OPTIMAL DESIGN OF PEER EFFECTS EXPERIMENTS WITH EXOGENOUS GROUP SELECTION

Hannah Bull

A thesis submitted in partial fulfilment of the degree of Master of Analysis and Policy in Economics

August 2017

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

Abstract

The best peer effects experiments are not necessarily randomised. In a peer effects experiment with exogenous group selection, accuracy may increase if group assignments are explicitly chosen to take into account the the distributions of individual and group characteristics. We propose five methods to choose experimental designs and demonstrate how these methods perform on simulated data. These methods improve the chances of accurately measuring contagion-type peer effects using linear models and hence better predict how outcomes may improve under alternative group assignments. Additionally, we propose algorithms to quickly find optimal group allocations in the case with leave-out-means or leave-out-variances as linear regressors.

Acknowledgements

Firstly, I would like to thank my supervisor Philipp Ketz for all the time spent introducing me to the world of simulations, theoretical econometrics and potential applications of data science in economics. I would also like to thank Antoine Chambaz for helping me think about causality in a more general way.

I owe much to my family and friends for supporting me over these last few years in Paris, even over long distances. Thank-you to my parents and sisters putting up with me being away all the time. Vielen lieben Dank an die Familie Henne für das zweite Zuhause. Merci beaucoup à mes amis extraordinaires, qui m'ont toujours encouragée à persévérer : Carmelo, François, Sophie, Karolina et Simon. Je tiens à remercier aussi Lucas pour l'inspiration infinie et pour l'aide indispensable.

Contents

1 Introduction

There are numerous methods developed for the purposes of variable selection. Imagine one has a data set with a huge number of variables, not all of which should be incorporated into a model, in order to prevent overfitting. Methods such as lasso, ridge regression and random trees all involve the selection of variables with the highest predictive power, whilst avoiding introducing too much bias. In such set-ups, one has a wide matrix of variables X and a vector of outcomes Y, and chooses an optimal combination or subset of the columns of X.

What if one did not have complete information on the outcome Y? Would it still be possible to choose some subset of X with high predictive power? What if the outcome were dependent on this choice of X? Is is possible to choose some subset of X such that the relationship between X and Y is strongest or most predictable?

These questions arise in the peer effects literature, where the experimenter must choose an allocation of individuals into groups, thus effectively choosing the distribution of each group. Considering all possible choices of allocations of individuals and all corresponding measures of group characteristics as possible variables, the experimenter's problem is one of variable selection. However, the twist is that the outcome Y is not known until after the experiment, and depends on the choice of X.

Intuitively, we know that high variation across the dependent variables is good, and can list some qualitative reasons:

- High variation in the dependent variable should increase the variation in the independent variable, making this variation easier to distinguish from noise or other factors
- If there are a wide range of values for the dependent variable, then we do not need to extrapolate far out of the range of the data in order to make predictions
- There are infinitely many different ways to account for group characteristics, so there should be as much variation within the groups as possible in order to cover as many points in the space of possibilities as possible

The aim of this paper is to precisely define this intuition, explain why it is true and then determine explicit methods to choose experimental designs. We rephrase the question of experimental design in peer effects experiments as one of prediction. Making the strong structural assumption of linearity, we propose various methods of choosing variables which best predict the outcome variable on unseen data.

The paper is structured as follows. Firstly we define the specific type of peer effects that we are interested in, then review numerous experiments identifying this type of peer effects under exogenous selection in the literature. We then state the problem of experimental selection in the most general way

possible, then proceed to add the structural assumptions necessary to explicitly define methods to best choose experiments. Finally, we use the developed methods to identify optimal experimental designs using simulated data.

1.1 Motivation

If the composition of peer groups has a measurable positive or a negative outcome of the individuals involved, and if one has some way of aggregating these positive and negative outcomes, then by rearranging the composition of groups, an optimal outcome can be attained. Rearranging the composition of groups in circumstances where, in any case, group compositions are decided upon by some higher authority is a relatively cheap policy, requiring the authority only to allocate groups in a different way.

Many times, experimenters are forced to rely on natural experiments, over whose design they have little control. Nevertheless, in some cases experimenters have the possibility to assign individuals to groups and to observe outcomes. In the few peer effects experiments conducted whereby the experimenter was able to choose the group allocation, the chosen design was largely intuitive and not necessarily optimal. In this paper, we consider different ways in which peer effects experimental designs can be chosen to best measure the effect of group characteristics on outcomes, as well as to best predict the outcome under a different allocation of individuals into groups.

Although this paper specifically considered the case of peer effects, similar ideas could be applied to any situation where the values of the dependent variables can be exogenously selected and where the distribution of the dependent variables matters for the outcome. For example, in an experiment aiming to predict the effect of the composition of vegetables, grains and meat in a diet on the chances of obesity, the experimenter may be able choose an optimal experimental design covering many combinations of these food items.

1.2 Definition of peer effects

Peer effects manifest themselves in different forms. To quote Shalizi and Thomas (2011):

Suppose that there are two friends named Ian and Joey, and Ian's parents ask him the classic hypothetical of social influence: "If your friend Joey jumped off a bridge, would you jump too?" Why might Ian answer "yes"?

1. *because Joey's example inspired Ian (social contagion/influence);*
2. *because Joey infected Ian with a parasite that suppresses fear of falling (biological contagion);*
3. *because Joey and Ian are friends on account of their shared fondness for jumping off bridges (manifest homophily, on the characteristic of interest);*

4. *because Joey and Ian became friends through a thrill-seeking club, whose membership rolls are publicly available (secondary homophily, on a different yet observed characteristic);*

5. *because Joey and Ian became friends through their shared fondness for roller-coasters, which was caused by their common thrill-seeking propensity, which also leads them to jump off bridges (latent homophily, on an unobserved characteristic);*

6. *because Joey and Ian both happen to be on the Tacoma Narrows Bridge in November 1940, and jumping is safer than staying on a bridge that is tearing itself apart (common external causation).*

In this paper, we are only interested in identifying peer effects of types 1 and 2: that is, contagion. If there is contagion, then individuals outcomes will be affected by the composition of the group. On the contrary, if there is no contagion, then a different allocation of individuals into group will noti necessarily affect the outcome. Specifically, we only consider experimental situations where individuals are sorted into groups by some authority. If we are only interested in the effect of the group composition on the outcome, then the latter effects should be identifiable under an experimental design through exogenous group selection based only on observable characteristics.

This set-up allows us to take into account for homophily and to avoid confusing homophily with contagion. The article Shalizi and Thomas (2011) shows that in situations where individuals are free to self-select into groups, homophily is indistinguishable from contagion under non-parametric assumptions. We will always assume that events such as those of type 6 are random and thus with large enough sample sizes, we can distinguish the effect of contagion from the noise of common external causation.

Examples of common situations involving allocations to peer groups include the sorting of students into classes, the allocation of apartments or rooms in a residence, selection of opponents in the initial round of competitions, the allocation of patients to shared hospital rooms or the allocation of medical staff to different hospital wards.

2 Review of experiments with exogenous selection

In this section, we discuss only on the part of the peer effects literature which attempts to identify the effect of group distributions on individual outcomes. We discuss structural assumptions in peer effects models, experimental design and optimisation of the outcome variable.

2.1 Parametric assumptions

Lack of sufficient data in peer effects scenarios is often compensated with strong parametric assumptions. One common assumption is that the outcome

of individuals is a linear function of their own characteristics as well as some characteristics of the distribution of their peer group.

This linear peer effects model is discussed in detail in Manski (1993) and in Angrist (2014). Manski (1993) discusses the tautological peer effects model, where the outcome variable is also used to measure the characteristics of groups. In this case, clearly the outcome is some result of the composition of the group, as they are both functions of the same variable. Angrist (2014) discusses the merits and drawbacks of experiments where, to avoid this tautology, all individual and group variables are measured at time t and the outcome variable is measured at time $t+1$.

Essentially, the linear peer effects model aims to estimate the parameters in the following equation:

$$y_{i,g} = \alpha + \beta\mu_{i,g} + \gamma x_{i,g} + \delta' w_{i,g} + u_{i,g}$$

where y is an outcome, e.g. a test score; μ is a group characteristic, such as the mean of a characteristic x within an individual's reference group, e.g. the average of past test scores in the reference group; w is a vector of characteristics that affect y, e.g. IQ, socioeconomic status; and u are unobserved characteristics that affect y. The peer effects in this model are captured in the coefficient β, which measures the impact of a group characteristic on the outcome for an individual in that group.

Usually, μ is the leave-out-mean, i.e. the mean of the peers in the group excluding the individual in question, or some other leave-out-group-characteristic. This is to avoid confounding between the effect of the group composition of peers and the effect of individual characteristics.

Caeyers and Fafchamps (2016) discusses the exclusion bias of linear models, due to the fact that an individual can never be in the same group as themselves, therefore group characteristics are always somehow correlated with individual characteristics. Feld and Zölitz (2017) discusses the bias due to measurement error. The fact that individual characteristics can be used twice - once to measure individual effects and once to measure group effects - means that measurement errors can be multiplicative. These two papers show how to account for such bias when applying the linear-in-individual-and-group characteristics model.

A key feature of the particular linear peer effects model involving only the mean or the leave-out-mean as a group characteristic is that the predicted average outcome y across individuals under any group allocation is identical when groups are the same size. Swapping two individuals in two different groups will decrease the leave-out-means in one group but increase the leave-out-means in the other group by the same amount. This model thus cannot be used to predict how different group allocations impact the average outcome of y. However, this linear model can be easily enriched by adding other terms such as the leave-out-variance or the leave-out-standard-deviation, or other non-linear terms measuring group characteristics. The addition of such terms alleviates the absence of policy implications.

As with all linear models of this form, the expectation of the error terms given the individual and group characteristics is assumed to be 0. In the case of peer effects models, this is particularly problematic, as there are infinitely many variables which measure the characteristics of groups, and these are all contained in the error term. Whilst the leave-out-mean and the leave-out-variance are considered to have an effect on the outcome, the leave-out-median or the leave-out-standard-deviation are expected to be 0, given the dependent variables. This assumption is unlikely to be true.

Lee (2007) proposes a linear model analogous to a spatial auto-regressive model, which allows for correlation between unobserved group characteristics and included dependent variables. This model relies on variation within group sizes in order to identify both the reciprocal effect of individuals on each other and the exogenous effect of group composition on the individual outcome.

Pinto (2010) proposes a semi-parametric method to identify peer effects in cases where individuals are exogenously sorted into groups, or alternatively there is an available instrument to separate endogenous selection from exogenous selection. The dependent variables are a single index of individual characteristics and a function of this index to measure peer quality.

Attempts at non-parametrically estimating peer effects suffer from lack of available or suitable data. Nevertheless, there are some methods which may be promising from the literature on network effects. A peer effect can be interpreted in a network context where interactions between individuals are modelled as connections between nodes. The existence or lack of a connection can have an impact on the entire network. Krackhardt (1988) proposes a method to identify whether or not the correlation between two individuals in the same group is statistically significant or not, called the quadratic assignment procedure. This method better predicted behaviour in the presence of peer effects in a monastery than OLS estimates.

2.2 Experimental design

Numerous peer effects experiments rely on natural experiments, where individuals are randomly allocated to groups conditional on a few observable characteristics. Such natural experiments occur when, for example, year groups are split into classes randomly, people are allocated to bedrooms or apartments randomly conditional on some preferences, such as going to bed early or late.

Two examples of natural experiments involving roommate allocation are Zimmerman (2003) and Carrell, Fullerton, and West (2009). Zimmerman (2003) uses a linear model to estimate the effect of being paired with a high, medium or low-SAT roommate at William College. Before entry into the residence, students must fill out a questionnaire regarding their personal living preferences. These questions are argued to be unrelated to the SAT score or academic performance of the roommate. The outcome variable (GPA) is assumed to be a linear combination of individual characteristics, one's own past SAT score and one's roommate's past SAT score. Low-SAT students are found to be negatively affected by other low-SAT students.

The natural experiment studied in Carrell, Fullerton, and West (2009) involved a random allocation of students into dormitories and squadrons at the United States Air Force Academy. Academic performance is modelled as a linear function of individual characteristics, the leave-out-mean of SAT scores of squadron peers and roommates' SAT scores. The leave-out-mean of squadron peers' SAT scores has a positive and significant effect on one's own academic outcome. This effect is potentially stronger for low-SAT individuals, suggesting that the average outcome of the cohort can be increased by choosing a different method to allocated students to groups.

The key problem with these two natural experiments is that there is only slight variation between the distribution of group characteristics across reference groups. If group allocation is essentially random, then the expected distribution of each of the groups is similar to the distribution of the population, and varies little across groups. For example, if students are randomly allocated to squadrons, then the average of student's past high school grades within each squadron is likely to be similar. This makes identification of the effect of the average of students' past high school grades within each group on individual performance difficult to measure, because the effect of such small variations in group averages may be indistinguishable from other effects or noise.

Moreover, if the estimated model from the group experiment suggests that the overall outcome can be improved by a different sorting, the distribution of the new group allocation may be completely different to the observed distributions under random allocation. For example, if the parameter estimates suggest that low-SAT students benefit from being with high-SAT peers, then a group allocation putting all the lowest-SAT students together with all the highest-SAT students will have a very different distribution to the relatively similar mixes of high, middle and low-SAT students under random selection. The model is thus extrapolated outside the range of observed group characteristics observed during the experiment with random selection.

Carrell, Sacerdote, and West (2013) conduct a follow-up experiment to Carrell, Fullerton, and West (2009), which does exactly this. The parameter estimates in Carrell, Fullerton, and West (2009) suggest that low-SAT students disproportionately benefit from being in squadrons with high-SAT peers. After running a second experiment grouping high-SAT students with low-SAT students and grouping mid-SAT students together, the overall impact on low-SAT students was found to be negative. This could be because:

- There was not enough variation in the dependent variables in order to correctly identify the direction of the effect
- The predicted outcome under the proposed group allocation was an extrapolation of model far outside the range of the dependent variables used to estimate the effect during the experiment
- There was some correlated cause, which could be some other group characteristic, but the high similarity across groups did not allow the experimenter to test for other possible measures of group characteristics

In order to avoid some of these issues, some peer effects experiments enforce variation in the group characteristics by not randomly allocating individuals into groups. In Duflo, Dupas, and Kremera (2011) and Booij, Leuven, and Oosterbeek (2015), the experimenters have the luxury of being in charge of group selection. Duflo, Dupas, and Kremera (2011) choose to randomly allocate students to classes in a selection of villages and to compare outcomes with those of students streamed into two levels in the remaining villages. The choice of streaming students into two levels, rather than three levels, or rather than some other group allocation rule is due to the specific question: whether streaming students by two levels is better than a random allocation. If, for example, the question were to identify the effect of the group mean on the outcome, then this is not necessarily the best experimental design.

Booij, Leuven, and Oosterbeek (2015) choose an experimental design wherein students are allocated into tutorial groups such that there are a mix of homogeneous, heterogeneous, high-level and low-level groups. The effect of group characteristics on outcomes is more likely to be identifiable in such experiments, because there is large enough variation across groups in order to draw conclusions about the benefits of streaming classes or the benefits of diversity within a group. The experimenters have access to the high-school results of individuals in three intervals (high, medium and low). They then choose the group allocation somewhat arbitrarily, assuring that there are groups with all high students, all low students, all medium students, and various mixes of high and low, high and medium, and medium and low students. Whilst this is no doubt a good way to ensure that there is large variation amongst the group variables of interest, it is again not necessarily the optimal way to choose groups for the experiment.

