



## Exploratory Mediation Analysis with Many Potential Mediators

Erik-Jan van Kesteren & Daniel L. Oberski

To cite this article: Erik-Jan van Kesteren & Daniel L. Oberski (2019) Exploratory Mediation Analysis with Many Potential Mediators, Structural Equation Modeling: A Multidisciplinary Journal, 26:5, 710-723, DOI: [10.1080/10705511.2019.1588124](https://doi.org/10.1080/10705511.2019.1588124)

To link to this article: <https://doi.org/10.1080/10705511.2019.1588124>



© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.



View supplementary material [↗](#)



Published online: 11 Apr 2019.



Submit your article to this journal [↗](#)



Article views: 1842



View related articles [↗](#)



View Crossmark data [↗](#)



# Exploratory Mediation Analysis with Many Potential Mediators

Erik-Jan van Kesteren<sup>id</sup> and Daniel L. Oberski<sup>id</sup>

*Utrecht University*

Social and behavioral scientists are increasingly employing technologies such as fMRI, smartphones, and gene sequencing, which yield ‘high-dimensional’ datasets with more columns than rows. There is increasing interest, but little substantive theory, in the role the variables in these data play in known processes.

This necessitates exploratory mediation analysis, for which structural equation modeling is the benchmark method. However, this method cannot perform mediation analysis with more variables than observations. One option is to run a series of univariate mediation models, which incorrectly assumes independence of the mediators. Another option is regularization, but the available implementations may lead to high false-positive rates.

In this article, we develop a hybrid approach which uses components of both filter and regularization: the ‘Coordinate-wise Mediation Filter’. It performs filtering conditional on the other selected mediators. We show through simulation that it improves performance over existing methods. Finally, we provide an empirical example, showing how our method may be used for epigenetic research.

**Keywords:** Mediation analysis, high-dimensional data, feature selection

## INTRODUCTION

Social and behavioral scientists are increasingly employing technologies such as fMRI, smartphones, and gene sequencing, which yield high-dimensional datasets with more variables than observations. These high-dimensional data are often intended to answer questions such as “*which areas of our brain are relevant for pain perception?*” (Atlas, Lindquist, Bolger, & Wager, 2014) and “*which genes mediate the effect of trauma on stress reactivity?*” (Houtepen et al., 2016). These are questions regarding exploratory mediation analysis (EMA).

Structural equation modeling (SEM) is the preferred method for mediation analysis with multiple mediators (Preacher & Hayes, 2008; Vanderweele & Vansteelandt, 2014). With this method, it is possible to determine to what extent specific  $M$  variables mediate the  $X \rightarrow Y$  effect conditional on the presence of other mediators in the model. However, this method fails when the data are *high-dimensional* when the variables under investigation outnumber the samples  $N$ . In this situation, the observed covariance matrix is rank-deficient, leading to linear dependence in the observed moments and, for the full mediation model, nonconvergence.

Several alternative methods for EMA have been proposed to deal with this issue. One option mentioned by Preacher and Hayes (2008) is to select relevant mediators from a series of univariate  $X \rightarrow M \rightarrow Y$  mediation models (e.g. Boca, Sinha, Cross, Moore, & Sampson, 2014; Liu et al., 2013). We call this the filter method, following the taxonomy of Guyon and Elisseeff (2003). Its main advantages are that it is simple to explain and run, requiring only  $P$  univariate path models. On the other hand, the “filter” method introduces bias through model misspecification: it

---

Correspondence should be addressed to Erik-Jan van Kesteren  
Department of Methodology and Statistics, Utrecht University Padualaan  
14, 3584 CH, Utrecht, Netherlands. E-mail: [e.vankesteren1@uu.nl](mailto:e.vankesteren1@uu.nl)

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/hsem](http://www.tandfonline.com/hsem).

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

takes into account only the *marginal* relationships of  $M$  with  $X$  and  $Y$ . A pitfall of this is that a variable useless by itself can be useful together with others (Guyon & Elisseeff, 2003). In other words, a certain mediator may be marginally irrelevant, but relevant conditional on another set of mediators.

Recently, another multivariate method was introduced by Serang, Jacobucci, Brimhall, and Grimm (2017). Their proposal was to perform EMA through regularized estimation of the full structural equation model: “XMed”. This method automatically shrinks small regression paths to 0, leading to a selection of potential mediators: mediators are variables for which both the  $X \rightarrow M$  path and the  $M \rightarrow Y$  path are nonzero after regularization. With this method, it is possible to detect mediators which are only relevant conditionally, while regularization resolves the identification issues of default SEM (Hastie, Tibshirani, & Wainwright, 2015). The disadvantage is that this method finds paths with a large effect rather than the desired subset of mediators: the regularization in XMed shrinks small  $\beta$  paths to 0, irrespective of the value of their associated  $\alpha$  paths – shrinkage is performed on all paths equally. This leads to inflated false-positive rates as reported by Serang et al. (2017) and Jacobucci, Brandmaier, and Kievit (2018). In summary, regularization methods do perform conditional estimation, but they select paths rather than mediators.

In this paper, we propose a hybrid approach to EMA which we call the “Coordinate-wise Mediation Filter” (CMF). This method combines advantages from both the filter and regularization methods: (a) it converges in case of high-dimensional data, (b) it takes into account mediator correlations, leading to the conditional selection of mediators, and (c) it selects based on mediation, not paths. CMF performs univariate filtering *conditional* on the other selected mediators by using an algorithm from regularized regression: cyclical coordinate descent on residuals (Breheny & Huang, 2011; Friedman, Hastie, & Tibshirani, 2009).

The remainder of the article is structured as follows: first, we provide relevant background on exploratory mediation analysis. Then, we outline the Coordinate-wise Mediation Filter as a hybrid method for mediator subset selection. Following this, we show through simulation where each of the discussed methods performs as well as SEM. In addition, we assess the performance of CMF relative to the other available methods in a high-dimensional simulation. Lastly, the CMF procedure is illustrated by applying it to the epigenetic process of trauma and stress reactivity.

### Exploratory mediation analysis

The fundamental goal of mediation analysis is to determine the process by which a variable  $X$  influences another variable  $Y$  (MacKinnon, Lockwood, & Williams, 2004). Exploratory mediation analysis (EMA) in particular is

used to explore a dataset for potential mediating variables (MacKinnon, 2008). In other words, EMA pertains to determining among multiple potential mediators which subset is most relevant. Through EMA, researchers can build theory and select variables of interest for further research into the process under investigation.

