the true distribution for any sample size and does not rely on asymptotic results of large samples. Hence, the test described in Definition 4.2 can be viewed as an exact test rather than an approximate, large-sample test.

When making inference with the HS method, we are looking for all possible θ 's that are plausible. We call this the Plausible Region of θ .

Definition 4.3 (The Plausible Region $\mathcal{R}(\alpha)$). The $1 - \alpha$ Plausible Region of θ is

$$\mathcal{R}(\alpha) = \{\theta \in \Theta : \pi(\theta) > \alpha\}$$

for $\alpha \in (0, 1)$.

Remark 4.2. Let θ^* denote the parameter vector of the same length as θ such that:

1. θ^* has the same value as θ^M for all parameters that are both in θ and θ^M .
2. For the parameters which are in θ but not in θ^M , the value in θ^* corresponds to the parameters having no impact (direct or indirect) on the variables W^M in the Mother Nature's Model.

A characteristic of the Plausible Region is that we expect it to contain the value of θ^* with probability $1 - \alpha$, that is

$$P[\theta^* \in \mathcal{R}(\alpha)] = 1 - \alpha.$$

$\mathcal{R}(\alpha)$ gives us a region of plausible parameter values, but it does not give us a particular value of θ which is deemed the most plausible. Neither does it enable us to say which of the several plausible models is the most plausible.

How these facts should be handled likely depends on the context. In some contexts, we may choose the θ with the highest significance value as our point estimate. In other cases, it may be wiser to refrain from drawing a conclusion and propose more research. Or it may be possible to choose a θ over others based on practical or theoretical considerations. For instance, if exposure to a pesticide is deemed highly dangerous according to some θ but not according to another, we may be best to assume that the most pessimistic θ is the correct one.

4.2 An algorithm for finding a plausible θ

The Plausible Region $\mathcal{R}(\alpha)$ can be found by performing a hypothesis test for each θ that we are interested in. In the simplest case, there are K different θ vectors ($\theta^1, \dots, \theta^K$) that we test with the following algorithm:

For $\theta^k \in (\theta^1, \dots, \theta^K)$:

$H_0 : f_P(\cdot | \theta^k)$ generated the data.

If $\pi(\theta^k) < \alpha \rightarrow$ reject H_0 , **else** do not reject H_0 .

For high-dimensional θ the above procedure becomes computationally costly, particularly if we want to consider the whole Θ . Therefore, it is desirable to develop a numerical method that can identify $\mathcal{R}(\alpha)$. No such method has yet been developed.

4.3 Density approximation techniques

To evaluate $\pi(\theta)$, we rely on methods of approximating the probability density function of the summary statistics based on simulated data. We have explored two alternatives: kernel density estimators (KDEs) and copula density estimators (CDEs). While the term “estimator” is established, a more fitting term for our application is “approximator”. This is since the data we use with the estimator is simulated data from the assumed data-generating process, such that the copula or kernel method approximates this probability density function.

We discuss both methods in the next sections, although CDEs are used in the simulations in sections 6, 7 and 8.

4.3.1 Kernel density estimators

The kernel density estimator is an established method of estimating an unknown probability density function based on observed data from this distribution.

Definition 4.4 (Kernel density estimator). Given d -dimensional datapoints $\{x_i\}, i = 1, \dots, n$, $x_i \in \mathbb{R}^d$ from an unknown probability density function f , the kernel density estimator $\hat{f} : \mathbb{R}^d \mapsto \mathbb{R}_+$ is

$$\hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^n K_H[(x_i - x)].$$

Here, $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$ with $K : \mathbb{R}^d \mapsto \mathbb{R}_+$. H is a symmetric, positive-definite $d \times d$ bandwidth matrix (sometimes called smoothing matrix). K and H can be chosen so that \hat{f}_H is $\hat{f}_H(x) \geq 0$ and $\int_X \hat{f}_H(x) dx = 1$, as is required from a probability density function. (Duong 2007)

There is an abundance of kernel functions to choose from. Most common, and the default option in most statistical software, is the Gaussian kernel: $K(x) = (2\pi)^{-d/2} \exp(-\frac{1}{2}x^t x)$. Generally, the choice of kernel has little impact on the performance of \hat{f}_H , while the bandwidth matrix H has a large influence. A natural measure of performance of \hat{f}_H is the mean integrated square error (MISE).

Definition 4.5 (MISE). MISE measures the average (over data-samples) squared distance between the estimated function and the true function:

$$MISE(\hat{f}_H) = E\left[\int_X (\hat{f}_H(x) - f(x))^2 dx\right] = \int_{\mathbb{R}^d} Bias[\hat{f}_H(d)]^2 dx + \int_{\mathbb{R}^d} Var(\hat{f}_H(x)) dx.$$

Remark 4.3. Based on the MISE, the optimal H matrix is $H_{MISE} = \arg \min_H MISE(\hat{f}_H)$, where H can be selected from the space of symmetric, positive definite $d \times d$ matrices. (Duong 2007)

MISE is in general impossible to express in closed form, so to find a suitable H one often relies on the asymptotic mean squared error (AMISE).

Definition 4.6 (AMISE). AMISE can be derived from the MISE via Taylor-series expansion:

$$AMISE(\hat{f}_H) = \frac{|H|^{-1/2}}{n} R(K) + \frac{\mu_2(K)^2 (\text{vech}(H)^t) \Psi_4 (\text{vech}(H))}{4}. \quad (5)$$

Here, $R(K) = \int_X K(x)^2 dx$; $\mu_2(K)I_d = \int_X x^t x K(x) dx$; and $\Psi_4 = \int_X \text{vech}[2\mathcal{H}_f - dg\mathcal{H}_f] \text{vech}[2\mathcal{H}_f - dg\mathcal{H}_f]^t dx$.

\mathcal{H}_f is equal to the Hessian matrix of f . dgA is the matrix A with all off-diagonal elements set to zero. $\text{vech}(H)$ is a $d(d+1)/2$ vector containing all entries in H on or below the diagonal, listed column by column. E.g. for $H = \begin{pmatrix} h_1 & h_2 \\ h_3 & h_4 \end{pmatrix}$, then $\text{vech}(H) = (h_1, h_3, h_4)$. (Wand & Jones 1996)

Remark 4.4. Based on AMISE, the optimal H matrix is $H_{AMISE} = \arg \min_H AMISE(\hat{f}_H)$.

As Wand & Jones (1996) notes, $AMISE(\hat{f}_H)$ depends on the unknown f , since Ψ_4 is defined in terms of the Hessian of f . In order to find H_{AMISE} , a common approach is to estimate f based on the data, and insert this estimate into $AMISE(\hat{f}_H)$. To this end, we note that Ψ_4 is a $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ matrix, the elements of which are the integrated density derivative functionals

$$\psi_r = \int_{\mathbb{R}^d} f^{(r)}(x) f(x) dx,$$

where $r = (r_1, \dots, r_d)$ and $|r| = \sum_{i=1}^d r_i$. Moreover,

$$f^{(r)}(x) = \frac{\partial^{|r|} f(x)}{\partial_{x_1}^{r_1} \cdots \partial_{x_d}^{r_d}},$$

with a $\psi_r \neq 0$ if and only if $|r|$ is an even integer. For instance, in the two-dimensional case,

$$\Psi_4 = \begin{pmatrix} \psi_{4,0} & 2\psi_{3,1} & \psi_{2,2} \\ 2\psi_{3,1} & 4\psi_{2,2} & 2\psi_{1,3} \\ \psi_{2,2} & 2\psi_{1,3} & \psi_{0,4} \end{pmatrix}.$$

Note that since f is a density, we have that

$$\psi_r = \int_X f^{(r)}(x) f(x) dx = E(f^{(r)}(x)). \quad (6)$$

Below we describe two common ways of estimating ψ_r : plug-in estimators and cross-validation estimators.

1. **Plug-in estimator:** A plugin-estimator uses a kernel estimator $\hat{\Psi}_{PI}$ of Ψ_4 in formula (5), so that the f which is used in the true ψ_r is replaced with $\hat{f}_G(x) = \frac{1}{n} \sum_j K_G(x - X_j)$. $\hat{\Psi}_{PI}$ has elements

$$\hat{\psi}_r(G) = E(\hat{f}_G^{(r)}) = \frac{1}{n^2} \sum_i^n \sum_j^n K_G^{(r)}(X_i - X_j)$$

with G as a pilot bandwidth matrix. Here we used result (6) in the second step.

According to Duong (2007), the choice of G is less important than the choice of H . It is common to use the diagonal matrix $G = g^2 I$, where g can be chosen based on several criteria, for instance the mean squared error of $\hat{\psi}_r$.

2. **Cross-validation estimators:** There are several types of cross-validation estimators: least squares, biased and smoothed cross-validation estimators. As the names suggest, all of them are based on leave-one-out cross-validation estimators. In the case of biased cross-validation estimators, we can either replace ψ_r with

$$\hat{\psi}_r^1(H) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n K_{2H}^{(r)}(X_i - X_j)$$

or with

$$\hat{\psi}_r^2(H) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n K_H^{(r)}(X_i - X_j).$$

See Sain, Baggerly & Scott (1994) for further details on these estimators, and also for more on the smoothed cross-validation estimator and the least-squares cross-validation estimator; the latter is based on minimizing an altogether different loss-function than AMISE.

Both plug-in estimators and cross-validation estimators can be used to identify diagonal and unrestricted matrices.

A benefit of KDEs is that they are very flexible and can be used to approximate complicated densities. Their non-parametric nature means that we impose few restrictions on the data. A negative aspect of the KDE method is that it is computationally demanding in higher dimensions.

KDEs are almost exclusively used for continuous variables, and several of the features outlined above rely on continuous data. KDEs may also be used with discrete and ordinal data as well, but performance is in general not as good, and different kernels and bandwidth matrices are preferred. For mixed data, consisting of both discrete and continuous data, there is less research, but the interested reader can consult Li & Racine (2003).

4.3.2 Copula estimators

Copulas are used to model the dependence between several random variables. The name copula comes from the Latin word copulare, which means “connect”, since the copula connects marginal distributions.

Definition 4.7 (Copula estimator). A copula estimator is a multivariate cumulative distribution function $C : [0, 1]^d \mapsto [0, 1]$ such that, by Schmidt (2006),

1. $C(u_1, \dots, u_d) = 0$ if $u_i = 0$ for any $i \in \{1, \dots, d\}$.
2. C has standard uniform univariate margins, so that $C(1, 1, \dots, u_i, \dots, 1) = u_i$ for all $u_i \in [0, 1]$ and all $i \in \{1, \dots, d\}$.
3. C is d -non-decreasing, so that for any d -rectangle $K \subseteq [0, 1]^d$, $\int_K C(k) dk \geq 0$.

Remark 4.5. Provided that the partial derivatives of C exist, the copula estimator of the probability density function is

$$c(u) = \frac{\partial^d C(u)}{\partial u_1 \cdots \partial u_d}.$$

Theorem 4.1 (Sklar's theorem). Mathematician Abe Sklar showed that for every multivariate cumulative distribution function F , there is a copula C such that

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)),$$

for all $x_i \in (-\infty, \infty), i \in \{1, \dots, d\}$. When all the F_i are continuous, the copula C is unique; otherwise it is uniquely defined on $\text{Range}(F_1) \times \dots \times \text{Range}(F_d)$. (Nielsen 2006, p 46)

Remark 4.6. If T_i is a strictly increasing function; $X_i, i \in \{1, \dots, d\}$ are continuous random variables; and if C is a valid copula for $X_i, i \in \{1, \dots, d\}$, then C is also a valid copula for $Y_1 = T_1(X_1), \dots, Y_d = T_d(X_d)$. (Nielsen 2006, p 25)

Definition 4.8 (The generalized inverse). The generalized inverse of F is

$$F^{\leftarrow}(t) = \{x : F(x) \geq t\}.$$

Remark 4.7. By definition 4.8, we can write a copula as

$$C(u_1, \dots, u_d) = F(F_1^{\leftarrow}(u_1), \dots, F_d^{\leftarrow}(u_d)).$$

This is the definition used in the simulation studies in sections 6, 7 and 8. (Schmidt 2006)

The question is what cumulative distribution functions F_i and F should be used. There are many alternatives. It is possible to estimate the marginal cumulative distribution functions with kernel estimators or to simply use the empirical marginal cumulative distribution function. Most common, however, are parametric cumulative distribution functions. While there is a risk of model misspecification, parametric distribution functions are computationally efficient. Some noteworthy parametric copulas are:

1. Gaussian copula:

The Gaussian copula is defined

$$C_R^G(u) = \Phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)),$$

with Φ_R as the multivariate standard normal distribution with covariance matrix R , and Φ the univariate standard normal distribution. It is due to Remark 4.6 that we only need to consider the standard normal distribution, as standardizing a variable is a strictly increasing function.

2. t-copula:

The t-copula is defined

$$C_{v,R}^t(u) = T_{v,R}(T_v^{-1}(u_1), \dots, T_v^{-1}(u_d)),$$

with $T_{v,R}$ the multivariate cumulative distribution function of the t-distribution with v degrees of freedom and correlation matrix R . T_v is the cumulative distribution function of the univariate t-distribution with v degrees of freedom; as v

increases the t-copula approximates the Gaussian copula. As we would expect, a t-copula is preferable to a Gaussian copula when events in the tails are more pronounced.

3. Archimedean copulas:

There is an abundance of Archimedean copulas. Their main characteristic is that they use a single parameter to model the dependence between the variables, no matter the dimension. At the heart of Archimedean copulas are so-called generator functions, g . These are continuous, strictly increasing and convex functions, $g : [0, 1] \times \theta \mapsto [0, \infty)$. It follows that $S = \sum_i^d g(u_i) \in [0, \infty)$, so that $g(S)^{-1} \in [0, 1]$. An example of an Archimedean copula is the Gumbel copula $g(t, \theta) = (-\log(t))^\theta$, and $g^{-1}(t) = \exp(-t^{1/\theta})$, so that

$$C^{Gu}(u_1, \dots, u_d) = \exp\left[\left(\sum_i^d [-\log(u_i)]^\theta\right)^{1/\theta}\right]$$

with $\theta \in [1, \infty)$. This assumes a non-negative correlation between the marginal distributions; other copulas make different assumptions about the correlation.

It is possible to compose a copula out of a single type of cumulative distribution function, or by using a mixture of distributions. The choices can be made based on theoretical or practical considerations. For example, when analyzing summary statistics such as log odds ratios, we know that they are expected to follow a normal distribution, and therefore a Gaussian copula or a t-copula makes sense. In the case of parametric copulas, the parameters can be estimated from the data. Goodness-of-fit measures can also be used to select the copula.

A central aspect of copulas is the modeling of the dependence structure between the variables. Pearson's r , which is a measure of linear correlation, can be selected as the correlation matrix for copulas composed of elliptical distributions, for instance the Gaussian copula and the t-copula. Outside of elliptical distributions, rank-based correlation such as Kendall's τ or Spearman's ρ are more suitable. Rank-based correlation is used in Archimedean copula. (Schmidt 2006)

Just as with KDEs, copula estimators are best suited to continuous variables. In particular, by Sklar's theorem, we know that there is not necessarily a unique copula for discrete data. Also, the fact that there are likely to be ties with discrete data makes the covariance more complicated to model. Copula estimators for discrete and mixed data is a field that has not been thoroughly investigated, but the interested reader can find more information in Genest & Neslehova (2007).

To the best of our knowledge, it is not common to use the copula estimator as a way of approximating a probability density in the way that we are interested in. The approach was conceived because the KDE approach is computationally untenable in higher dimensions. Compared with KDEs, the great benefit of the copula approach is that it is highly computationally efficient, in particular with a parametric copula. The danger of using copulas in this way is that it is sensitive to model misspecification, which is less of a problem with KDEs.

4.4 Correcting for multiple studies

This section is based on Agresti & Coull (1998).

As stated in section 4.2, inference with the HS method can be viewed as a hypothesis test where H_0 is that $f_P(w|\theta)$ generated the data. H_0 is not rejected when

$$\pi(\theta) \geq \alpha \Leftrightarrow \cap_{j=1}^J \pi_j(\theta) \geq \alpha.$$

When $J > 1$, this hypothesis test has to be adapted to the fact that several tests are being performed. Otherwise, if the studies are independent of each other, we have that under H_0

$$P(\pi(\theta) > \alpha) = (1 - \alpha)^J < (1 - \alpha).$$

There are several ways of avoiding this. One option is Bonferroni correction, which sets the confidence level to $1 - \alpha/J$. The test then becomes

$$\cap_{j=1}^J \pi_j(\theta) > \alpha/J,$$

which yields a test on the desired confidence level, approximately.

Bonferroni correction is problematic when used with a kernel or copula density estimator. As J increases, $1 - \alpha/J$ becomes more fine-grained, and more simulations are required for the density estimator to make the distinction between different confidence levels. Therefore, if J is large and the data is in a high dimension, Bonferroni correction becomes extremely computationally costly.

Another option is to consider the statistic

$$L = \sum_{j=1}^J I(\pi_j(\theta) > \alpha).$$

Under H_0 , $L \sim \text{Bin}(J, p_0)$, with $p_0 = 1 - \alpha$. Based on the observed L value, we can calculate an estimation \hat{p} of p_0 and reject H_0 if $\hat{p} \neq p_0$ on significance level α . There are several ways of performing this hypothesis test.

One option is the Clopper-Pearson exact confidence interval. With l denoting the observed value of L , the lower limit of the confidence interval is $\omega_l = \inf\{p : P(\text{Bin}(J, p) \leq l) > \alpha/2\}$ and the upper limit is $\omega_u = \sup\{p : P(\text{Bin}(J, p) \geq l) > \alpha/2\}$. The confidence interval is given by

$$CI_{CP} = (\omega_l, \omega_u).$$

The Clopper-Pearson interval contains the true p with at least probability $1 - \alpha$. With small J , it has too great a coverage.

An alternative is to calculate a $1 - \alpha$ Wald confidence interval

$$CI_{Wa} = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{J}},$$

with $\hat{p} = l/J$ and z_a the $1 - a$ quantile of the standard normal distribution. The Wald confidence interval relies on a normal approximation of the binomial distribution which

is reasonable for large J , but not for small J . In most applications of the HS method that we can imagine, it may not be a wise choice.

Several modifications of the Wald confidence interval have been proposed to better handle situations with small J . A simple modification of the Wald confidence interval was suggested by Agresti & Coull (1998), although they did not invent it. We add two successes and two failures to our data and calculate a standard Wald confidence interval based on the modified test data. This simple modification makes the Wald test work much better when there are a small number of trials.

A more elaborate improvement of the Wald confidence interval was proposed by Edwin B. Wilson. It replaces the estimated variance with the null variance, and the upper and lower limits of the confidence intervals are: $\omega_u^* = \inf\{p : (\hat{p} - p_0)/\sqrt{[p_0(1 - p_0)]/J} \geq z_{\alpha/2}\}$ and $\omega_l^* = \sup\{p : (\hat{p} - p_0)/\sqrt{[p_0(1 - p_0)]/J} \leq -z_{\alpha/2}\}$, with the confidence interval

$$CI_{Wi} = (\omega_l^*, \omega_u^*).$$

In Figure 2, we compare the performance of the Wilson, Coull-Agresti and Clopper-Pearson tests for 1, ..., 100 trials. As we can see, there is little difference between the tests for more than 40 trials, but for fewer trials, there are large oscillations and no method is consistently better than the others. Therefore, the choice of test reasonably depends on the number of studies included in an analysis.

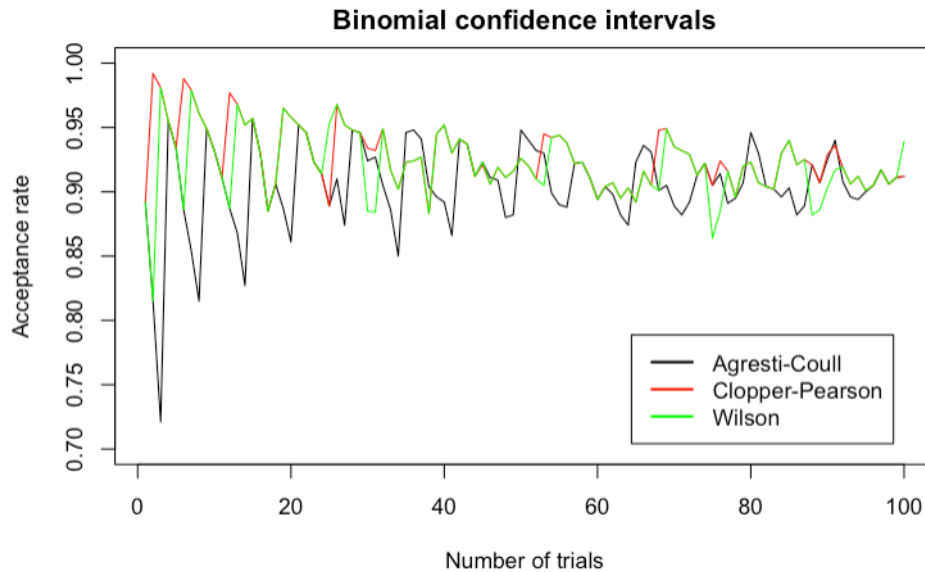


Figure 2: A graph comparing the proportion of time that three different binomial test confidence intervals contain the true parameter $p = 0.9$. The proportions are based on 1000 repetitions per number of trials.

4.5 Likelihood-based inference

The sampling-based inference method outlined in section 4.1-4.3 is used in the simulation studies in sections 6, 7 and 8. As noted, this method has limitations. Going forward, it is desirable to develop a more sophisticated method and to ground this method in probability theory.

One possible direction is to use likelihood-based inference. A benefit of this option is that methods that resemble the HS method have already been developed within the likelihood framework, both frequentist and Bayesian.

Based on J studies and approximate sampling distributions $\hat{f}_j(s_j(w)|f_P(w|\theta)), j \in \{1, \dots, J\}$ as in (4), we define the approximate likelihood as

$$\hat{l}(\theta) = \prod_{j=1}^J \hat{f}_j(s_j(w_j)|f_P(w|\theta)).$$

The approximate maximum likelihood estimator is

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{l}(\theta).$$

Assuming that standard likelihood theory is applicable, $\hat{\theta}$ has the asymptotic distribution

$$\hat{\theta} \overset{apx}{\sim} N(\theta, \hat{\Sigma}),$$

with $\hat{\Sigma}$ being the estimator of the covariance matrix Σ . Assuming a large number of summary statistics, we could make inference on θ based on the asymptotic distribution, to produce a point estimate of $\hat{\theta}$ as well as a $1 - \alpha$ confidence region.

However, it is far from obvious how such an inference method can be properly developed. Compared to standard situations where maximum likelihood estimation is used, there are some important differences with the HS method. Some of these were highlighted in section 3, but they are worth repeating in this context:

1. **Small sample size:**

While we can use a wide range of studies with the HS method, we cannot in general expect to have so many studies that large-scale asymptotics can be used with reliable results.

2. **Summary statistics:**

We need to be able to make reliable inference from summary statistics. This is certainly possible if the summary statistics are sufficient, meaning that for the sufficient statistic $t(x)$, we have that $f(x|t(x))$ is independent of the θ we are estimating. However, this is not often likely to be the case. As far as we know, little research has been done on how to make inference based on summary statistics in similar cases. Proponents of Approximate Bayesian Computation (discussed in the next section) have considered it to some extent.

3. **The sampling process:**

The confidence region for the estimate $\hat{\theta}$ is valid under the assumption that the

data used to perform inference is sampled from a population. It is unclear what the population is and how data was sampled for the summary statistics that are included in an analysis with the HS method. The result of a particular study depends not only on the participants who were sampled but also on the experimental design, association measures and many other details of the study. Are we sampling from all possible study designs or some subset? This question is similar to the one we asked regarding the Random Effects model.

4. Asymptotics of what?

Related to the question about the sampling process, is how approximations based on asymptotic results should be interpreted. Is it the number of people that are being sampled in studies that goes to infinity, or is it rather the number of studies that go to infinity?

4.6 Similar methods

In this section, we look at two methods that share similarities with the HS method: a method suggested by Diggle & Gratton and Approximate Bayesian Computation. Neither of these two methods is meant to be used for data from heterogeneous sources in the way that the HS method is. Rather, they have been developed to handle complicated likelihoods. We also briefly describe sensitivity analysis and imputation, which share similarities with the HS method.

4.6.1 Diggle & Gratton's method

Diggle & Gratton (1984) suggest a method for approximating the maximum likelihood estimates of the parameters in an “implicit model”, that is a model of the data-generating process, in situations where this model cannot be expressed analytically. Diggle & Gratton estimate the maximum likelihood estimator of θ in an implicit model from simulated realizations from such an implicit model, according to the following algorithm. For a model $g(\cdot|\theta)$ and observed data y :

1. For $k = 1, \dots, K$:
 - a) Pick θ^k from the K -vector of possible θ vectors: $(\theta^1, \dots, \theta^K)$.
 - b) Generate $y_j^* \sim g(y|\theta^k)$, $j = 1, \dots, m$.
 - c) Based on y_j^* , $j = 1, \dots, m$ create a kernel-density estimator $\hat{g}(y|\theta^k)$, which serves as an approximation of $g(y|\theta^k)$
2. Approximate $\hat{\theta}_{ML}$ by $\max_{\theta^k} \hat{g}(y|\theta^k)$

Diggle & Gratton also suggest a numerical method for finding the optimal θ instead of being reliant on a vector of θ values.