2.3 The socially optimal allocation

In many cases, the goal of the social planner is considered to be the average outcome across individuals, or the average outcome across individuals of a certain type. Carrell, Sacerdote, and West (2013) particularly care about the average outcome of low-SAT individuals, and Duflo, Dupas, and Kremera (2011) and Booij, Leuven, and Oosterbeek (2015) care about the overall predicted average outcome as well as the outcome for students in different quantiles. The optimal group allocation is chosen according to the predicted effect on the average outcome, or the average outcome for a particular subset of the population.

If the socially optimal group allocation is the solution to a maximisation problem of an average, then nonlinear terms must be included in the model in order to conclude that one group allocation is better than another. For example, Booij, Leuven, and Oosterbeek (2015) includes both the leave-out-mean and the leave-out-standard-deviation in order to estimate better group allocations.

3 Generalised problem statement

In a perfect world, the experimenter would assign a set of individuals to groups, measure the outcome under the group allocation, go back in time, then reassign the individuals to groups, and again measure the outcome. This process would continue until the experimenter understood which group allocations lead to the best outcomes.

Clearly this is an impossible experimental design, but nevertheless, the effect of assigning individuals to groups in a particular way can be measured from feasible situations. When individuals are randomly allocated to groups or are allocated to groups based on observable characteristics, then measurable outcomes can be attributed to group composition, and not just due to common causes or reasons related to endogenous self-selection. We show this by first framing the problem as a non-parametric structural equations model.

Let $W = (W_1, ..., W_n)$ be a $n \times c$ matrix of c individual characteristics of n individuals. Let $A = (A_1, ..., A_n)$ be the attribution of the n individuals amongst g groups. Let $Y = (Y_1, ..., Y_n)$ be the outcome variable. A is always a partition of the set of individuals, that is, $|A| = n$ and $A_i \in \{1, ..., g\}$ for $1 \leq i \leq n$. No individual is a member of two different groups.

Additionally, we often consider equally-sized groups, that is, $n = kg$ where k is the number of individuals in each group. Then we also have that for $1 \leq j \leq g$:

$$|\{1 \leq i \leq n : A_i = j\}| = k.$$

There exist sources of randomness (U_W, U_A, U_Y) and deterministic functions f_W, f_A, f_Y such that

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(A, W, U_Y) \end{aligned}$$

The counterfactual situation is as follows: Given a rule $r : \mathcal{W} \to \mathcal{A}$:

$$\begin{aligned} W &= f_W(U_W) \\ A &= r(W) \\ Y(r) &= f_Y(r(W), W, U_Y) \end{aligned}$$

The goal is to find the optimal rule:

$$r^* = \text{argmax}_r \mathbb{E}(\phi(Y(r)))$$

where $\phi : \mathcal{Y} \to \mathbb{R}$ is the utility function we wish to maximise. For example, one may want to maximise the sum of the entries of $Y(r)$, equivalently defining

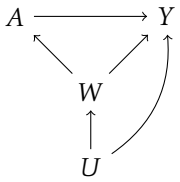
$$\phi : Y(r) \mapsto \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} Y(r).$$

We assume that the function ϕ is known.

3.1 Identification

Using Pearl's structural causal diagrams (Pearl (2009)), we show that the effect of A on Y is identifiable under an experimental design where $A = r(W)$. The assumptions are as in figure 1. The outcome Y depends on A, W and unobserved characteristics U. As A is determined solely by observed characteristics W, all paths between A and Y containing an arrow into A are blocked by W, which is observed. By the back-door criterion, the causal effect of A on Y is identifiable (Theorem 3.3.2 in Pearl (2009)).

Figure 1: Effect of individual and group characteristics on outcome



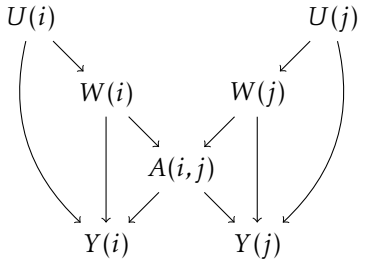
Thus, in order to interpret $\mathbb{E}(Y|A_2, W) - \mathbb{E}(Y|A_1, W)$ as the exogenous effect of a group allocation A_2 as opposed to another group allocation A_1 on the outcome Y, the key assumption is:

$$(A \perp\!\!\!\perp U) \mid W.$$

A rule r is chosen such that $A = r(W)$, thus this conditional independence assumption holds.

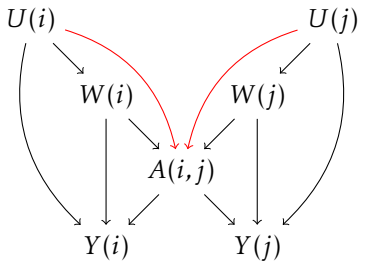
Again using Pearl's causal diagrams, we can look at the specific mechanism in the case of individuals i and j (Figure 2). Unobservable characteristics U influence observable characteristics W, which in turn are used to choose a group allocation A. The outcome Y is the result of unobserved and observed individual and group characteristics. All of the back-door paths between A and Y are blocked by the observed W, and so the effect of the group composition A on Y is identifiable.

Figure 2: Peer effects with exogenous group selection (Shalizi and Thomas (2011))



In the case of endogenous group selection, where individuals may self-select into groups, unobservable characteristics may be responsible for both the outcome and the group allocation. In this case, it is impossible to identify the effect of group composition on the outcome without stronger assumptions, as it may be confounded with unobservables.

Figure 3: Peer effects with endogenous group selection (Shalizi and Thomas (2011))



3.2 Optimal experiments and social optima

There are two key questions:

1. How should a rule r be chosen under experimental conditions in order to best predict the outcome $Y(r')$ or $\phi(Y(r'))$ under an alternative rule r'?
2. How should a rule r^* be chosen in order to maximise $\mathbb{E}(\phi(Y(r^*)))$?

We refer to a rule satisfying the first question as an *experimentally* optimal rule and the rule satisfying the second question as a *socially* optimal rule.

Under ideal experimental conditions, an rule r is chosen, then $Y(r)$ is measured and used to accurately estimate the parameters of the model. Using this information, an estimation for $\mathbb{E}(Y(r'))$ and $\mathbb{E}(\phi(Y(r')))$ for any rule $r' \in \mathcal{R}$ can be made. Using the partial ordering given by the estimation of $\mathbb{E}(\phi(Y(r)))$, a

rule r_1 can be judged to be socially better, worse or equivalent to another rule r_2. A socially optimal rule r^* in the finite space $\mathcal{R}$ can thus always be found.

Of these two questions, the first is the more complex. It is not at all obvious how to choose r in order to best predict the outcome under any other rule r'. The mean squared error is one of many ways of measuring the accuracy of a model, and we use this measure where relevant.

Definition 1 (Mean squared error)**.** *Let $\hat{Y}(r)$ denote the estimated outcome of a model $\mathcal{M}$ whose parameters were estimated using an experimental design r. Let $Y(r)$ be the true outcome. The mean squared error of this model is:*

$$\text{MSE}(\mathcal{M}, r) = \frac{1}{n}(Y(r) - \hat{Y}(r))'(Y(r) - \hat{Y}(r))$$

Definition 2 (Mean squared prediction error)**.** *Let $\hat{Y}(r', r)$ denote the estimated outcome of a model $\mathcal{M}$ applied to data from groups chosen under rule r', but whose parameters were estimated using an experimental design r. Let $Y(r')$ be the true outcome. The mean squared error of prediction of this model is:*

$$\text{MSPE}(\mathcal{M}, r', r) = \frac{1}{n}(Y(r') - \hat{Y}(r', r))'(Y(r') - \hat{Y}(r', r))$$

More precisely, the first question is: how can we choose r to reduce the expected mean squared error of prediction under another group assignment rule r'?

The second question is purely computational. We want to find an optimal $r^* \in \mathcal{R}$ which maximises $\mathbb{E}(\phi(Y(r^*)))$. Given a finite space and a known ordering, a maximum can always be found. In large finite spaces such as $\mathcal{R}$, clearly one needs clever algorithms which are capable of finding an optimal or near optimal solution in a reasonable computational time. Coming up with such algorithms is clearly a difficult problem which lies outside the scope of this paper, and we satisfy ourselves with the knowledge that such a maximum exists.

3.3 Rule selection

The number of ways of dividing n distinct individuals into g groups of k individuals is

$$\frac{n!}{(k!)^g g!}.$$

Allowing for variation in group size, the number of ways of dividing n individuals into g groups is the Stirling number of the second kind, given by

$$\left\{ \frac{n}{g} \right\} = \frac{1}{g!} \sum_{i=0}^{g} (-1)^{g-i} \binom{g}{i} i^n.$$

In general, the number of partitions of a set of size n is the Bell-number B_n, where

$$B_n = \sum_{i=0}^{n} \left\{ \frac{n}{i} \right\}.$$

Even for small n and g, the number of combination is very large. For example, the number of ways of dividing 20 people into 5 groups of 4 people is 2,546,168,625! The space $\mathcal{R}$ of rules $r : \mathcal{W} \to \mathcal{A}$ is finite, but very large.

3.4 Classes of rules

We can also consider rules r as a realisation of a random variable R, for example the outcome group allocation when individuals are randomly sorted into g groups.

Instead of choosing a rule r in $\mathcal{R}$, one can choose a random variable $R : \mathcal{R} \to \mathcal{R}$. The goal can then be interpreted as the optimal choice of R under expectation.

$$R^* = \text{argmin}_R \mathbb{E}(\phi(Y(R)))$$

Some examples of R are:

- Sorting n individuals randomly into g groups of k individuals.
- Sorting n individuals with characteristic $x \in \mathbb{R}$ into g groups, by first sorting them into two levels $x \leq x'$ and $x > x'$, then further sorting the individuals in each level into a and b number of groups respectively, where $a + b = g$.
- Sorting n individuals with characteristic $x \in \mathbb{R}$ into g groups of k individuals in order of x.

4 Linear-in-group-characteristics models

A common structural assumption in peer-effects experiments is a linear-in-group-characteristics assumption such as

$$Y = f_Y(A, W, U_Y) = \beta_0 + l_1(A, W)\beta_1 + l_2(W)\beta_2 + U_Y \tag{1}$$

where l_1 is a linear function of group characteristics, l_2 is a linear function of individual characteristics and where $\mathbb{E}(U_Y|A, W) = 0$.

For example, Manski's linear-in-means model in Manski (1993) takes the following functional form:

$$Y = f_Y(A, W, U_Y) = \beta_0 + \mu(A, W)\beta_1 + x(W)\beta_2 + U_Y$$

where $x(W)$ are the individual characteristics of person i in group g and $\mu(A, W)$ is the mean (or leave-out-mean) of some characteristics of the group.

In this section, we consider all linear-in-group-characteristics models, which includes all models with leave-out-means, leave-out-variances, higher-order terms, non-linear terms, individual characteristics etc. as dependent variables.

We assume that the error terms are homoscedastic and independent of A and W. We assume without loss of generality that all observed data has been demeaned, omitting a constant term that can be later added if required.

4.1 A socially optimal rule

In this section, we suppose that the social planner wishes to optimise the average outcome Y. This is of course equivalent to maximising the sum of the entries of Y.

We firstly show if the outcome is a linear function of the mean or leave-out-mean of one or more individual characteristics, as well as individual characteristics, then the expected outcome under all group allocations is equivalent.

Proposition 1. *Assume* $|\{1 \leq i \leq n : A_i = j\}| = k$*, i.e each group has the same number of individuals. Let*

$$Y(r) = X(r)\beta + U_Y$$

where X is a $p \times n$ matrix whose columns correspond to either means or leave-out-means of an observed variable within the allocated group of an individual, or an individual characteristic, under a group allocation $r \in \mathcal{R}$. Let

$$\phi : Y(r) \mapsto \begin{bmatrix} 1 & 1 & ... & 1 \end{bmatrix} Y(r).$$

Then $\mathbb{E}(\phi(Y(r_1))) = \mathbb{E}(\phi(Y(r_2)))$ *for all* $r_1, r_2 \in \mathcal{R}$.

Proof.

$$\mathbb{E}(\phi(Y(r))) = \sum_{s=1}^{n} (\beta_1 x_{s,1}(r) + \beta_2 x_{s,2}(r) + ... + \beta_p x_{s,p}(r)) \tag{2}$$

where β_j is the j-th entry of β and $x_{i,j}$ is the (i, j)-th entry of X. We show that for all $r_1, r_2 \in \mathcal{R}$ and for all $j \in \{1, ..., p\}$,

$$\sum_{s=1}^{n} \beta_j x_{s,j}(r_1) = \sum_{s=1}^{n} \beta_j x_{s,j}(r_2). \tag{3}$$

By assumption, $x_{s,j}(r)$ is a vector of individual characteristics, group means or leave-out-means. If $x_{s,j}(r)$ is a vector of individual characteristics, equation 3 holds as the individual characteristics do not depend on the group allocation by experimental design.

If $x_{s,j}(r)$ is the mean or the leave-out-mean of a characteristic c_s for an individual s, then

$$\sum_{s=1}^{n} \beta_j x_{s,j}(r) = \sum_{s=1}^{n} \beta_j c_s$$

which does not depend on r. Thus equation 3 holds. □

Suppose $x_{s,1}(r)$ in equation 2 is the variance or the leave-out-variance. Then $\mathbb{E}(\phi(Y(r)))$ does depend on the group allocation rule r, because the sum of $x_{s,1}(r)$ across individuals depends on the group allocation. It is possible to swap the group allocation of two individuals and increase the variances and leave-out-variances of both the groups concerned, leaving the variance of other groups unchanged.

Suppose the first column of X is the group variance or leave-out-variance and the rest of the columns of X are individual characteristics, group means or group leave-out-means. If we wish to maximise $\mathbb{E}(\phi(Y(r)))$, then we must maximise $\sum_{s=1}^{n} x_{s,1}(r)$ if we think that β_1 is positive and minimise this expression if we think that β_1 is negative. The social optimal thus entirely depends on the knowledge of whether heterogeneity (as measured by the variance) within groups is good or bad. In order to determine the social optimum, we only need to know the sign of the coefficient on the variance or leave-out-variance.

Note that it is of course possible to simultaneously increase the sum of the group leave-out-variances across individuals and decrease the sum of the group leave-out-standard-deviations across individuals. In general

$$a_1^2 + ... + a_n^2 < b_1^2 + ... + b_n^2$$

does not imply that

$$|a_1| + ... + |a_s| < |b_1| + ... + |b_s|.$$

The choice of including leave-out-variances instead of leave-out-standard-deviations in a linear model can be somewhat arbitrary, and so the interpretation of the coefficient on any of these terms should be interpreted with care. One should also keep in mind that by assumption, in a linear model with leave-out-variances instead of leave-out-standard-deviations, the expectation of the latter given the former is 0. This is unlikely to be true.

4.2 Choosing an experimentally optimal rule

How should a rule r be chosen under experimental conditions in order to best estimate the parameters of the linear functions l_1 and l_2? Under the assumption that $\mathbb{E}(U_Y|A, W) = 0$, the estimates of the linear coefficients using ordinary least squares estimation are unbiased. The goal is thus to find estimates which also have low variance.