An example application of EMA is the research by Ammerman et al. (2018), who investigated how childhood maltreatment leads to suicidal behavior. They defined 46 potential mediators, including psychological counseling, closeness to parents, and self-esteem. The authors did not test a fully specified mediation model about the precise relations of each of these variables to childhood maltreatment and suicidal behavior. Instead, this study was exploratory, identifying which variables were the most relevant targets for future research. Indeed, the authors conclude that the study highlights factors that may be potential targets for risk assessment and for treatment among adolescents with a history of childhood maltreatment.

### Univariate mediation analysis and the filter method

A common framework for *univariate* mediation analysis is a system of regression equations (Equation (1); MacKinnon et al., 2004). The system is displayed graphically in Figure 1. In the present paper, we consider only the case where the data from  $X$ ,  $M$ , and  $Y$  are continuous and their relations are linear. For nonlinear discrete extensions to mediation analysis, see Hayes and Preacher (2010) and Hayes and Preacher (2014), respectively. For further details, refer to the reviews by MacKinnon, Fairchild, and Fritz (2007) and Preacher (2015).

$$M = \mu_M + \alpha X + e_M$$

$$Y = \mu_Y + \tau X + \beta M + e_Y \quad (1)$$

Under the standard assumptions of linear SEM, the parameter estimates of this system may be used to determine whether  $M$  is a mediator — a dichotomous decision. There are several ways to make this decision, usually based on a quantity of interest  $q$  and a measure of uncertainty (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). For example,  $q$  may represent the size of the indirect effect through the product of its coefficients  $q_{\text{prod}} = \alpha\beta$ , and uncertainty measures for  $q_{\text{prod}}$  can be obtained using asymptotic

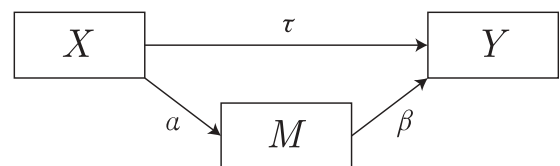


FIGURE 1 Graphical representation of the system of Equation (1). For clarity, the residuals are not shown.

standard error methods (e.g., Olkin & Finn, 1995; Sobel, 1986) or bootstrapping (Preacher & Hayes, 2008).

Combining the quantity of interest  $q$  with an uncertainty estimate and a specified alpha level yields a dichotomous decision criterion based on a  $p$ -value. We call this *univariate decision function*  $\mathcal{D}$ : a function that maps the data of  $X$ ,  $M$ , and  $Y$  to a binary decision of whether  $M$  should be considered a mediator (1) or not (0).

$$\mathcal{D} : (\mathbf{x}, \mathbf{m}, \mathbf{y}) \mapsto \{0, 1\}$$

Note that any function that follows this specification can be considered a decision function, regardless of complexity. An example of higher complexity decision functions is given by VanderWeele (VanderWeele, 2015, p. 46), who states exposure-outcome confounding should by default be controlled for when testing for mediation. The decision function encodes the researcher's definition of mediation: a product of coefficients decision function with a  $p$ -value cutoff of 0.1 will lead to different results than an exposure-outcome controlled decision function with a stricter cutoff.

This decision function framework thus provides a convenient abstraction, highlighting a key advantage for mediation analysis methods: if the choice of decision functions is flexible, a method is adaptable to the specific needs of a researcher. If researchers want to follow the recommendation of VanderWeele (2015), they can do so by adding an  $XM$  interaction term into the decision function.

While these decision functions are univariate, EMA is an inherently multivariate procedure, requiring analysis of multiple indirect effects. To perform EMA, a researcher can apply their chosen decision function to each mediator separately, through  $P$  different mediation models as in Figure 1. This filter method will result in a subset of relevant mediators. However, the implicit assumption is that the  $M$  variables are independent of one another. In other words, the selected subset will not include mediators that are relevant only conditionally on another mediator.

### Multivariate mediation analysis and xmed

To make mediation decisions multivariately, Preacher and Hayes (2008) recommend the SEM approach. In this approach, the quantities of interest  $q_1, \dots, q_P$  and their uncertainty are estimated directly from a multiple mediation model as in Figure 2. A decision can then be made for each individual  $M_p$  based on its multivariately estimated quantity  $q_p$ . Unlike the filter method, this approach estimates  $q_p$  conditional on the other  $P - 1$  quantities, so that marginally irrelevant true mediators may still be detected.

However, the SEM approach is unavailable in the case of high-dimensional data because SEM parameters are estimated from observed covariances. High dimensional data ( $P > N$ ) leads to a  $P \times P$  observed covariance matrix of at most rank  $N$ , meaning a linear dependence exists among elements. If

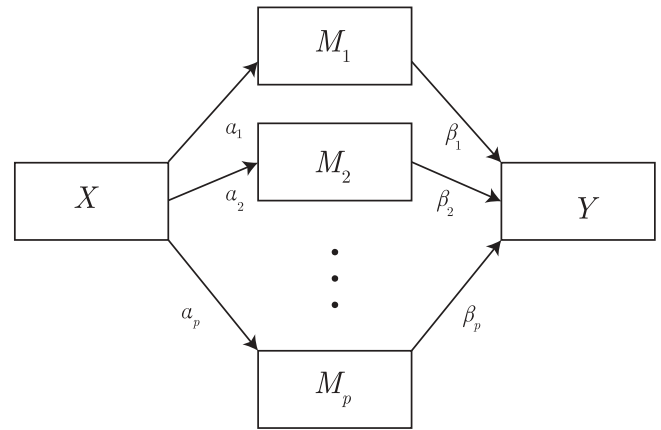


FIGURE 2 Exploratory mediation analysis with a set of  $p$  potential mediators  $M$ . For clarity, we omitted the  $P(P + 1)/2$  parameters belonging to the residuals of  $M$  and their covariances, as well as the residual variance of  $Y$ .

dependent elements are mapped to separate parameters in the SEM model, an infinite number of solutions exist for the same log-likelihood, so there is no maximum likelihood solution. This is the case in the full mediation model. As an alternative intuitive explanation, it is possible to view the  $M \rightarrow Y$  part of the mediation model as a high-dimensional multiple regression, where ordinary least squares (OLS) estimates are unavailable because the covariance matrix cannot be inverted (Hastie et al., 2015).