While Diggle & Gratton's method is similar to the HS method in the sense that both methods rely on simulating from a complicated statistical model, there are differences in

the imagined applications. Using a large-sample approximation of the likelihood makes less sense in a situation where there are few data sources to use, and it is not clear how the method works in situations where the only data available are non-sufficient summary statistics.

A suggestion for future research is to look more closely at applications of the Diggle & Gratton method, and to investigate if the numerical method they suggest is relevant for the HS method.

4.6.2 Approximate Bayesian Computation

This section is based on Beaumont (2010).

Just like the method by Diggle & Gratton, Approximate Bayesian Computation (ABC) is developed to handle situations where the likelihood function of our statistical model is complicated. But as suggested by the name, ABC is used within a Bayesian framework, meaning that we do not consider the parameters in the model to be fixed; instead, we think of them as random variables. Bayesian inference is centered around the concept of the posterior distribution of the parameters θ , given data x and a prior distribution of θ , $\gamma(\theta)$:

$$p(\theta|x) = \frac{p(x|\theta)\gamma(\theta)}{p(x)}.$$

The numerator $p(x) = \int p(x|\theta)\gamma(\theta)d\theta$ is a normalizing constant that can often be ignored. It remains to evaluate $p(x|\theta)$ and $\gamma(\theta)$. If $p(x|\theta)$ is too complex to work with analytically, Monte Carlo simulation can be used.

There are several ABC methods, that differ from each other mostly concerning details of the simulation scheme. The simplest ABC method works in the following way, given observed data y and a fixed ϵ :

1. Repeat until N θ^* are accepted:
 - a) Sample θ^* from $\gamma(\theta)$
 - b) Simulate $y^* \sim p(y|\theta^*)$
 - c) If $\rho(s(y^*), s(y)) < \epsilon$, then θ^* was accepted, otherwise discarded

Here, $\rho(.,.)$ is a distance function, e.g. Euclidian distance, and $s(.)$ is a summary statistic.

The outcome of the algorithm is a sample of parameter values from an approximation of the posterior distribution $p(\theta|y)$. A kernel-density estimator can be used to visualize this distribution, and to give mean, mode or median values, as well as measures of dispersion, such as credible intervals.

The ABC method resembles the HS method in the sense that both are based on simulations and both often rely on making inference based on summary statistics. This means that ABC users have dealt with the problem of making inference based on summary statistics that are not necessarily sufficient. The solutions proposed by Beaumont (2010) and Joyce & Marjoram (2008) is to include many jointly sufficient summary

statistics. However, if many statistics are included, the simulation algorithm becomes inefficient due to the curse of dimensionality, and the estimate of θ may be plagued by stochastic noise. To deal with this, Hamilton (2005) suggests that we weigh the statistics according to their association with the parameters in the model. Joyce & Marjoram (2008) suggest that statistics are selected based on how they contribute to the posterior.

For future research, it may be relevant to look further into how statistics are selected in ABC. In case the HS method is used within a Bayesian framework, the inference procedure may end up looking very much like the ABC method.

4.6.3 Sensitivity analysis and imputation

Sensitivity analysis is a broad range of methods where the aim is to investigate how sensitive the conclusions of a statistical analysis are to the assumptions behind the analysis. A common application is to test how the conclusion of an analysis changes depending on the potential effects of an unmeasured confounder. For instance, if we have a measure of the association between cancer and resin exposure, we may wonder if the observed association is due to the smoking habits of the participants. Even if there is no information on smoking habits, we can model the possible relationship between resin exposure and smoking, and explore how extreme the assumptions need to be for the apparent association between resin and cancer to be explained by the smoking habits of the participants exposed to resin. (Greenland 1998)

Imputation is a method used to handle missing data. When there is missing data, applying traditional statistical methods directly to the observed data may lead to bias. The idea behind imputation techniques is to handle the missing data in such a way that our estimations become unbiased. (Rubin 1996) Sometimes, imputation can be considered to be a special type of sensitivity analysis, because we can investigate how extreme the missing data would had to have been to change the conclusion of our investigation.

Imputation can be directly applied with the HS method, for instance when we want to model the effect of non-response bias, publication bias and other types of bias due to missing data. Similarly, if we are lacking observations related to a possible confounder in our Postulated Model, then the Plausible Interval for the parameter related to the confounder can be considered the outcome of a sensitivity analysis, since the interval tells us under what values for this parameter our model remains plausible.

5 Computational aspects of the inference

In this section, we describe some of the technical details in performing inference according to the algorithm described in section 4.

5.1 Implementation of the kernel density estimator

Although the kernel density estimator plays a minor role in the simulation studies, it was implemented and used. We used the R package **ks**, which includes the function **kde**, that produces a kernel density estimator based on simulated data. A Gaussian kernel is default, and we did not attempt to use any other kernel.

The **ks** packages lets us estimate the bandwidth matrix H with several different methods. We used a plug-in estimator and three types of cross-validation estimators: biased, smoothed, and least-squares estimators. Each of these estimators comes with different options. For instance, in the case of plug-in estimators, we can choose how to select g in the pilot matrix $G = g^2 I_d$, and whether the input data should be scaled or sphered. While we explored the different options, no interesting results came out of this in the simulation studies described in sections 6, 7 and 8.

5.2 Implementation of the copula estimator

In the simulation studies in sections 6, 7 and 8 we implemented our own copula density estimator in the following way:

1. Collect d -vectors of summary statistics $s_i, i \in \{1, \dots, I\}$ from I simulated studies in a matrix S , such that each row corresponds to a simulated study and each columns corresponds to a particular statistic. Insert the observed statistics s^* to S , such that S is a $(I + 1) \times d$ matrix.
2. Calculate the empirical distribution function for each column in S . Transform S by applying the inverse of the cumulative probability density function of the standard normal distribution, $\Phi^{-1}(x)$, for each cell in S .
3. Set $R = Cor(S)$ with the correlation being Pearson's r or Spearman's ρ .
4. The row in S corresponding to the observed statistic s^* is now transformed into an observation from a $N(0_d, R)$ distribution. Call this row s' .
5. Evaluate $\phi_R(s')$ and check whether it belongs to the $1 - \alpha$ highest density region of the distribution.

5.3 Correction for multiple studies

To approximate the plausible level $1 - \alpha$ when there is more than one study, we implemented the tests by Clopper-Pearson, Coull-Agresti and Wilson. In the simulation studies in sections 6 and 7, we are using 1, 5, 20, 100 research studies with plausible level

0.9, and the Coull-Agresti test is expected to be closest to the true confidence region with these particular numbers. Therefore we implemented the Coull-Agresti test via the R package **fastR2**, and the function **wilson.ci**. The name **wilson.ci** suggests that we estimate a Wilson confidence interval. This is because in the literature, the Coull-Agresti and Wilson tests are used interchangeably.

6 Simulation Study I

We have undertaken two simulation studies to confirm that the HS method performs as expected, and to answer some of the relevant questions regarding its performance. In the first simulation study, the following questions are answered:

1. **Is a true model accepted 100(1 − α)% of the time?**
This question is important to investigate since we claim that the plausible interval $\mathcal{R}(\alpha)$ should contain the true θ in 100(1 − α)% of the time.
2. **What is the effect of the number J of studies and the number n of participants in a study on the inference?**
We would expect more studies with more participants to yield a narrower $\mathcal{R}(\alpha)$. If this were not the case, it would seem futile to include more data in an analysis.
3. **Are we able to extract information about variables that are not directly observed in any study; in particular, can we extract information about unmeasured confounders?**
This question is interesting concernign the question of what data can be used to make inference on our Postulated Model. If we can gain information about variables that are not directly observed, it makes the method much more useful.
4. **Is inference possible for heterogeneous/homogeneous populations?**

To answer these questions, we have performed simulations with two data-generating models. In both cases, the variables are X , Y and Z , where Y is an outcome of interest. X impacts Y and Z impacts both X and Y . Hence, Z is a confounder. We assume that the available research studies report the association between X and Y , and do not take Z into account. Hence, Z is an unmeasured confounder.

Next, we describe the data-generating models, and then describe how the simulations were performed.

6.1 The data-generating models

6.1.1 Mother Nature's Model 1

For the participants in a study, the true data-generating model (that is, the Mother Nature's Model) is:

$$Z^B \sim Be(g(0))$$

$$X^B \sim Be(g(-0.5 + Z^B))$$

$$Y^B \sim Be(g(-1 + X^B + Z^B + X^B Z^B))$$

where $g(a) = [1 + \exp(-a)]^{-1}$

The corresponding Postulated Model 1 is:

$$Z^B \sim Be(g(c_7))$$

$$X^B \sim Be(g(c_5 + c_6 Z^B))$$

$$Y^B \sim Be(g(c_1 + c_2 X^B + c_3 Z^B + c_4 X^B Z^B))$$

As we can see, the Postulated Model 1 has the same form as the Mother Nature's Model 1; the only difference is that the parameter values are unknown.

6.1.2 Mother Nature's Model 2

For the participants in a study, the true data-generating model (that is, the Mother Nature's Model) is:

$$Z^N \sim Norm(mean = 0, sd = 1)$$

$$X^N \sim Norm(mean = 0 + Z^N, sd = 1)$$

$$Y^N \sim Norm(mean = 0 + X^N + Z^N + X^N Z^N, sd = 1)$$

The corresponding Postulated Model 2 is:

$$Z^N \sim Norm(mean = d_9, sd = d_{10})$$

$$X^N \sim Norm(mean = d_6 + d_7 Z^N, sd = d_8)$$

$$Y^N \sim Norm(mean = d_1 + d_2 X^N + d_3 Z^N + d_4 X^N Z^N, sd = d_5)$$

As we can see, the Postulated Model 2 has the same form as the Mother Nature's Model 2; the only difference is that the parameter values are unknown.

6.2 The simulation algorithm

We assume that $J \in \{1, 5, 20, 100\}$ studies with $n \in \{10, 100, 100\}$ participants have investigated the same research question with the same statistical analysis and sampling procedure. Below, we describe the simulation procedure for Postulated Model 1:

1. Set K coefficient vectors $c^k, k \in \{1, \dots, K\}$ that we wish to test.
2. Use the true data-generating model to generate J datasets $\mathbf{w}_j, j \in \{1, \dots, J\}$, where each dataset consists of X^B, Z^B and Y^B values for n participants. Estimate the corresponding summary statistic $s_j, j \in \{1, \dots, J\}$, where s_j is the intercept and slope in a logistic regression model of Y^B given X^B .
3. For $k \in \{1, \dots, K\}$:

- a) Use parameter vector c^k with Postulated Model 1 to create 1000 simulated datasets $\mathbf{w}_{k,i}^*, i \in \{1, \dots, 1000\}$, where each dataset contains X^B , Z^B and Y^B values for n participants. For each dataset, estimate the intercept and slope $s_{k,i}^*$ in a logistic regression model of Y^B given X^B .
 - b) Use $\{s_{k,i}^*\}, i \in \{1, \dots, 1000\}$ to create a density estimator of $\pi(c^k)$.
 - c) If $\pi(c^k) > \alpha$, then c^k is deemed plausible, otherwise it is deemed implausible.
4. Repeat step (2)-(3) 100 times and calculate the proportion of simulations in which $c^k, k = 1, \dots, K$ was deemed plausible.

The above algorithm is repeated for all possible combinations of $n \in \{10, 100, 100\}$ participants and $J \in \{1, 5, 20, 100\}$ studies.

The simulation procedure is used to investigate one parameter in the parameter vector at a time. When investigating a particular parameter, the K parameter vectors are such that all parameters except for one are fixed at the true value, according to the Mother Nature's Model. The value of the remaining parameter varies across the vectors. For instance, we could investigate c_1 by testing the following vectors, where only the first digit changes:

$$(-2, 1, 1, 1, 0, 1, 0)$$

$$(-1, 1, 1, 1, 0, 1, 0)$$

$$(0, 1, 1, 1, 0, 1, 0)$$

By plotting the results in a graph, we can see how often a given coefficient vector c^k is deemed plausible.

For Postulated Model 2, the simulations are performed analogously, meaning that we use Mother Nature's Model 2 and Postulated Model 2, and the statistic which is reported is the intercept and slope from a linear regression model of Y^N given X^N .

6.3 Computational details of the inferential procedure

In order to evaluate $\pi(\cdot)$, we need to determine which density estimation method to use. As pointed out in section 5, we have used both kernel density estimators and copula density estimators with no notable difference. The results presented were produced with the copula estimator, using a Gaussian copula.

The Gaussian copula was used for several reasons. Firstly, we know that the statistics we are estimating are expected to follow a normal distribution: this is the case both for the maximum likelihood estimators of the coefficients in a linear regression model and the maximum likelihood estimates of the log odds ratio from a logistic regression model. This has been confirmed by graphical analysis. Secondly, we could see that the copula estimator was performing similarly to how the kernel density estimator performed. No further analysis of the behavior of the copula density estimator has been performed, and no comparison has been done to other potential copula estimators, such as the t-copula.

In the presented results, the correlation coefficient used is Pearson’s r . It made sense to use r since it is calculated based on values converted into values from a Gaussian distribution, and they are assumed to originally be normally distributed.

While we could have made a more thorough investigation into the details of the kernel density estimator and copula estimator, the main purpose of the current study is to see if the HS method works. Optimizing the method can be the subject of another study.

6.4 Simulations results

Graphs illustrating the results of all simulations are found in section 10. We refer to these in the text below.

1. Is the true model accepted $100(1 - \alpha)\%$ of the time?

For both the Postulated Model 1 and Postulated Model 2, we can see on all of the graphs that the true parameter value is accepted $100(1 - \alpha)\%$ of the time. In the simulation study, we set $\alpha = 0.1$, and it is indeed the case that the true model is deemed plausible in around 90% of the simulations.

Table 1 shows the proportion of simulations when the true coefficient vector was deemed plausible depending on the number of research studies included in the analysis. As we can see, both Postulated Model 1 and Postulated Model 2 perform as expected given that we used the Coull-Agresti method to correct for the number of studies.

This means that the Plausible Region $\mathcal{R}(\alpha)$ seems to work as expected with a wide range of number of studies, and for data that is both binary and continuous.

No. studies	Coull-Agresti	Postulated Model 1	Postulated Model 2
1	0.9	0.91	0.89
5	0.92	0.92	0.92
20	0.87	0.88	0.86
100	0.9	0.89	0.89

Table 1: This table shows the proportion of the simulations in which the true vector was deemed plausible, depending on the number of research studies included in the analysis. The reported proportions are averages over all possible numbers of participants n , for a given number of studies. The column Coull-Agresti shows the expected proportions using a Coull-Agresti test, based on a simulation study.

2. What is the effect of the number of participants?

The impact of the number of participants is large, in particular for Postulated Model 1. With $n = 1000$ participants and $J = 20$ studies we can make precise inference for all parameters except for c_5 . We discuss this in section 6.5. With $n = 10$ and $J = 20$, the

inference is much less precise. In particular for parameters related to the unmeasured confounder Z^B , it is hard to say what the true value of the parameters is based on the graphs. In section 6.5.1, we return to these difficulties. Having $n = 100$ participants gives a marked improvement from $n = 10$, but there is still a probability that extreme values for the parameters related to Z^B are deemed plausible.

For Postulated Model 2, the effect of having more participants is also strong, but not as apparent as for Postulated Model 1. With $n = 10$ participants, it is very difficult to make inference on the variance parameters d_5 , d_8 and d_{10} , and for parameter d_4 . With $n = 100, 1000$ participants, we can make quite exact inference for all parameters except for d_5 , which is discussed in section 6.5.2.

In summary, the effect of the number of participants is large. In particular, inference is greatly improved if we are able to avoid studies with only a few participants.

3. What is the effect of the number of studies?

A greater number of studies lead to more precise inference, but the effect is not as large as the effect of the number of participants. For both Postulated Model 1 and 2, a single study with $n = 1000$ participants lets us make inference which is approximately as precise as the inference we can make with $J = 100$ studies with $n = 100$ participants. This suggests that even though all studies in our simulation are exploring the same research question, the effect of including more studies is different from the effect of pooling participants from different studies.

There are exceptions to the general trend that the number of studies J has a limited impact on inference. This is coefficient d_5 , the variance parameter for Y^N . With $J = 1, 5, 20$, high values of d_5 and extreme values of c_5 are always deemed plausible for all n values. When $J = 100$, however, the pattern changes and inference becomes reasonable. This is because we are performing a double-sided hypothesis test, meaning that we are rejecting the null hypothesis if the coefficient vector is accepted too often. When $J \leq 20$, this cannot happen, but when $J = 100$ it can happen. Of course, the fact that the problem disappears with so many research studies is of little use in practical scenarios.

Taking into consideration the effect of the number of participants and the number of research studies, it seems wise to focus on making inference from fewer studies with many participants. At least, this is true for studies that are investigating similar research questions. In realistic scenarios, studies differ from each other in relevant ways, and it is wise to include many of them because they provide different pieces of information.

4. Is inference related to unmeasured confounders possible?

We can make inference on the parameters related to unmeasured confounders, although the precision of the inference is not as good as for parameters related to measured variables. With $n = 1000$ participants the general level of precision of the inference is so high, that we can make inference on parameters related to the unmeasured confounders Z^B and Z^N . This is clear with $J = 20$ studies in Figures 10.9 and 10.24.

With $n = 10$, the possibilities of making inference are much weaker, in particular for Z^B in Postulated Model 1. As we can see in Figure 10.7 with $n = 10$ participants and $J = 20$ studies, we cannot say much at all about c_6 and c_7 . For c_3 and c_4 , the possibilities of inference are better: we can at least say that the true value is likely positive. With $J = 100$ studies, we can feel confident in this conclusion, as we see in Figure 10.10. It is not surprising that inference of c_3 and c_4 is better, since these parameters directly impact the outcome Y^B whereas c_6 and c_7 have an indirect impact.

With $n = 100$ participants there is a marked improvement from $n = 10$ participants. With $J = 20$ studies and $n = 100$ participants, except for the fact that extreme values for c_4 and c_6 are at times deemed plausible, we can make quite good estimations for the parameters related to Z^B . See Figure 10.8.

In summary, it is possible to make inference from unmeasured confounders. To do so, it is of utmost importance to have studies with many participants, in particular when dealing with binary data.

5. Is inference possible for heterogeneous/homogeneous data?

In Postulated Model 1, by letting c_4 and c_6 have high or low values, we are making the data homogeneous or heterogeneous. If c_6 is high, it means that the observations that had $Z^B = 1$ likely have $X^B = 1$ as well. Again, holding everything else constant, if we give c_4 a high value, then observations that had $X^B = Z^B = 1$ almost certainly have $Y = 1$. When both c_4 and c_6 are high, it means we have a group where if $Z^B = 1$, then both X^B and Y^B tend to be 1. To investigate the effect of heterogeneity, we have used a modified Mother Nature's Model 1, with $c_4 = c_6 = 3$. We performed simulations with $J = 5$ research studies and $n = 10, 100, 1000$ participants. The results can be found in Figures 10.13-10.15.

The values of all parameters are more difficult to estimate with this setup compared to when $c_4 = c_6 = 1$. In particular, c_4 and c_6 are difficult to estimate. Even with $n = 1000$ participants, in most simulations we can only gather that the c_4 and c_6 are non-negative.

This is not surprising given that with this setup, it is hard to separate the effect of X^B and Z^B , since if $Z^B = 1$ then almost certainly $X^B = 1$. Moreover, with $c_4 = c_6 = 3$ there are very few participants for which $X^B = 1$ but $Y^B = 0$. This leads to an unstable estimator of the log odds ratio, for reasons described in section 6.5.1.

6.5 Difficulties in the simulations

Problems encountered in Simulation Study 1 have been mentioned above. Here we describe them more in-depth.

6.5.1 Postulated Model 1

We have seen that with Postulated Model 1, there are difficulties in making inference on parameter c_5 . It is also generally difficult to make inference on the model when $n = 10$. Moreover, in a situation when the true values of c_4 and c_6 are 3, inference was more difficult, in particular for c_4 and c_6 .

This is not surprising in light of properties of the odds ratio statistic. When X^B and Y^B are binary variables, the data can be presented in a 2×2 table as

	$X = 1$	$X = 0$
$Y = 1$	n_{11}	n_{10}
$Y = 0$	n_{01}	n_{00}

When performing logistic regression, the log odds ratio is estimated using the maximum likelihood method. The estimate follows an approximate normal distribution with variance $1/n_{11} + 1/n_{10} + 1/n_{01} + 1/n_{00}$. (Agresti 2014, p 70; Held & Bov 2014, p 97-98)

However, the normal distribution breaks down when the number of observations is small, or when there are only a few observations per cell in the 2×2 table; less than 5 observations per cell is the traditional rule of thumb. The estimates of the intercept and slope in the logistic regression model may then start to oscillate between high and low values, and it becomes difficult to detect the statistics reported in the research studies as extreme relative to the simulated statistics.

This is the root of the problems that occur. When c_5 takes on very high or low values, very few or very many participants may have $X^B = 1$. When $c_4 = c_6 = 3$, participants with $X^B = 1$ and $Y^B = 0$ may be practically non-existent. When $n = 10$, we have very few observations in general, making the log odds ratio statistic unreliable, in particular when c_6 and c_7 take on extreme values so that almost all or almost no participant has $X^B = 1$.

More than a limitation of the HS method, however, this is a known limitation of odds ratios and of the logistic regression model that is often used to model odds ratios. These considerations show that users of the HS method have to be careful when analyzing binary data with few observations; when probabilities modeled are very large or small; and when the data contains only a few observations from one of the categories.

We suspected that the problems are strengthened by the copula estimator. Since this estimator ranks all input values and performs a Gaussian transform, it would seem similar to a situation where the simulated values are oscillating between high and low, the statistics from the research studies are given middle ranks, and hence be deemed plausible. Then a kernel density estimator may be wiser to use. While this is an issue to have in mind, the problems were not solved by implementing a kernel density estimator instead of a copula estimator in this particular case.

6.5.2 Postulated Model 2

For Postulated Model 2, we encounter problems estimating d_4 with $n = 10$ participants, and d_5 with any number of participants. Both problems are related to the variance of Y^N . Extreme values on d_4 yield oscillating Y^N values, and with $n = 10$ this leads to oscillating intercepts and slopes from the linear regression model. Similarly, since d_5 is the variance of Y^N , high d_5 values leads to a wide range of Y^N values, resulting in a similarly varying collection of simulated intercepts and slopes. This means that the intercept and slope from the real studies are not extreme relative to the simulated statistics, even though the parameter values are completely wrong.

As we noted before, the problem disappears with $J = 100$, due to the Coull-Agresti test. In practical situations, this is not very helpful as we seldom have that number of identical studies. Regarding d_5 , this problem can be solved easily if we use information regarding the variance of Y^N . If we let the research studies report this and we also recreate this statistic in our simulations, inference on d_5 is quite precise.

When applying the HS method to real data, we may have to work in this manner: first try to make inference based on available data; and when inference is too imprecise, we have to look for more data or request more research.

7 Simulation Study II

In the second simulation study, our goal is to answer the following questions:

- i. **Can we make inference based on studies that investigate different research questions and report different statistics, as long as the investigations are related to the Postulated Model?**

It is vital for the method that this is possible. If we are unable to make inference based on studies that are reporting on the association between different variables using different statistics, the HS method has a much more limited use than if it is possible.

- ii. **Can we make inference with an overparameterized Postulated Model?**

We claimed in section 3 that if the Postulated Model contains all important variables from the Mother Nature's Model, then the estimates of the parameters related to those variables should not be biased because we also included variables that are not part of the Mother Nature's Model. Rather, we should be able to see that the unnecessary parameters have little or no impact on the outcome. This has to be confirmed.

- iii. **Can we make inference when a study reports results from statistical analyses which are flawed?**

If our Postulated Model can generate the same data as the true data-generating model, then it should not matter for the HS method if the statistical method that is used to analyze the data is flawed, for instance in the sense that it gives biased estimates. As long as we follow the same procedure, our results should be the same if we have a plausible Postulated Model.

- iv. **Can we make inference when the variables in the Postulated Model are following different distributions (binary, continuous, categorical)?**

In realistic scenarios, different variables follow different distributions, so the HS method must be able to handle this.