Although individual characteristics W are fixed, A depends on the experimental design r, as $A = r(W)$. The variance and the co-variance of the parameters depends on the variance and co-variance of W and A. Let

$$X = \begin{bmatrix} l_1(W) & l_2(A, W) \end{bmatrix}$$

and

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Then

$$Y = X\beta + U_Y$$

and

$$\mathrm{Var}(\hat{\beta}) = (X'X)^{-1} X \mathbb{E}(U_Y U_Y') X' (X'X)^{-1}.$$

By the homoskedasticity assumption,

$$\mathrm{Var}(\hat{\beta}) = (X'X)^{-1}\sigma^2$$

where σ^2 is the variance of the error terms U_Y. Note that U_Y is independent of X by assumption, so $\mathrm{Var}(\hat{\beta})$ depends only on the distribution of X. Choosing the values for X is essentially choosing the values for $\mathrm{Var}(\hat{\beta})$.

The expected mean squared error of a linear model with homoscedastic errors is identical for all rules. Let $X(r)$ denote the matrix of individual and group characteristics under rule r and let $\hat{\beta}(r)$ be the estimate of β from such an experiment.

$$\begin{aligned}
\mathbb{E}(\mathrm{MSE}(\mathcal{M}, r)) &= \mathbb{E}\left(\frac{1}{n}(Y(r) - \hat{Y}(r))'(Y(r) - \hat{Y}(r))\right) \\
&= \frac{1}{n}\left(\mathbb{E}((X(r)\beta - X(r)\hat{\beta}(r))'(X\beta - X\hat{\beta}(r))) + \mathbb{E}(U_Y U_Y')\right) \\
&= \frac{1}{n}\left(\mathbb{E}((\beta - \hat{\beta}(r))'X(r)'X(r)(\beta - \hat{\beta}(r))) + n\sigma^2\right) \\
&= \frac{1}{n}\mathrm{Tr}\left(\mathrm{Var}(\hat{\beta}(r))(X(r)'X(r))\right) + \sigma^2 \\
&= \frac{1}{n}\mathrm{Tr}\left(\sigma^2(X(r)'X(r))^{-1}(X(r)'X(r))\right) + \sigma^2 \\
&= \frac{p+n}{n}\sigma^2
\end{aligned}$$

where p is the number of parameters.

This does not depend on the entries of X. Therefore, the expected mean squared error cannot be used to distinguish different group allocation rules.

However, different group allocation rules lead to different expected mean squared errors of prediction. Let r_T denote a new group allocation rule, which was not used to estimate the parameters (i. e. $\{Y(r_T), X(r_T)\}$ is the test set).

$$\begin{aligned}
\mathbb{E}(\text{MSPE}(\mathcal{M}, r_T, r)) &= \mathbb{E}\left(\frac{1}{n}(Y(r_T) - \hat{Y}(r_T, r))'(Y(r_T) - \hat{Y}(r_T, r))\right) \\
&= \frac{1}{n}\left(\mathbb{E}((X(r_T)\beta - X(r_T)\hat{\beta}(r))'(X(r_T)\beta - X(r_T)\hat{\beta}(r))) + \mathbb{E}(U_Y U_Y')\right) \\
&= \frac{1}{n}\left(\mathbb{E}((\beta - \hat{\beta}(r))'X(r_T)'X(r_T)(\beta - \hat{\beta}(r))) + n\sigma^2\right) \\
&= \frac{1}{n}\text{Tr}\left(\text{Var}(\hat{\beta}(r))(X(r_T)'X(r_T))\right) + \sigma^2
\end{aligned}$$

The value of this last expression depends on both the data $X(r_T)$ and $\text{Var}(\hat{\beta}(r))$.

In general, it is not possible to choose an $r \in \mathcal{R}$ which minimises the expected mean squared prediction error for arbitrary test data $X(r_T)$, as we discuss in the next section.

4.3 Positive semidefinite case

We show that if $\text{Var}(\hat{\beta}(r')) - \text{Var}(\hat{\beta}(r))$ is a positive semidefinite matrix, then the expected mean squared prediction error using experimental design r' is greater than or equal to the expected mean squared prediction error using experimental design r.

Proposition 2. *Suppose* $\text{Var}(\hat{\beta}(r')) - \text{Var}(\hat{\beta}(r))$ *is a positive semidefinite matrix and let* r_T *be any group allocation rule. Then* $\mathbb{E}(\text{MSPE}(\mathcal{M}, r_T, r')) \geq \mathbb{E}(\text{MSPE}(\mathcal{M}, r_T, r))$.

Proof. $X(r_T)'X(r_T)$ is a positive semidefinite matrix, and the trace of the product of two positive semidefinite matrices is greater than or equal to zero. Thus we have:

$$\begin{aligned}
&\text{Tr}\left((\text{Var}(\hat{\beta}(r')) - \text{Var}(\hat{\beta}(r)))(X(r_T)'X(r_T))\right) \geq 0 \\
\implies &\text{Tr}\left(\text{Var}(\hat{\beta}(r'))(X(r_T)'X(r_T))\right) \geq \text{Tr}\left(\text{Var}(\hat{\beta}(r))(X(r_T)'X(r_T))\right) \\
\implies &\text{Tr}\left(\text{Var}(\hat{\beta}(r'))(X(r_T)'X(r_T))\right) + \sigma^2 \geq \text{Tr}\left(\text{Var}(\hat{\beta}(r))(X(r_T)'X(r_T))\right) + \sigma^2 \\
\implies &\mathbb{E}(\text{MSPE}(\mathcal{M}, r_T, r')) \geq \mathbb{E}(\text{MSPE}(\mathcal{M}, r_T, r))
\end{aligned}$$

□

Using this result, we can now define a partial ordering on any subset of rules $\mathcal{R}_{PSD} \subset \mathcal{R}$ such that for any $r_1, r_2 \in \mathcal{R}_{PSD}$, either $\text{Var}(\hat{\beta}(r_2)) - \text{Var}(\hat{\beta}(r_1))$ is positive semidefinite or $\text{Var}(\hat{\beta}(r_1)) - \text{Var}(\hat{\beta}(r_2))$ is positive semidefinite. Note that $\text{Var}(\hat{\beta}(r_2)) - \text{Var}(\hat{\beta}(r_1))$ is positive semidefinite if and only if $(X(r_2)'X(r_2))^{-1} - (X(r_1)'X(r_1))^{-1}$ is positive semidefinite, as $\text{Var}(\hat{\beta}) = (X'X)^{-1}\sigma^2$. We define the following partial ordering on $\mathcal{R}_{PSD}$.

Definition 3. *Let r_1 and r_2 be elements of the rule space $\mathcal{R}_{PSD}$. Define $r_2 \preceq r_1$ if and only if $(X(r_2)'X(r_2))^{-1} - (X(r_1)'X(r_1))^{-1}$ is a positive semidefinite matrix.*

By proposition 2, the ordering given in definition 3 says that a rule r_1 is better than a rule r_2 if the expected mean squared error using parameters estimated with rule r_1 is less than or equal to the expected mean squared error using parameters estimated with rule r_2. We can easily identify the optimal experimental design rule in $\mathcal{R}_{PSD}$, which gives the lowest mean squared prediction errors, by looking at the values of $(X'X)^{-1}$ under different group allocations.

We now identify various properties of this partial ordering. We will later use these properties to define possible partial orderings on the whole of $\mathcal{R}$, which are equivalent to $\preceq$ on any subset $\mathcal{R}_{PSD}$ as previously defined.

The following proposition is used without proof, but the proofs of these features of semidefinite matrices can be found in many linear algebra textbooks.

Proposition 3 (Well-known properties of symmetric positive semidefinite matrices)**.** *Let $A, B \in \mathbb{R}^n \times \mathbb{R}^n$ be symmetric positive semidefinite matrices with i-th largest eigenvalue denoted by λ_i:*

1. *A^{-1} is symmetric positive semidefinite*
2. *$\lambda_i(A) \geq 0$ for all i*
3. *$\mathrm{Tr}(A) = \sum_i \lambda_i(A)$*
4. *$\mathrm{Tr}(A^{-1}) = \sum_i \frac{1}{\lambda_i(A)}$*
5. *If $A - B$ is a positive semidefinite matrix, then $\lambda_i(A) \geq \lambda_i(B)$ for all i*

Lemma 1. *Let $\mathrm{Var}(\hat{y}_i(r))$ denote the variance of the i-th entry of the vector $\hat{Y}$, where this estimate is attained using the estimated coefficients $\hat{\beta}$ from an experimental design r. Then $r_2 \preceq r_1$ if and only if $\mathrm{Var}(\hat{y}_i(r_2)) \geq \mathrm{Var}(\hat{y}_i(r_1))$ for all $1 \leq i \leq n$ and for any data X (test or training data).*

Proof. Note that $\mathrm{Var}(\hat{y}_i) = x_i \mathrm{Var}(\hat{\beta}) x_i'$, where x_i is the i-th row of X.

($\implies$) Assume $r_2 \preceq r_1$. Then $\mathrm{Var}(\hat{\beta}(r_2)) - \mathrm{Var}(\hat{\beta}(r_1))$ is a positive semidefinite matrix. Thus $x_i'(\mathrm{Var}(\hat{\beta}(r_2)) - \mathrm{Var}(\hat{\beta}(r_1)))x_i \geq 0$ for any x_i and the result follows.

($\impliedby$) Assume $\mathrm{Var}(\hat{y}_i(r_2)) \geq \mathrm{Var}(\hat{y}_i(r_1))$ for all $1 \leq i \leq n$ and for any data X. By definition $x_i \mathrm{Var}(\hat{\beta}(r_2)) x_i' \geq x_i \mathrm{Var}(\hat{\beta}(r_1)) x_i'$ for all $1 \leq i \leq n$ and for any data X. Therefore $x_i(\mathrm{Var}(\hat{\beta}(r_2)) - \mathrm{Var}(\hat{\beta}(r_1)))x_i' \geq 0$ for any x_i and $\mathrm{Var}(\hat{\beta}(r_2)) - \mathrm{Var}(\hat{\beta}(r_1))$ is a positive semidefinite matrix. Thus $r_2 \preceq r_1$. □

Lemma 2. *Suppose $r_2 \preceq r_1$. Let r' be any other group allocation rule. Then $\hat{\beta}(r')'(X(r_2)'X(r_2))\hat{\beta}(r') \leq \hat{\beta}(r')'(X(r_1)'X(r_1))\hat{\beta}(r')$.*

Proof. As $r_2 \preceq r_1$, $(X(r_2)'X(r_2))^{-1} - (X(r_1)'X(r_1))^{-1}$ is a positive semidefinite matrix. Moreover $X(r')'X(r')$ is a positive semidefinite matrix. Thus,

$$
\begin{aligned}
&\operatorname{Tr}\Big(((X(r_2)'X(r_2))^{-1} - (X(r_1)'X(r_1))^{-1})(X(r')'X(r'))\Big) \geq 0 \\
\implies &\operatorname{Tr}\Big((X(r_2)'X(r_2))^{-1}(X(r')'X(r'))\Big) \geq \operatorname{Tr}\Big((X(r_1)'X(r_1))^{-1}(X(r')'X(r'))\Big) \\
\implies &\operatorname{Tr}\Big((X(r_2)'X(r_2))(X(r')'X(r'))^{-1}\Big) \leq \operatorname{Tr}\Big((X(r_1)'X(r_1))(X(r')'X(r'))^{-1}\Big) \\
\implies &\operatorname{Tr}\Big((X(r_2)'X(r_2))\sigma^2(X(r')'X(r'))^{-1}\Big) \leq \operatorname{Tr}\Big((X(r_1)'X(r_1))\sigma^2(X(r')'X(r'))^{-1}\Big) \\
\implies &\operatorname{Tr}\Big((X(r_2)'X(r_2))\operatorname{Var}(\hat{\beta}(r'))\Big) \leq \operatorname{Tr}\Big((X(r_1)'X(r_1))\operatorname{Var}(\hat{\beta}(r'))\Big) \\
\implies &\hat{\beta}(r')'(X(r_2)'X(r_2))\hat{\beta}(r') \leq \hat{\beta}(r')'(X(r_1)'X(r_1))\hat{\beta}(r')
\end{aligned}
$$

The third line is due to proposition 3.

□

Lemma 3. *Suppose $r_2 \preceq r_1$. Then*

$$\operatorname{Tr}((X(r_2)'X(r_2))) \leq \operatorname{Tr}((X(r_1)'X(r_1)))$$

and

$$\operatorname{Det}((X(r_2)'X(r_2))) \leq \operatorname{Det}((X(r_1)'X(r_1))).$$

Proof. We have that $(X(r_2)'X(r_2))^{-1} - (X(r_1)'X(r_1))^{-1}$, $(X(r_1)'X(r_1))^{-1}$ and $(X(r_1)'X(r_1))^{-1}$ are symmetric positive semidefinite matrices. Thus,

$$
\begin{aligned}
&\operatorname{Tr}((X(r_2)'X(r_2))^{-1} - (X(r_1)'X(r_1))^{-1}) \geq 0 \\
\implies &\operatorname{Tr}(X(r_2)'X(r_2))^{-1}) \geq \operatorname{Tr}(X(r_1)'X(r_1))^{-1}) \\
\implies &\operatorname{Tr}(X(r_2)'X(r_2))) \leq \operatorname{Tr}(X(r_1)'X(r_1))).
\end{aligned}
$$

The i-th largest eigenvalue of $(X(r_2)'X(r_2))^{-1}$ is always greater than or equal to the i-th largest eigenvalue of $(X(r_1)'X(r_1))^{-1}$ (proposition 3). As the determinant is the product of the eigenvalues, $\operatorname{Det}((X(r_2)'X(r_2))^{-1}) \geq \operatorname{Det}((X(r_1)'X(r_1))^{-1})$. The determinant of an inverse of a matrix is the inverse of the determinant, which implies the result.

□

We have shown that if $(X(r_2)'X(r_2))^{-1} - (X(r_1)'X(r_1))^{-1}$ is a positive semidefinite matrix, then we can say that experimental design r_1 estimates the parameters of the linear model with more accuracy and provides better predictions than experimental design r_2. The expected variance of the model parameters using the experimental design r_1 is always lower than using the experimental design r_2 by lemma 1. The expected mean squared prediction error is also lower by proposition 2.

Unfortunately, for not all choices of rules $r_1, r_2 \in \mathcal{R}$ will it be true that $(X(r_2)'X(r_2))^{-1} - (X(r_1)'X(r_1))^{-1}$ and $\operatorname{Var}(\hat{\beta}(r_2)) - \operatorname{Var}(\hat{\beta}(r_1))$ are positive semidefinite matrices. In this case, it is not obvious to say whether or not rule r_1 is better or worse than r_2 using the criteria of the mean squared prediction error.