XMed is an adjustment to the SEM method that not only allows for high-dimensional data, but it also automatically selects a subset of mediators without an explicit decision function. The estimation method for XMed is RegSEM (Jacobucci, Grimm, & McArdle, 2016), which applies regularization to a chosen subset of model parameters in a structural equation model. This shrinkage is determined by the hyperparameter  $\lambda$  along with the penalization function  $P(\cdot)$  in the objective function of RegSEM:

$$F_{regsem} = F_{ML} + \lambda P(\cdot)$$

where  $\cdot$  is a vector of parameters.

In XMed specifically, shrinkage is applied to the vectors of  $\alpha$  ( $x \rightarrow M$ ) and  $\beta$  ( $M \rightarrow y$ ) parameters. Subset selection of the mediators occurs through the chosen regularization method; the penalty function  $P(\cdot)$  is the LASSO penalty, the  $\ell_1$  norm of the chosen parameter vector:  $P(\cdot) = \|\cdot\|_1$ . Depending on the value of  $\lambda$ , the LASSO penalty shrinks the smallest of the chosen parameters to 0 during estimation. This immediately forms the decision rule: for a potential mediator  $M_p$ , if  $\alpha_p$  or  $\beta_p$  equals 0, then the estimated indirect effect  $\alpha\beta_p$  is 0, thus  $M_p$  is not considered to be a true mediator.

A well-known algorithm for computing the LASSO solution, which can also be applied in SEM, is coordinate-wise conditioning or coordinate descent: the conditional solution is well-known and easy to find, in SEM the

maximum likelihood estimates, and the penalized solution is found by cyclically updating and soft-thresholding the conditional solution for each parameter in turn, until convergence (Hastie et al., 2015).

A sequential combination of the ideas of filtering and regularization was proposed by Zhang et al. (2016) in a three-step approach called HIMA. First, in the screening step, the authors marginally filter irrelevant potential mediators based on the  $M \rightarrow Y$  relations. Second, the remaining  $M \rightarrow Y$  paths are estimated with regularization. Lastly, the test step performs the joint significance test as introduced by Baron and Kenny (1986) with Bonferroni correction on the remaining mediators.

The main disadvantage of these methods is that there is a pertinent difference between (a) penalized estimation of the paths and (b) finding mediators. For XMed, a relatively small  $\alpha_p$  path will be shrunk to 0 before stronger  $\alpha$  paths, regardless of the strength of its associated  $\beta_p$  path. This holds for HIMA too, since in the selection stage it considers only  $\beta$  paths. Thus, these methods do not target *mediators with strong indirect effects*  $\alpha\beta$ , but *intermediate variables with strong  $\alpha$  or  $\beta$  paths*. Even though these methods do work conditionally, they make the implicit assumption that the mediators also have the strongest  $X \rightarrow M$  and  $M \rightarrow Y$  paths, which need not be so.

Rephrasing this in terms of decision functions, the regularization methods exclude variables which have a relatively weak covariance with  $X$  or  $Y$ . However, this decision criterion only partially captures theoretically plausible mediators: true mediators may exist for which the covariance with  $X$  or  $Y$  is relatively weak, but the indirect effect  $\alpha\beta$  is relatively strong. The regularization methods will thus underperform in the presence of noise variables which are not mediators, but which strongly covary with either  $X$  or  $Y$ . We illustrate this in the simulation section.

In conclusion, to perform EMA, (a) the SEM method is optimal but unavailable for high-dimensional data, (b) the filter method is simple and flexible but does not select mediators conditionally, and (c) regularization methods do proper conditioning but are estimating paths rather than selecting mediators.

## COORDINATE-WISE MEDIATION FILTER

We propose a hybrid method, the Coordinate-wise Mediation Filter (CMF), which contains both theory-driven decision functions and conditional estimation of the quantity of interest. Like the filter method, CMF applies a decision function to each of the mediators,

but it performs this task conditional on the set of currently selected mediators. The procedure is similar to cyclical coordinate descent, the algorithm underlying regularization procedures in various software implementations – but differs in that mediation rather than separate regression paths are explicitly identified as the target. A key component of this algorithm is the use of residuals to remove dependency among the coordinates (Hastie et al., 2015). CMF generalizes this idea to mediator selection with arbitrary objective functions.

The CMF implementation consists of two components: an inner algorithm, which handles feature selection using the decision function  $\mathcal{D}$  through coordinate descent, and an outer algorithm, which performs random starts, feature subsampling, and subsequent aggregation. The combined procedure can be characterized as a stochastic coordinate descent algorithm. The following two sections give a detailed outline of the inner and outer algorithm.

### Inner algorithm

First, we initialize a vector of length  $P$  which contains the current mediator selection in the form of 0 and 1 values – the starting values. A *step* is then as follows: for each potential mediator  $M_p$ , create a data matrix  $\mathbf{M}_*$ , which contains all the mediators currently selected, excluding the variable  $M_p$  under consideration. Then, perform the decision function  $\mathcal{D}$  on the parts of  $\mathbf{x}$  and  $\mathbf{y}$  orthogonal to (conditional on) this matrix. This conditioning is performed through calculating the *residuals* of  $\mathbf{x}$  and  $\mathbf{y}$  with respect to  $\mathbf{M}_*$ :

$$\begin{aligned} \mathbf{r}_x &= \mathbf{x} - \mathbf{M}_*(\mathbf{M}_*'\mathbf{M}_*)^{-1}\mathbf{M}_*'\mathbf{x} \\ \mathbf{r}_y &= \mathbf{y} - \mathbf{M}_*(\mathbf{M}_*'\mathbf{M}_*)^{-1}\mathbf{M}_*'\mathbf{y} \end{aligned}$$

The decision function is thus performed as  $\mathcal{D}(\mathbf{r}_x, \mathbf{M}_p, \mathbf{r}_y)$ , leading to a binary decision whether mediator  $p$  selected, conditional on  $\mathbf{M}_*$ .

The inner algorithm is run continuously, randomly ordering the choice of  $p$  in each iteration. It stops either when the mediator selection does not change from one step to the next or when the prespecified maximum number of iterations is reached. The resulting program, shown in Algorithm 2.0.1, is a binary, randomized form of cyclical coordinate descent similar to those in Hastie et al. (2015). The randomization improves stability for very high-dimensional data (Nesterov, 2012). Richtárik and Takáč (2014) show that this method attains relatively fast convergence even with a billion variables in a sparse regression situation.