7.1 The Mother Nature's Model 3

In order to answer the questions, we use the following Mother Nature's Model 3:

$$Z^W \sim Be(g(0))$$

$$W^W \sim Exp(1)$$

$$X^W \sim Be(-1 + Z^W + 0.5W^W)$$

$$\{Q_1, Q_2, Q_3\} \sim Multi(1/3, 1/3)$$

$$Y^W \sim Be(-4.5 + X^W + Z^W + X^W Z^W + 0.5W^W + X^W W^W + 2Q_2 - Q_3)$$

The Postulated Model 3 is the same as the Mother Nature's Model 3, except that it has an additional variable W_2^W :

$$Z^W \sim Be(e_1)$$

$$W^W \sim Exp(e_2)$$

$$W_2^W \sim N(mean = e_3, sd = e_4)$$

$$X^W \sim Be(g[e_5 + e_6 Z^W + e_7 W^W])$$

$$\{Q_1, Q_2, Q_3\} \sim Multi(e_8, e_9)$$

$$Y^W \sim Be(g[e_{10} + e_{11} X^W + e_{12} Z^W + e_{13} X^W Z^W + e_{14} W^W + e_{15} X^W W^W + e_{16} W_2^W + e_{17} Q_2 + e_{18} Q_3])$$

7.2 The simulation algorithm

There are 5 studies reporting on variables related to Postulated Model 3. Each study has 1000 participants each. However, all the studies investigate different research questions and they are reporting different measures of association between different variables in the Mother Nature's Model 3.

Study 1: Intercept and slope in a logistic regression model of Y^W given X^W .

Study 2: Intercept and slope in a logistic regression model of Y^W given W^W .

Study 3: Intercept and slope in a logistic regression model of X^W given W^W .

Study 4: Mean values of X^W and W^W .

Study 5: Coefficients from a linear regression of Y^W given the $\{Q_1, Q_2, Q_3\}$ variables. However, the authors behind this study treated $\{Q_1, Q_2, Q_3\}$ not as three different binary variables. Instead they created a new variable Q such that $Q = 1$ if $Q_1 = 1$; $Q = 2$ if $Q_2 = 1$; and $Q = 3$ if $Q_3 = 1$. Such an analysis is not sensible, since Y is binary, so performing a linear regression in the sense described above should give a biased estimate.

The simulations were performed according to the algorithm described in section 5. The only difference is that this time every study reported a unique statistic, and the same statistic was calculated when performing simulations to recreate the same study. Also, we only created $J = 5$ research studies with $n = 1000$ participants.

7.3 Computational details of the inferential procedure

Just as in Simulation Study I, all the summary statistics are continuous and can be expected to follow a normal distribution – visual inspection confirmed this. Hence, the plausibility $\pi(\cdot)$ can be modeled with a Gaussian copula estimator with Pearson's r as a correlation coefficient. A kernel density estimator was also implemented, with very similar results. Our preference for the copula estimator is, again, due to its computational efficiency.

7.4 Simulations results

Results from the estimation of all parameters can be found in Figures 10.28 and 10.29.

i. Can we make inference based on studies that investigate different questions and report different association measures?

We can make precise inference on all parameters in the model. In particular, we can make inference on the unmeasured confounder Z^W . Compared with Simulation Study I, with $n = 1000$ participants and $J = 5$ studies, the intervals of plausible parameter values are broader. This could be because the Postulated Model is more comprehensive and also because the studies are reporting different association measures.

ii. Can we make inference with an overparameterized model?

The value of coefficient e_{16} , describing the association between Y^W and W_2^W , which is deemed plausible most often is 0. For the other variables related to W_2^W , namely e_3 and e_4 , all values are deemed plausible $100(1 - \alpha)\%$ of the time. This is what we expect to see, since W_2^W does not affect on any variable in the true data-generating model.

iii. Can we make inference when a study reports results from a statistical analysis that is flawed?

Since we can make inference on all parameters in the model even though study 5 is included in the analysis, the answer is yes. In particular, we are able to make inference on parameters e_8 , e_9 , e_{17} and e_{18} , which are all related to $\{Q_1, Q_2, Q_3\}$. The only study in which these variables are reported is in the flawed study 5, so we can make inference based on the flawed study.

iv. Can we make inference when the variables are following different types of distributions (binary, continuous, categorical)?

Since our analysis was successful in all other senses, it is clearly possible to make inference in a situation where the different variables are following different distributions.

8 An analysis on the association between Paraquat exposure and Parkinson's Disease

In this section, we apply the HS method to data regarding the association between exposure to the pesticide Paraquat and Parkinson's Disease. While the analysis is still a work in progress, it illustrates how the HS method can be applied to real data.

Parkinson's Disease (PD) is a neurological illness that affects approximately 1% of the US population. Almost all PD patients are over 50 years of age, and men are more likely than women to contract the disease. It is also believed that exposure to pesticides can increase the risk of contracting PD. However, the evidence is mixed and comes mostly from epidemiological studies (Tanner 2009).

Paraquat dichloride (hereafter Paraquat) is a herbicide that is commonly used in farming to control weeds. Handling Paraquat can be very dangerous and potentially lethal. Therefore, Paraquat products can only be used by certified applicators in the USA. Other countries, including the EU, have forbidden Paraquat use altogether (EPA 2021).

We investigated a meta-analysis of studies estimating the association between exposure to Paraquat and risk of PD by Ntzani et. al. (2013), and set out to perform an analysis of the same data, using the HS method. This decision was made after a suggestion from the Environmental Protection Agency (EPA) in the USA, which is a collaborator in the development of the HS method.

In the next section, we briefly describe the meta-analysis by Ntzani et. al. After that, we describe how the analysis was performed using the HS method. We then discuss some difficulties that we encountered during the analysis.

8.1 The meta-analysis by Ntzani et. al.

The meta-analysis by Ntzani et. al. is based on eight research studies estimating the association between exposure to Paraquat and PD. The studies have many similarities: all report the association in terms of an odds ratio; all used participants from the USA; and most studies had study participants with similar characteristics in terms of age, sex and other relevant variables. Seven studies are case-control studies, the remaining one is a cohort study.¹

Three studies are based on the same participant data, namely Gatto (2009), Wang (2011) and Costello (2008). All three studies use data from the same three counties in California, but measure the association between different types of Paraquat exposure and PD. The obvious dependence between the results from these studies is not taken into account by Ntzani et. al.

Two studies are based on data from the Agricultural Health Study cohort. Kamel (2006) uses data from the whole cohort, whereas Tanner (2011) performs a case-control study based on a sample from the cohort.

¹The studies are Gatto (2009), Wang (2011), Costello (2008), Kamel (2006), Tanner (2011), Firestone (2010), Dhillon (2008) and Tanner (2009).

While the studies share similarities, there are also potentially important differences between them, for example:

- The definition of what it means to be “exposed to Paraquat” differs between the studies. For instance, in Dhillon (2008), exposure is to have “personally applied or mixed” Paraquat at least once; in Costello (2008), a person is considered exposed if they live within a particular distance of a site where Paraquat is used.
- All studies except one uses PD as the health outcome, whereas Tanner (2009) uses Parkinsonism as the outcome. Parkinsonism is a diagnosis with slightly different symptoms and is more common than PD.
- The reported odds ratios provide different information, since they are adjusted for different variables. For example, Firestone (2010) adjusts for smoking, sex, age and ethnicity whereas Dhillon (2008) adjusts for no other variable.

Despite the differences between the studies, Ntzani et. al. use a Fixed-Effect model to summarize the findings of the studies. This means that all studies are considered to estimate the same intervention effect, and differences are merely due to sampling error.

As we can see in Figure 3, the only information used from each study is an estimate of the odds ratio (with corresponding confidence interval) describing the association between Paraquat exposure and PD. The exception is Costello (2008) from which two odds ratio estimates were used: one odds ratio measures the association between exposure to Paraquat and PD, and the other odds ratio measures the association between exposure to both Paraquat and the pesticide Rotenone, and PD. In several of the studies, more than one statistic measuring the effect of Paraquat on PD is reported; why certain statistics were selected in favor of others is unknown to us.

8.2 Plan for performing an analysis

In order to perform an analysis of the eight research studies with the HS method, we pursued the following steps:

1. Construct a Postulated Model, $f_P(w|\theta)$.
2. Perform simulations to recreate the reported summary statistics, using different possible values of θ .
3. Once a plausible θ is found, use the Postulated Model to calculate association measures of interest.

In the next subsections, we describe each of these steps.

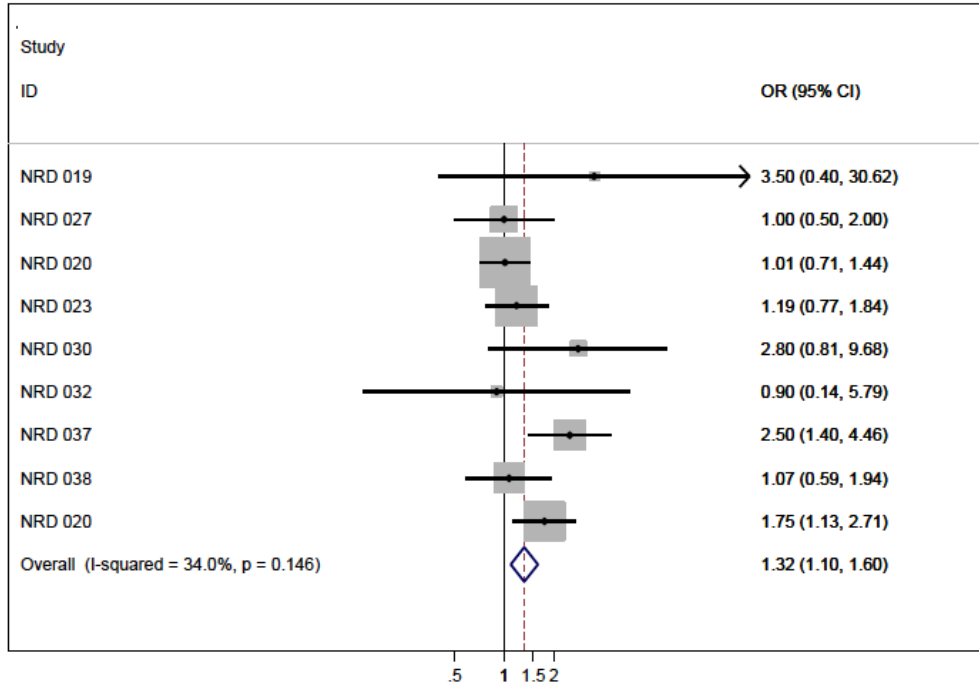


Figure 3: A forest plot of the reported odds ratios, from the meta-analysis by Ntzani et al. The leftmost column contains the id numbers of the eight studies, and the rightmost column has point estimates of the odds ratio, with a 95% confidence interval.

8.3 Step 1. Construct a Postulated Model

As previously explained, a Postulated Model should be constructed based on scientific expertise. Due to time limitations, we were not able to construct our Postulated Model of Parkinson's Disease this way. Instead we constructed a Postulated Model based on layman's knowledge and common sense.

When deciding what variables to include, we paid attention to the information provided by the research studies. Fortunately, all eight research studies had tables with summary statistics of the participants' distribution of age, sex, smoking habits, education level, and many other variables relevant for the PD risk. We limited ourselves to the following initial model:

Postulated Model A

$$State \sim Multinom(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$$

$$Age \sim Normal(Mean = \theta_6 + \theta_7^{State}, Var = \theta_8)$$

$$Sex \sim Be(g[\theta_{20}])$$

$$Smoker \sim Be(g[\theta_9])$$

$$Farmer \sim Be(g[\theta_{10}])$$

$$Paraquat^b \sim Be(g[\theta_{11} + \theta_{12}Age + \theta_{13}Sex + \theta_{14}Farmer + \theta_{15}^{State}])$$

If $Paraquat^b = 1$:

$$Paraquat^c \sim Exp(\theta_{16} + \theta_{17}Age + \theta_{18}Sex + \theta_{19}Farmer + \theta_{20}^{State})$$

If $Paraquat^b = 0$:

$$Paraquat^c = 0$$

$$Parkinson \sim Be(g[\theta_{21} + \theta_{22}Age + \theta_{23}Sex + \theta_{24}Farmer + \theta_{25}Paraquat^b])$$

We used the variable *State* because Paraquat may be used to different extents in different states and the study participants were sampled from different states; we used *Age*, *Sex* and *Smoker* (ever/never) because these variables are strongly correlated with PD. The variable *Farmer* is a latent variable and the name is arbitrary.

All eight studies report on all the variables used in the model, except for the latent variable *Farmer*.

Paraquat exposure is measured with two variables. $Paraquat^b$ is binary and determines whether a person has been exposed at all to Paraquat. $Paraquat^c$ is continuous and determines how much Paraquat exposure a person had. It is the binary variable that is associated with the risk of PD. The continuous variable is used to determine if a person has been exposed to Paraquat by the definition of Paraquat exposure used in the different studies. This is detailed further in Section 8.4.

This is the biggest flaw in Postulated Model A. If Paraquat is associated with the risk of PD, then more Paraquat exposure should be associated with an increase or decrease in the PD risk. By that reasoning, it would be more reasonable to let $Paraquat^c$ be associated with the PD risk, instead of $Paraquat^b$. However, we chose to treat Paraquat exposure as a binary variable in order to make our analysis easier to compare with the study by Ntzani et. al., since it treats Paraquat exposure as binary.

We return to how this and the many other flaws of Postulated Model A can be remedied at the end of this section.

8.4 Step 2. Recreate the summary statistics and find a plausible θ

To recreate the summary statistics reported in a study, we followed the experimental design described in each of the studies. Since most studies were case-control studies, the procedures shared many similarities. Here is an outline of how the simulations were performed for study j and a particular θ :

1. For $i \in \{1, \dots, 1000\}$
 - a) Generate a large number of individuals living in the state where participants in study j were sampled from.
 - b) Randomly select the n_j PD cases and m_j matched controls that participated in study j .

- c) Based on the n_j cases and m_j controls, calculate the number of cases and controls exposed to Paraquat, the mean age, the proportion of males and the proportion of those who have ever smoked. Save the vector of statistics as s_i^* .
- 2. Use $\{s_i^*\}, i = 1, \dots, 1000$ to create $\pi_j(\theta)$.
- 3. Evaluate $\pi_j(\theta) > \alpha$.

This procedure was repeated for all seven case-control studies. The remaining study was a cohort study that reported the age distribution in a slightly different manner. The algorithm for recreating this study was very similar to the above algorithm, except that no sampling took place and the statistics calculated were the number of people exposed to Paraquat among cases and controls; the proportion of participants below age 50; the proportion of smokers; and the proportion of males.

Whether a simulated person is considered exposed to Paraquat in step 1(c), is determined by their value on the *Paraquat^c* variable. If this variable has a higher value than the threshold value, the person is considered exposed. The threshold values were set to reflect the fact that different studies had different definitions of Paraquat exposure. The threshold values should ideally be set by scientific experts, and our decisions were likely flawed.

The method for finding a plausible θ was to make guesses based on common sense and trial and error. It was possible to make educated guesses for almost all parameters. For instance, it makes sense that the mean age for the population in question is around 70; that the impact of age on PD risk is positive; and that tobacco smoking has a slightly negative association with the PD risk. It was also helpful to plot graphs of the kernel density estimator and to print quantiles from the marginal distribution functions of the copula estimator. Clearly, the process would be smoother with the help of scientific expertise and a numerical method for exploring the relevant θ .

A full investigation of $\mathcal{R}(\alpha)$ was not performed, due to time limitations. Instead, we paid particular attention to the parameter θ_{20} , which describes the association between exposure to Paraquat and the risk of PD in terms of a log odds ratio. We focused on how much θ_{20} could vary while keeping the rest of θ fixed.

8.4.1 Computational details of the inferential procedure

The association between Paraquat exposure and PD is presented with an odds ratio in all eight studies. All studies also present the number of cases and controls that have been exposed to Paraquat. In most cases, the number of exposed was rather small – in several cases less than 5 exposed among the cases and/or controls. We saw in Simulation Study I that this yields unstable estimators of the odds ratio. For this reason, we chose not to recreate the odds ratio in our simulations; instead the numbers of exposed cases and controls were recreated. This means that we are dealing with discrete data, and as we saw in section 4, it is unclear whether several of the features of copulas and kernel density estimators apply to discrete data. How this affected the performance of the inference has not been further explored.

To approximate $\pi(\theta)$, both kernel density estimator and Gaussian copula estimator were tested, with equivalent results. The results presented in the thesis are based on the Gaussian copula estimator, a choice motivated by its computational efficiency.

8.4.2 The effect of adding new statistics and studies

We started by replicating two statistics from a single study: namely the number of exposed PD cases and the number of exposed controls in Dhillon (2008). We then tried to add more studies and more statistics, related to age, sex and smoking.

This had the following impact on the Postulated Model:

1. Adding new studies meant that the interval of values that the coefficients in our model could take was reduced. In particular, when only including Dhillon's study in the analysis, keeping all other parameters fixed, θ_{20} could vary in the interval $[-2, 4]$. This means that the Postulated Model is consistent with both a strongly negative and strongly positive association between Paraquat exposure and PD. With all eight studies included in the analysis, the interval of plausible values for θ_{20} had shrunk to $[-0.1; 0.3]$. An important factor in how the width of the plausible interval changed was the number of participants in a study. More participants meant a narrower interval. These results confirm what we saw in Simulation Study I regarding the number of studies and the number of participants. In particular, Dhillon (2008) uses rather few participants, 184 in total. Only 4 cases and 1 control had been exposed to Paraquat, making it very hard to make precise inference. This partly explains why the plausible interval was so wide with only this study in the analysis.
2. Adding new statistics meant that we had to make changes to the model. As an example, when we only recreated statistics related to Paraquat exposure, all states had the same mean age and proportion of males. When we added statistics of age distribution and proportion of males to the analysis, we had to change the parameters so that different states had different mean ages and proportions of males. This illustrates that by adding new types of information to our analysis, we can make more precise information about the parameters in our Postulated Model.

8.5 Step 3: Calculating a measure of association

Once we had a plausible θ , we used the Postulated Model A with θ to calculate a measure of association. There are many interesting measures that we could calculate. For instance, we could calculate the PD risk conditional on Paraquat exposure and use this to give an odds ratio, risk ratio or risk difference. We could make predictions for the numbers of PD cases in a certain region and time period. These possibilities would certainly not have been open to us if all we had was an estimate of an odds ratio, which would have been the case if we had performed a standard meta-analysis.

The measure we chose to calculate was the prevalence of PD in the whole population at various age intervals, and conditional on having been exposed or not exposed to

Paraquat. The figures are plotted in Figure 4.

The value of θ_{20} used to create this graph is 0.2, which was picked simply for the reason that it is in the interval $[-0.1, 0.3]$. The other parameters in θ could likely also have taken other values, with θ remaining plausible. Had we been able to explore the full $\mathcal{R}(\alpha)$, we would have been able to calculate an interval of plausible PD prevalence. In a real application of the HS method, this is what we would like to do.

Along with the PD prevalence estimated from our Postulated Model, we also plotted the observed PD prevalence in a research study by Tanner (2010). As we can see, the PD prevalence according to our Postulated Model is much higher than the observed prevalence. The fact that our Postulated Model is consistent with a PD prevalence so far from the observed prevalence, indicates that our Postulated Model is flawed. This should not come as a surprise. We constructed the model without consulting scientific expertise and in the inferential procedure, we did not recreate statistics of PD prevalence. Our Postulated Model must be improved, and one way of doing this is to make sure it is consistent with observed data regarding PD prevalence.

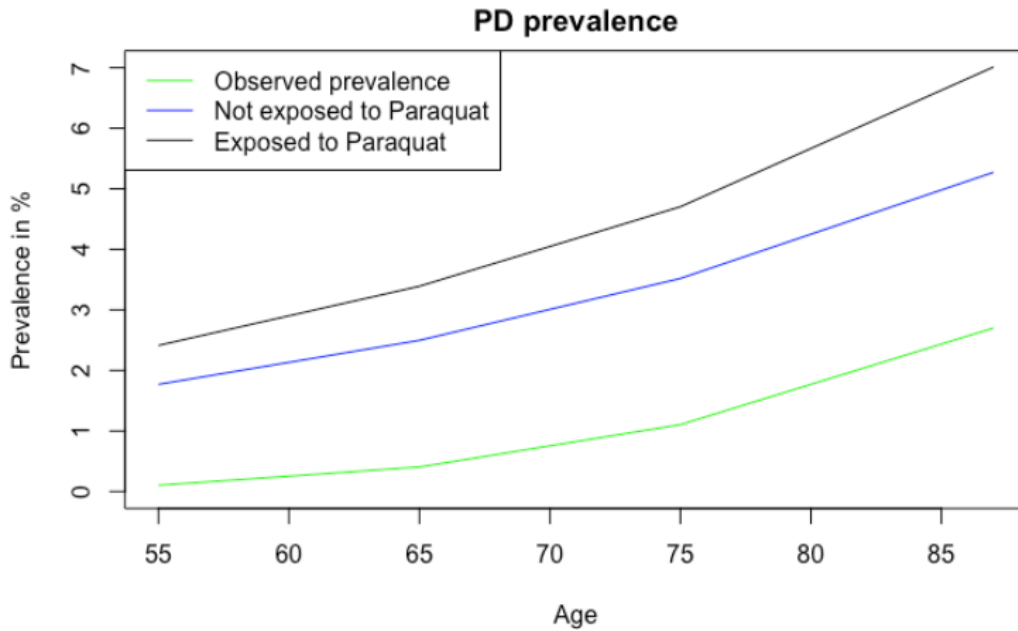


Figure 4: Graph showing the prevalence of PD according to Postulated Model A, for people exposed and not exposed to Paraquat. The θ used to generate the graphs was deemed plausible based on eight research studies, with $\theta_{20} = 0.2$. As a comparison, we also plot the observed prevalence of PD in a research study by Tanner (2010).

8.6 Adding a study of PD prevalence

To improve Postulated Model A, we included the results of the study by Tanner (2010) in the analysis. Tanner (2010) reports the PD prevalence among men and women across several age intervals in four counties in California – the same counties from which Gatto (2009), Wang (2011) and Costello (2008) sampled participants.

The result of the Tanner (2010) study was recreated by using our Postulated Model to generate the same number of people with the relevant age interval residing in the four counties, and calculating the proportions of PD cases. This was repeated 1000 times and a copula estimator was approximated. Of course, the Postulated Model still had to be consistent with the original eight studies.

To recreate the observed PD prevalence, we lowered the coefficient θ_{16} , describing the baseline risk of PD. We were also forced to model the effect of age on the PD risk with a linear spline. Whether a linear spline is the right choice is something that should be decided by consulting scientific expertise. A natural cubic spline would probably work better. This resulted in a new Postulated Model:

Postulated Model B

$$State \sim Multinom(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$$

$$Age \sim Normal(Mean = \theta_6 + \theta_7^{State}, Var = \theta_8)$$

$$Sex \sim Be(g[\theta_{20}])$$

$$Smoker \sim Be(g[\theta_9])$$

$$Farmer \sim Be(g[\theta_{10}])$$

$$Paraquat^b \sim Be(g[\theta_{11} + \theta_{12}Age + \theta_{13}Sex + \theta_{14}Farmer + \theta_{15}^{State}])$$

$$\text{If } Paraquat^b = 1:$$

$$Paraquat^c \sim Exp(\theta_{16} + \theta_{17}Age + \theta_{18}Sex + \theta_{19}Farmer + \theta_{20}^{State})$$

$$\text{If } Paraquat^b = 0:$$

$$Paraquat^c = 0$$

$$Parkinson \sim Be(g[\theta_{21} + \theta_{22}(Age - \theta_{23}) + \theta_{24}(Age - \theta_{25})_+ + \theta_{26}(Age - \theta_{27})_+ + \theta_{28}Sex + \theta_{29}Farmer + \theta_{30}Paraquat^b])$$

Here, $(a)_+ = \max(a, 0)$.

We used Postulated Model B to estimate the PD prevalence again. The result is shown in Figure 5, and as we can see, our model is now much more consistent with the observed data.

This is an illustration of how an analysis should proceed with the HS method: we should make sure that our Postulated Model is consistent with all relevant data, and continually make improvements so that this is the case.