Suppose the variances of the two $\hat{\beta}$ under different rules are

$$\text{Var}(\hat{\beta}(r_1)) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\text{Var}(\hat{\beta}(r_2)) = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}.$$

Note that $\text{Var}(\hat{\beta}(r_2)) - \text{Var}(\hat{\beta}(r_1))$ is clearly not a positive semidefinite matrix. Suppose we have two sets of test data

$$X(r_T) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$X(r_t) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

Then

$$\begin{aligned} \mathbb{E}(\text{MSPE}(r_T, r_1)) &= \text{Tr}(X(r_T)'X(r_T)\text{Var}(\hat{\beta}(r_1))) + 2\sigma^2 = 2 + 2\sigma^2 < \\ \mathbb{E}(\text{MSPE}(r_T, r_2)) &= \text{Tr}(X(r_T)'X(r_T)\text{Var}(\hat{\beta}(r_2))) + 2\sigma^2 = 2.5 + 2\sigma^2 \end{aligned}$$

but

$$\begin{aligned} \mathbb{E}(\text{MSPE}(r_t, r_1)) &= \text{Tr}(X(r_T)'X(r_T)\text{Var}(\hat{\beta}(r_1))) + 2\sigma^2 = 5 + 2\sigma^2 > \\ \mathbb{E}(\text{MSPE}(r_t, r_2)) &= \text{Tr}(X(r_T)'X(r_T)\text{Var}(\hat{\beta}(r_2))) + 2\sigma^2 = 4 + 2\sigma^2. \end{aligned}$$

Also, using test data $X(r_T)$,

$$1 = \text{Var}(\hat{y}_1(r_1)) \leq \text{Var}(\hat{y}_2(r_2)) = 2$$

but

$$2 = \text{Var}(\hat{y}_1(r_1)) > \text{Var}(\hat{y}_2(r_2)) = 0.5.$$

The mean squared error of prediction using estimates from experimental design rule r_1 is lower than that from using design rule r_2 on the test data set $X(r_T)$ but the mean squared error of prediction using estimates from experimental design rule r_1 is higher than that from using design rule r_2 on the test data set $X(r_t)$. Moreover, neither rule r_1 nor rule r_2 gives consistently lower variances estimates of the entries of Y. Unfortunately, this means that the mean squared error of prediction cannot be used to define a partial ordering on $\mathcal{R}$ in the same way used to define a partial ordering on $\mathcal{R}_{PSD}$ as in definition 3.

In peer effects experiments, there may be a trade-off between the variance of two entries of the vector $\hat{\beta}$, leading to variance matrices such as $\text{Var}(\hat{\beta}(r_1))$

and $\text{Var}(\hat{\beta}(r_2))$ above. For example, if both group mean and group variance are determinants of Y, then a good strategy to increase the variance across group means can be a poor strategy to increase the variance across group variances. Imagine one is required to sort students into classes knowing only their GPA. A good way to increase the variance across group means is to sort students into GPA quantiles. Yet such classes would be relatively homogeneous. Students would have similar GPAs to those of their classroom peers. The estimated coefficient on the leave-out-mean would have low variance but the estimated coefficient on the leave-out-variance would have high variance. A different group allocation strategy would decrease the variation in the leave-out-means across groups but may increase the variation in the leave-out-variances.

4.4 Five methods of ranking experimental rules

We now suggest a range of possible partial orderings on $\mathcal{R}$. The first 4 partial orderings only use the information of the data X. The final partial ordering requires some prior estimate of β.

Before listing these five methods, we begin by noting the obvious experimental choice when the objective is to predict an outcome for a given rule r. We note that of the mean squared error of prediction is lowest and identical to the mean squared error when the both are computed using the same data $X(r)$. As ordinary least squares estimation by definition reduces the mean squared error, if the experimenter knows that the purpose of the experiment is to predict the outcome for data $X(r)$, the best experiment uses data $X(r)$ to estimate the parameters of the model. In this section, we consider the more general case, where the experimenter wishes to best predict outcomes using any data X.

4.4.1 Ordering by the variance of one coefficient of interest

Suppose that one only cares about the coefficient of one linear term of the model, say β_i, the i-th entry of β. This could occur when the socially optimal group allocation merely depends on the sign or the value of a particular term, such as in the case where the social planner wishes to maximise the average outcome and the model includes as group characteristics the leave-out-mean and the leave-out-variance. The socially optimal group allocation can be determined as soon as the sign of the coefficient of the leave-out-variance is known. This is discussed in section 4.1.

As the estimates of β are unbiased, we simply wish to reduce the variance of the estimate of β_i.

$$\text{Var}(\hat{\beta}_i) = (X'X)^{-1}_{i,i}\sigma^2$$

where $(X'X)^{-1}_{i,i}$ is the i-th diagonal of $(X'X)^{-1}$.

Definition 4. *Define the partial ordering $\preceq_{1Ci}$ on $\mathcal{R}$ to be such that $r_2 \preceq_{1Ci} r_1$ if and only if* $\text{Var}(\hat{\beta}_i(r_2)) \geq \text{Var}(\hat{\beta}_i(r_1))$*, or equivalently, if and only if* $(X(r_2)'X(r_2))^{-1}_{i,i} \geq (X(r_1)'X(r_1))^{-1}_{i,i}$.

Clearly, on $\mathcal{R}_{PSD}$, this partial ordering is equivalent to that in definition 3, as the diagonal entries of a symmetric positive semi-definite matrix are non-negative.

4.4.2 Ordering by trace

Suppose one wishes to simply minimise the sum of the variances of each of the entries of $\hat{\beta}$, i. e. the trace of $\text{Var}(\hat{\beta})$.

Definition 5. *Define the partial ordering $\preceq_{Tb}$ on $\mathcal{R}$ to be such that $r_2 \preceq_{Tb} r_1$ if and only if* $\text{Tr}(\text{Var}(\hat{\beta}(r_2))) \geq \text{Tr}(\text{Var}(\hat{\beta}(r_1)))$*, or equivalently if and only if*

$$\text{Tr}((X(r_2)'X(r_2))^{-1}) \geq \text{Tr}((X(r_1)'X(r_1))^{-1}).$$

This partial ordering is cleaarly equivalent to the partial ordering in definition 3 on $\mathcal{R}_{PSD}$.

One may also wish to increase the variance of each of the dependent variables. On $\mathcal{R}_{PSD}$, this is equivalent by lemma 3.

Definition 6. *Define the partial ordering $\preceq_{Tx}$ on $\mathcal{R}$ to be such that $r_2 \preceq_{Tx} r_1$ if and only if* $\text{Tr}(X(r_2)'X(r_2)) \leq \text{Tr}(X(r_1)'X(r_1))$.

By lemma 3, this partial ordering is equivalent to the partial ordering in definition 3 on $\mathcal{R}_{PSD}$, but in general on all of $\mathcal{R}$ it is not true that $\preceq_{U\ Tb}$ is equivalent to $\preceq_{U\ Tx}$, as in general it is not true that

$$\text{Tr}(A) \geq \text{Tr}(B) \iff \text{Tr}(A^{-1}) \geq \text{Tr}(B^{-1}).$$

However, this partial ordering using the trace of the variance of $\hat{\beta}$ or X does not the covariances between the estimates of different entries of β or the different columns of X are not taken into account. This is problematic, because the variance of $\hat{y}_i$ is a function of both the variance and the covariance of $\hat{\beta}$ and X, and so the precision of the estimates of the outcome is not necessarily a decreasing function in the trace of $\text{Var}(\hat{\beta})$ or an increasing function in the trace of $X'X$.

4.4.3 Ordering by determinant or Shannon entropy

Similarly, by lemma 3, we could also define a partial ordering using the determinant of $X'X$, which is equivalent to the partial ordering in definition 3 on $\mathcal{R}_{PSD}$.

Definition 7. *Define the partial ordering $\preceq_D$ on $\mathcal{R}$ to be such that $r_2 \preceq_D r_1$ if and only if* $\text{Det}(X(r_2)'X(r_2)) \leq \text{Det}(X(r_1)'X(r_1))$.

The rationale for defining an ordering on $\mathcal{R}$ this way is at first unclear. As a first observation, consider a simple scenario with two regressors.

$$X'X = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

We have $\text{Det}(X'X) = ac - b^2$. Increasing the variance ($= a$ or c) of a variable increases the determinant but increasing the covariance ($= b$) between the two variables is penalised. Unlike the ordering from definition 6, the covariance between regressors affects the ordering.

More rigorously, we can justify the use of the determinant of the variance matrix of X in finding a partial ordering on $\mathcal{R}$ using information theory.

Information can be measured by Shannon's entropy H.

Definition 8. *[Shannon entropy] The Shannon entropy of a discrete random variable Z with probability mass function P is given by*

$$H(Z) = \text{E}(\text{I}(Z)) = \mathbb{E}(-\ln(\text{P}(Z)))$$

where I *is the information.*

Rare events thus have high entropy and high information.

For example, say a population of 400 has a characteristic x which follows a standard normal distribution $\mathcal{N}(0,1)$. These individuals are divided into groups of 20. We observe that for some of the groups, the leave-out-mean is around 3. This is unlikely to occur, as there are relatively few group configurations which place individuals with very high values for x all in the same group. On the other hand, there are many group configurations which give values for the leave-out-mean at around 0, the population mean. A value of 3 for the leave-out-mean thus contains more information than a value of 0. For example, we know that the variance within a group with a leave-out-mean of around 3 must be very low, as it is highly unlikely that an individual has a characteristic x much larger than 3. In a group with leave-out-mean 0, a much more probable outcome, there are many different possibilities for the leave-out-variance.

Having more information about the distribution of the group is a desirable characteristic for an experiment aiming to identify the effect of the distribution of the group on the individuals in that group. Therefore, it makes sense to increase the entropy of the matrix of dependent variables X.

The differential entropy of a continuous multivariate normal distribution is

$$H(X) = \frac{p}{2}\ln(2\pi e) + \frac{1}{2}\ln(\text{Det}(X'X))$$

where p is the number of columns of X. This is an increasing function in $\text{Det}(X'X)$. The partial ordering $\preceq_D$ is thus equivalent to a partial ordering

inherited from the entropy of X under a different experimental design rules. Unfortunately this monotone relationship between the determinant of the variance matrix of X relies on the assumption that X follows a multivariate normal distribution. The determinant of the variance matrix does not increase with the Shannon entropy for all distributions. Nevertheless, this monotone relationship suggests that the determinant of the matrix $X'X$ measures some kind of desirable attribute for a group allocation.

Singh et al. (2003) discusses the use of the determinant of $X'X$ as an estimate of the entropy of X. This paper also suggests a better estimate of the entropy of X for the more general case where X does not follow a multivariate normal distribution. The method calculates the average distance between K nearest neighbours in order to estimate the probability or a point falling in a particular region of the space. This estimator could also be used to compute Shannon entropy.

The idea behind Shannon entropy is that group configurations which are unlikely to occur contain more information. We thus wish to choose group allocations leading to values of X which are unlikely to occur. We can estimate the most likely values for the variables measuring group characteristics, and then use a distance measure between these values and the group characteristic values under alternate group selection rules.

4.4.4 Ordering by symmetric Kullback-Leibler divergence

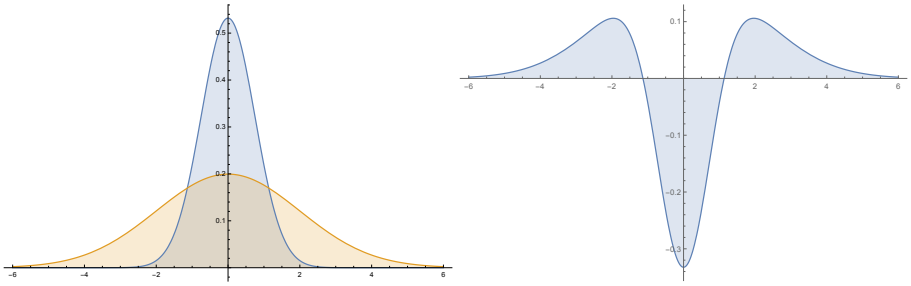
Suppose $R_M : \mathcal{R} \to \mathcal{R}$ is a random variable as described in section 3.4 corresponding to sorting individuals randomly into groups (R_M for *RandoM*). We can then estimate the the distribution of $X(R_M)$. For example, for a group characteristic g, usually $\mathbb{E}(g(R_M))$ will be close to the characteristic calculated for the whole population. If g is the leave-out-mean of an individual characteristic I, then $\mathbb{E}(g(R_M))$ will be the mean of I in the whole population.

We thus wish to choose a group allocation rule which leads to group characteristics having improbable values. This can be achieved by choosing r' such that the distribution of $X(r')$ is 'far away' from the distribution of $X(R_M)$. The notion of 'far away' requires a distance measure on the distribution of X. There are many possible measures of distance between distributions, such as Kullback-Leibler divergence, Rényi's divergence, Heillinger distance, energy distance etc.. As an example, we choose to measure the statistical difference between distributions by symmetric Kullback-Leibler divergence, due to its relationship with Shannon entropy.

Definition 9. *[Kullback-Leibler divergence] The Kullback-Leibler divergence of a random variable Z from a random variable X is defined as*

$$D_{\mathrm{KL}}(Z\|X) = \mathbb{E}\left(P(Z)\ln\frac{P(Z)}{P(X)}\right).$$

Figure 4: Kullback-Leibler divergence of a normal distribution with mean 0 and variance 2 from a normal distribution with mean 0 and variance 0.75. The Kullback-Leibler divergence is the area of the difference between the probability distribution functions.



Definition 10. *[Symmetric Kullback-Leibler divergence] The symmetric Kullback-Leibler divergence of between random variables Z and X is defined as*

$$D_{\mathrm{SKL}}(Z,X) = D_{\mathrm{SKL}}(X,Z) = D_{\mathrm{KL}}(Z\|X) + D_{\mathrm{KL}}(X\|Z).$$

We use the symmetric Kullback-Leibler divergence as opposed to the non-symmetric Kullback-Leibler divergence because it is a metric on the space of X. The distance from $X(r)$ to $X(r')$ is identical to the distance from $X(r')$ to $X(r)$.

Definition 11. *Define the partial ordering $\preceq_{SKL}$ on $\mathcal{R}$ to be such that $r_2 \preceq_{SKL} r_1$ if and only if $D_{\mathrm{SKL}}(X(r_2), X(R_M)) \leq D_{\mathrm{SKL}}(X(r_1), X(R_M))$.*

This is a formal definition of the intuitive idea that groups should be as diverse as possible in all of the dimensions of X. The variance and the covariance of the dependent variables should be far away from that which occurs 'naturally', when groups are chosen at random.

In the case where the rows of X follow a multivariate normal distribution, the Kullback-Leibler divergence is given by:

$$D_{\mathrm{KL}}(X(r)\|X(R_M)) = \frac{1}{2}\left(\mathrm{Tr}\left((X(R_M)'X(R_M))^{-1}X(r)'X(r)\right) - p + \ln\left(\frac{\mathrm{Det}(X(R_M)'X(R_M))}{\mathrm{Det}(X(r)'X(r))}\right)\right).$$

The symmetric Kullback-Leibler divergence is thus given by:

$$D_{\mathrm{SKL}}(X(r), X(R_M)) = \frac{1}{2}\left(\mathrm{Tr}\left((X(R_M)'X(R_M))^{-1}X(r)'X(r)\right) + \mathrm{Tr}\left((X(r)'X(r))^{-1}X(R_M)'X(R_M)\right) - 2p\right).$$

A high symmetric Kullback-Leibler divergence on multivariate normal data basically says that the expected mean squared prediction error using estimates $\hat{\beta}$ from an experiment using a design R_M on test data $X(r)$ should be high, and the expected mean squared prediction error using estimates $\hat{\beta}$ from an experiment using a design r on test data $X(R_M)$ should be also be high. The symmetric Kullback-Leibler divergence between identical distributions is 0.

This is *not* equivalent to the partial ordering in definition 3 on $\mathcal{R}_{PSD}$. Here we wish to *increase* the mean squared prediction error when testing out estimates on test data chosen using a rule R_M. The assumption here is that the experimental design rule R_M is the worst design rule, and that design rules improve with the distance from this rule, as measured by the symmetric Kullback-Leibler divergence. This assumption is not necessarily true, as can be seen by the fact that this ordering is not equivalent to the partial ordering in definition 3. Nevertheless, this partial ordering may be useful if the goal is to design an experiment to provide very different information from a counterfactual situation where groups are chosen randomly.