Algorithm 1	Inner CMF algorithm
1: $\text{scale}(x)$ ; $\text{scale}(\mathbf{M})$ ; $\text{scale}(y)$	
2: $P \leftarrow \text{ncol}(\mathbf{M})$	▷ number of mediators
3: $\text{decvec} \leftarrow 0_1, 0_2, \dots, 0_P$	▷ initialise 0/1 decision vector
4: <b>repeat</b>	
5: <b>for</b> $p$ in $1:P$ <b>do</b>	
6: $\mathbf{M}_* \leftarrow \mathbf{M}_{[:, \text{decvec} \& !p]}$	▷ selected mediators excluding $p$
7: $\mathbf{r}_x \leftarrow \mathbf{x} - \mathbf{M}_*(\mathbf{M}_*'\mathbf{M}_*)^{-1}\mathbf{M}_*'\mathbf{x}$	▷ residual of $\mathbf{x}$
8: $\mathbf{r}_y \leftarrow \mathbf{y} - \mathbf{M}_*(\mathbf{M}_*'\mathbf{M}_*)^{-1}\mathbf{M}_*'\mathbf{y}$	▷ residual of $\mathbf{y}$
9: $\text{decvec}[p] \leftarrow \mathcal{D}(\mathbf{r}_x, \mathbf{M}_{[:, p]}, \mathbf{r}_y)$	▷ decision function
10: <b>until</b> $\text{decvec} == \text{decvec}_{\text{prev}}$	▷ convergence when decvec is stable

### Outer algorithm

The value of the decision vector resulting from the inner algorithm depends to some extent on the starting values, due to the discrete nature of its coordinates. Therefore, the algorithm is embedded in an outer loop that performs multiple random starts. After aggregating the results from the different starts, the decision vector of length  $P$  is continuous: each element  $p$  in this vector signifies the proportion of times the potential mediator  $M_p$  was selected by the inner algorithm. These proportions or *empirical selection probabilities*, naturally lead to a mediator ranking. This ranking can then again be dichotomized using a cutoff score.

The second essential part in the outer algorithm is *feature sampling*. With feature sampling, the inner algorithm will loop over only  $\lceil \sqrt{P} \rceil$  potential mediators at each iteration. This procedure is similar to how the random forest decorrelates its trees (Breiman, 2001). Zhang, Zhao, Zhang, and Wei (2017) show in a sparse regression setting that feature sampling improves and stabilizes the performance of feature selection. Furthermore, there are links between feature sampling and shrinkage: for linear regression, considering only  $\lceil \sqrt{P} \rceil$  variables during training is equivalent to ridge regression on the standardized predictors. This generalizes to more complex methods such as GLM (Wager, Wang, & Liang, 2013). Feature sampling in the CMF algorithm thus takes on the crucial role of regularization.

The entire CMF procedure is implemented in the R package *cmfilter*, available from <https://github.com/vankesteren/cmfilter>. An example analysis with specific hyperparameters and cutoff score determination is described in the application section to this paper, with accompanying R code in the supplementary material.

The CMF method addresses the most important issues associated with both filter and regularization methods: it

conditions on the other mediators while simultaneously being flexible to the choice of theoretically relevant decision functions. In the next section, we investigate the performance of CMF through simulation.

## SIMULATIONS

This section is subdivided into two parts. The first part aims to show empirically the theoretical advantages and disadvantages of SEM, filter, XMed, HIMA, and CMF. We simulate specific conditions which are theoretically challenging for some but not all methods. The results from the first section are aimed at generating an understanding of the theoretical background in the present paper.

The second part is aimed at simulating real-world performance in a controlled high-dimensional situation. The results from this section indicate to what extent the CMF method outperforms its rival methods in practice, in addition to providing an anchor for the expected absolute level of performance in terms of false positives and true positives in such a situation.

All the simulations were run on R version 3.5.0 (R Core Team, 2018). The full environment used for the simulations is shown in [Appendix A](#).

### Theoretical conditions

The goal of this section is to illustrate when each method performs adequately and when it does not. Two situations are of particular interest: (a) suppression through correlation among mediators, and (b) noise in the  $\alpha$  and  $\beta$  paths, overshadowing a potential mediator. Filter methods are likely to underperform in terms of power in the first case, as the effect of a mediator is dependent on another and marginally invisible. In the second case, the regularization methods are theorized to under-perform because the  $\alpha$  and  $\beta$  paths are regularized independently whereas it is their combination that indicates mediation.

The data were controlled to behave according to the population, i.e., the data was transformed to exhibit the exact correlation matrix implied by the data-generating model (see [Appendix B](#)). In each simulation, we show the power and false discovery rates of the three methods in 100 simulated datasets of 400–600 observations. The decision function under consideration for the filter, SEM, and CMF methods was the Sobel test (Sobel, 1986), one of the most common tests in the product of coefficients category (MacKinnon et al., 2002). For these tests, any variable with a  $p$ -value below .1 was considered to be a mediator. The SEM and filter methods were implemented using the *lavaan* package (Rosseeel, 2012), and CMF was implemented using the accompanying *cmfilter* package. For XMed, the *regsem* package (Jacobucci et al., 2016) was used with cross-validation was to find the optimal penalty parameter,



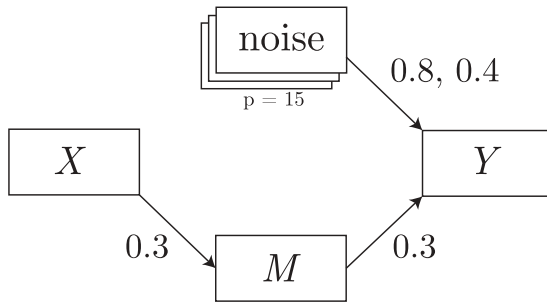


FIGURE 5 Data-generating model for the simulation of noise in the  $\beta$  paths.

TABLE 3

Selection Rates of Each Mediator in 100 Simulated Datasets Where the Noise Variables (2–16) Have a Nonzero Relation with the  $Y$  Variable.  $M$  Is the True Mediator, Dot Indicates 0

Method	M	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
SEM	99	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Filter	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
XMed	92	5	5	6	6	3	4	2	3	4	2	3	4	5	6	3
HIMA	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
CMF	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

The results of this combined simulation, displayed in Table 4 show again that CMF performs at the benchmark level. An interesting quantity for the imperfect methods is the positive predictive value (PPV): the probability that a mediator selected by a method truly mediates the effect of  $X$  on  $Y$ . For filter and XMed methods, the PPV is lowered through either a relatively low true positive rate (power) or a high false-positive rate (type-I error).