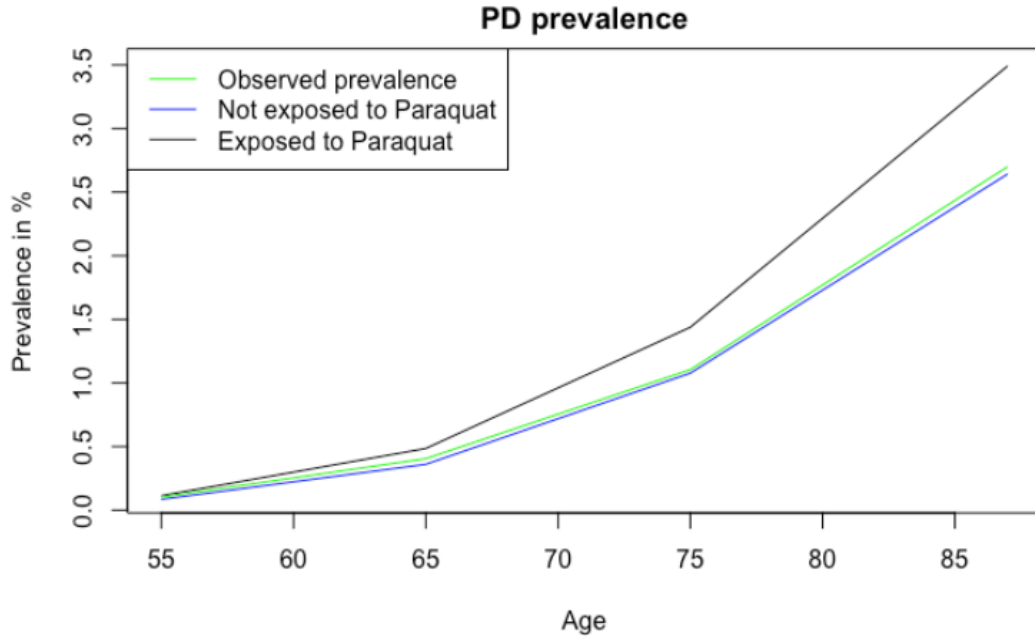


Figure 5: Graph showing the prevalence of PD according to our updated Postulated Model B. The θ used to generate the graphs was deemed plausible based on nine research studies, with $\theta_{20} = 0.2$. As a comparison, we also plot the observed prevalence of PD in a research study by Tanner (2010).

8.7 Comparison with the Ntzani et. al. study

In the meta-analysis by Ntzani et. al., the point estimate of the odds ratio is 1.32 with a 95% confidence interval [1.1; 1.6]. A naive “plausible interval” for the parameter θ_{20} is $[-0.1; 0.3]$, which can be interpreted as an interval of plausible odds ratios [0.91; 1.35]. This interval is naive because it ignores how the other parameters in θ can vary.

A direct comparison of the confidence interval of the Ntzani et. al. study and the “plausible interval” based on the HS method is difficult for several reasons. First of all, the Ntzani et. al. study calculates a 95% confidence interval whereas we calculated a 90% plausible interval, which has a different interpretation. Secondly, θ_{20} is the log odds ratio adjusted for age, sex, smoking and an unknown variable that we call Farmer. The odds ratio that Ntzani et. al. report is an average of odds ratios that are individually adjusted for different variables.

Nevertheless, the point estimate that Ntzani et. al. report is within the plausible interval of our study, suggesting that the conclusions of the two analyses are not completely inconsistent.

8.8 Some suggested improvements

The analysis of the association between Paraquat and PD is far from finished. Here we list some of the improvements that should be made before finalizing the analysis. A general point regarding these suggestions is that the eight research studies included in the Ntzani et. al. study provides us with a lot of highly relevant information about the participants that is related to our Postulated Model. With the HS method we can use all of this information, which would have been impossible with a standard meta-analytical method.

1. Paraquat exposure on a continuous scale

As discussed, it makes sense to measure Paraquat exposure and its effect on the PD risk on a continuous scale. The studies included in our analysis reports different types of Paraquat exposure, and it would be advantageous to use this information. This requires help from scientific experts, both to understand how Paraquat exposure should be measured and how different amounts of Paraquat exposure may impact the risk of PD.

2. The relationship between Paraquat and other pesticides

Most studies report on the association between PD and other pesticides besides Paraquat. It would be useful to include such pesticides in the model. This way, we may find out how harmful Paraquat is relative to other pesticides; if there are interaction-effects; and if the apparent effect of Paraquat is due to confounding from other pesticides.

3. More comprehensive PD definition

The PD diagnosis is given when several symptoms show. Several studies report on each of these symptoms. We could incorporate this into our model, by adding the symptoms as variables. This may help us discover new patterns. For instance, Paraquat exposure may be related specifically to one of the symptoms but not to the others.

4. Improvement of Copula estimator

We are using summary statistics which are both discrete and continuous, while using a Gaussian copula. We have not explored what effect this has on our estimation.

5. Including different studies

In Simulation Study I, we saw that using many studies that are investigating the same question has little benefit. The studies included in the Ntzani et. al. meta-analysis are similar to each other. Therefore, it would make sense to look for studies investigating different questions and include them in our analysis, just as we did with the Tanner (2010) study to make the Postulated Model more reasonable.

6. Account for biases

Epidemiological studies are prone to biases that we may model in our simulations.

For instance, all studies have non-responses that may bias the results, and most studies rely on the information given by the study participants, meaning that there is a risk of self-report bias.

9 Conclusions

In this thesis, we introduce a new method for making inferences based on data from heterogeneous sources. We have given intuitive reasons to believe that the method can be useful and we have shown that the method works in two simulation studies. While the method can potentially be used in many different contexts, we limit our attention to situations where we currently would use meta-analytical methods.

The conclusion that we draw from this thesis is that more research is warranted to further develop the HS method. The topics of future investigations can broadly be divided into two categories: theoretical and practical. Below we describe some of the most pertinent questions.

Theoretical investigations

In sections 3 and 4, we mentioned some topics that require further development. We repeat them here:

1. **What is the sample?**

It is unclear whether the data should be considered a sample from a population (and in that case what sort of population) or if we should simply consider it as facts about the world which our Postulated Model can either be consistent or inconsistent with.

2. **Inference based on summary statistics**

To the extent that we only include summary statistics in a study, there is the question of how we can make inference based on such statistics most efficiently. In particular, how reliable is the inference when the statistics are not sufficient?

3. **Bayesianism or frequentism?**

In light of the answers to the above two questions, we can choose to set the HS method in a frequentist or Bayesian framework.

4. **Is likelihood-based inference possible?**

In light of the answers to the previous questions, we may attempt to develop a likelihood-based method of inference. This could be superior to the method we have used in this thesis, as we would be able to point to one value of the parameter θ as the most plausible one.

Practical investigations

1. **Efficient calculation of $\mathcal{R}(\alpha)$**

The way that the Plausible Region $\mathcal{R}(\alpha)$ is currently calculated is slow and inconvenient; an algorithm that does this efficiently would be needed.

2. **Use of scientific expertise**

As pointed out, the design of any Postulated Model should be performed with

subject-matter experts. An interesting question is how this may work in practice: there is the possibility that the cooperation may be more difficult than we imagine and that the fruits of the cooperation may be more valuable than we imagine. For instance, subject-matter experts may be able to nail down a much smaller interval that a parameter value could reasonably take than we can on our own. But the opposite could also be true.

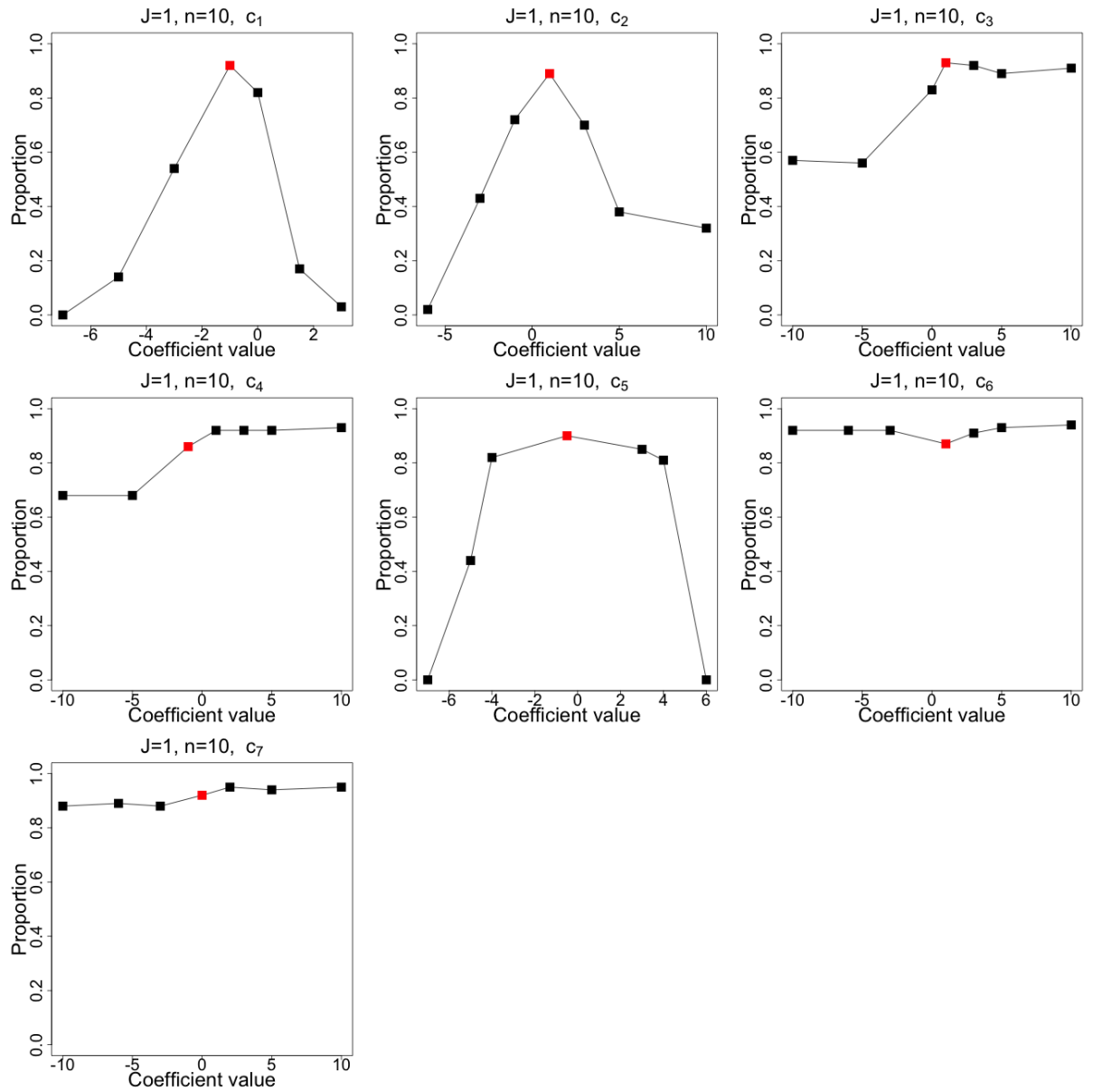
3. **Suitable area of application**

What type of deficiencies in the HS method are acceptable largely depends on the application. For instance, the method is much more time-consuming to apply than traditional methods of meta-analysis, meaning that we most likely only want to apply it to questions of great scientific value. Similarly, we have seen that in some scenarios, the interval of plausible parameter values can be wide. This is a big problem if we need a very precise measure of an intervention effect; it is less of a problem if all we need to know is if the effect is outside of a particular interval, e.g. if the effect is positive and not negative.

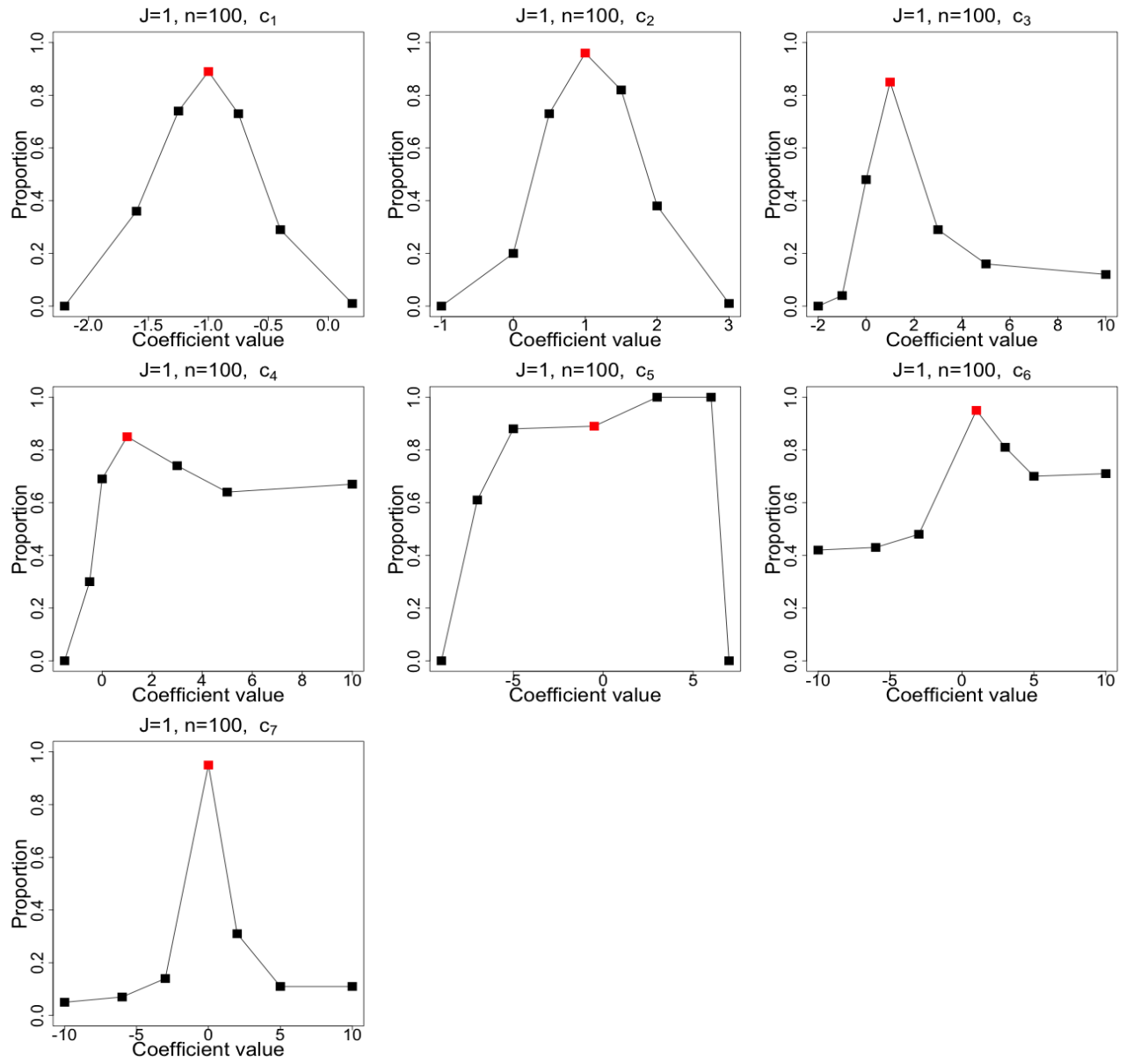
10 Graphs

The following pages show graphs from Simulation Study 1 and 2. In each graph, the x-axis shows the parameter values tested and the y-axis shows the proportion of the simulations that a parameter value was deemed plausible. The red dot in each graph represents the true parameter value, from the Mother Nature's Model.

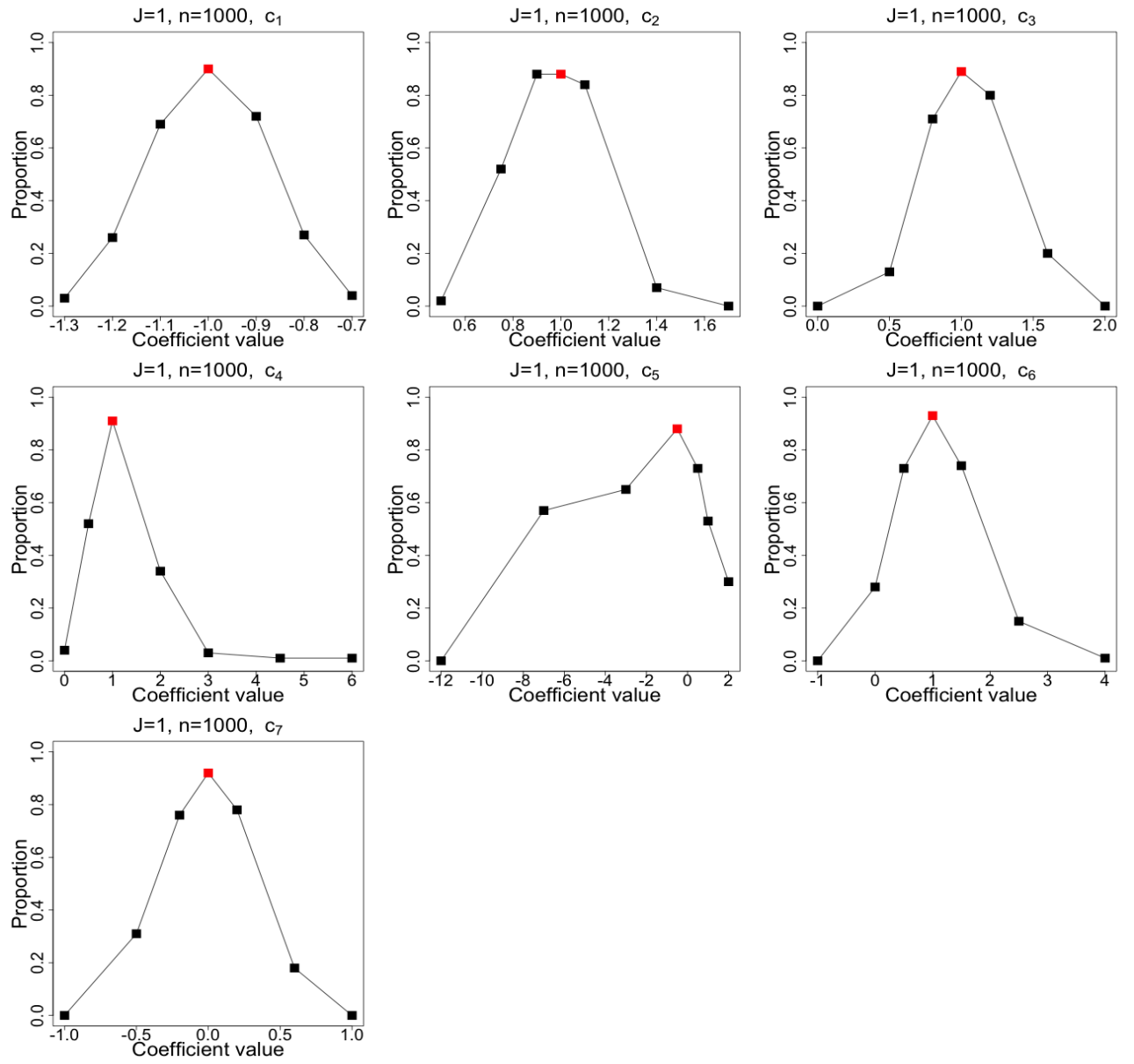
10.1 Postulated Model 1, $J=1$, $n=10$



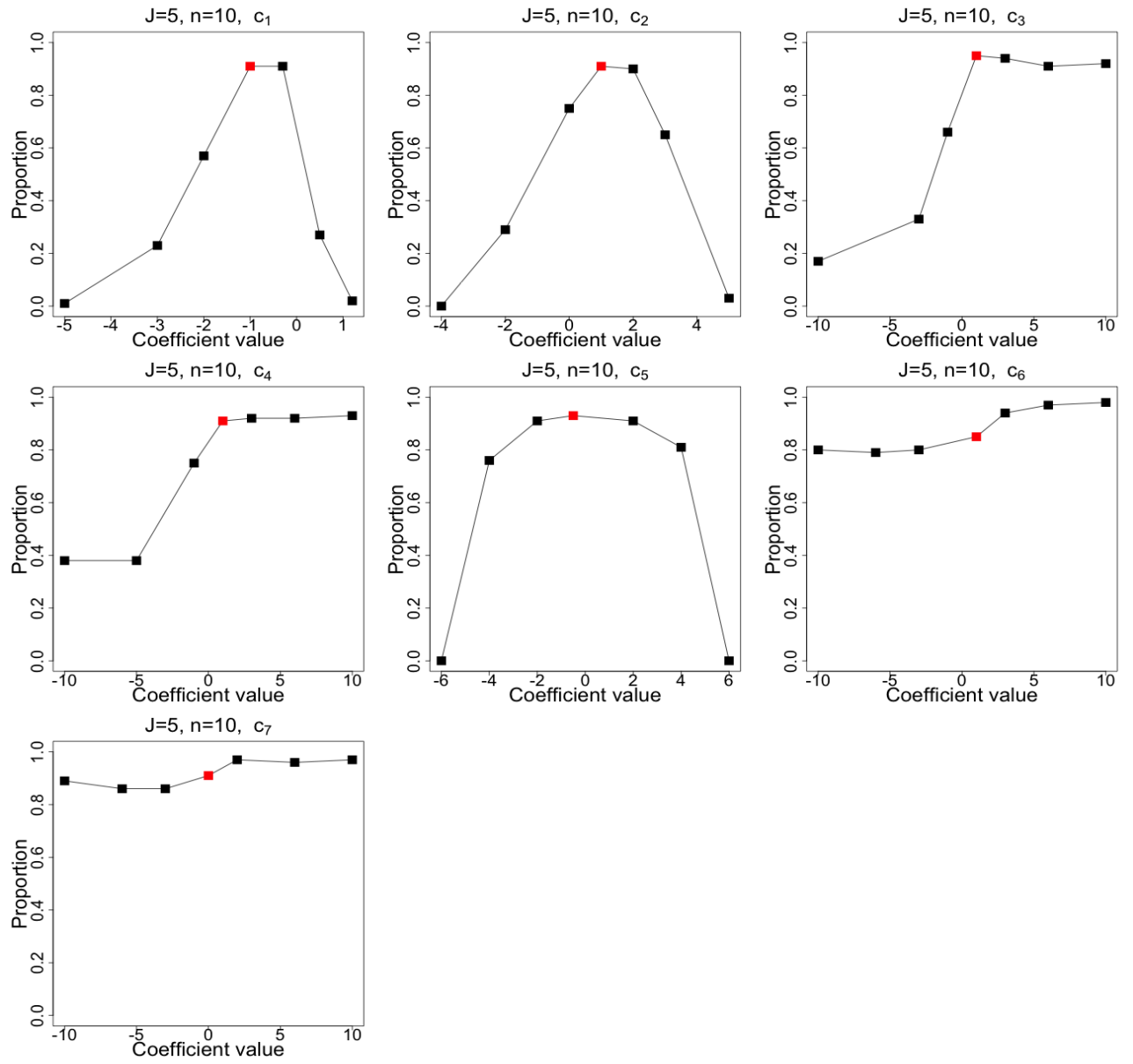
10.2 Postulated Model 1, $J=1$, $n=100$



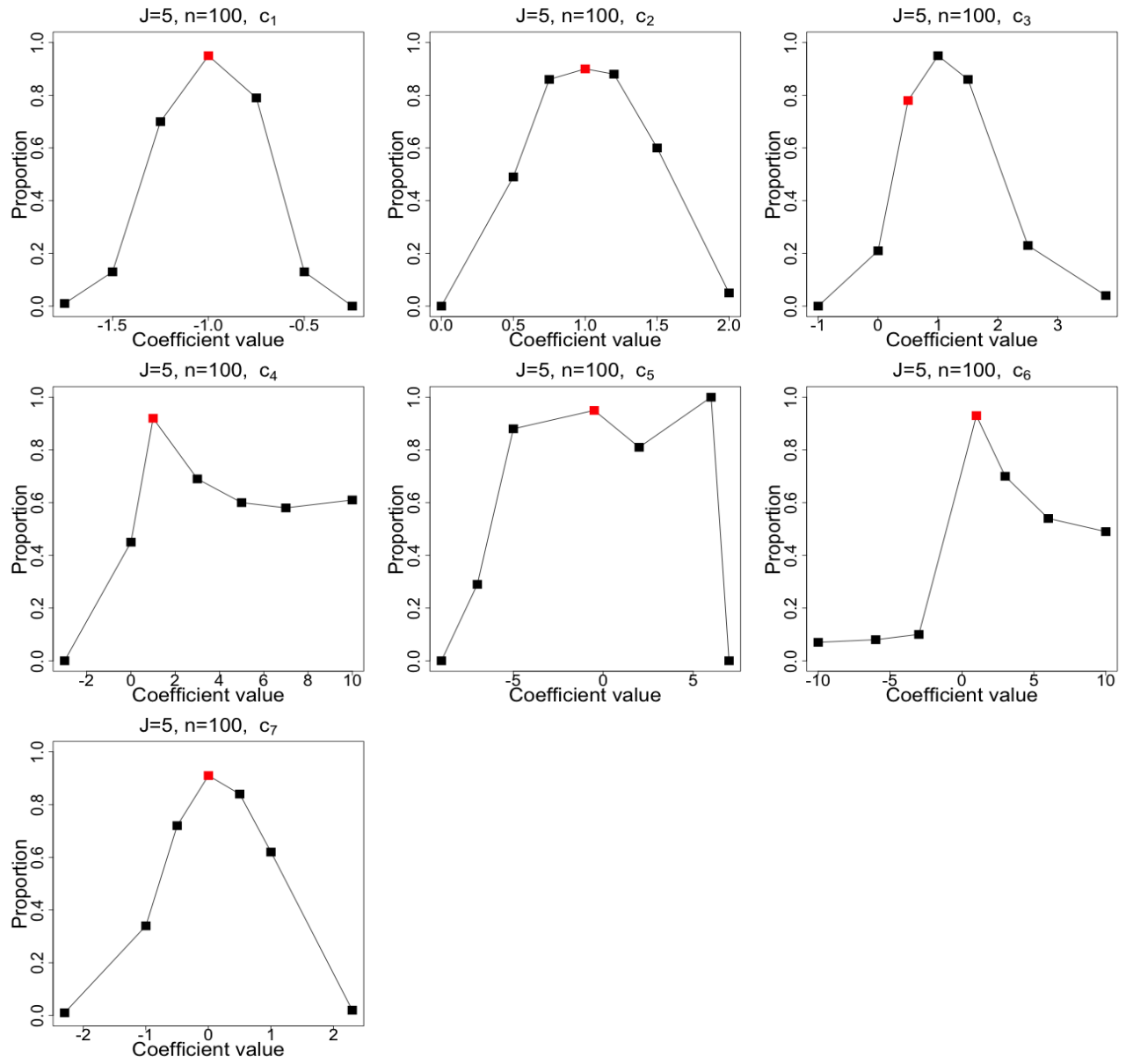
10.3 Postulated Model 1, $J=1$, $n=1000$