We report the values of the symmetric Kullback-Leibler divergence supposing that X follows a multivariate normal distribution. Boltz, Debreuve, and Barlaud (2007) proposes another approximation of the Kullback-Leibler divergence using K nearest neighbours, which does not rely on the normality assumption, but we do not use this method due to convergence issues of this estimator.

4.4.5 Ordering by the expected variance of the outcome

It is not only the information of X which matters, but the strength of the relationship between X and Y. Unfortunately, this relationship is immeasurable, because we do not have access to the data Y prior to running the experiment. Nevertheless, if we have some prior information about β, then we can use this information in order to estimate Y. This final partial ordering on $\mathcal{R}$ supposes that we have some prior estimate of β.

Intuitively, it makes sense to care about the variance of Y. We wish to conduct an experiment from which we are later capable of predicting Y under a wide variety of scenarios. Therefore, we hope to cover a large enough range of group allocations to predict a wide range of the outcome variable. Estimating Y by $\hat{Y}$, we have

$$\mathbb{E}(\text{Var}(\hat{Y})) = \mathbb{E}(\text{Var}(X\hat{\beta}))$$

Increasing the variance of Y is thus equivalent to increasing the variance of $X\hat{\beta}$. This requires some estimate of β! We hence define a partial ordering on $\mathcal{R}$ using the variance of $X\hat{\beta}$.

Definition 12. *Suppose $\hat{\beta}$ is an estimate of β. Define the partial ordering $\preceq_V$ on $\mathcal{R}$ to be such that $r_2 \preceq_V r_1$ if and only if $\hat{\beta}'X(r_1)'X(r_1)\hat{\beta} \geq \hat{\beta}'X(r_2)'X(r_2)\hat{\beta}$.*

By lemma 2, this partial ordering is equivalent to that in definition 3 on $\mathcal{R}_{PSD}$ if $\hat{\beta}$ is an OLS estimate of β under a group allocation $r' \in \mathcal{R}$.

We show that the expected order of experiments does not depend on the distribution of the error terms.

Proposition 4. *Suppose the errors U_{Ya} follow a distribution with variance σ_a^2 and the errors U_{Yb} follow a distribution with variance σ_b^2. Let r_1 and r_2 be design rules. Let $\hat{\beta}(r)_a$ denote an OLS estimate of β under an experimental design rule r with errors U_{Ya} and $\hat{\beta}(r)_b$ denote an OLS estimate of β under an experimental design rule r with errors U_{Yb}. Then if*

$$\mathbb{E}\left(\hat{\beta}(r_1)_a' X(r_T)' X(r_T) \hat{\beta}(r_1)_a\right) \geq \mathbb{E}\left(\hat{\beta}(r_2)_a' X(r_t)' X(r_t) \hat{\beta}(r_2)_a\right)$$

then

$$\mathbb{E}\left(\hat{\beta}(r_1)_b' X(r_T)' X(r_T) \hat{\beta}(r_1)_b\right) \geq \mathbb{E}\left(\hat{\beta}(r_2)_b' X(r_t)' X(r_t) \hat{\beta}(r_2)_b\right).$$

This is to say that the expected partial order $\preceq_V$ on $\mathcal{R}$ is independent of the distribution of the error terms.

Proof.

$$\begin{aligned}
&\mathbb{E}\left(\hat{\beta}(r_1)_a' X(r_T)' X(r_T) \hat{\beta}(r_1)_a\right) \geq \mathbb{E}\left(\hat{\beta}(r_2)_a' X(r_t)' X(r_t) \hat{\beta}(r_2)_a\right) \\
&\implies \operatorname{Tr}(\sigma_a^2 (X(r_1)' X(r_1))^{-1} X(r_T)' X(r_T)) \geq \operatorname{Tr}(\sigma_a^2 (X(r_1)' X(r_1))^{-1} X(r_t)' X(r_t)) \\
&\implies \operatorname{Tr}(\sigma_b^2 (X(r_1)' X(r_1))^{-1} X(r_T)' X(r_T)) \geq \operatorname{Tr}(\sigma_b^2 (X(r_1)' X(r_1))^{-1} X(r_t)' X(r_t)) \\
&\implies \mathbb{E}\left(\hat{\beta}(r_1)_b' X(r_T)' X(r_T) \hat{\beta}(r_1)_b\right) \geq \mathbb{E}\left(\hat{\beta}(r_2)_b' X(r_t)' X(r_t) \hat{\beta}(r_2)_b\right)
\end{aligned}$$

□

This means that this partial ordering on $\mathcal{R}$ also does not depend on the distribution of the error terms.

We can justify this partial order in two ways. The variance of Y is an approximation of mutual information between X and Y under some assumptions, and the variance of Y drives the coefficient of determination, given the error terms.

The mutual information between Y and X is a measure of the amount of information on Y provided by X. This is used for the purposes of variable selection in Rossi et al. (2006), for example.

Definition 13. *[Conditional entropy](Rossi et al. (2006)) The conditional entropy of a random variable Y given X is defined as*

$$H(Y|X) = \mathbb{E}\left(\ln \frac{P(Y)}{P(Y,X)}\right).$$

Definition 14. *[Mutual information](Rossi et al. (2006)) The mutual information between two random variables X and Y is defined as*

$$\mathrm{I}(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

The mutual information is the amount of additional information provided by Y on top of the amount of information on Y provided by X. It can also be considered as the reduction of uncertainty of Y from X. If Y and X are independent, then $\mathrm{I}(X,Y)=0$, and if $Y=f(X)$ then then $\mathrm{I}(X,Y)=\infty$. We thus want to maximise $\mathrm{I}(X,Y)$.

The term $H(Y|X)$ is a function of the error terms, because it is the information of the part of Y excluding X. Assuming that the error terms are independent, the only part of $\mathrm{I}(X,Y)$ which changes under different design rules is $H(Y)$.

If Y were to follow a normal distribution, the entropy of Y would be increasing in the variance of Y.

$$H(Y)=\frac{1}{2}\ln\left(2\pi e\mathrm{Var}(Y)\right)$$

Although Y does not necessarily follow a normal distribution - the assumption required for the mutual information of Y and X to be monotone in the variance of Y, another classic measure of the strength of the relationship between X and Y is also monotone in the variance of Y - the coefficient of determination.

The coefficient of determination is an estimate of

$$\Phi=1-\frac{\sigma^2}{\mathrm{Var}(Y)}.$$

This is increasing in the variance of Y. We can rewrite this as

$$\Phi=\frac{\beta'X'X\beta}{\beta'X'X\beta+U_Y'U_Y}$$

and estimate it by

$$R^2=\frac{\hat{\beta}'X'X\hat{\beta}}{\hat{\beta}'X'X\hat{\beta}+\hat{\sigma}^2}.$$

Assuming that $\hat{\sigma}^2$ is identical for all experiments, this is a strictly increasing function in $\hat{\beta}'X'X\hat{\beta}$. A partial ordering on this estimation of the R^2 value is thus identical to that in definition 12.

4.5 Complete information case

Imagine that we had complete information about Y and X for multiple group allocation rules. We could use this information to evaluate the experimentally optimal rule using the following partial ordering on $\mathcal{R}$.

Definition 15. *Suppose $\hat{\beta}(r)$ is the OLS estimate of β from an experiment with group allocation rule r and suppose $\hat{\sigma}^2(r)$ is the sum of the squared residuals. Define the partial ordering $\preceq_{R2}$ on $\mathcal{R}$ to be such that $r_2\preceq_{R2}r_1$ if and only if*

$$\frac{\hat{\beta}(r_1)'X(r_1)'X(r_1)\hat{\beta}(r_1)}{\hat{\beta}(r_1)'X(r_1)'X(r_1)\hat{\beta}(r_1)+\hat{\sigma}(r_1)^2}\geq\frac{\hat{\beta}(r_2)'X(r_2)'X(r_2)\hat{\beta}(r_2)}{\hat{\beta}(r_2)'X(r_2)'X(r_2)\hat{\beta}(r_2)+\hat{\sigma}(r_2)^2}$$

We firstly show that the expected order of group allocation rules using this partial ordering on $\mathcal{R}$ does not depend on the variance of the error terms if the error terms follow a normal distribution with mean 0. This is not obvious, due to the bias of the R^2-value as an estimate of Φ.

Proposition 5. *Suppose the errors U_{Ya} follows a distribution $\mathcal{D}_a \sim \mathcal{N}(0, \sigma_a^2)$ and the errors U_{Yb} follows a distribution $\mathcal{D}_b \sim \mathcal{N}(0, \sigma_b^2)$. Let r_1 and r_2 be design rules. Then if*

$$\mathbb{E}\left(\frac{\hat{\beta}(r_1)'X(r_1)'X(r_1)\hat{\beta}(r_1)}{\hat{\beta}(r_1)'X(r_1)'X(r_1)\hat{\beta}(r_1) + \hat{\sigma}_a(r_1)^2}\right) \geq \mathbb{E}\left(\frac{\hat{\beta}(r_2)'X(r_2)'X(r_2)\hat{\beta}(r_2)}{\hat{\beta}(r_2)'X(r_2)'X(r_2)\hat{\beta}(r_2) + \hat{\sigma}_a(r_2)^2}\right)$$

then

$$\mathbb{E}\left(\frac{\hat{\beta}(r_1)'X(r_1)'X(r_1)\hat{\beta}(r_1)}{\hat{\beta}(r_1)'X(r_1)'X(r_1)\hat{\beta}(r_1) + \hat{\sigma}_b(r_1)^2}\right) \geq \mathbb{E}\left(\frac{\hat{\beta}(r_2)'X(r_2)'X(r_2)\hat{\beta}(r_2)}{\hat{\beta}(r_2)'X(r_2)'X(r_2)\hat{\beta}(r_2) + \hat{\sigma}_b(r_2)^2}\right).$$

This is to say that the expected partial order $\preceq_{R2}$ on $\mathcal{R}$ is independent of the variance of the error terms, when the error terms are normally distributed.

Proof. Cramer (1987) shows that the expected value of R^2 in a linear model with normal errors is given by:

$$\mathbb{E}(R^2)(\lambda) = \exp(-\lambda/2)\left(\sum_{j=0}^{+\infty} \frac{(\lambda/2)^j}{j!}\left(\frac{(p-1)/2 + j}{(n-1)/2 + j}\right)\right)$$

where $\lambda = \beta' X' X \beta / \sigma^2$. We show that this function is increasing in λ, and hence monotone decreasing in σ^2.

$$\mathbb{E}(R^2)'(\lambda) = \exp(-\lambda/2)\left(-\frac{1}{2}\sum_{j=0}^{+\infty} \frac{(\lambda/2)^j}{j!}\left(\frac{(p-1)/2 + j}{(n-1)/2 + j}\right) + \sum_{j=0}^{+\infty}(j+1)\frac{(\lambda/2)^j}{2(j+1)!}\left(\frac{(p-1)/2 + j + 1}{(n-1)/2 + j + 1}\right)\right)$$

$$\mathbb{E}(R^2)'(\lambda) \equiv \exp(-\lambda/2)g(\lambda)$$

$$
\begin{aligned}
g(\lambda) &= \left(-\frac{1}{2} \sum_{j=0}^{+\infty} \frac{(\lambda/2)^j}{j!} \left(\frac{(p-1)/2+j}{(n-1)/2+j} \right) + \sum_{j=0}^{+\infty} (j+1) \frac{(\lambda/2)^j}{2(j+1)!} \left(\frac{(p-1)/2+j+1}{(n-1)/2+j+1} \right) \right) \\
&= \sum_{j=0}^{+\infty} \frac{(\lambda/2)^j}{j!} \left(\frac{(p-1)/2+j+1}{(n-1)/2+j+1} - \frac{(p-1)/2+j}{(n-1)/2+j} \right) \\
&= \sum_{j=0}^{+\infty} \frac{(\lambda/2)^j}{j!} \left(\frac{2j+p+1}{2j+n+1} - \frac{2j+p-1}{2j+n-1} \right) \\
&= \sum_{j=0}^{+\infty} \frac{(\lambda/2)^j}{j!} \left(\frac{2(n-p)}{(2j+n+1)(2j+n-1)} \right) \\
&= 2(n-p) \sum_{j=0}^{+\infty} \frac{(\lambda/2)^j}{j!} \left(\frac{1}{(2j+n+1)(2j+n-1)} \right)
\end{aligned}
$$

As $n > p$ and $\lambda \geq 0$, $g(\lambda) \geq 0$ and thus $\mathbb{E}(R^2)'(\lambda)$. This means that the expected R^2 value is increasing in λ. As λ is monotone decreasing in σ^2, the expected rank of R^2-values is identical for any variance of the distribution of the normal error terms.

□

This is not necessarily true for all distributions of the error term. However, for large enough n, the bias of R^2 as an estimate of Φ will be extremely small in comparison to the R^2-value itself, having next to no effect on the expected partial order $\preceq_{R2}$ on $\mathcal{R}$.

Srivastava, Srivastava, and Ullah (1995) computes a first order expansion ($O(n^{-1})$) of the large sample bias of the coefficient of determination as

$$
B(R^2) = \frac{1-\Phi}{n} \left(p + \Phi(2\Phi - 1) + \Phi(1-\Phi)\gamma \right)
$$

where p is the number of columns of X and where γ is the excess kurtosis.

We can thus write a first order approximation of the expected value of R^2 by:

$$
\mathbb{E}(R^2)(\Phi, \gamma) = \frac{1-\Phi}{n} \left(p + \Phi(2\Phi - 1) + \Phi(1-\Phi)\gamma \right) + \Phi.
$$

For large enough n, the bias is small and so the order $\preceq_{R2}$ on $\mathcal{R}$ will not change under different distributions of error terms, so long as the error terms are independent of X.

The value added of using the R^2-value to create a partial ordering requiring knowledge of Y compared a partial ordering using a good estimate of β such as in definition 12 is thus very low, as the R^2 value is essentially increasing in $\hat{\beta}' X' X \hat{\beta}$.

We could also use the mutual information between X and Y to compute the experiment with the strongest relationship between X and Y. Any other variable selection method requiring knowledge of Y such as lasso regression, ridge regression or random trees could also be applied here. We do not elaborate on these possibilities, as the main focus is to consider the case where Y is unknown.

5 Simulation study

In this section we simulate two linear models, the first with an individual characteristic and the leave-out-mean of each peer group, and the second with an individual characteristic, the leave-out-mean and the leave-out-variance of each peer group. We always consider 400 individuals split into 20 groups of 20 people.

All code used for the simulations is available at
`https://github.com/hannahbull/peer_effects`.

5.1 Group allocation rule algorithms

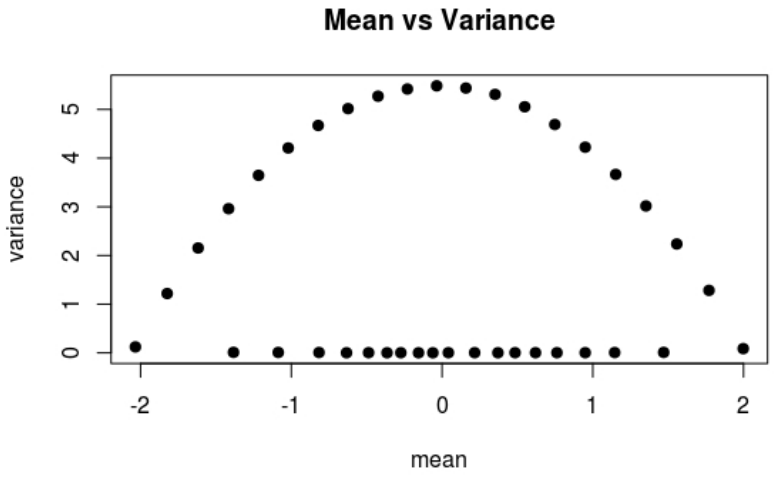
Although there are finitely many group allocations, it is computationally necessarily to create group allocation algorithms to generate ranges of the measures of group characteristics unlikely to occur naturally. We construct group allocations using the three algorithms defined in this section. The intention of each of these algorithms is to maximise the variance in the means or the variances, or both at the same time.