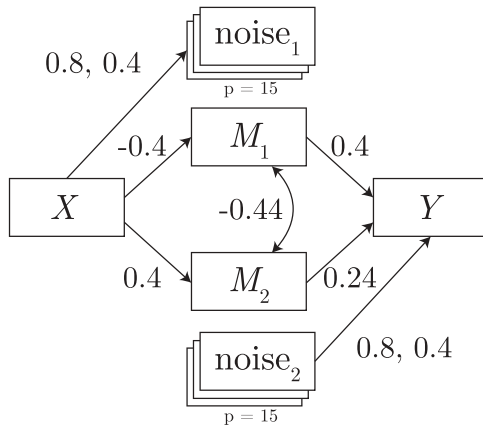


FIGURE 6 Data-generating model for the simulation of suppression with noise in the  $\alpha$  and  $\beta$  paths.

TABLE 4

True Positive Rates, False-Positive Rates, and Positive Predictive Values (PPV) of the Combined Suppression and Noise Simulation. The PPV Indicates the Probability that a Mediator Selected by the Method Is a True Mediator

Method	Power M	Power M <sub>2</sub>	FPR	PPV
SEM	0.99	0.99	0.00	1.00
Filter	1.00	0.00	0.00	1.00
XMed	0.88	0.87	0.10	0.37
HIMA	1.00	0.00	0.00	1.00
CMF	1.00	1.00	0.00	1.00

### Interim conclusion

While the considered data-generating mechanisms are very specific, the differences in performance between the methods can be exacerbated and diminished by altering the parameter values while preserving the structure. Overall, CMF is the only method that performs as well as the baseline in all of these data-generating mechanisms. Together, they show that this method is robust to boundary cases where other methods may fail. This is a valuable property of a mediator selection method, because these situations may occur simultaneously, with no way to test them in real-world datasets. In the next part, we explore how well the CMF method performs in high-dimensional circumstances, where the baseline optimal SEM method cannot work.

### High-dimensional mediation simulation

In this section, we compare the performance of the available EMA methods in a simplified high-dimensional situation. Due to the wide nature of the dataset ( $p = 1000$ ), the benchmark default SEM method is unavailable.

### Simulation setup

Following one of the high-dimensional simulation conditions of Zhang et al. (2016), the dataset consists of 100 samples and 1000 potential mediators. These mediators are generated in four uncorrelated blocks: one block with true mediators ( $M$ ), one with noise variables related to  $X$  ( $A$ ), one noise block covarying with  $Y$  ( $B$ ), and one large white noise block without any covariance ( $I$ ). The general structure can be found in Figure 7. For each of the simulations, this structure was created as a sparse block matrix using the Matrix package (Bates & Maechler, 2017), after which multivariate normal data was generated using the sparseMVN package (Braun, 2018). Specific data generation and simulation R code can be found in the supplementary materials.

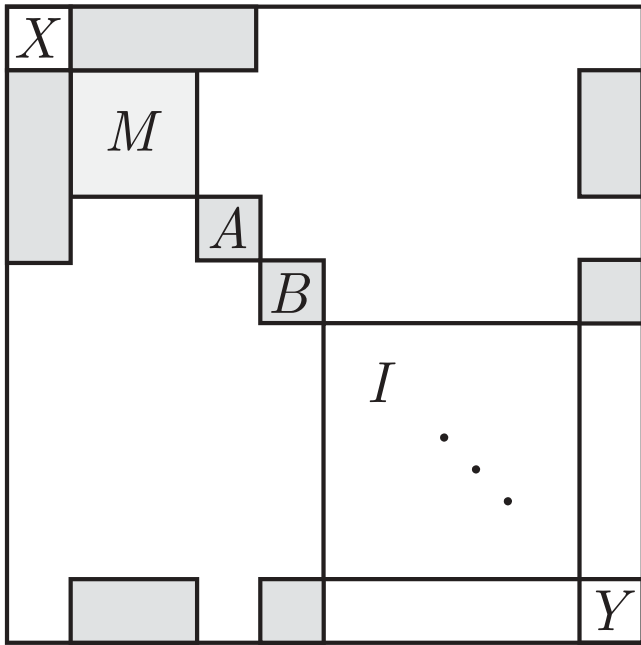


FIGURE 7 General covariance structure for the high-dimensional performance simulation. In the white sections of the matrix, there is no covariance. The true mediator block  $M$  is related to both  $X$  and  $Y$ , whereas the correlating noise blocks are related to either  $X$  (block  $A$ ) or  $Y$  (block  $B$ ). The largest block is the identity matrix block  $I$ , which generates only unrelated noise variables. Note that unlike the illustrative simulations, these data favor the filter method: there is no suppression or excessive interdependence of potential mediators. Therefore, the filter method is the benchmark in this simulation. The XMed method was omitted from this simulation because it requires estimation of the full SEM model before regularizing: it would need to be adjusted to work with high-dimensional data.

## Results

The results are displayed in Table 5. The CMF method has the highest true positive rate, and a medium false-positive rate, leading to a similar positive predictive value (PPV) to the filter method. In other words, the mediators selected by CMF are as likely to be true mediators as those selected by the benchmark filter method. As true positive rates and false-positive rates can be adjusted by the choice of alpha level, we conclude that the CMF method also performs at the benchmark level in this high-dimensional situation.

TABLE 5

True Positive Rates, False-Positive Rates, and Positive Predictive Values for the High-Dimensional Data Simulation. Note that XMed Failed to Run As-Is for the Simulated Datasets, as It Required Running the Full SEM Model before Regularizing

	Power	Type I Error	PPV
CMF	0.265	0.003	0.507
Filter	0.241	0.002	0.512
HIMA	0.069	0.009	0.032

## APPLICATION TO EPIGENETIC DATA

In this section, we show how the CMF method can be used for exploratory mediation analysis in a real-world setting. Aside from the results shown here, the full R syntax is available in the supplementary material.

Houtepen et al. (2016) researched which locations in the genome are likely to mediate the relation between childhood trauma and stress reactivity later in life. In order to identify the genomic locations, they measured methylation at CpG sites using array-based technology. In a discovery sample, they found a location of interest which they subsequently researched further and related to functional changes in the human prefrontal cortex.

Here, we re-analyze the original discovery sample dataset to investigate whether CMF yields different potentially relevant locations compared to the correlational filter analysis of the original authors.