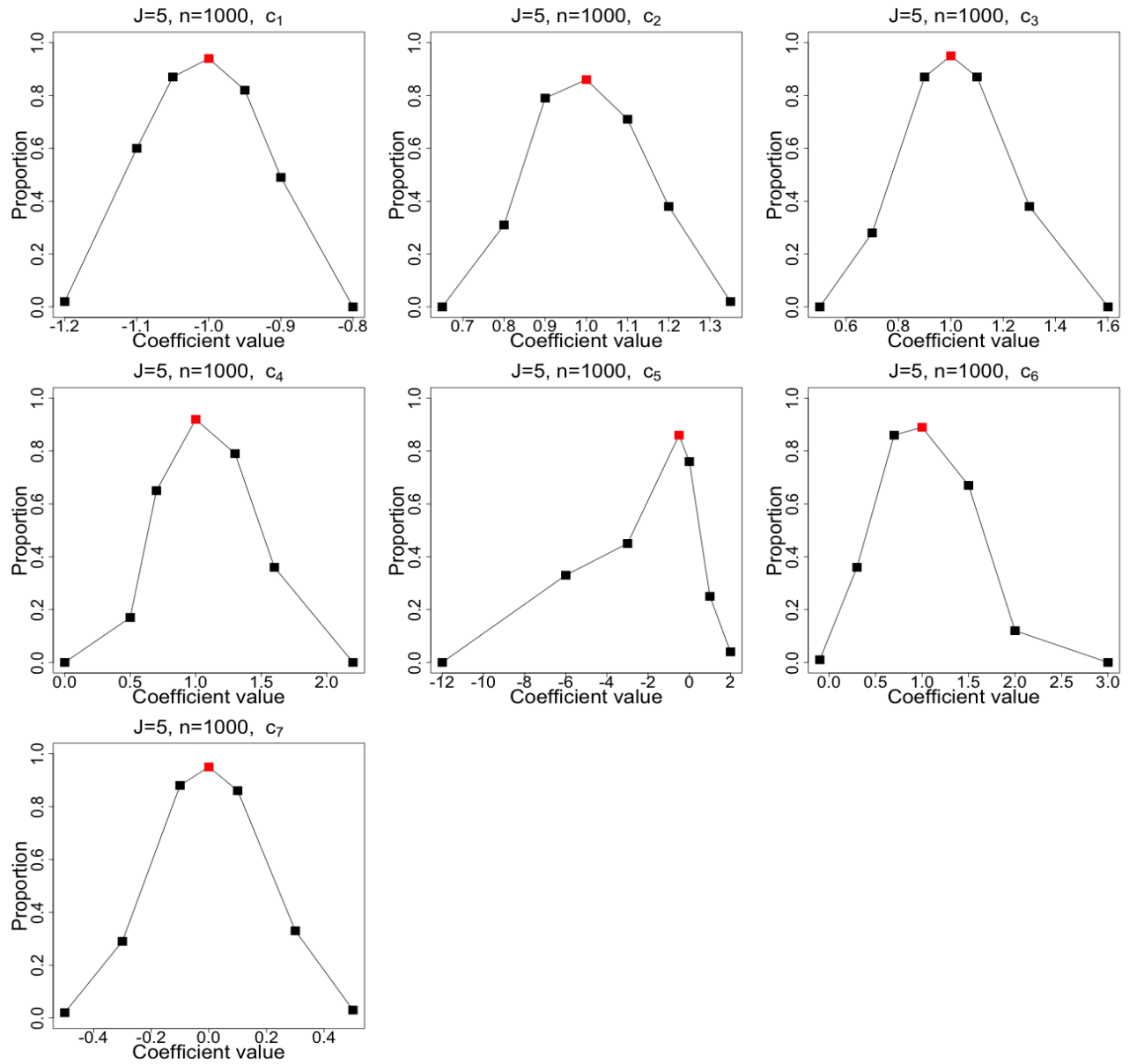
10.4 Postulated Model 1, $J=5$, $n=10$



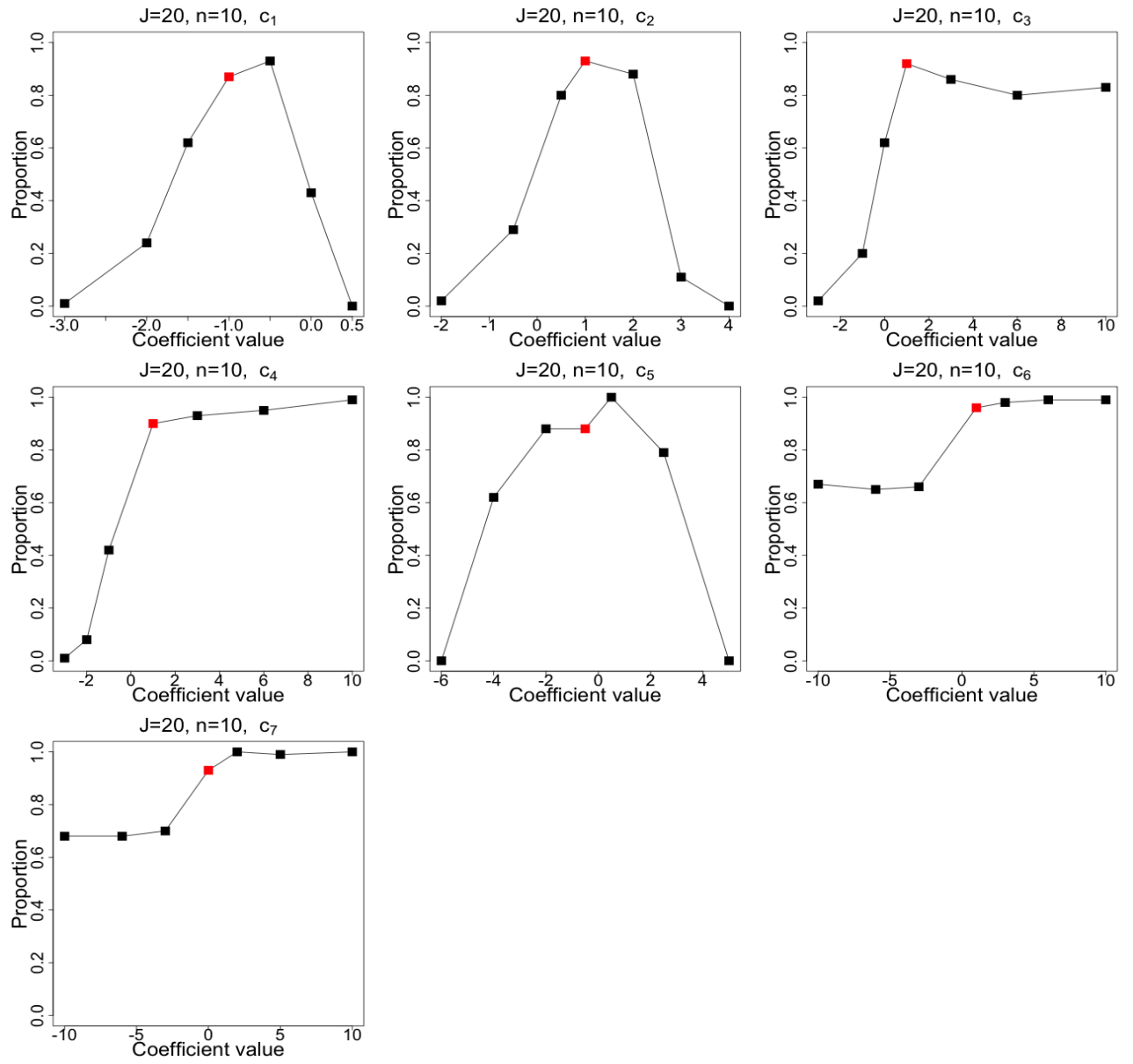
10.5 Postulated Model 1, $J=5$, $n=100$



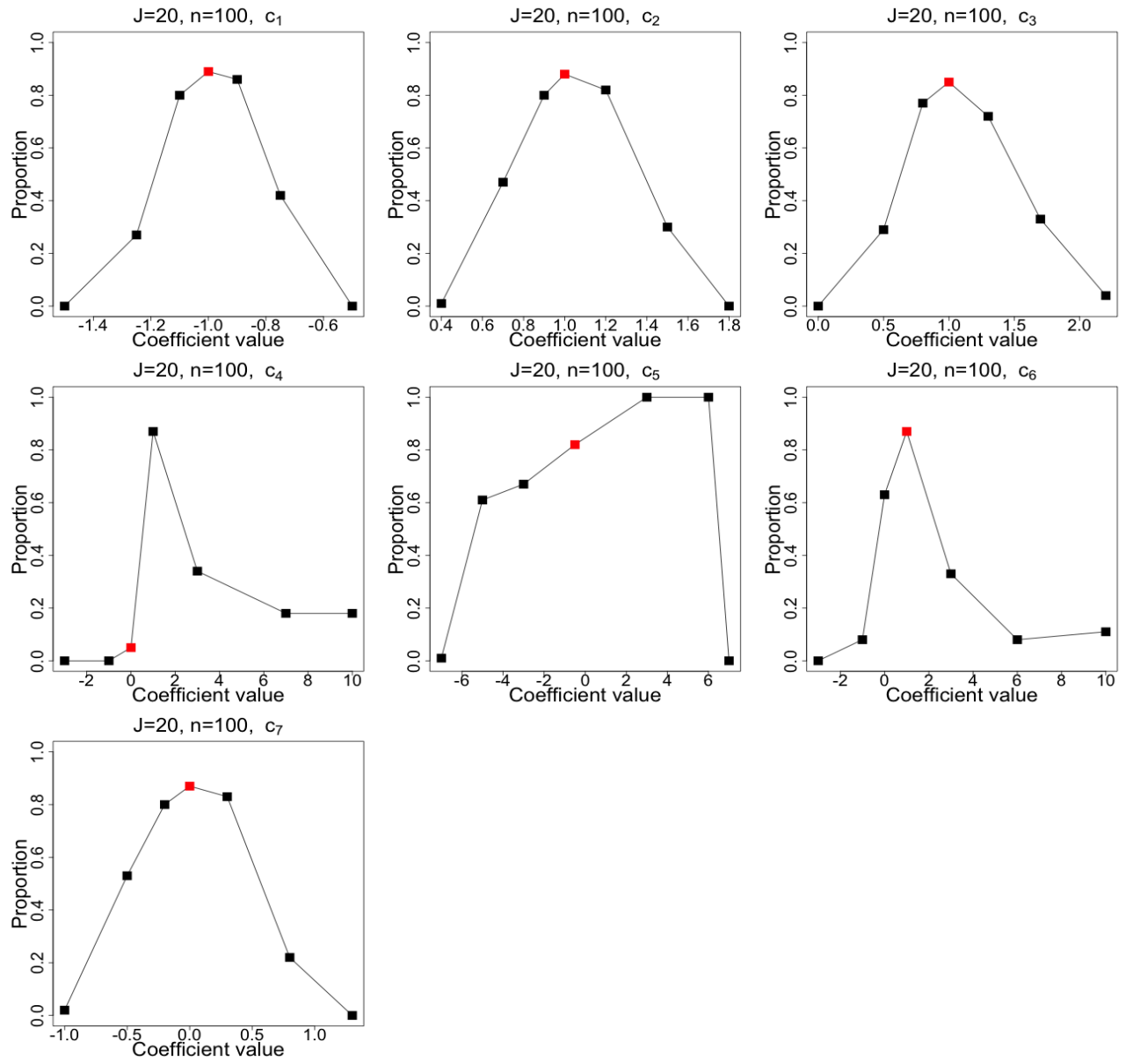
10.6 Postulated Model 1, $J=5$, $n=1000$



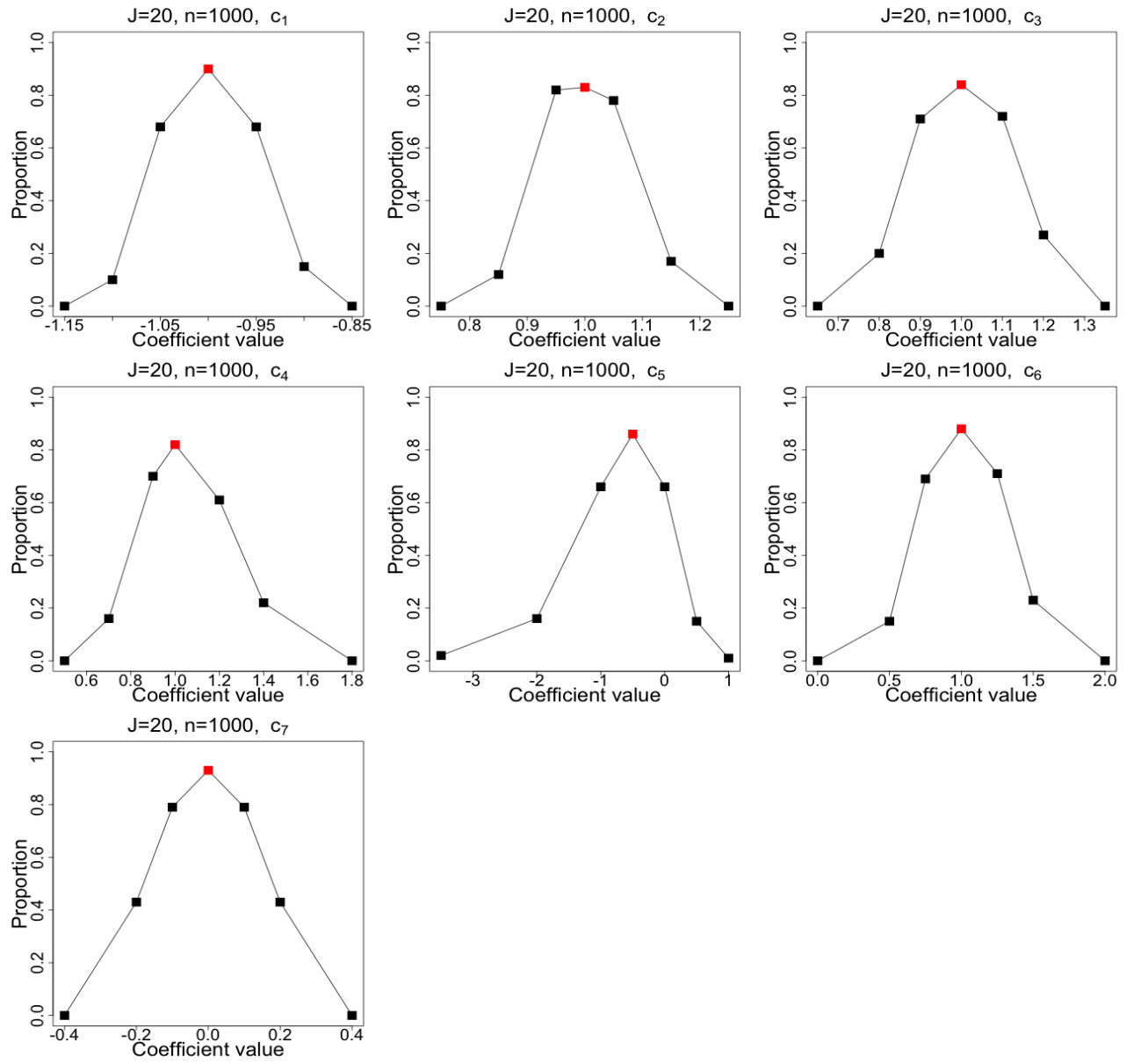
10.7 Postulated Model 1, $J=20$, $n=10$



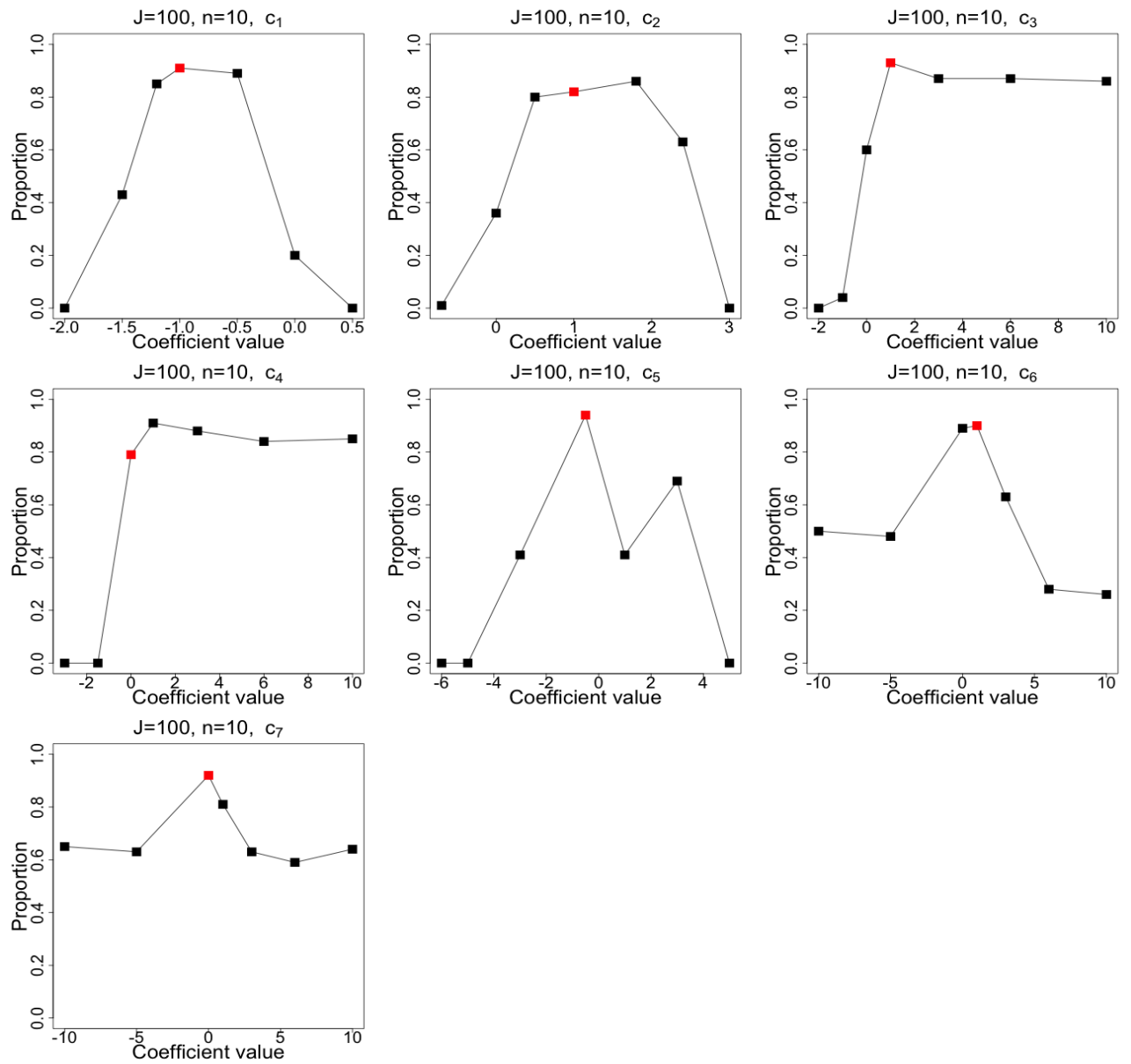
10.8 Postulated Model 1, $J=20$, $n=100$



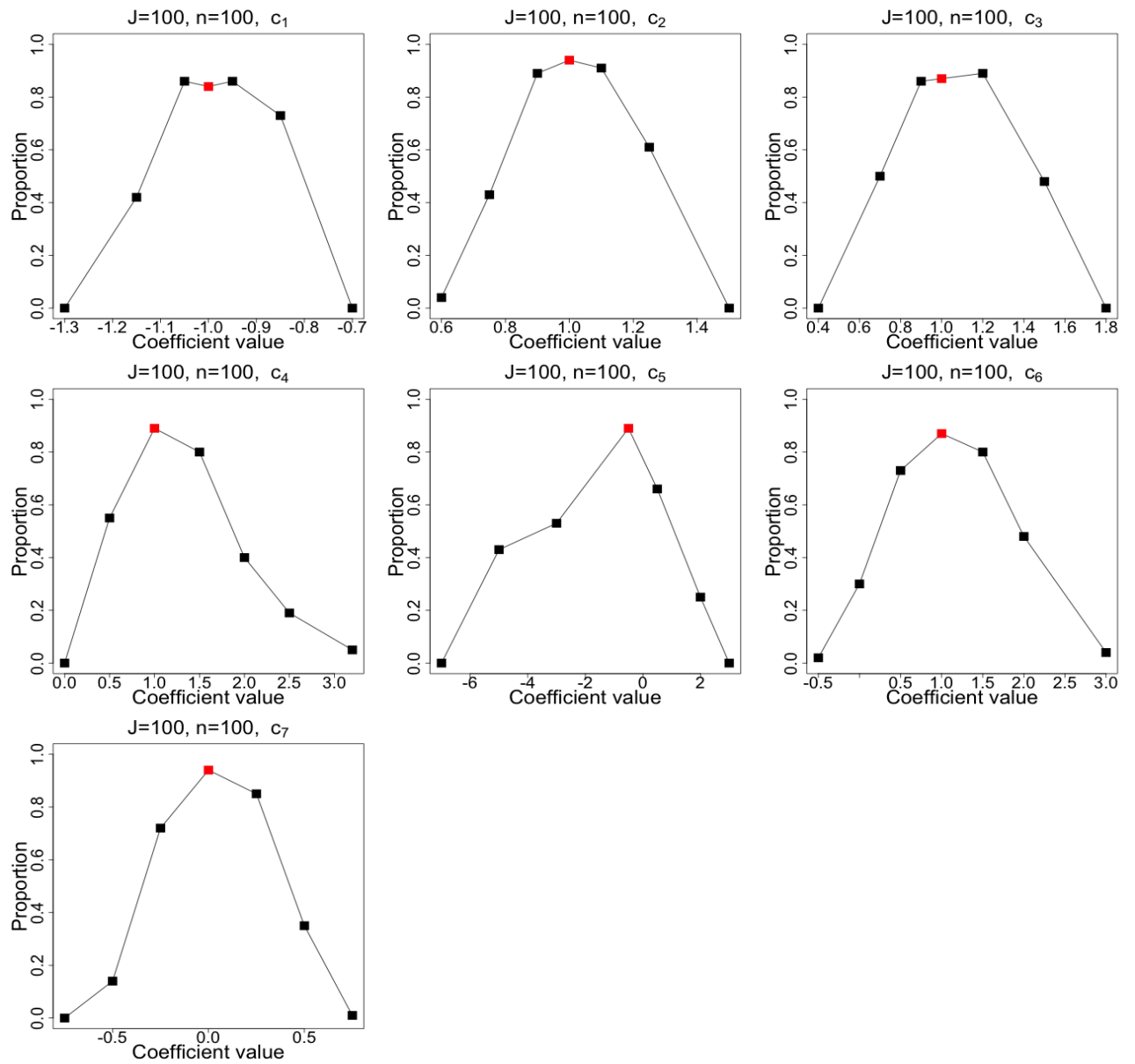
10.9 Postulated Model 1, $J=20$, $n=1000$