We reassign groups multiple times according to these three algorithms. Each random draw involves re-choosing the groups according to the specified algorithm with parameter l.

Figure 5 shows the points at the edge of the space of group means and group variances of an individual characteristic which follows a standard normal distribution. The highest possible variance is attained by putting all of the individuals with characteristic furthest from the mean into the same group. The mean of this group is thus around 0, as all the highest and the lowest individuals are in the same group. Other high values for the variance can be attained by putting varying proportions of the highest and the lowest individuals in the same group. The lowest possible value for the mean occurs when the 20 lowest individuals are in the same group. This group of course has low variance. Similarly, the highest possible value for the group mean occurs when the highest 20 individuals are in the same group. Other values of the group mean but with low variance can be attained by allocating similar 20 individuals to the same group.

Ideally, we aim to fill as much of the space within these points as possible. However, choosing groups randomly tightly concentrates all of the points around mean 0 and variance 1, as in figure 6. Alternative algorithms provide higher variation.

Figure 5: Boundary of values for the group means and group variances for an individual characteristic following a standard normal distribution



5.1.1 Streaming algorithm

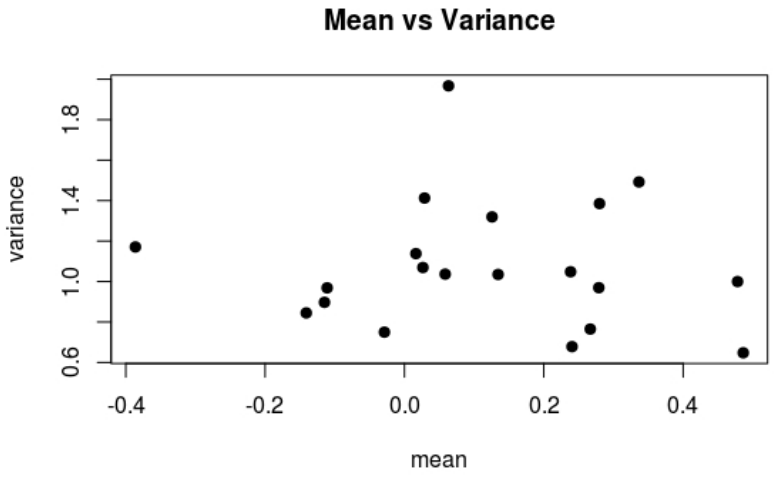
This algorithm aims to maximise the variance between the leave-out-means of each observation. We call this the streaming algorithm, because it is equivalent to a rule which could be used to stream students in classes by ability.

We first order the population by the value of the individual characteristic x, and then divide the population into l approximately equally sized levels, then choose 20 people randomly from each level to put in each group. Once there are fewer than 20 people in each level, the remaining individuals are sorted in order of x then split into groups of 20. When $l = 1$ and $n = 400$, this amounts to choosing the groups at random. When $s = 20$, this corresponds to grouping the individuals by ability. When $s = 2$, this corresponds to separating the students into two levels: 'higher than the median of x' and 'lower than the median of x', then randomly subdividing these two divisions into groups of 20.

Presumably, choosing individuals randomly will not be the best experiment, because the variation in the means of the groups will be low. But, for example, it is not at all obvious whether the above algorithm where $l = 2$ performs better or worse than when $l = 20$.

When groups are chosen randomly ($l = 1$), then the leave-out-means are slightly negatively correlated with the individual characteristic. This is because the group mean for a higher individual is likely to be lower than their high individual characteristic, and the group mean for a lower individual is likely to be higher than their low individual characteristic. When individuals

Figure 6: Group means and variances for a random allocation of individuals with an individual characteristic following a standard normal distribution



are streamed into more levels, the leave-out-means becomes strongly positively correlated with the individual characteristics. This can be seen in figure 7.

5.1.2 Streaming by variance algorithm

This algorithm aims to maximise the variance between the leave-out-variances of each observation. It is similar to the previous algorithm, but streams individuals into levels by squared difference to the mean of the individual characteristic x, instead of by the value of x.

We order the population by the squared difference to the mean of x. We then divide this population into l approximately equal-sized levels. The algorithm is then identical to the streaming algorithm. When $l = 1$ and $n = 400$, this amounts to choosing the groups at random. When $s = 20$, this corresponds to putting the highest and the lowest x individuals in the same group, then putting the next highest and the next lowest individuals in the same group, etc.. The final group is the 20 individuals closest to the mean. When $s = 2$, this corresponds to separating the students into two levels: 'far from the mean of x' and 'close to the mean of x', then randomly subdividing these two divisions into groups of 20.

Figure 8 displays some examples of the mean and variance of groups chosen using this algorithm.

5.1.3 Mean and variance buckets algorithm

Here we aim to maximise both the variance in leave-out means and the variance in leave-out-variances of the observations, assuring that the covariance between the explanatory variables is not too high.

One potential algorithm to do this consists of ordering the population by value of the individual characteristic x and dividing the population into l approximately equally sized levels. Then 20 individuals are chosen from some random $a \leq l$ number of levels, as equally as possible. For example, if $a = 3$ then 6 individuals are chosen from one of the l levels, and 7 individuals are chosen from two other of the l levels. This process is repeated until all individuals are divided into groups of 20.

For example, when $l = 1$, groups are chosen randomly. When $l = 2$, this corresponds to separating the students into two divisions: 'higher than the median of x' and 'lower than the median of x'. The groups are either made up of 20 randomly chosen individuals in the bottom division, 20 randomly chosen individuals in the top division, or 10 randomly chosen individuals from each division.

Figure 9 displays some examples of the mean and variance of groups chosen using this algorithm.

5.2 Linear model with leave-out-means

We simulate the following model:

$$y_{i,g} = x_{i,g} + 0.2\overline{x}_{g-i} + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0,1)$.

We test this with three different distributions of x, a standard normal distribution, a uniform distribution between 0 and 1 and a skewed distribution: 1 minus a beta distribution with parameters $\alpha = 1$ and $\beta = 5$. We draw group allocations using streaming by levels, where the number of levels ranges from 1 to 20.

The true $\beta = (1, 0.2)'$. We also use two incorrect estimates of β in order to compute the expected variance of y. These are $\hat{\beta}_1 = (0.5, 0.5)'$ and $\hat{\beta}_1 = (1, -0.2)'$.

Due to the functional form of the peer effects model, all rules in $\mathcal{R}$ are optimal, as changing the allocation of the individuals amongst groups does not change the average outcome of the individuals.

5.2.1 Uniform distribution

The results in table 1 show that randomly sorting individuals is not the best experimental design, but is also not necessarily the worst design. Moreover, contrary to intuition, maximising the variance of the leave-out-means is not

necessarily the best way to reduce the variance of the coefficient on the leave-out-means. This is because the covariance between the individual characteristic and the leave-out-means is high (see figure 7).

There are essentially two types of orderings of the best experimental designs. The partial orderings by minimising the variance of the leave-out-mean, minimising the trace of $(X'X)^{-1}$ and maximising the determinant of $X'X$ all agree that the best experiment involves streaming by 2 levels, and the worst experiment involves streaming by 20 levels. This is because all of these orders heavily weight the covariance of the individual characteristic and the leave-out-mean. The group allocation rule by streaming into two levels is a good compromise between having some variance in the leave-out-means, but not too much covariance between the individual characteristic and the leave-out-means.

The partial ordering by maximising the trace of $X'X$, maximising the symmetric Kullback-Leibler divergence and the expected variance of $\hat{Y}$ estimating β with its true value or the slightly incorrect value $\hat{\beta}_1$ all agree that the best experiment involves dividing individuals into 20 levels by the individual characteristic. The second best experiment involves dividing the group into 10 levels in order of the individual characteristic, then further randomly dividing each of these 10 levels into 2 groups. This is because these orderings are less affected by covariance between the individual characteristic and the leave-out-means.

The rank of experiments using expected variance of Y under an incorrect estimate of β by $\hat{\beta}_2$ is interestingly quite similar to that by minimising the variance of the leave-out-mean, minimising the trace of $(X'X)^{-1}$ or maximising the determinant of $X'X$. The incorrect $\hat{\beta}_2$ mistakes the sign of the leave-out-mean, estimating the absolute covariance between the individual characteristic and the group characteristic too highly.

Whilst a slighly incorrect estimate of β may not lead to a different ordering compared to that using the true β, a very incorrect estimate of β can lead to a completely different partial ordering when compared to that using the true β.

5.2.2 Normal distribution

The simulation results are provided in table 2 and the order of the best and the worst experimental designs is very similar to the case where the individual characteristic follows a uniform distribution (table 1).

5.2.3 Skewed distribution

The simulation results are provided in table 3 and the order of the best and the worst experimental designs is very similar to the case where the individual characteristic follows a uniform distribution (table 1).

5.3 Linear model with leave-out-means and leave-out-variances

We simulate the following model:

$$y_{i,g} = x_{i,g} + 0.2\overline{x}_{g-i} + 0.2\widetilde{x}_{g-i} + \varepsilon$$

where $\overline{x}_{g-i}$ is the leave-out-mean of individual i in group g, $\widetilde{x}_{g-i}$ is the leave-out-variance of individual i in group g and $\varepsilon \sim \mathcal{N}(0,1)$.

We test this with three different distributions of x, a standard normal distribution, a uniform distribution between 0 and 1 and a skewed distribution: 1 minus a beta distribution with parameters $\alpha = 1$ and $\beta = 5$. We draw group allocations using the streaming by variance algorithm and the mean and variance buckets algorithm.

The true $\beta = (1, 0.2, 0.2)'$. We also use two incorrect estimates of β in order to compute the expected variance of y. These are $\hat{\beta}_1 = (0.5, 0.5, 0.5)'$ and $\hat{\beta}_1 = (1, 0.2, -0.2)'$.

5.3.1 Uniform distribution

The different partial orderings all agree that an experimental design using the streaming by variance algorithm is best or quite good when $l = 7$ or when $l = 11$ (except for the partial ordering using the incorrect $\hat{\beta}_2$), as can be seen in table 4. Unfortunately, the sum of the leave-out-variances is comparatively lowest when $l = 11$. The experimentally optimal design can thus be the socially worst design if the coefficient of the leave-out-variance is positive.

Moreover, if the coefficient of the leave-out-variance is positive, the socially optimal experiment occurs at $l = 2$ when using the streaming by variance algorithm. This is a poor experimental design according to all of the different partial orderings (except for the partial ordering using the incorrect $\hat{\beta}_2$). If the coefficient of the leave-out-variances is negative, then the socially optimal experiment occurs when $l = 11$, which is also experimentally ideal according to most of the partial orders on group allocation rules.

For the mean and variance buckets algorithm, low numbers of levels except 1, the random allocation, tend to be the best experiments, as reported in table 5. Unfortunately here, the socially optimal experiments are very different from the optimal rule choice if the coefficient on the leave-out-variances is positive.

According to some of the partial orderings, the streaming by variance algorithm works best, but according to other partial orderings, the mean and variance buckets algorithm works best. A ranking comparing these two algorithms is provided in table 6.

5.3.2 Normal distribution

The simulation results are provided in table 7 and the order of the best and the worst experimental designs is similar to the case where the individual characteristic follows a uniform distribution. However, the streaming by variance al-

gorithm almost consistently performs better than the mean and variance buckets algorithm.

5.3.3 Skewed distribution

The results are provided in table 8. Here, the streaming by variance algorithm is basically always better than the mean and variance algorithm. However, the different partial orderings disagree on the best experiment. Whilst $l = 11$ for the the streaming by variance algorithm is a poor experiment in terms of minimising the variance of the parameter on the left-out-variance, this is the optimal experiment according to alternative partial orderings. The optimal choice of experimental design thus depends on the objectives of the experiment.

6 Conclusion

Peer effects experiments with exogenous selection can be improved through non-random selection of group allocations. Nevertheless, it is not at all obvious how to identify an experimentally optimal group allocation in the case of a linear model with multiple dependent variables. There are many possible nonequivalent orderings on group allocation rules, due to the trade-off between the variance and covariance of individual and group characteristics.

Each of the five listed partial orderings on the space of group allocations has merits and faults. One must choose a criteria with which the best experimental design should be determined. Even if not the best criteria, the experimental design is likely to be significantly better than a random group allocation. Depending on the objectives of the experiment, such as testing a group allocation which could be socially optimal, or reducing the variance of a particular parameter, the optimal experimental design can be determined under classic assumptions for estimation using ordinary least squares.

The socially optimal group allocation can sometimes be identified under some loose hypotheses about the model parameters. For example, in a linear model with leave-out-means and leave-out-variances, the group allocation leading to the predicted maximum average outcome can be identified by simply assuming the sign of the coefficient on the leave-out-variance. Predicting the socially optimal group allocation prior to running the experiment can be very helpful in choosing an experimental design.

This paper is of course only a first step towards how one could improve experiments through non-random selection of characteristics. We make many strong structural assumptions such as linearity and independence of error terms, which are unlikely to be true when applied to real peer effects data. Nevertheless, similar ideas can be applied under relaxed hypotheses. Some further questions are thus:

1. Under a linear model with errors which are not independent of X, are there methods to choose experimentally and socially optimal group allocations?

2. Using different structural assumptions, such as varying group sizes as in Lee (2007), or non-parametric models, can we also choose optimal experimental designs?

3. What if one chooses the incorrect causal variable, for example variance instead of standard-deviation? How can incorrect predictions from this incorrect assumption be best avoided?

Another area of interest is in creating algorithms to identify the experimentally optimal or socially optimal group allocation. We propose three possible methods in this paper applicable to a regression using leave-out-means or leave-out-variances, but much work could be done to identify computationally fast methods of optimising the search for a maximum in the finite but large space of group allocation rules.

The ideas on dependant variable selection in this article are applicable outside of peer effects experiments. In any experiment where one has the opportunity to distribute values of what will be used as a dependent variable, there could be a method to choose the experimentally optimal or the socially optimal distribution. Variable selection to achieve greater accuracy is possible with just information about the data X, without the outcome Y. Much work is yet to be done to understand how data X should be chosen under varying constraints and under varying structural and non-parametric hypotheses.