## Dataset and preprocessing

The dataset of the discovery sample was obtained from ArrayExpress, the data repository of the European Bioinformatics Institute: <https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-77445>. The sample consists of 85 healthy individuals. The  $X$  variable is a score on a childhood trauma questionnaire, and the  $Y$  variable is the increase in cortisol after a stress test defined as an increase in the area under the curve (iAUC). The 385 884 potential mediators  $M$  were taken from the analysis of DNA methylation in the blood, with default preprocessing. From the available respondent characteristics, age and sex were considered to be confounders. For full details of the dataset, see Houtepen et al. (2016).

Before analysis,  $X$ ,  $Y$ , and  $M$  were residualized with respect to their intercept, age, and sex. Since the number of  $M$  variables was so large, the last preprocessing step was a straightforward univariate filter. For this, the top 1000 potential mediators in terms of their absolute product of correlations with  $X$  and  $Y$  were retained. For more details, see the preprocessing R code in the supplementary materials.

## Analysis and results

The CMF algorithm was performed using the centered  $X$  and  $Y$  and the 1000 potential mediators  $M$ . The Sobel test with a  $p$ -value of 0.1 was used as the decision function  $\mathcal{D}$  and 10 000 iterations with random starts were run to ensure the stability of the results. After inspecting the scree plot of the selection rates, the cutoff for selection was set to 0.075. The resulting selection rates and selected cg locations in the genome are shown in Figure 8.

These locations were annotated using the BioConductor package FDb.InfiniumMethylation.hg18 (Triche, 2014) to find the nearest protein-coding gene. The shortened

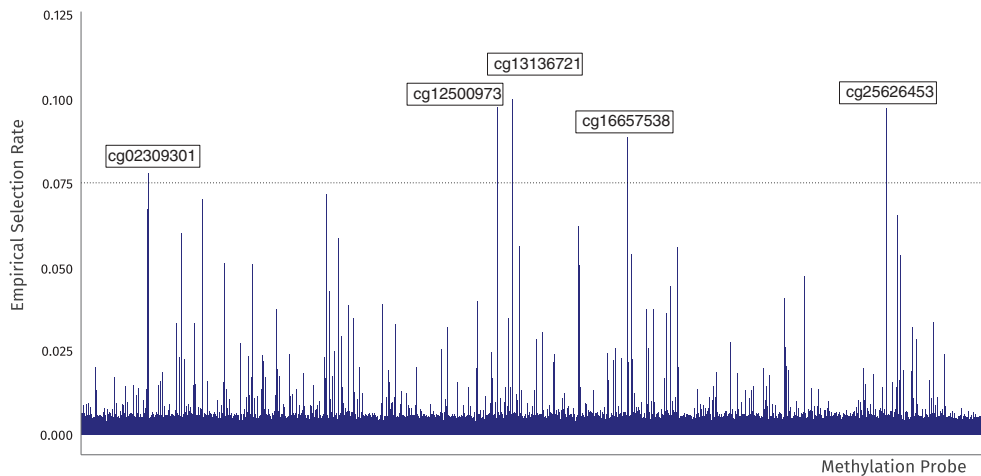


FIGURE 8 Selection rates of the potential mediators in the methylation dataset.

TABLE 6  
Annotation of the Selected Mediators from the CMF Algorithm

Probe	Gene	Description
cg16657538	ZSCAN30	Involved in transcriptional regulation
cg25626453	PRRC2A	Associated with the age-at-onset of diabetes
cg02309301	ARGLU1	Associated with sexual development
cg13136721	RPTOR	Involved in regulation of cell growth and survival
cg12500973	HNRNPF	Involved in regulation of mRNA

descriptions were summarized from the GeneCards database (Safran et al., 2002). The result is shown in Table 6.

Inspecting and comparing these results more closely, two of the locations identified by CMF have been previously associated with development throughout the lifespan: PRRC2A and ARGLU1. These two locations are also in the top 10 of the lists generated by HIMA and filter. In addition, the RPTOR gene has been associated with cell growth and survival – development on a cellular level. Relative to other sites, this last location does not have a strong correlation with either childhood trauma ( $r = 0.186$ ) or stress reactivity ( $r = 0.233$ ), but due to its conditional indirect effect, it is deemed relevant by both CMF and HIMA. The ZSCAN30 gene has a small marginal correlation with stress reactivity ( $r = 0.096$ ) which lowers its rank for both the filter and HIMA methods. However, due to its strong correlation with childhood trauma ( $r = 0.347$ ) and its conditional relevance this site is still high on the list for CMF.

In conclusion, CMF has overlapped with other methods but can identify relevant locations that other methods may miss. Further research using replication samples could focus on exploring whether and how methylation at these locations may alter stress reactivity after childhood trauma.

## DISCUSSION

Structural equation modeling, the benchmark method for exploratory mediation analysis, is unavailable in the case of high-dimensional data. Several alternative methods exist, but in the current paper, we have shown through simulations that these underperform in situations with specific dependence among mediators, noise variables related to either  $X$  or  $Y$ , or a combination thereof. Taking these situations into account, we have introduced CMF, a hybrid algorithmic method to identify from a set of potential mediators the most likely true mediators.

CMF improves upon the existing methods by combining the estimation method from regularized regression with the theory-based decision functions from classic mediation analysis. It extends EMA with theoretically relevant decision functions to the high-dimensional case. As a full package including software implementation, it is flexible to the choice of decision function, robust in the tested situations, and it scales to multiple processor cores.

Besides its role as a novel method for EMA, CMF contributes several ideas to the statistical literature. It shows that the use of cyclically calculated residuals is applicable beyond regression into the territory of structural equation modeling. In addition, its performance is greatly improved by feature subsampling, which has regularizing effects on the estimated parameters and thus on the mediator selections. CMF is an example of how combining a deterministic algorithm with a stochastic outer component can lead to adequate performance.

One result of the approach taken in this paper is that there is no formal proof of convergence, and the algorithm may take a long time to stabilize. In addition, the complications introduced in the outer loop make a determination of the cutoff for selection nontrivial. In general, the algorithm will output a top- $N$  vector of most selected mediators, and potential options for

deciding which cutoff to take are a visual inspection of the scree plot or a form of parallel analysis (Horn, 1965). In addition, the error rates (type I and type II errors) are not analytically defined and have a complex relation with the alpha level of the base decision function. This could be investigated empirically in the future.

For this work, we only considered direct feature selection on the set of  $M$  variables. Another solution is projecting the available features onto a low-dimensional space before or during estimation. Feature selection can then be performed in this space, leading to variable importance upon reprojection to the original space. Examples are PCA, PLS, or the directions of mediation method by (Chén, Crainiceanu, Ogburn, Caffo, & Wager, 2017). However, we chose to exclude these methods because they do not select mediators, but rather linear combinations of all mediators.