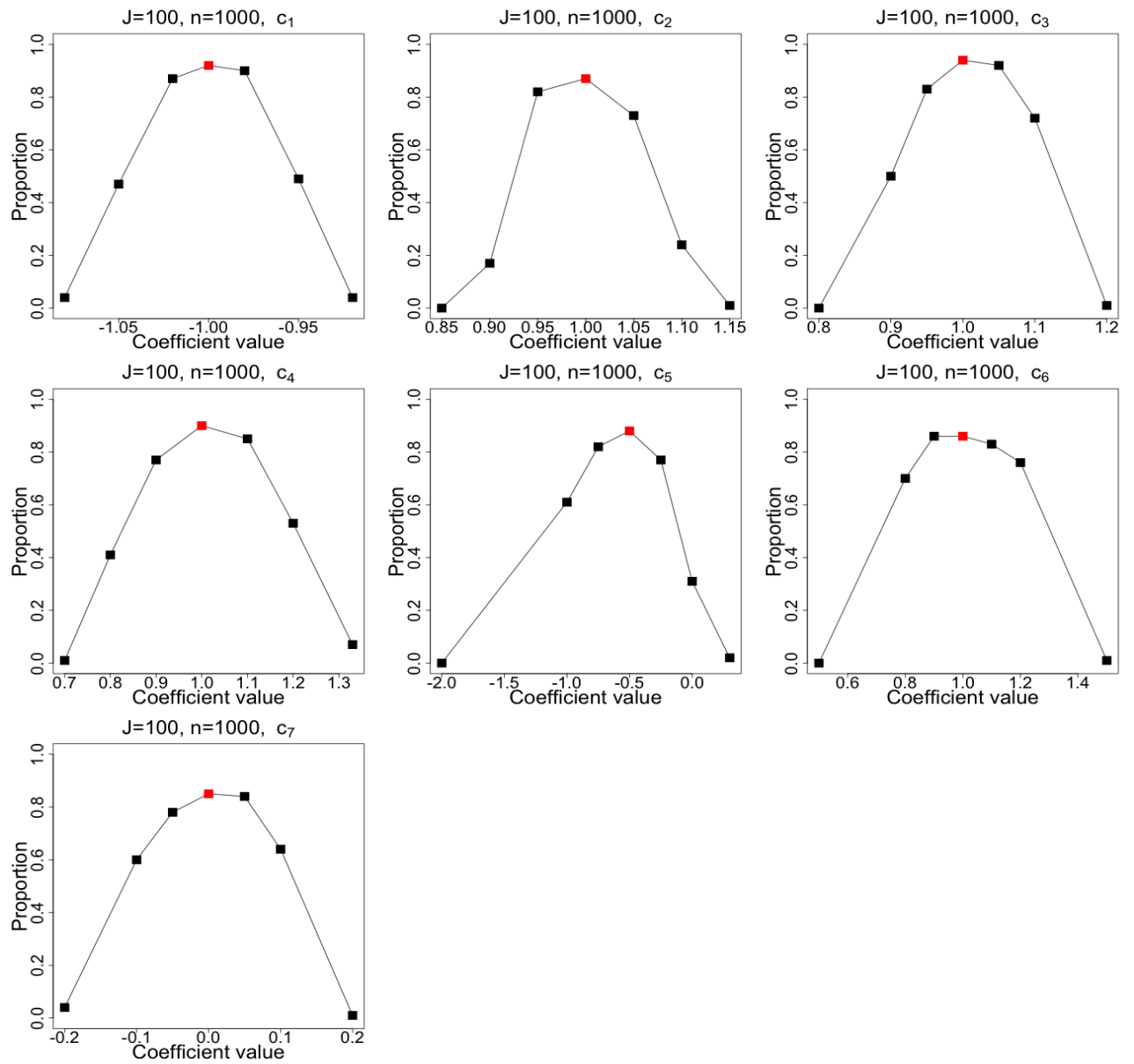
10.10 Postulated Model 1, $J=100$, $n=10$



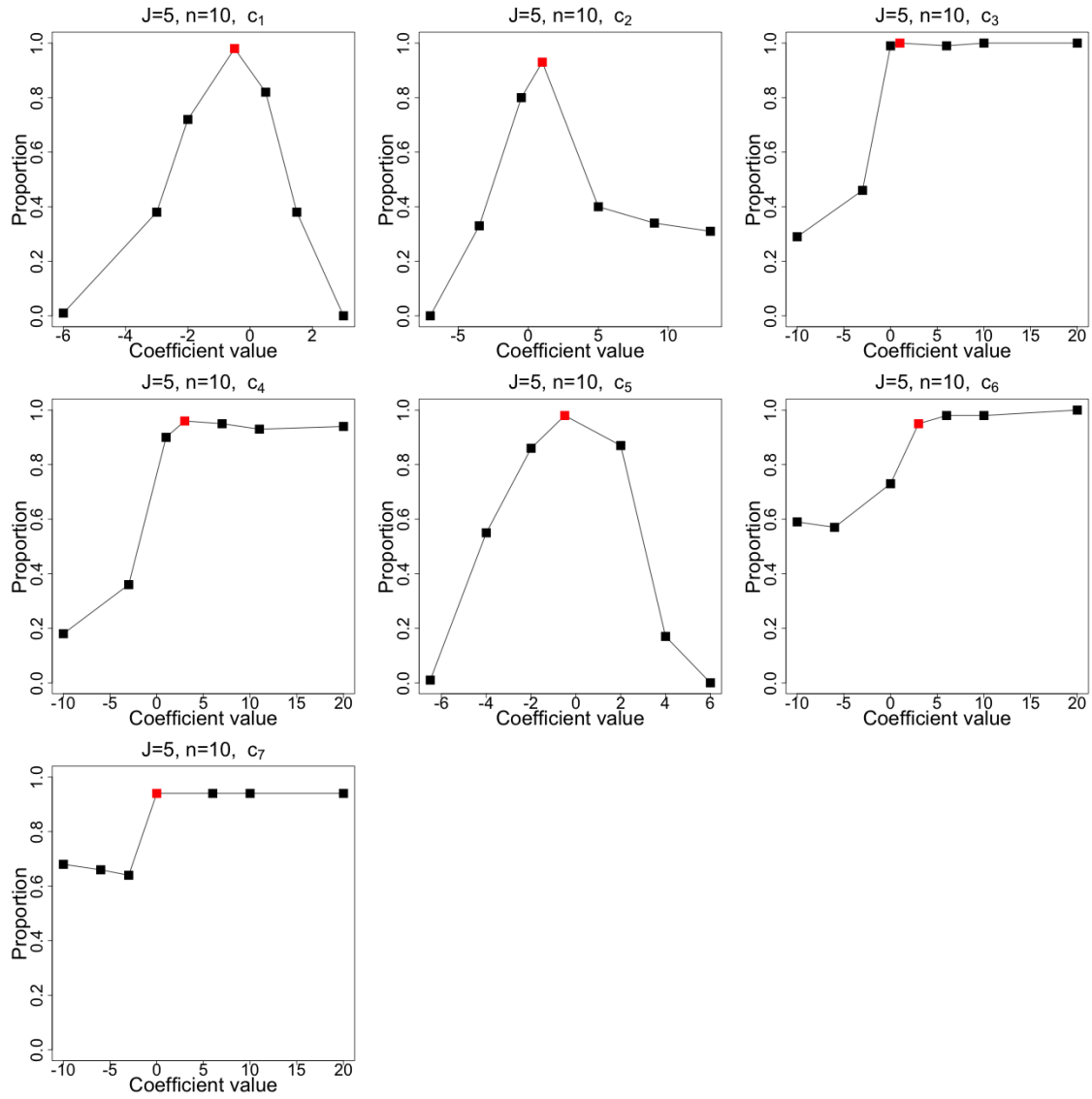
10.11 Postulated Model 1, $J=100$, $n=100$



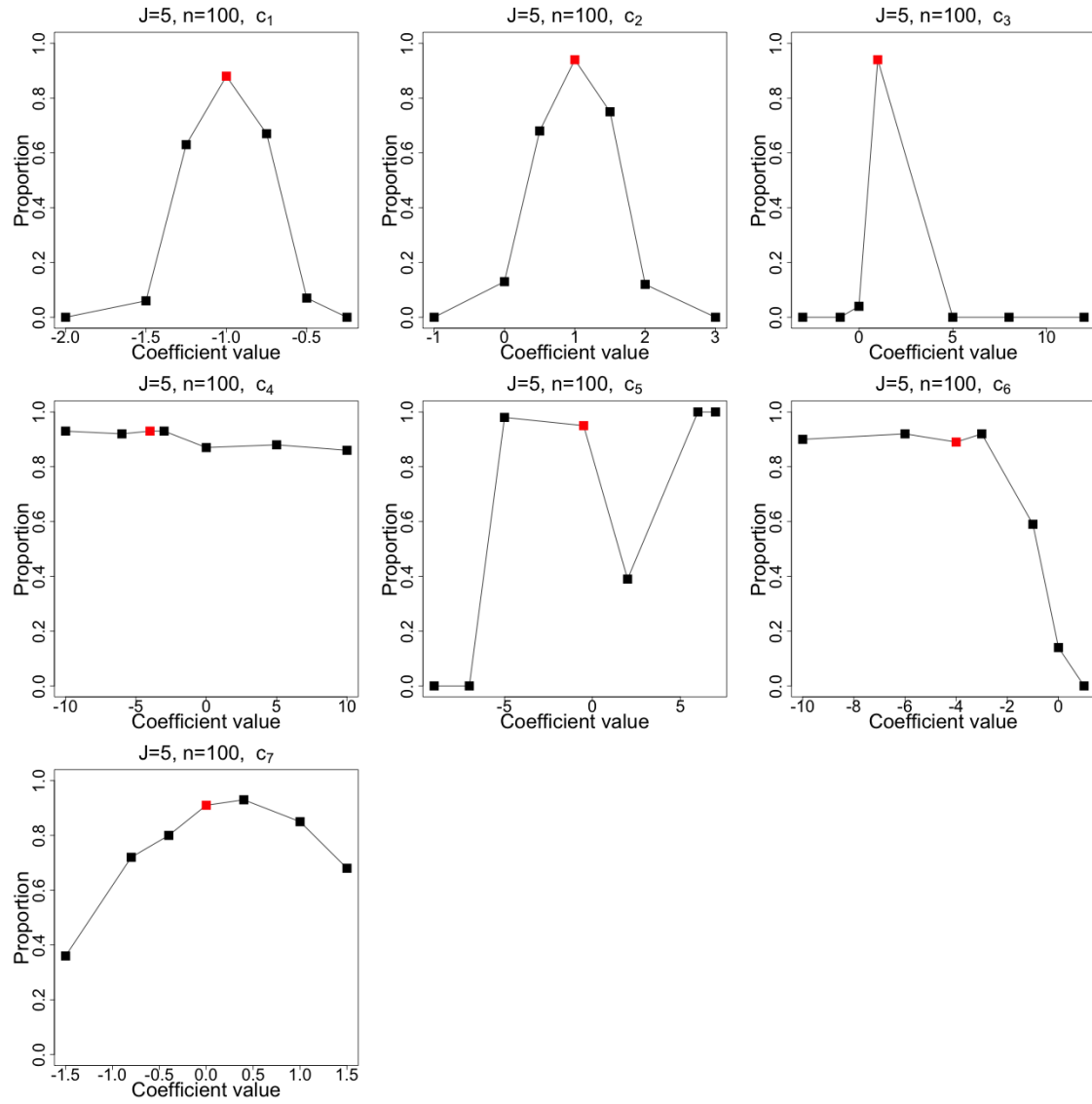
10.12 Postulated Model 1, $J=100$, $n=1000$



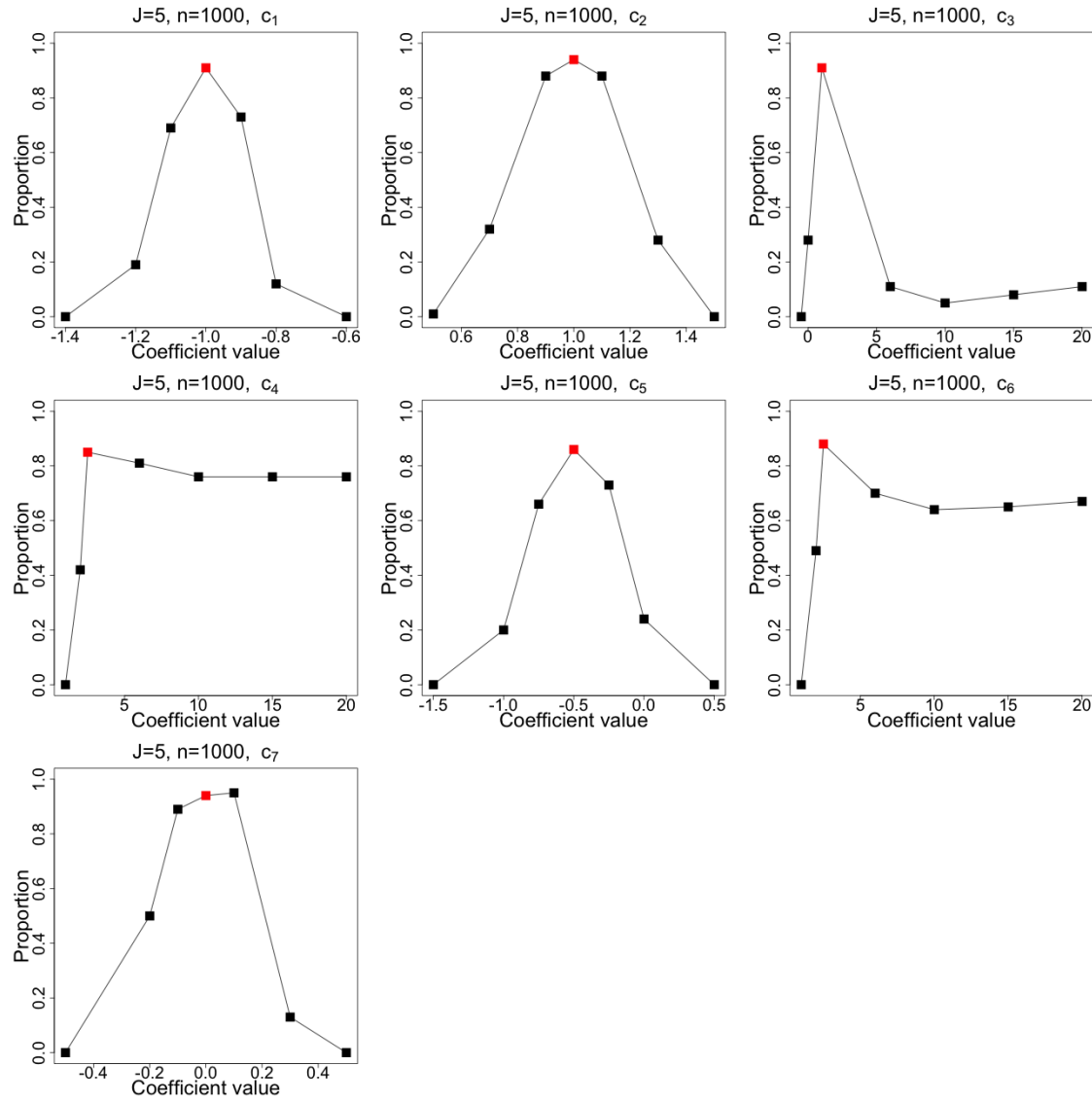
10.13 Postulated Model 1, $J=5$, $n=10$, $c_4 = c_6 = 3$



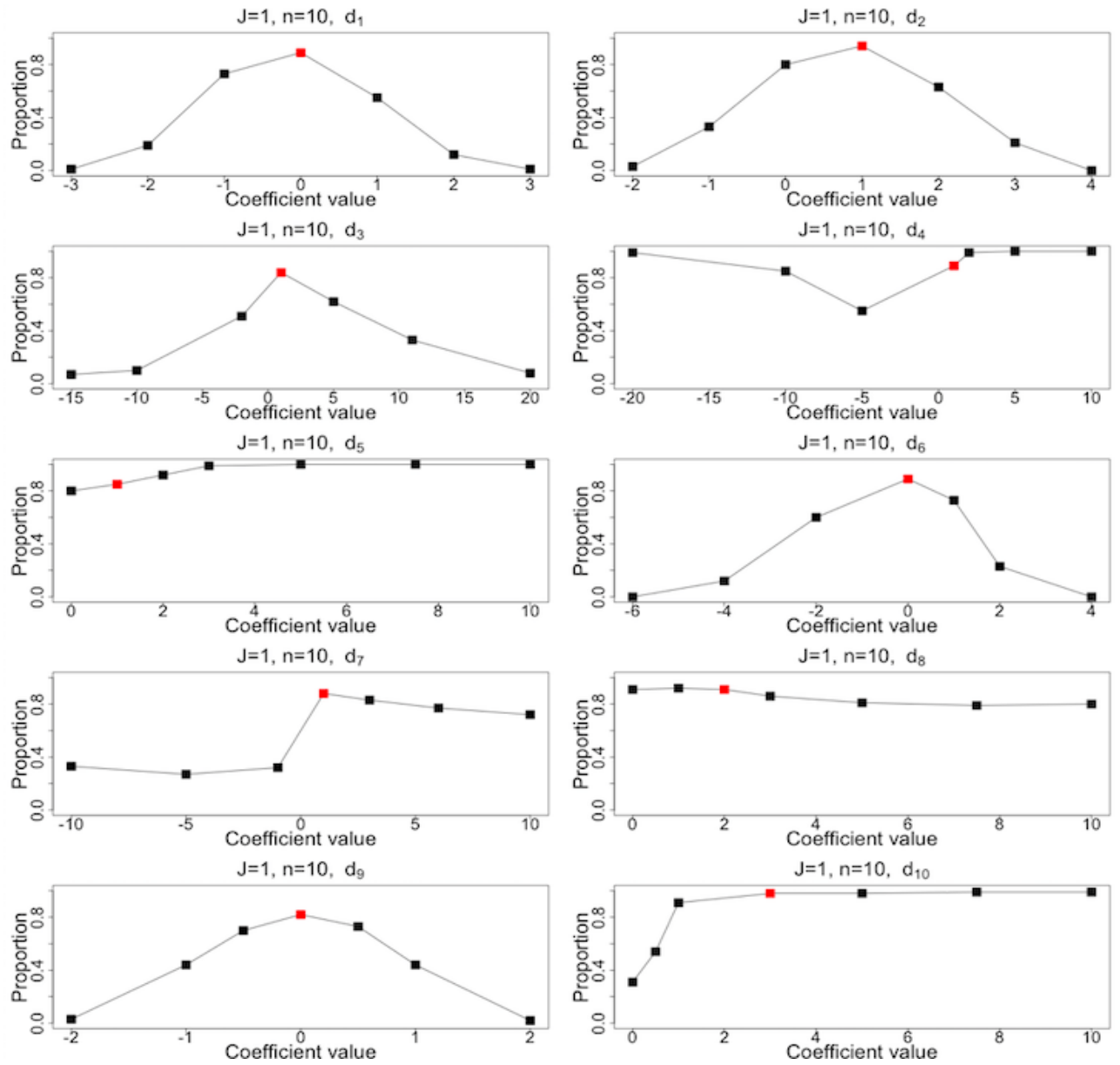
10.14 Postulated Model 1, $J=5$, $n=100$, $c_4 = c_6 = 3$



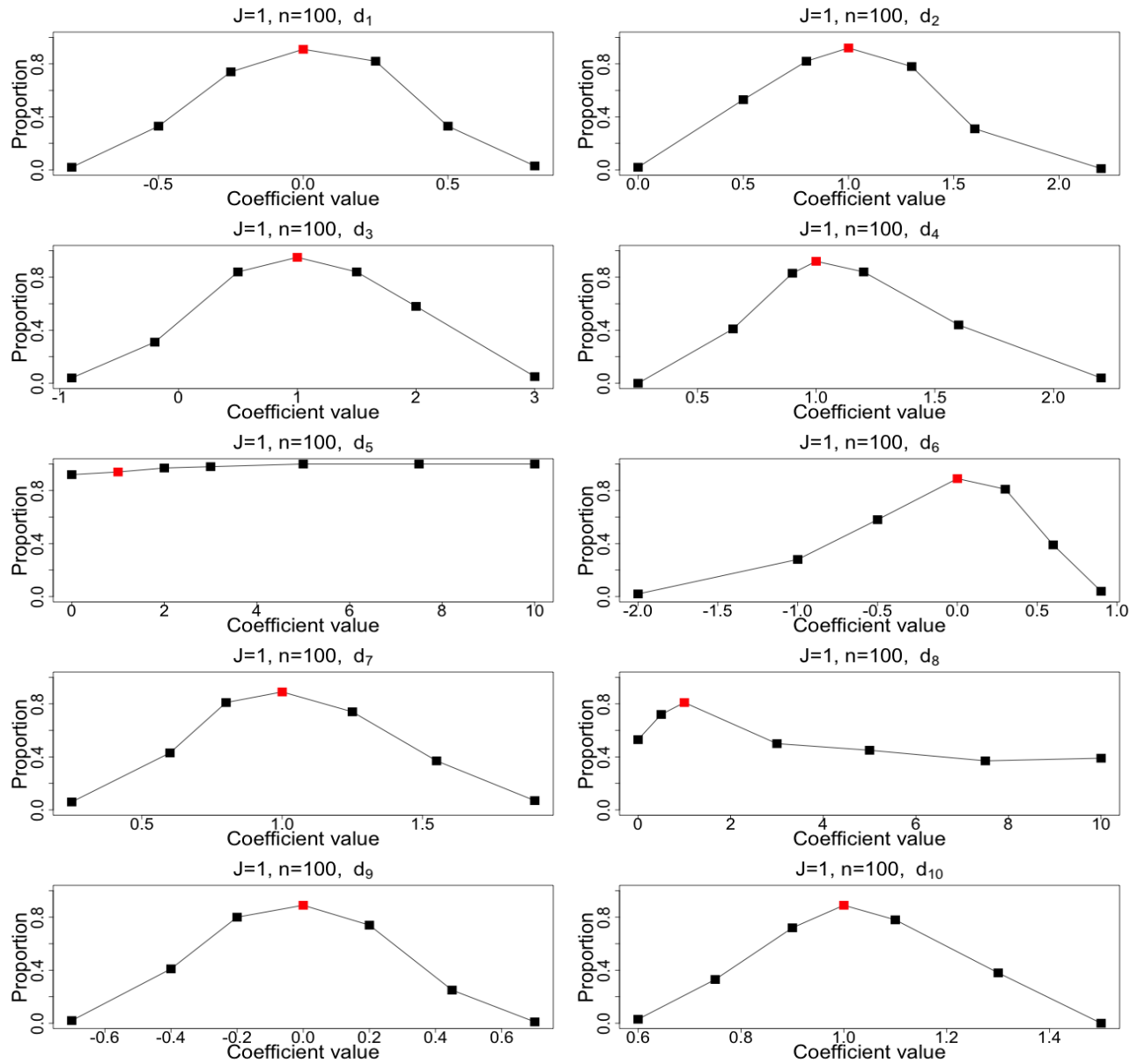
10.15 Postulated Model 1, $J=5$, $n=1000$, $c_4 = c_6 = 3$