Figure 7: Streaming into 1, 2, 7 and 20 levels by the value of an individual characteristic following a demeaned uniform distribution

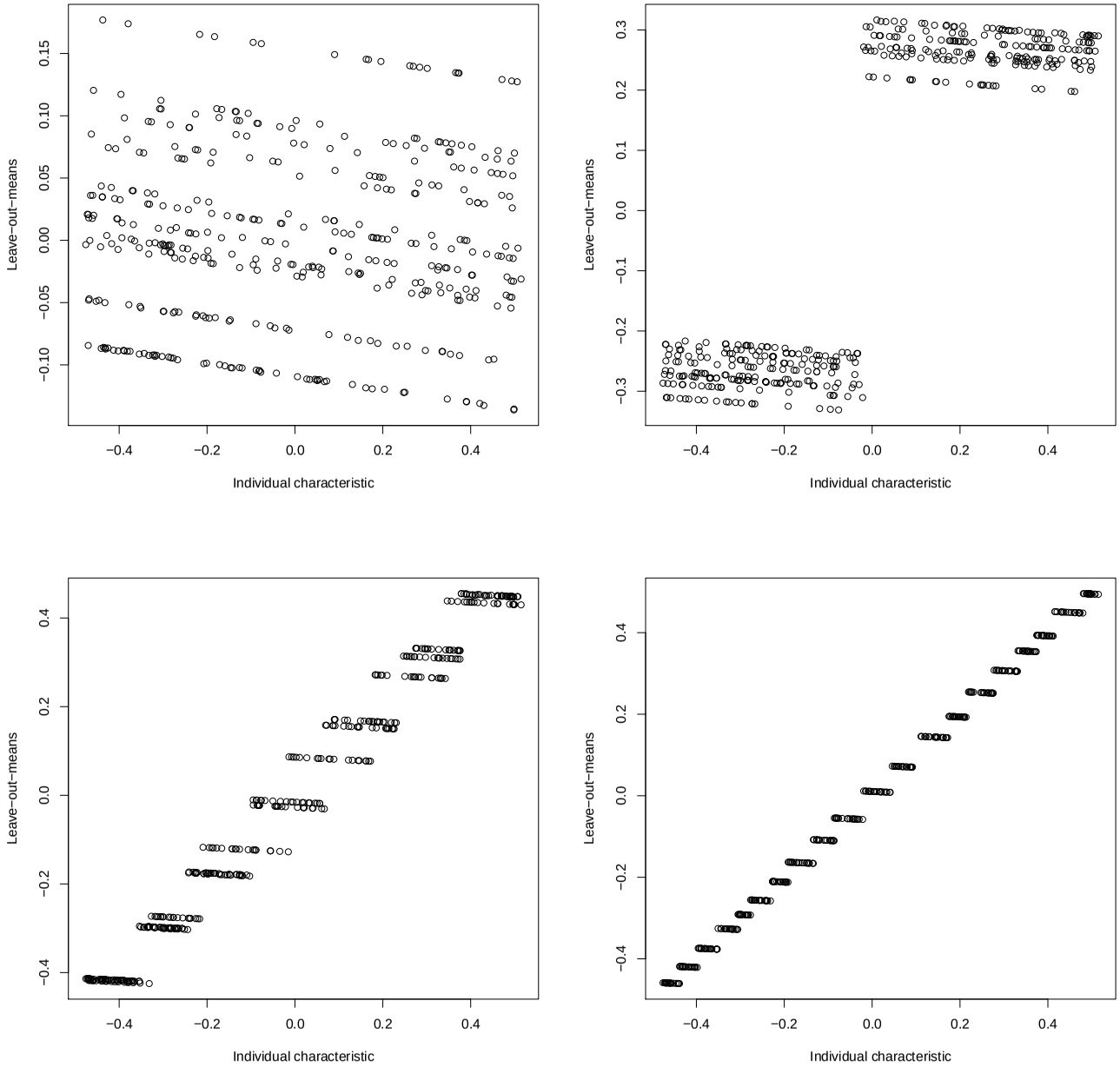


Figure 8: Streaming by variance into 2, 3, 10 and 20 levels by the value of an individual characteristic following a uniform distribution (values are demeaned)

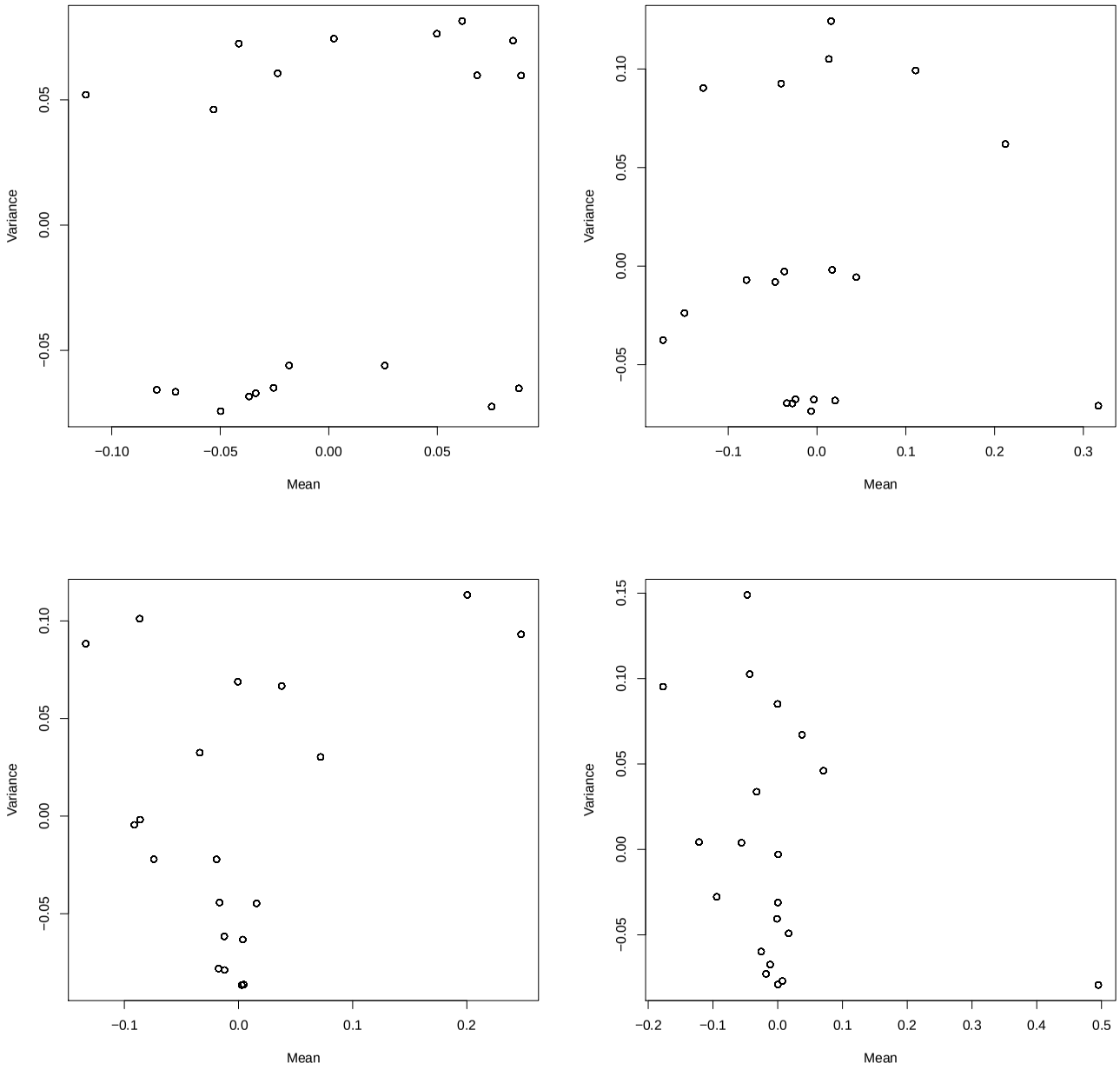


Figure 9: Mean and variance buckets algorithm using 2, 3, 10 and 20 levels by the value of an individual characteristic following a uniform distribution (values are demeaned)

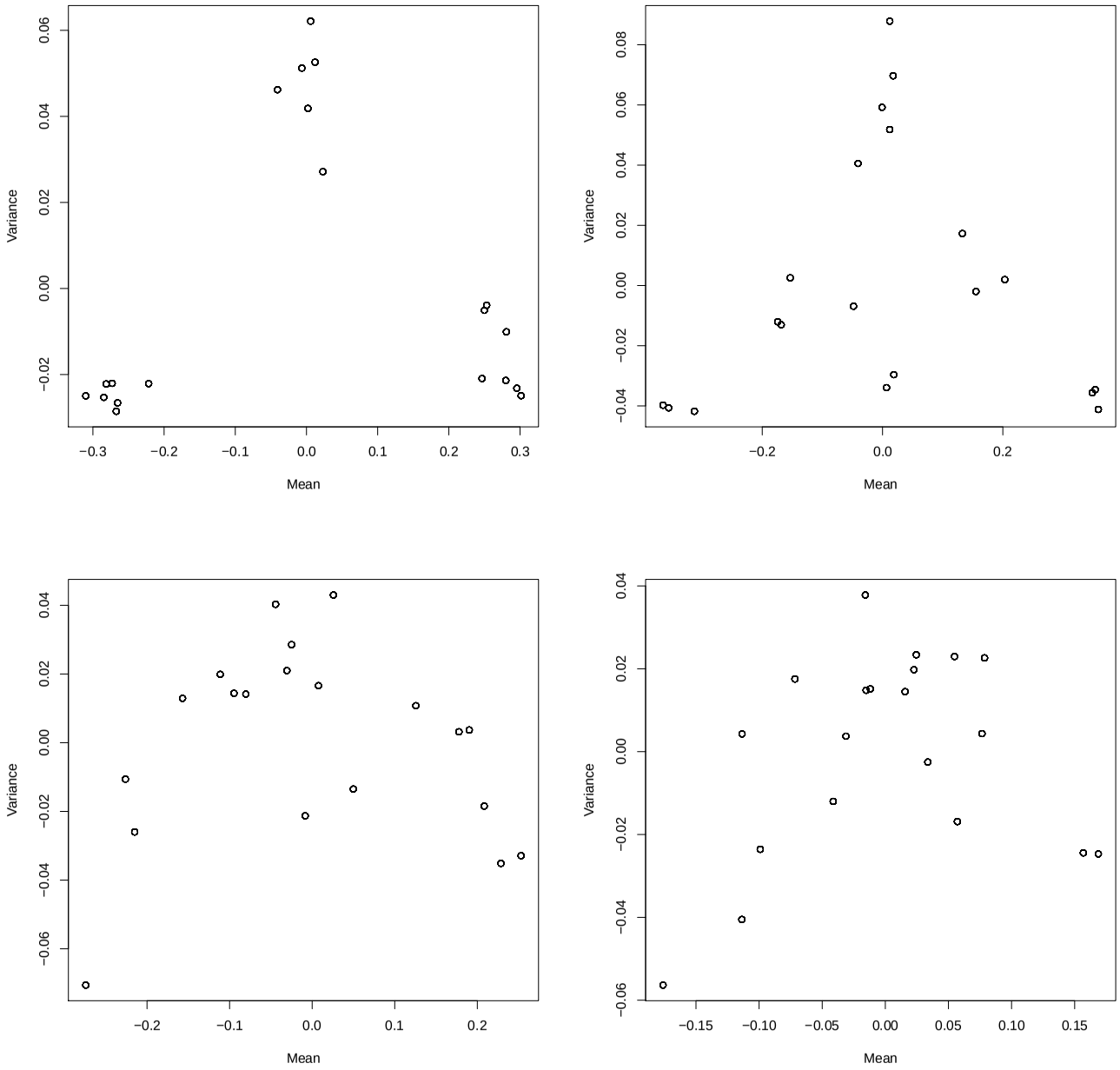


Table 1: 500 random draws, uniform distribution, streaming algorithm

Number of levels	Variance of the left-out-mean parameter	Trace of $(X'X)^{-1}$	Trace of $(X'X)$	Determinant of $(X'X)$	SKL-divergence of X from $X(R_M)$	Expected variance of Y, true β	Expected variance of Y, incorrect β ($\hat{\beta}_1$)	Expected variance of Y, incorrect β ($\hat{\beta}_2$)
1	6	3	20	5	20	20	20	1
2	1	1	19	1	19	19	19	2
3	2	2	18	2	18	18	18	3
4	3	4	17	3	17	17	17	4
5	7	7	14	7	14	14	14	7
6	5	6	15	6	15	15	15	6
7	11	11	10	11	10	10	10	11
8	10	10	11	10	11	11	11	10
9	8	8	13	8	13	13	13	8
10	19	19	2	19	2	2	2	19
11	17	17	4	17	4	4	4	17
12	18	18	3	18	3	3	3	18
13	16	16	5	16	5	5	5	16
14	15	15	6	15	6	6	6	15
15	14	14	7	14	7	7	7	14
16	13	13	8	13	8	8	8	13
17	12	12	9	12	9	9	9	12
18	9	9	12	9	12	12	12	9
19	4	5	16	4	16	16	16	5
20	20	20	1	20	1	1	1	20

Table 2: 500 random draws, normal distribution, streaming algorithm

Number of levels	Variance of the left-out-mean parameter	Trace of $(X'X)^{-1}$	Trace of $(X'X)$	Determinant of $(X'X)$	SKL-divergence of X from $X(R_M)$	Expected variance of Y, true β	Expected variance of Y, incorrect β $(\hat{\beta}_1)$	Expected variance of Y, incorrect β $(\hat{\beta}_2)$
1	11	6	20	10	20	20	20	1
2	1	1	19	1	19	19	19	2
3	2	2	18	2	18	18	18	3
4	3	3	17	3	17	17	17	4
5	5	5	15	5	15	15	15	6
6	4	4	16	4	16	16	16	5
7	9	10	11	9	11	11	11	10
8	8	9	12	8	12	12	12	9
9	6	7	14	6	14	14	14	7
10	19	19	2	19	2	2	2	19
11	18	18	3	18	3	3	3	18
12	17	17	4	17	4	4	4	17
13	16	16	5	16	5	5	5	16
14	15	15	6	15	6	6	6	15
15	14	14	7	14	7	7	7	14
16	13	13	8	13	8	8	8	13
17	12	12	9	12	9	9	9	12
18	10	11	10	11	10	10	10	11
19	7	8	13	7	13	13	13	8
20	20	20	1	20	1	1	1	20

Table 3: 500 random draws, left-tailed distribution, streaming algorithm

Number of levels	Variance of the left-out-mean parameter	Trace of $(X'X)^{-1}$	Trace of $(X'X)$	Determinant of $(X'X)$	SKL-divergence of X from $X(R_M)$	Expected variance of Y, true β	Expected variance of Y, incorrect β ($\hat{\beta}_1$)	Expected variance of Y, incorrect β ($\hat{\beta}_2$)
1	11	5	20	11	20	20	20	1
2	1	1	19	1	19	19	19	2
3	2	2	18	2	18	18	18	3
4	3	3	17	3	17	17	17	4
5	5	6	15	5	15	15	15	6
6	4	4	16	4	16	16	16	5
7	9	10	11	9	11	11	11	10
8	8	9	12	8	12	12	12	9
9	6	7	14	6	14	14	14	7
10	18	18	3	18	3	3	3	18
11	19	19	2	19	2	2	2	19
12	17	17	4	17	4	4	4	17
13	16	16	5	16	5	5	5	16
14	15	15	6	15	6	6	6	15
15	14	14	7	14	7	7	7	14
16	13	13	8	13	8	8	8	13
17	12	12	9	12	9	9	9	12
18	10	11	10	10	10	10	10	11
19	7	8	13	7	13	13	13	8
20	20	20	1	20	1	1	1	20