Our coordinate-wise mediation filter bears resemblance to a class of metaheuristic algorithms in the SEM literature for specification search (Marcoulides & Falk, 2018). These algorithms perform an exploratory search for the optimal model based on overall model fit, e.g., the BIC objective. CMF could be considered specification search where the objective is not overall model fit but mediation analysis: it is targeted towards determining whether a specific variable is relevant to a process rather than searching for the optimal model. In addition, CMF performs regularization required for high-dimensional data. In the future, other specification search strategies could be implemented for EMA, but they each need to be adjusted to incorporate both a specific mediation objective and regularization.

Future research should focus on embedding mediation analysis theory directly in penalization procedures for these datasets, either in a classical estimation setting (Zhao & Luo, 2016) or using Bayesian estimation with shrinkage priors (Erp, Oberski, & Mulder, 2018). More generally, enriching structural equation models beyond EMA with embedded feature selection mechanisms will enable social and behavioral scientists to develop and test theories on the novel, high-dimensional datasets.

## ACKNOWLEDGMENTS

We thank Marco Boks, Yves Rosseel, Katrijn van Deun, Milica Miočević, and Ayoub Bagheri for their helpful suggestions at various stages of this work.

## FUNDING

This work was supported by the Netherlands Organization for Scientific Research (NWO) under Grant number 406.17.057.

## ORCID

Erik-Jan van Kesteren  <http://orcid.org/0000-0003-1548-1663>

Daniel L. Oberski  <http://orcid.org/0000-0001-7467-2297>

## REFERENCES

- Ammerman, B. A., Serang, S., Jacobucci, R., Burke, T. A., Alloy, L. B., & McCloskey, M. S. (2018). Exploratory analysis of mediators of the relationship between childhood maltreatment and suicidal behavior. *Journal of Adolescence*, 69, 103–112. doi:10.1016/j.adolescence.2018.09.004
- Atlas, L. Y., Lindquist, M. A., Bolger, N., & Wager, T. D. (2014). Brain mediators of the effects of noxious heat on pain. *Pain*, 155, 1632–1648. doi:10.1016/j.pain.2014.05.015
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182. doi:10.1037/0022-3514.51.6.1173
- Bates, D., & Maechler, M. (2017). Matrix: Sparse and dense matrix classes and methods. Retrieved from <https://cran.r-project.org/package=Matrix>
- Boca, S. M., Sinha, R., Cross, A. J., Moore, S. C., & Sampson, J. N. (2014). Testing multiple biological mediators simultaneously. *Bioinformatics*, 30, 214–220. doi:10.1093/bioinformatics/btt633
- Braun, M. (2018). sparseMVN: Multivariate normal functions for sparse covariance and precision matrices. R package version. Retrieved from <https://cran.r-project.org/package=sparseMVN>
- Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for non-convex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5, 232–253. doi:10.1214/10-AOAS388
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. doi:10.1023/A:1010933404324
- Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., & Wager, T. O. R. D. (2017). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, (September), 1–16. doi:10.1093/biostatistics/kxx040
- Erp, S. V., Oberski, D. L., & Mulder, J. (2018). Shrinkage Priors for Bayesian Penalized Regression. *OSF Preprint*, 1–39. 10.31219/osf.io/cg8fq
- Friedman, J., Hastie, T., & Tibshirani, R. (2009). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–24.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research (JMLR)*, 3, 1157–1182. doi:10.1016/j.aca.2011.07.027
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations* (pp. 362). Boca Raton, FL: CRC Press. doi:10.1201/b18401-1
- Hayes, A. F., & Preacher, K. J. (2010). Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. *Multivariate Behavioral Research*, 45, 627–660. doi:10.1080/00273171.2010.498290
- Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, 67, 451–470. doi:10.1111/bmsp.12028

- Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185. doi:10.1007/BF02289447
- Houtepen, L. C., Vinkers, C. H., Carrillo-Roa, T., Hiemstra, M., Lier, P. A., Van, Meeus, W., ... Boks, M. P. M. (2016). Genome-wide DNA methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nature Communications*, 7, 10967. doi:10.1038/ncomms10967
- Jacobucci, R., Brandmaier, A. M., & Kievit, R. A. (2018). Variable selection in structural equation models with regularized MIMIC models. *PsyArXiv Preprint*, 1–40. doi:10.17605/OSF.IO/BXZJF
- Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling*, 23, 555–566. doi:10.1080/10705511.2016.1154793.Regularized
- Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., ... Feinberg, A. P. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31, 142–147. doi:10.1038/nbt.2487
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Routledge.
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation Analysis. *Annual Review of Psychology*, 58, 593–614. doi:10.1146/annurev.psych.58.110405.085542
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104. doi:10.1037/1082-989X.7.1.83
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99–128. doi:10.1207/s15327906mbr3901
- Marcoulides, K. M., & Falk, C. F. (2018). Model specification searches in structural equation modeling with r. *Structural Equation Modeling*, 25, 484–491. doi:10.1080/10705511.2017.1409074
- Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization: A Publication of the Society for Industrial and Applied Mathematics*, 22, 341–362. doi:10.1137/100802001
- Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, 118, 155–164. doi:10.1037/0033-2909.118.1.155
- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66, 825–852. doi:10.1146/annurev-psych-010814-015258
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40, 879–891. doi:10.3758/BRM.40.3.879
- R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org/>
- Richtárik, P., & Takáč, M. (2014). Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144, 1–38. doi:10.1007/s10107-012-0614-z
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36. doi:10.18637/jss.v048.i02
- Safran, M., Solomon, I., Shmueli, O., Lapidot, M., Shen-Orr, S., Adato, A., ... Lancet, D. (2002). GeneCards™ 2002: Towards a complete, object-oriented, human gene compendium. *Bioinformatics*, 18, 1542–1543. doi:10.1093/bioinformatics/18.11.1542
- Serang, S., Jacobucci, R., Brimhall, K. C., & Grimm, K. J. (2017). Exploratory Mediation analysis via regularization. *Structural Equation Modeling*, 24, 733–744. doi:10.1080/10705511.2017.1311775
- Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. *Sociological Methodology*, 16, 159–186. doi:10.2307/270922
- Triche, T. (2014). *FDb.InfiniumMethylation.hg18: Annotation package for Illumina Infinium DNA methylation probes*.
- VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction* (pp. 706). New York, NY: Oxford University Press.
- Vanderweele, T. J., & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2, 95–115. doi:10.1515/em-2012-0010.Mediation
- Wager, S., Wang, S., & Liang, P. S. (2013). Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, 17, 351–359.
- Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., ... Liu, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32, 3150–3154. doi:10.1093/bioinformatics/btw351
- Zhang, J., Zhao, Z., Zhang, K., & Wei, Z. (2017). A feature sampling strategy for analysis of high dimensional genomic data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.
- Zhao, Y., & Luo, X. (2016). *Pathway lasso: Estimate and select sparse mediation pathways with high dimensional mediators*. Retrieved from <http://arxiv.org/abs/1603.07749>