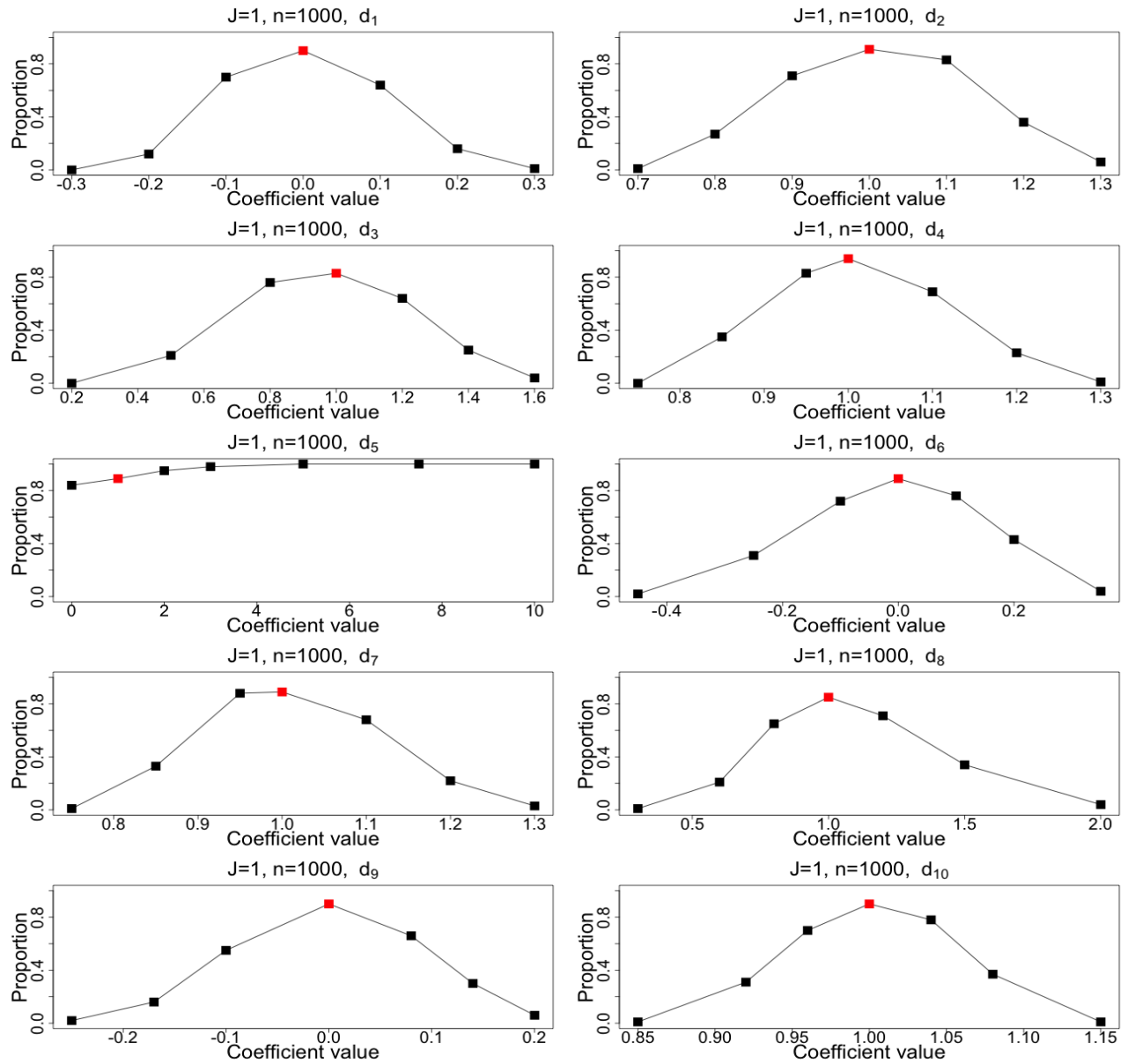
10.16 Postulated Model 2, $J=1$, $n=10$



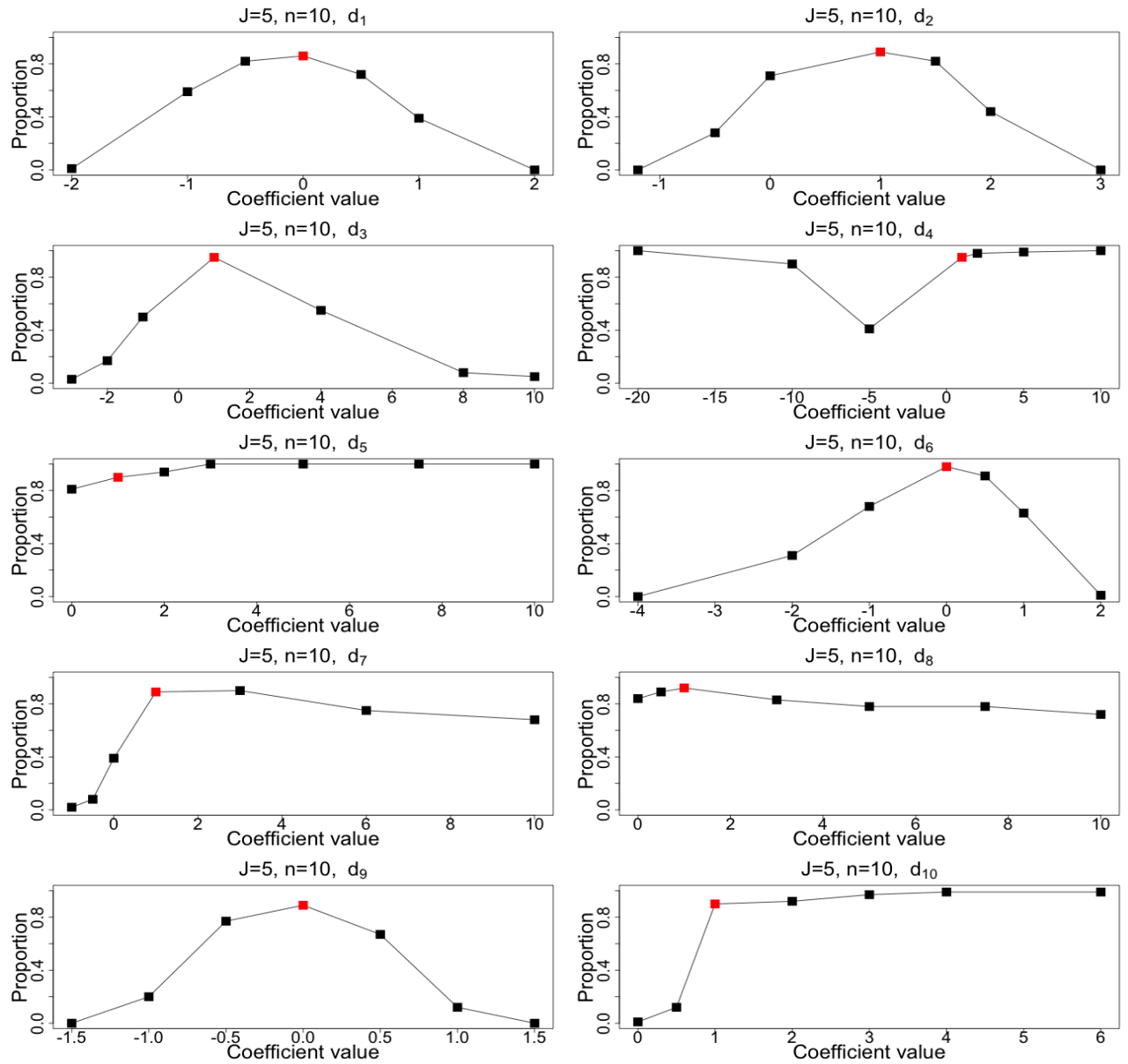
10.17 Postulated Model 2, $J=1$, $n=100$



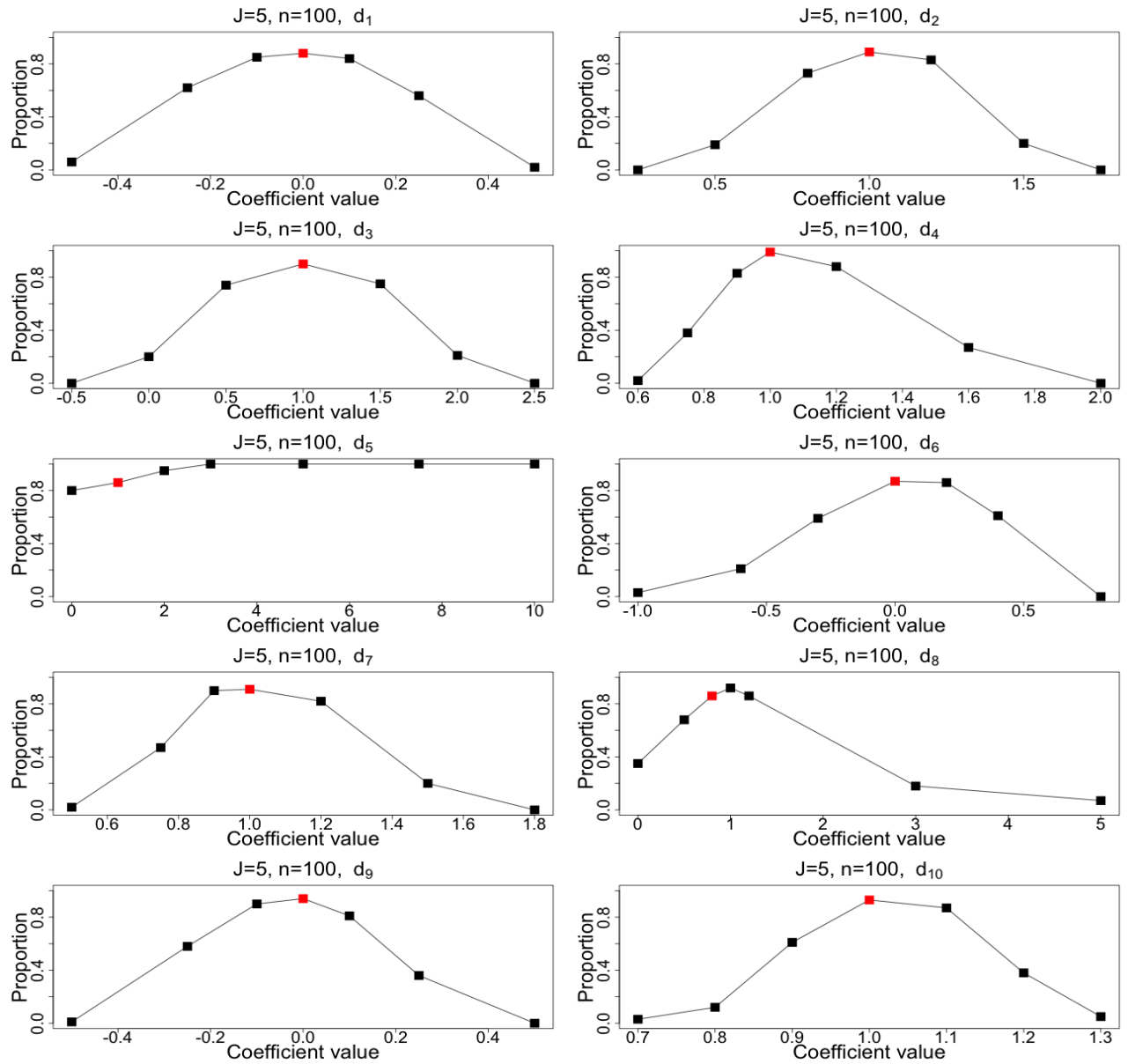
10.18 Postulated Model 2, $J=1$, $n=1000$



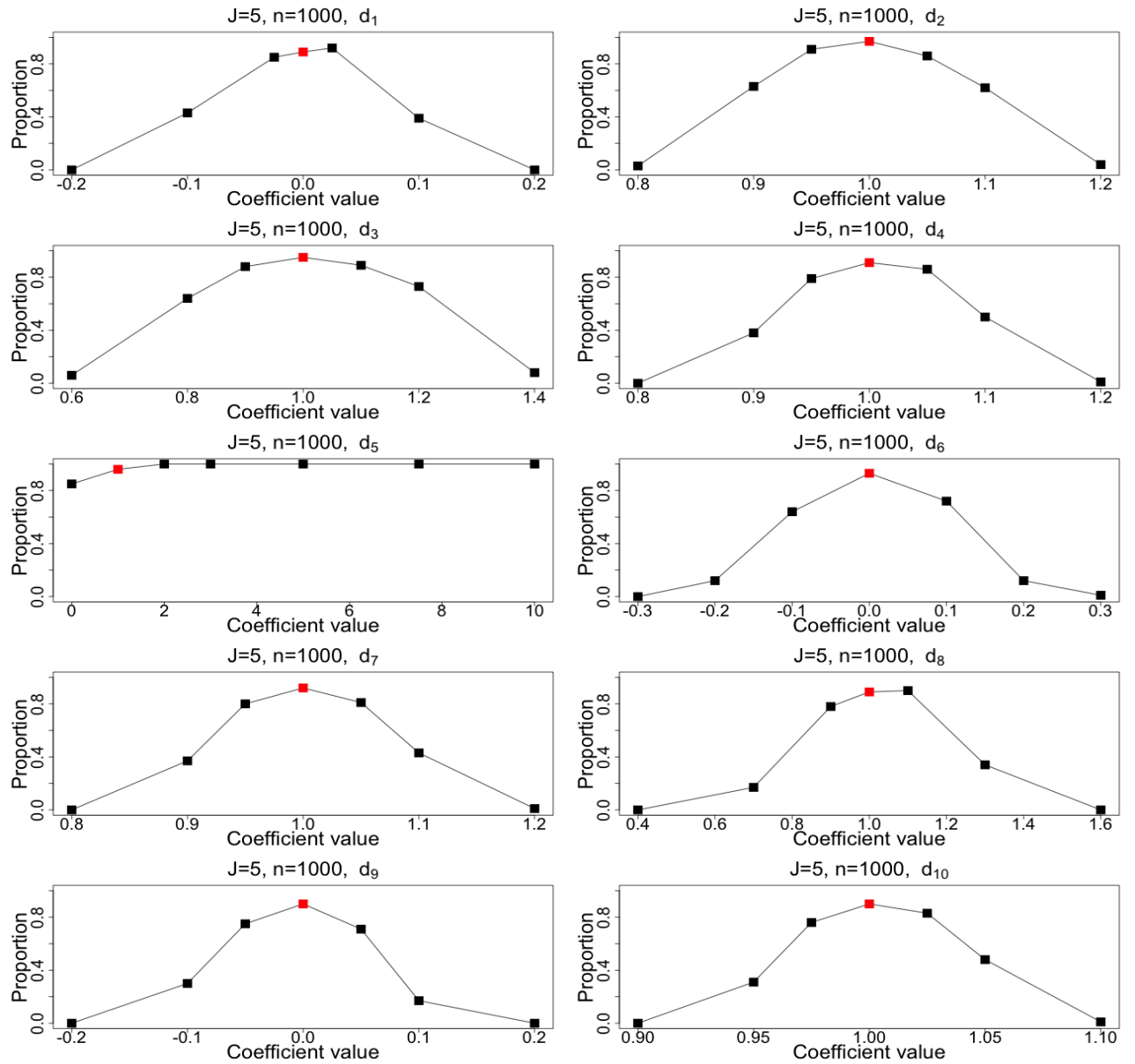
10.19 Postulated Model 2, $J=5$, $n=10$



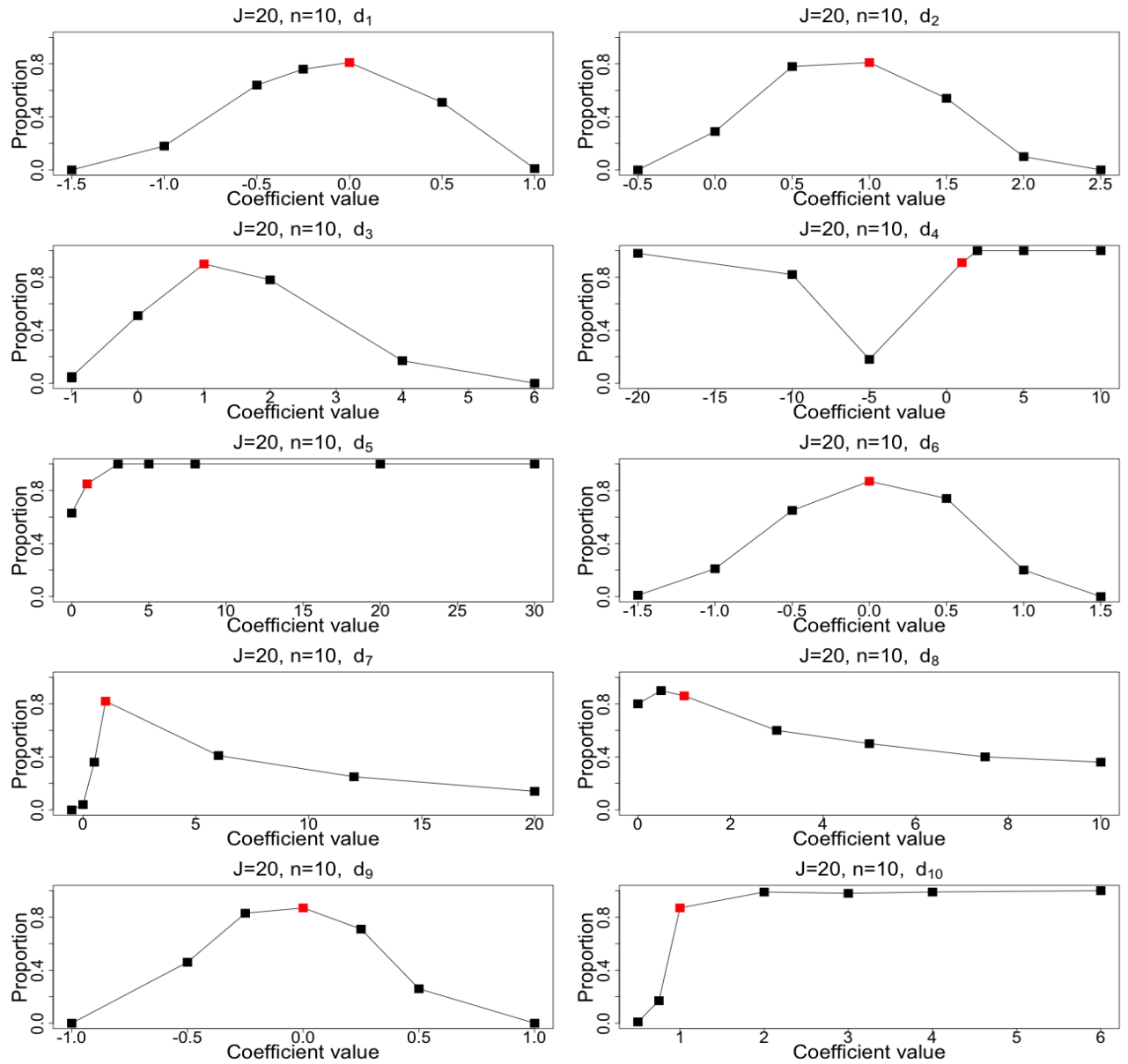
10.20 Postulated Model 2, J=5, n=100



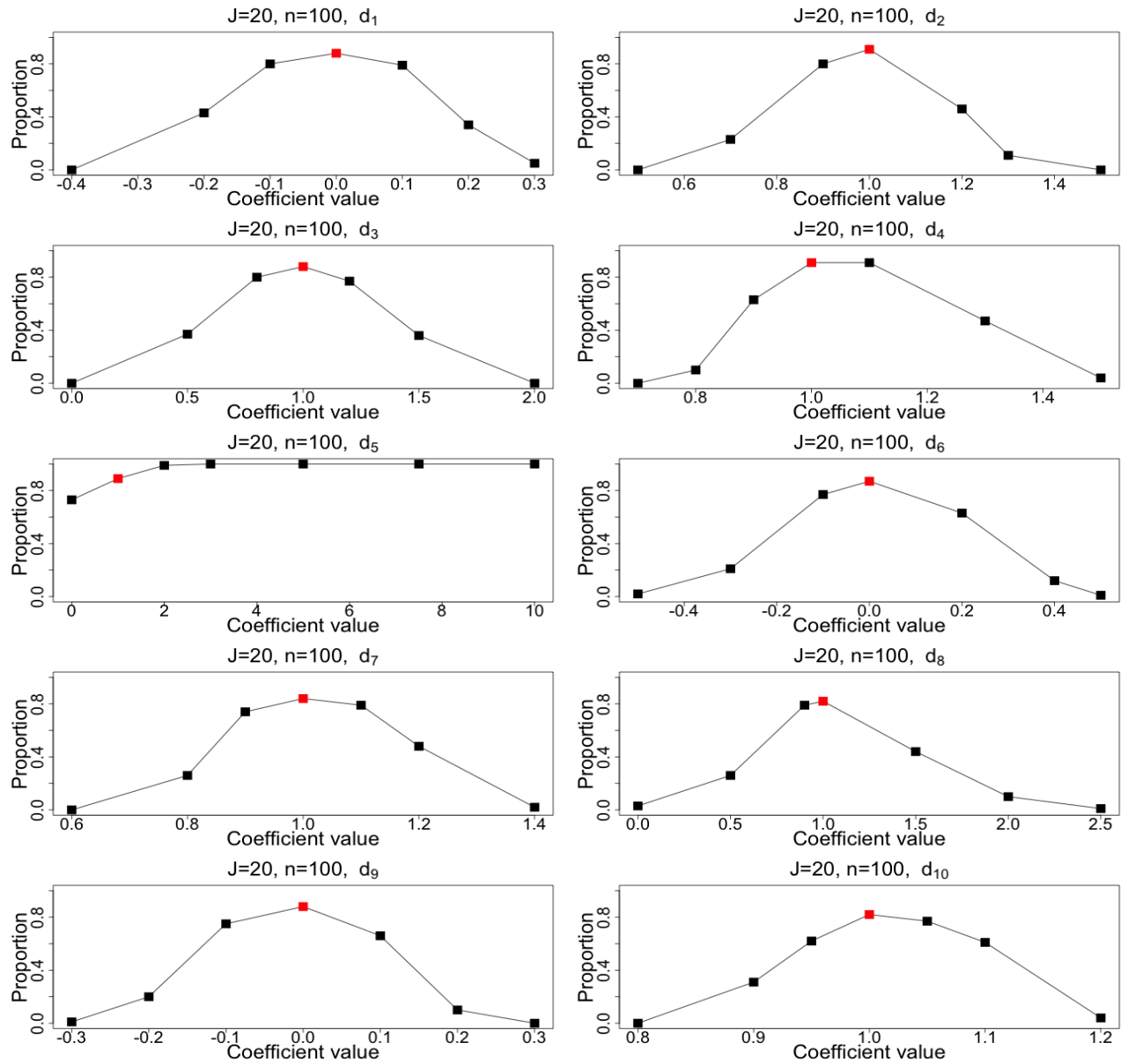
10.21 Postulated Model 2, $J=5$, $n=1000$



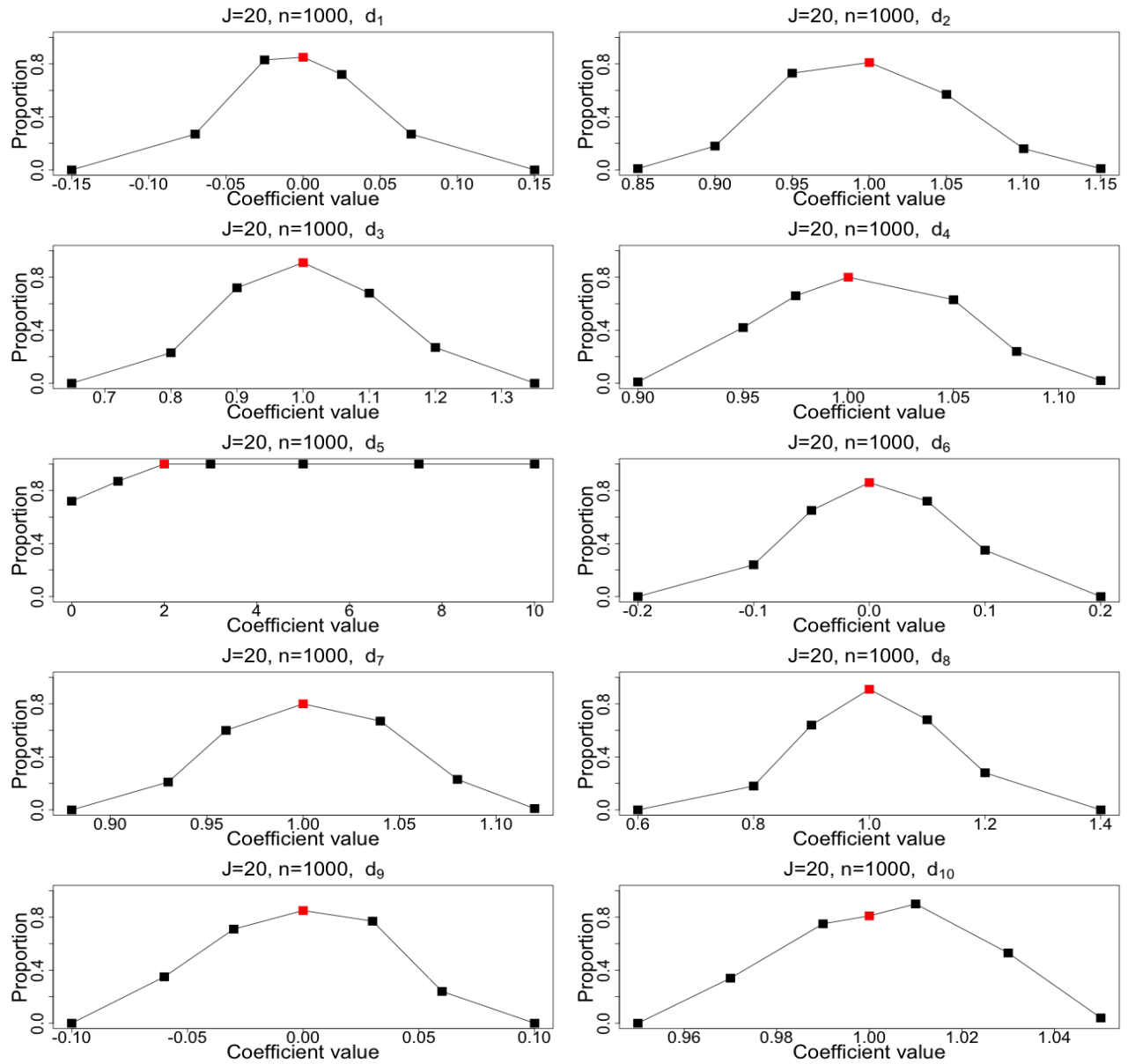
10.22 Postulated Model 2, $J=20$, $n=10$