Table 4: 500 random draws, uniform distribution, streaming by variance algorithm

Number of levels	Variance of the left-out-variance parameter	Trace of $(X'X)^{-1}$	Trace of $(X'X)$	Determinant of $(X'X)$	SKL-divergence of X from $X(R_M)$	Expected variance of Y, true β	Expected variance of Y, incorrect β $(\hat{\beta}_1)$	Expected variance of Y, incorrect β $(\hat{\beta}_2)$	Sum of leave-out-variances
1	20	20	20	20	20	20	20	4	3
2	18	19	19	19	19	19	19	3	1
3	6	13	15	15	17	12	12	7	7
4	5	18	18	18	14	18	18	1	2
5	4	17	17	17	12	17	17	2	4
6	2	7	14	14	10	13	13	6	6
7	1	1	6	1	2	6	5	14	15
8	3	2	9	7	6	8	8	11	12
9	7	11	13	12	16	11	11	8	8
10	10	16	16	16	18	16	16	5	5
11	9	4	1	3	1	1	1	20	20
12	8	3	2	2	3	2	2	19	19
13	12	5	3	4	4	3	3	18	18
14	13	6	4	5	5	4	4	17	17
15	17	9	5	6	7	5	6	16	16
16	15	8	7	8	8	7	7	15	14
17	14	10	8	9	11	9	9	13	13
18	16	12	10	10	13	10	10	12	11
19	19	15	12	13	15	15	15	10	10
20	11	14	11	11	9	14	14	9	9

Table 5: 500 random draws, uniform distribution, mean and variance buckets algorithm

Number of levels	Variance of the left-out-variance parameter	Trace of $(X'X)^{-1}$	Trace of $(X'X)$	Determinant of $(X'X)$	SKL-divergence of X from $X(R_M)$	Expected variance of Y, true β	Expected variance of Y, incorrect β ($\hat{\beta}_1$)	Expected variance of Y, incorrect β ($\hat{\beta}_2$)	Sum of leave-out-variances
1	20	20	20	20	20	20	20	1	1
2	6	6	1	5	2	1	1	20	20
3	2	2	2	2	1	2	2	19	19
4	1	1	3	1	3	3	3	18	18
5	3	3	4	3	4	4	4	17	17
6	4	4	5	4	5	5	5	16	16
7	5	5	6	6	6	6	6	15	15
8	7	7	7	7	7	7	7	14	14
9	8	8	8	8	8	8	8	13	13
10	9	9	9	9	9	9	9	12	12
11	10	10	10	10	10	10	10	11	11
12	11	11	11	11	11	11	11	10	10
13	12	12	12	12	12	12	12	9	9
14	13	13	13	13	13	13	13	8	8
15	14	14	14	14	14	14	14	7	7
16	15	15	15	15	15	15	15	6	6
17	16	16	16	16	16	16	16	5	5
18	17	17	17	17	17	17	17	4	4
19	18	18	18	18	18	18	18	3	3
20	19	19	19	19	19	19	19	2	2

Table 6: 500 random draws, uniform distribution, streaming by variance algorithm and mean and variance buckets algorithm

Algorithm	Number of levels	Variance of left-out-variance parameter	Trace of $(X'X)^{-1}$	Trace of $(X'X)$	Determinant of $(X'X)$	SKL-divergence of X from $X(R_M)$	Expected variance of Y, true β	Expected variance of Y, incorrect β ($\hat{\beta}_1$)	Expected variance of Y, incorrect β ($\hat{\beta}_2$)	Sum of leave-out-variances
Streaming by variance	1	40	40	40	40	40	40	40	4	4
Streaming by variance	2	18	20	34	26	22	38	38	3	1
Streaming by variance	3	6	13	23	17	17	21	20	15	17
Streaming by variance	4	5	18	31	25	14	37	34	1	2
Streaming by variance	5	4	17	29	23	12	33	30	2	5
Streaming by variance	6	2	7	22	14	10	22	21	13	15
Streaming by variance	7	1	1	9	1	2	9	8	30	31
Streaming by variance	8	3	2	13	7	6	13	13	25	27
Streaming by variance	9	7	11	20	12	16	18	18	18	20
Streaming by variance	10	10	16	25	19	19	26	26	10	13
Streaming by variance	11	9	4	2	3	1	2	2	39	39
Streaming by variance	12	8	3	4	2	3	4	4	37	37
Streaming by variance	13	12	5	5	4	4	5	5	36	36
Streaming by variance	14	13	6	6	5	5	6	6	35	35
Streaming by variance	15	17	9	8	6	7	8	9	33	33
Streaming by variance	16	15	8	10	8	8	11	11	31	30
Streaming by variance	17	14	10	12	9	11	14	14	29	28
Streaming by variance	18	16	12	15	10	13	16	16	27	25
Streaming by variance	19	19	15	18	13	15	24	24	24	23
Streaming by variance	20	11	14	17	11	9	23	23	23	22
Mean and var buckets	1	40	40	40	40	40	40	40	4	4
Mean and var buckets	2	25	25	1	21	20	1	1	40	40
Mean and var buckets	3	21	21	3	16	18	3	3	38	38
Mean and var buckets	4	20	19	7	15	21	7	7	34	34
Mean and var buckets	5	22	22	11	18	23	10	10	32	32
Mean and var buckets	6	23	23	14	20	24	12	12	28	29
Mean and var buckets	7	24	24	16	22	25	15	15	26	26
Mean and var buckets	8	26	26	19	24	26	17	17	22	24
Mean and var buckets	9	27	27	21	27	27	19	19	21	21
Mean and var buckets	10	28	28	24	28	28	20	22	20	19
Mean and var buckets	11	29	29	26	29	29	25	25	19	18
Mean and var buckets	12	30	30	27	30	30	27	27	17	16
Mean and var buckets	13	31	31	28	31	31	28	28	16	14
Mean and var buckets	14	32	32	30	32	32	29	29	14	12
Mean and var buckets	15	33	33	32	33	33	30	31	12	11
Mean and var buckets	16	34	34	33	34	34	31	32	11	10
Mean and var buckets	17	35	35	35	35	35	32	33	9	9
Mean and var buckets	18	36	36	36	36	36	34	35	8	8
Mean and var buckets	19	37	37	37	37	37	35	36	7	7
Mean and var buckets	20	38	38	38	38	38	36	37	6	6

Table 7: 500 random draws, normal distribution, streaming by variance algorithm and mean and variance buckets algorithm

Algorithm	Number of levels	Variance of left-out-variance parameter	Trace of $(X'X)^{-1}$	Trace of $(X'X)$	Determinant of $(X'X)$	SKL-divergence of X from $X(R_M)$	Expected variance of Y, true β	Expected variance of Y, incorrect β ($\hat{\beta}_1$)	Expected variance of Y, incorrect β ($\hat{\beta}_2$)	Sum of leave-out-variances
Streaming by variance	1	40	38	40	40	40	40	40	10	6
Streaming by variance	2	19	35	19	26	23	30	25	6	4
Streaming by variance	3	18	17	18	15	18	19	19	7	18
Streaming by variance	4	13	36	17	21	16	23	17	3	3
Streaming by variance	5	4	37	15	20	12	25	14	2	2
Streaming by variance	6	10	18	14	12	15	16	10	5	16
Streaming by variance	7	14	5	10	5	7	6	4	29	33
Streaming by variance	8	6	8	7	8	9	11	7	12	27
Streaming by variance	9	2	20	8	11	11	17	9	4	14
Streaming by variance	10	1	40	1	23	6	22	11	1	1
Streaming by variance	11	17	3	9	3	1	1	1	37	39
Streaming by variance	12	15	1	5	2	2	2	2	35	37
Streaming by variance	13	12	2	4	1	3	4	3	33	35
Streaming by variance	14	9	4	2	4	4	7	5	30	32
Streaming by variance	15	5	6	3	6	5	9	6	27	30
Streaming by variance	16	3	7	6	7	8	12	8	21	28
Streaming by variance	17	8	9	11	9	10	15	12	18	25
Streaming by variance	18	11	13	13	10	13	20	15	13	22
Streaming by variance	19	16	21	16	13	17	28	21	9	15
Streaming by variance	20	7	24	12	14	14	27	18	8	10
Mean and var buckets	1	40	38	40	40	40	40	40	10	6
Mean and var buckets	2	34	19	23	25	20	3	16	40	40
Mean and var buckets	3	24	12	20	18	19	5	13	39	38
Mean and var buckets	4	21	10	21	16	21	8	20	38	36
Mean and var buckets	5	20	11	22	17	22	10	22	36	34
Mean and var buckets	6	22	14	24	19	24	13	23	34	31
Mean and var buckets	7	23	15	25	22	25	14	24	32	29
Mean and var buckets	8	25	16	26	24	26	18	26	31	26
Mean and var buckets	9	26	22	27	27	27	21	27	28	24
Mean and var buckets	10	27	23	28	28	28	24	28	26	23
Mean and var buckets	11	28	25	29	29	29	26	29	25	21
Mean and var buckets	12	29	26	30	30	30	29	30	24	20
Mean and var buckets	13	30	27	31	31	31	31	31	23	19
Mean and var buckets	14	31	28	32	32	32	32	32	22	17
Mean and var buckets	15	32	29	33	33	33	33	33	19	13
Mean and var buckets	16	33	30	34	34	34	34	34	20	12
Mean and var buckets	17	35	31	35	35	35	35	35	17	11
Mean and var buckets	18	36	32	36	36	36	36	36	16	9
Mean and var buckets	19	37	33	37	37	37	37	37	14	8
Mean and var buckets	20	38	34	38	38	38	38	38	15	7

Table 8: 500 random draws, skewed distribution, streaming by variance algorithm and mean and variance buckets algorithm

Algorithm	Number of levels	Variance of left-out-variance parameter	Trace of $(X'X)^{-1}$	Trace of $(X'X)$	Determinant of $(X'X)$	SKL-divergence of X from $X(R_M)$	Expected variance of Y, true β	Expected variance of Y, incorrect β ($\hat{\beta}_1$)	Expected variance of Y, incorrect β ($\hat{\beta}_2$)	Sum of leave-out-variances
Streaming by variance	1	40	40	40	40	40	38	38	2	2
Streaming by variance	2	5	15	38	26	23	40	40	3	3
Streaming by variance	3	1	2	25	16	17	27	27	15	15
Streaming by variance	4	6	10	27	21	16	35	35	13	12
Streaming by variance	5	7	8	24	17	20	28	28	18	16
Streaming by variance	6	3	4	21	13	14	23	23	20	19
Streaming by variance	7	2	1	9	1	8	11	12	32	31
Streaming by variance	8	4	3	15	2	12	17	17	28	26
Streaming by variance	9	12	11	18	11	19	18	18	25	23
Streaming by variance	10	15	14	17	15	18	16	16	23	24
Streaming by variance	11	19	19	1	14	1	1	1	40	40
Streaming by variance	12	18	18	2	12	4	2	2	39	39
Streaming by variance	13	17	17	3	10	5	3	3	38	38
Streaming by variance	14	16	16	4	8	2	4	4	37	37
Streaming by variance	15	14	13	5	6	3	5	5	36	36
Streaming by variance	16	9	6	6	4	6	6	6	33	33
Streaming by variance	17	8	5	10	3	7	9	9	31	32
Streaming by variance	18	10	7	11	5	9	10	10	29	30
Streaming by variance	19	11	9	14	7	11	14	14	24	27
Streaming by variance	20	13	12	13	9	10	12	11	27	28
Mean and var buckets	1	40	40	40	40	40	38	38	2	2
Mean and var buckets	2	31	31	8	25	15	8	8	35	35
Mean and var buckets	3	23	23	7	20	13	7	7	34	34
Mean and var buckets	4	20	20	12	18	21	13	13	30	29
Mean and var buckets	5	21	21	16	19	22	15	15	26	25
Mean and var buckets	6	22	22	19	22	24	19	19	22	22
Mean and var buckets	7	24	24	20	23	25	20	20	21	21
Mean and var buckets	8	25	25	22	24	26	21	21	19	20
Mean and var buckets	9	26	26	23	27	27	22	22	17	18
Mean and var buckets	10	27	27	26	28	28	24	24	16	17
Mean and var buckets	11	28	28	28	29	29	25	25	14	14
Mean and var buckets	12	29	29	29	30	30	26	26	12	13
Mean and var buckets	13	30	30	30	31	31	29	29	11	11
Mean and var buckets	14	32	32	31	32	32	30	30	10	10
Mean and var buckets	15	33	33	32	33	33	31	31	9	9
Mean and var buckets	16	34	34	33	34	34	32	32	8	8
Mean and var buckets	17	35	35	34	35	35	33	33	7	7
Mean and var buckets	18	36	36	35	36	36	34	34	6	6
Mean and var buckets	19	37	37	36	37	37	36	36	5	5
Mean and var buckets	20	38	38	37	38	38	37	37	4	4

References

Angrist, Joshua D. 2014. "The perils of peer effects." *Labour Economics* 30:98–108.

Boltz, Sylvain, Eric Debreuve, and Michel Barlaud. 2007. "kNN-based high-dimensional Kullback-Leibler distance for tracking." In *Image Analysis for Multimedia Interactive Services, 2007. WIAMIS'07. Eighth International Workshop on*, 16–16. IEEE.

Booij, Adam S, Edwin Leuven, and Hessel Oosterbeek. 2015. "Ability peer effects in university: Evidence from a randomized experiment."

Caeyers, Bet, and Marcel Fafchamps. 2016. *Exclusion bias in the estimation of peer effects.* Technical report. National Bureau of Economic Research.

Carrell, Scott E, Richard L Fullerton, and James E West. 2009. "Does your cohort matter? Measuring peer effects in college achievement." *Journal of Labor Economics* 27 (3): 439–464.

Carrell, Scott E, Bruce I Sacerdote, and James E West. 2013. "From natural variation to optimal policy? The importance of endogenous peer group formation." *Econometrica* 81 (3): 855–882.

Cramer, Jan Solomon. 1987. "Mean and variance of R2 in small and moderate samples." *Journal of econometrics* 35 (2-3): 253–266.

Duflo, Esther, Pascaline Dupas, and Michael Kremera. 2011. "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya." *The American Economic Review* 101 (5): 1739–1774.

Feld, Jan, and Ulf Zölitz. 2017. "Understanding peer effects: on the nature, estimation, and channels of peer effects." *Journal of Labor Economics* 35 (2): 387–428.

Krackhardt, David. 1988. "Predicting with networks: Nonparametric multiple regression analysis of dyadic data." *Social networks* 10 (4): 359–381.

Lee, Lung-fei. 2007. "Identification and estimation of econometric models with group interactions, contextual factors and fixed effects." *Journal of Econometrics* 140 (2): 333–374.

Manski, Charles F. 1993. "Identification of endogenous social effects: The reflection problem." *The review of economic studies* 60 (3): 531–542.

Pearl, Judea. 2009. *Causality.* Cambridge university press.

Pinto, CC d X. 2010. "Semiparametric Estimation of Peer Effects in Classrooms: Evidence for Brazilian Schools in 2003." *Trabalho apresentado nos seminários de UCIrvine School of Social Sciences.*

Rossi, Fabrice, Amaury Lendasse, Damien François, Vincent Wertz, and Michel Verleysen. 2006. "Mutual information for the selection of relevant variables in spectrometric nonlinear modelling." *Chemometrics and intelligent laboratory systems* 80 (2): 215–226.

Shalizi, Cosma Rohilla, and Andrew C Thomas. 2011. "Homophily and contagion are generically confounded in observational social network studies." *Sociological methods & research* 40 (2): 211–239.

Singh, Harshinder, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. 2003. "Nearest neighbor estimates of entropy." *American journal of mathematical and management sciences* 23 (3-4): 301–321.

Srivastava, Anil K, Virendra K Srivastava, and Aman Ullah. 1995. "The coefficient of determination and its adjusted version in linear regression models." *Econometric reviews* 14 (2): 229–240.

Zimmerman, David J. 2003. "Peer effects in academic outcomes: Evidence from a natural experiment." *The Review of Economics and statistics* 85 (1): 9–23.