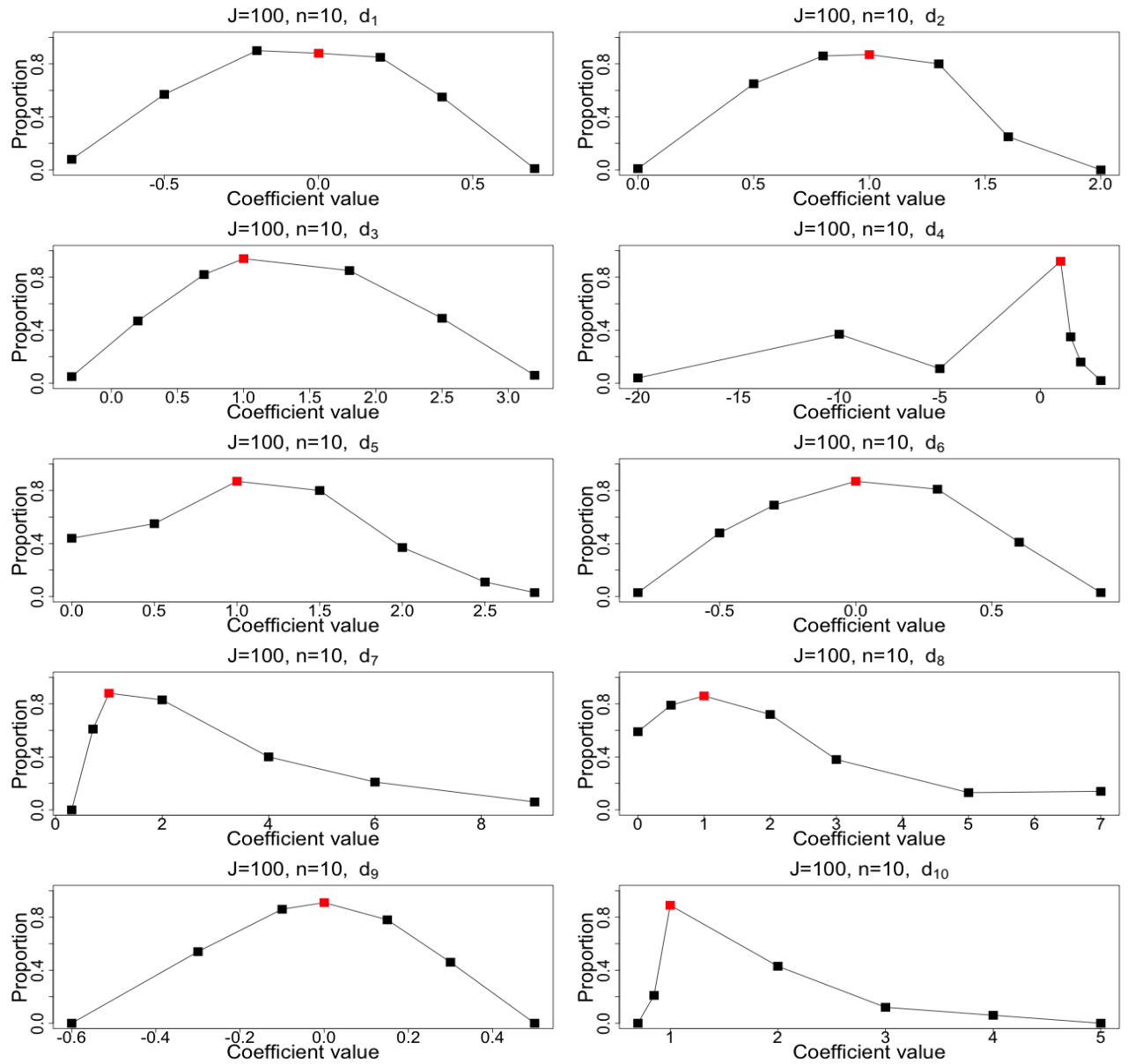
10.23 Postulated Model 2, $J=20$, $n=100$



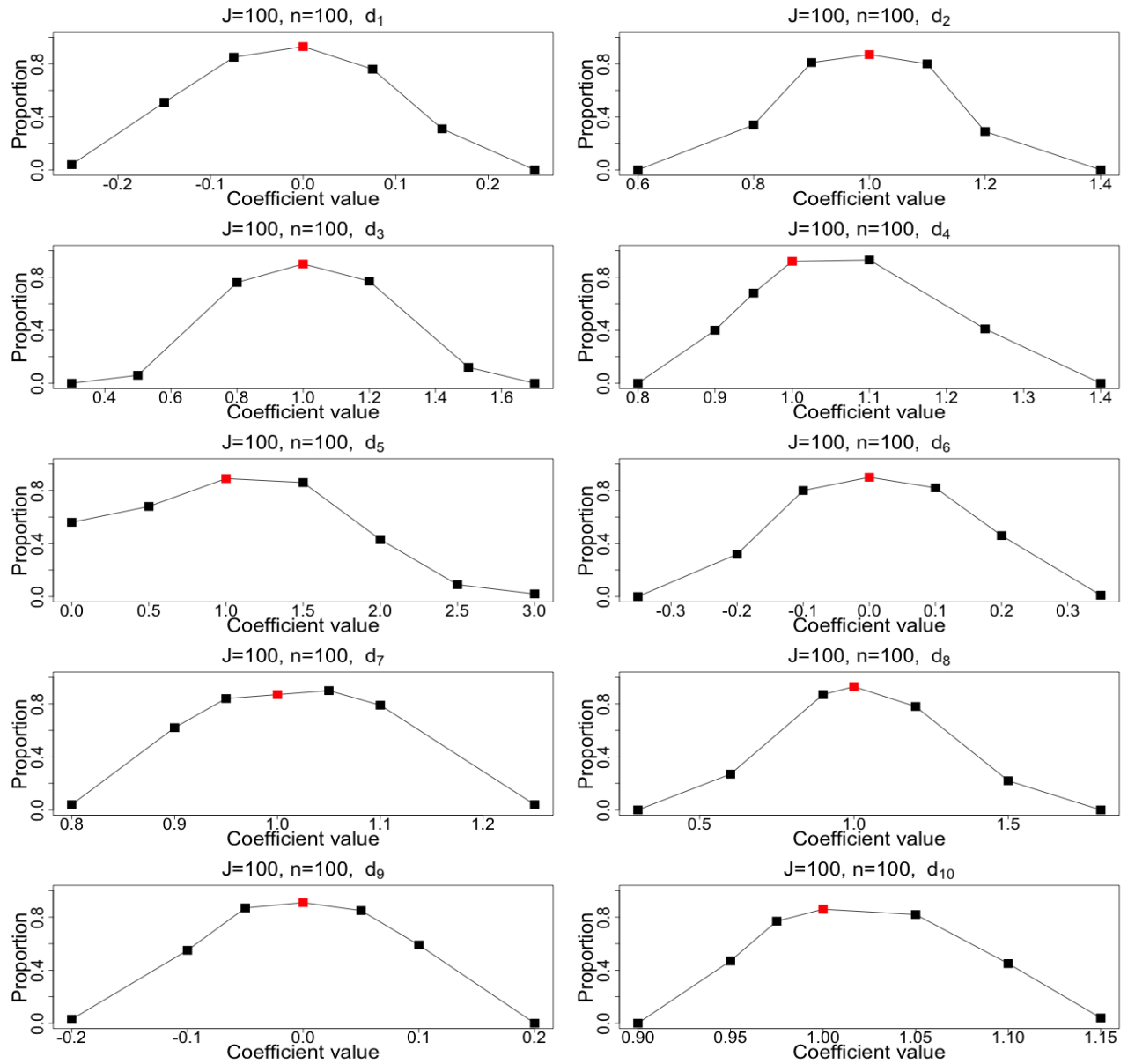
10.24 Postulated Model 2, J=20, n=1000



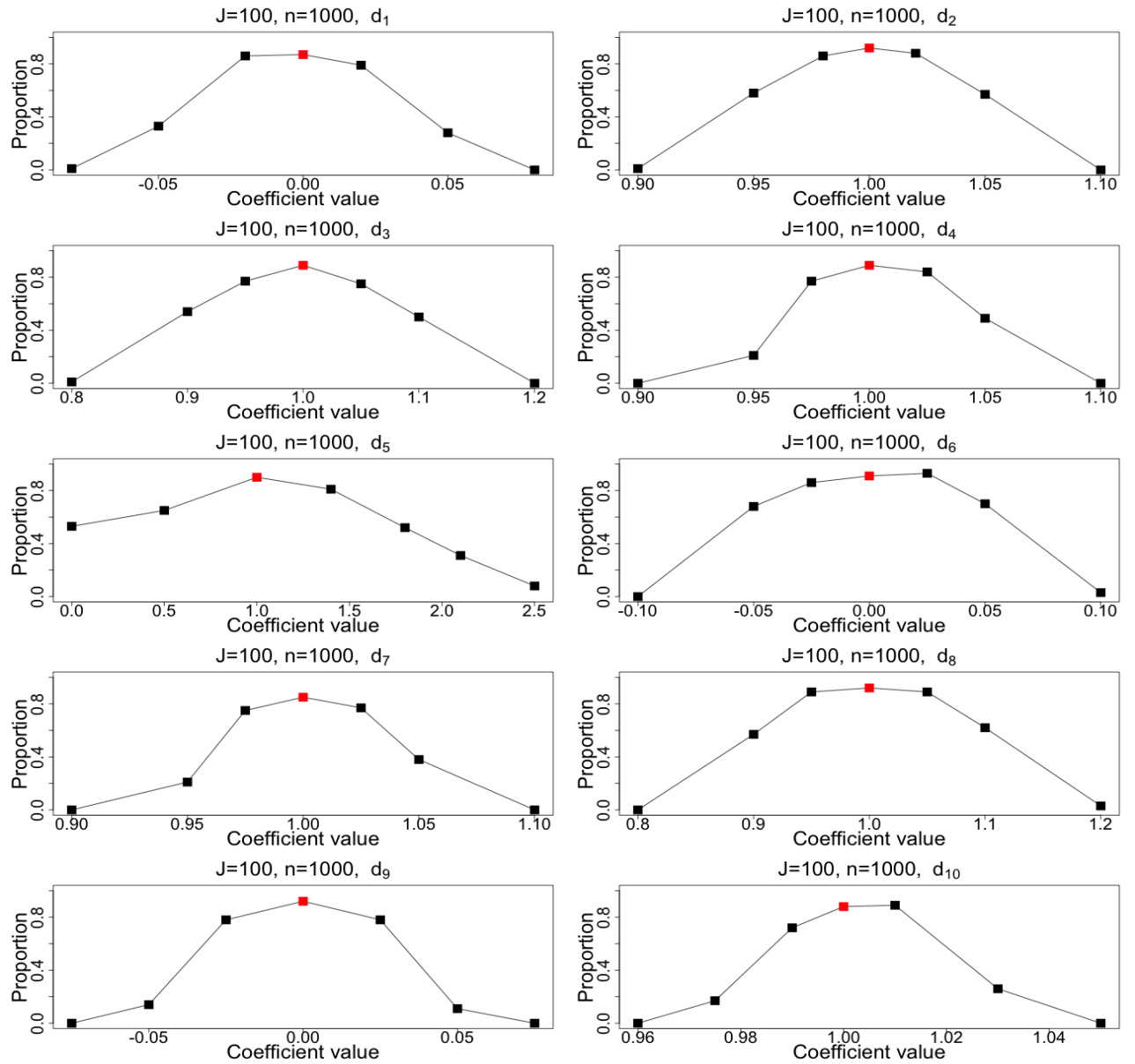
10.25 Postulated Model 2, $J=100$, $n=10$



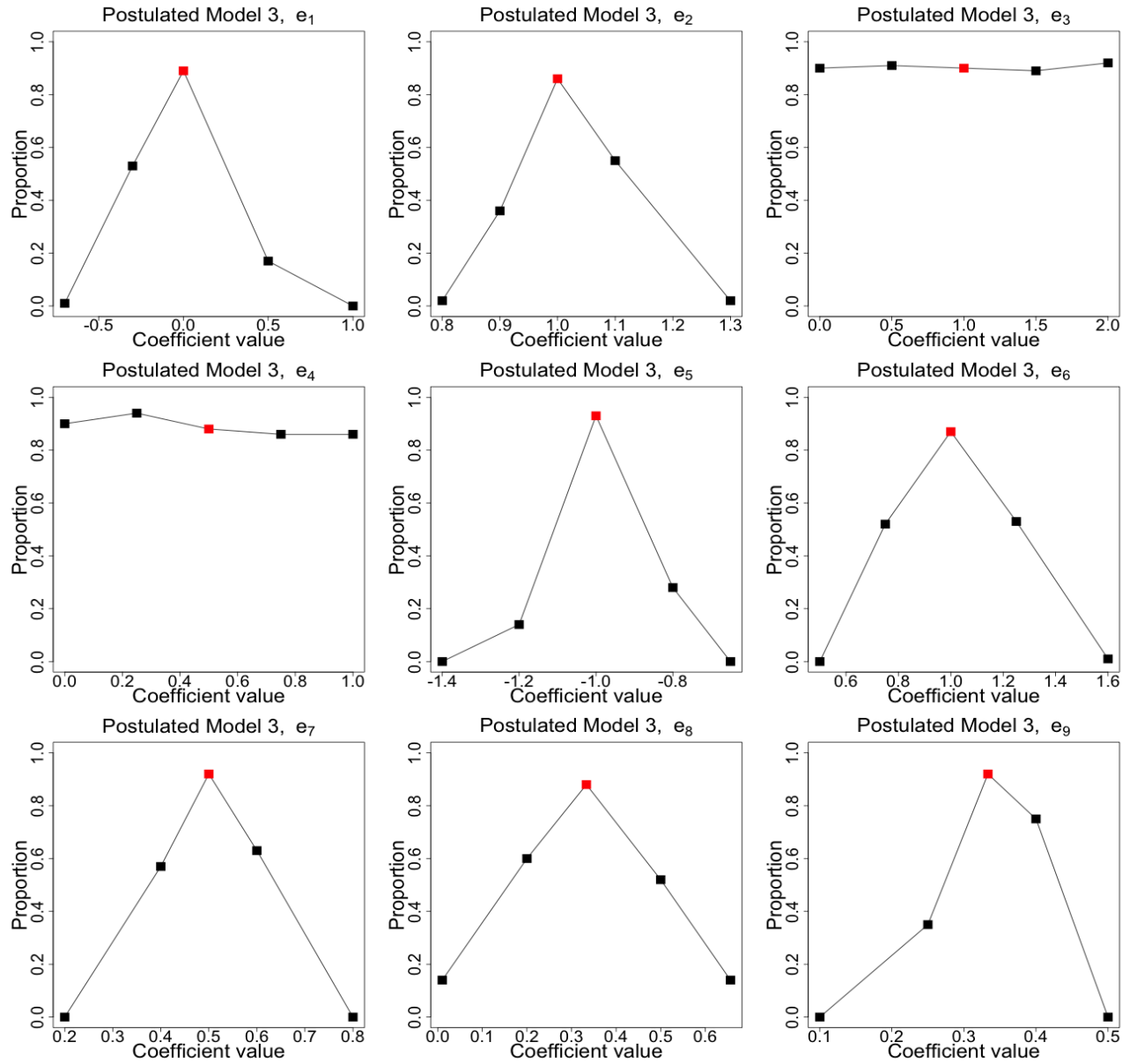
10.26 Postulated Model 2, J=100, n=100



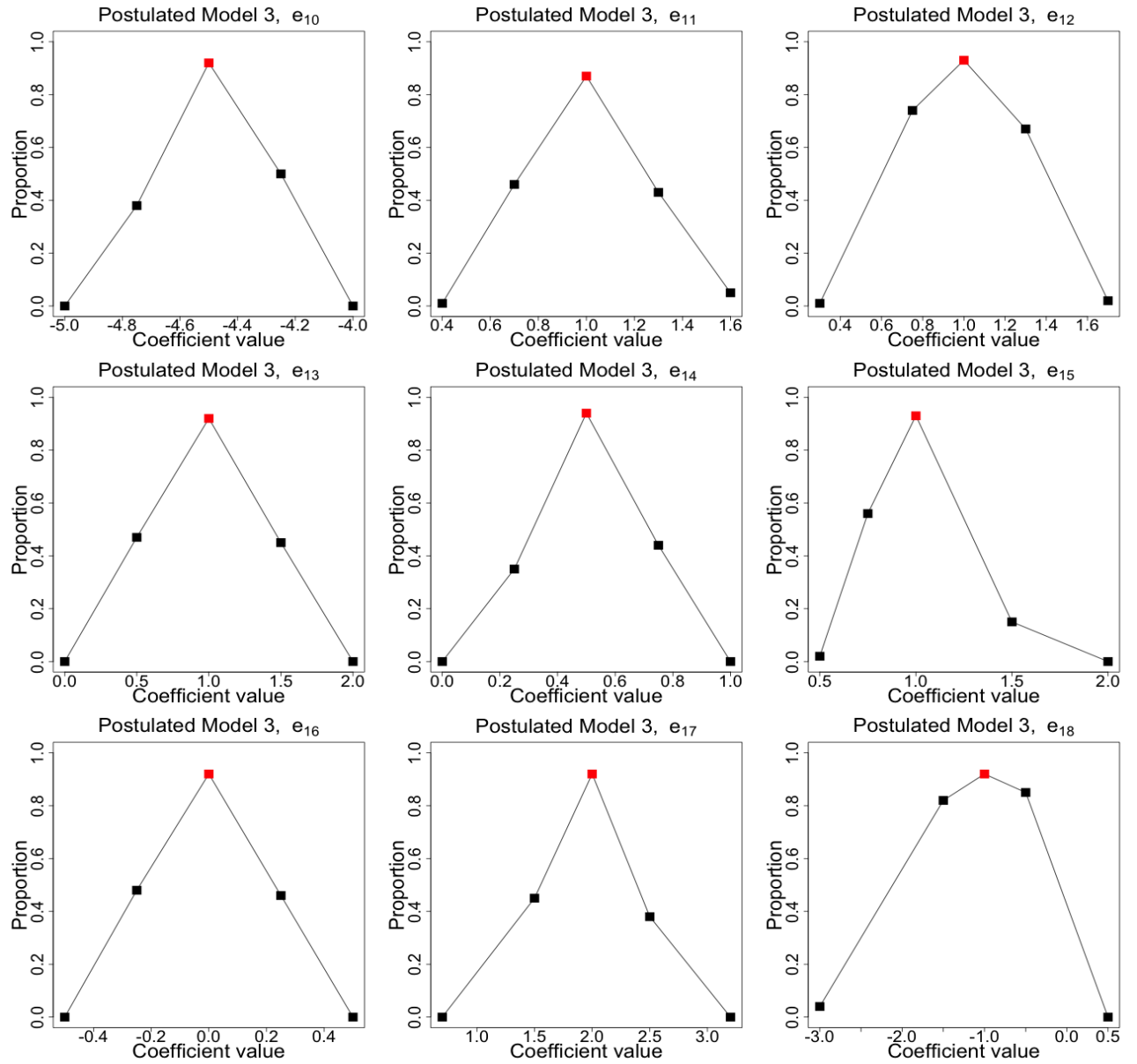
10.27 Postulated Model 2, J=100, n=1000



10.28 Postulated Model 3



10.29 Postulated Model 3



11 Sources

- Agresti, Alan & Coull, Brent A. (1998). "Approximate Is Better than 'Exact' for Interval Estimation of Binomial Proportions" in *The American Statistician*, vol. 52, no. 2: p 119-126.
- Agresti, Alan (2012). *Categorical Data Analysis*, 3rd edition. John Wiley Sons.
- Beaumont, Mark A. (2010). "Approximate Bayesian Computation in Evolution and Ecology" in *The Annual Review of Ecology, Evolution, and Systematics*, vol. 41: p 379–406.
- Bodnar, Olha; Link, Alfred; Arendack, Barbora; Possolo, Antonio; Elster, Clemens. (2016). "Bayesian estimation in random effects meta-analysis using a non-informative prior" in *Statistics in Medicine*, vol. 36: p 378–399.
- Borenstein, Michael, et. al. (2007). *Meta-analysis: Fixed Effect vs. Random Effect*. Web source: meta-analysis.com. Accessed: February 4 2021.
- Borenstein, Michael, et. al. (2009). *Introduction to Meta-analysis*. John Wiley & Sons.
- Costello, Sarah, et. al. "Parkinsons Disease and Residential Exposure to Maneb and Paraquat From Agricultural Applications in the Central Valley of California" in *American Journal of Epidemiology*, vol. 169, no. 8: p 919-926.
- Davey J; Turner R.M; Clarke M.J.; Higgins J. (2011). "Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis" in *BMC Medical Research Methodology*, vol. 11: p 160.
- DerSimonian, Rebecca & Laird, Nan (1986). "Meta-Analysis in Clinical Trials" in *Controlled Clinical Trials*, vol 7: p 177–188.
- Dhillon, Amanpreet S. et al. (2008). "Pesticide/Environmental Exposures and Parkinson's Disease in East Texas" in *Journal of Agromedicine*, vol. 13, no. 1: p 37–48.
- Diggle, Peter J. & Gratton, Richard J. (1984). "Monte Carlo Methods of Inference for Implicit Statistical Models" in *Journal of the Royal Statistical Society*, vol 46, no. 2: p 193–227.
- Duong, Tarn (2007). "ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in R" in *Journal of Statistical Software*, vol. 21, issue 7: p 1–16.
- EPA (Environmental Protection Agency) (2021). Web page: epa.gov/ingredients-used-pesticide-products/paraquat-dichloride. Accessed April 17 2022. Page last updated August 21 2021.

- Firestone, Jordan A. et. al. (2010). “Occupational Factors and Risk of Parkinsons Disease: A Population-Based CaseControl Study” *American Journal of Industrial Medicine*, vol. 53, no. 3: p 217-223.
- Gatto, Nicole M. (2009). “Well-Water Consumption and Parkinsons Disease in Rural California” in *Environmental Health Perspectives*, vol. 117, no. 12: p 1912–1918.
- Glass, Gene V. (1977). “Integrating Findings: The Meta-Analysis of Research” in *Review of Research in Education*, vol. 5: p 351–379.
- Greenland, Sander (1996). “Basic Methods for Sensitivity Analysis of Biases” in *International Journal of Epidemiology*, vol 25, no 6: p 1107–1116.
- Hamilton Grant, et. al. (2005). “Bayesian estimation of recent migration rates after a spatial expansion” in *Genetics*, vol. 170: p 409–417.
- Hedges, Larry V. (1992). “Meta-Analysis” in *Journal of Educational Statistics* , vol. 17, no. 4: p 279–296.
- Held, Leonard & Bove, Daniel (2014): *Applied Statistical Inference: Likelihood and Bayes*. Springer.
- Higgins, Julian P. T. & Thompson, Simon G. (2002a). “Quantifying heterogeneity in a meta-analysis” in *Statistics in Medicine*, vol. 21: p 1539-1558.
- Higgins, Julian P. T. & Thompson, Simon G. (2002b). “How should meta-regression analyses be undertaken and interpreted?” in *Statistics in Medicine*, vol. 21: p 1559–1573.
- Hu, Dapeng; O’Connor, Annette M; Chong, Wang; Sargeant, Jan M.; Winder, Charlotte B. (2020). “How to conduct a Bayesian Network meta-analysis” in *Frontiers in Veterinary Science*, vol 7: Article 271.
- Joyce Paul & Marjoram Paul (2008). “Approximately sufficient statistics and Bayesian computation” in *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1: Article 26.
- Kamel, Freya. et. al. (2006). “Pesticide Exposure and Self-reported Parkinsons Disease in the Agricultural Health Study” in *American Journal of Epidemiology*, vol. 165, no. 4: p 364–374.
- Meyer-Baron, Monica, et. al. (2015). “Meta-analysis on occupational exposure to pesticides – Neurobehavioral impact and doseresponse relationships” in *Environmental Research*, vol. 136: p 234–245.
- Nelsen, Roger B. (2006). *An introduction to copulas*. 2nd Edition. Heidelberg: Springer.

Ntzani, Evangelia E, et. al. (2013). “Literature review on epidemiological studies linking exposure to pesticides and health effects”. EFSA supporting publication 2013:EN-497.

Rubin, Donald B. (1996). “Multiple Imputation After 18+ Years” in Journal of the American Statistical Association, vol 91, no 434: p 473–489.

Sain, Stephan R. et. al. (1994). “Cross-Validation of Multivariate Densities” in Journal of the American Statistical Association, vol 89, no 427, p 807–817.

Tanner, Caroline M. et. al. (2009). “Occupation and Risk of Parkinsonism: A Multicenter Case-Control Study” in Archives of Neurology, vol. 66, no. 9: p 1106–1113.

Tanner, Caroline M. et. al. (2011) “Rotenone, Paraquat, and Parkinsons Disease” in Environmental Health Perspectives, vol. 119, no. 6: p 866–872.

Tanner, Caroline M. (2014) “Californias Parkinsons Disease Registry Pilot Project – Coordination Center and Northern California Ascertainment”. A report prepared for the U.S. Army Medical Research and Materiel Command.

Thompson, Simon G & Higgins, Julian (2002). “How should meta-regression analyses be interpreted and undertaken?” in Statistics in Medicine, vol. 21: p 1559-1573.

Wang, Anthony, et al. (2011). “Parkinsons disease risk from ambient exposure to pesticides” in European Journal of Epidemiology, vol. 26, no. 7: 547-555.

Wand, M. P. & Jones, M. C. (1994). “Multivariate plug-in bandwidth selection” in Computational Statistics, vol. 9: p 97–116.