## APPENDIX A. R ENVIRONMENT USED

---

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows ≥ 8 × 64 (build 9200)
##
## Matrix products: default
##
## locale:
## [ 1] LC_COLLATE=Dutch_Netherlands.1252 LC_CTYPE=Dutch_Netherlands.1252
## [ 3] LC_MONETARY=Dutch_Netherlands.1252 LC_NUMERIC=C
## [ 5] LC_TIME=Dutch_Netherlands.1252
##
## attached base packages:
## [ 1] stats4 parallel stats graphics grDevices utils datasets
## [ 8] methods base
##
## other attached packages:
## [ 1] Massign_1.1.0
## [ 2] firatheme_0.1.0
## [ 3] FDb.InfiniumMethylation.hg18_2.2.0
## [ 4] org.Hs.eg.db_3.6.0
## [ 5] TxDb.Hsapiens.UCSC.hg18.knownGene_3.2.2
## [ 6] pbapply_1.3-4
## [ 7] MASS_7.3-50
## [ 8] HIMA_1.0.7
## [ 9] ncvreg_3.10-0
## [10] regsem_1.1.2
## [11] Rcpp_0.12.18
## [12] lavaan_0.6-2
## [13] cmfilter_0.2.1
## [14] magrittr_1.5
## [15] forcats_0.3.0
## [16] stringr_1.3.1
## [17] dplyr_0.7.6
## [18] purrr_0.2.5
## [19] readr_1.1.1
## [20] tidyr_0.8.1
## [21] tibble_1.4.2
## [22] ggplot2_3.0.0
## [23] tidyverse_1.2.1
## [24] GenomicFeatures_1.32.1
## [25] AnnotationDbi_1.42.1
## [26] Biobase_2.40.0
## [27] GenomicRanges_1.32.6
## [28] GenomeInfoDb_1.16.0
## [29] IRanges_2.14.10
## [30] S4Vectors_0.18.3
## [31] BiocGenerics_0.26.0
```

---

(Continued)

(Continued)

---

```
## [ 32] glmnet_2.0-16
## [ 33] foreach_1.4.4
## [ 34] Matrix_1.2-14
##
## loaded via a namespace (and not attached):
## [ 1] nlme_3.1-137
## [ 3] matrixStats_0.54.0
## [ 5] bit64_0.9-7
## [ 7] progress_1.2.0
## [ 9] rprojroot_1.3-2
## [11] backports_1.1.2
## [13] DBI_1.0.0
## [15] colorspace_1.3-2
## [17] mnormt_1.5-5
## [19] prettyunits_1.0.2
## [21] bit_1.1-14
## [23] cli_1.0.0
## [25] xml2_1.2.0
## [27] rtracklayer_1.40.4
## [29] pbivnorm_0.6.0
## [31] Rsamtools_1.32.2
## [33] XVector_0.20.0
## [35] htmltools_0.3.6
## [37] rlang_0.2.1
## [39] rstudioapi_0.7
## [41] bindr_0.1.1
## [43] BiocParallel_1.14.2
## [45] GenomeInfoDbData_1.1.0
## [47] stringi_1.1.7
## [49] SummarizedExperiment_1.10.1
## [51] plyr_1.8.4
## [53] blob_1.1.1
## [55] lattice_0.20-35
## [57] haven_1.1.2
## [59] knitr_1.20
## [61] codetools_0.2-15
## [63] XML_3.98-1.15
## [65] evaluate_0.11
## [67] Rttf2pt1_1.3.7
## [69] gtable_0.2.0
## [71] assertthat_0.2.0
## [73] iterators_1.0.10
## [75] memoise_1.1.0

bitops_1.0-6
lubridate_1.7.4
doParallel_1.0.11
httr_1.3.1
tools_3.5.0
R6_2.2.2
lazyeval_0.2.1
withr_2.1.2
tidyselect_0.2.4
extrafontdb_1.0
compiler_3.5.0
rvest_0.3.2
DelayedArray_0.6.4
scales_1.0.0
digest_0.6.15
rmarkdown_1.10
pkgconfig_2.0.1
extrafont_0.17
readxl_1.1.0
RSQLite_2.1.1
jsonlite_1.5
RCurl_1.95-4.11
munSELL_0.5.0
yaml_2.2.0
zlibbioc_1.26.0
grid_3.5.0
crayon_1.3.4
Biostrings_2.48.0
hms_0.4.2
pillar_1.3.0
biomaRt_2.36.1
glue_1.3.0
modelr_0.1.2
cellranger_1.1.0
papaJa_0.1.0.9709
broom_0.5.0
GenomicAlignments_1.16.0
bindrcpp_0.2.2
```

---

## APPENDIX B. COVARIANCE MATRICES FOR THE ILLUSTRATIVE SIMULATIONS

TABLE B1  
The Marginal Covariance Matrix for the Data of the First Illustrative Simulation (Suppression)

$X$	$M_1$	$M_2$	$Y$
1.00	-0.40	0.40	-0.06
-0.40	1.00	-0.60	0.26
0.40	-0.60	1.00	0.00
-0.06	0.26	0.00	1.00

TABLE B2  
The Marginal Covariance Matrix for the Data of the Third Illustrative Simulation

$X$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$	$M_6$	$M_7$	$M_8$	$M_9$	$M_{10}$	$M_{11}$	$M_{12}$	$M_{13}$	$M_{14}$	$M_{15}$	$M_{16}$	$Y$
1.00	0.30	-0.80	-0.80	0.80	-0.40	-0.40	0.40	0.40	0.40	-0.40	0.40	0.40	0.40	-0.40	-0.40	0.40	0.09
0.30	1.00	-0.24	-0.26	0.33	-0.17	-0.14	0.15	0.18	0.18	-0.24	0.12	0.14	0.19	-0.23	-0.15	0.17	0.30
-0.80	-0.24	1.00	0.65	-0.64	0.31	0.29	-0.27	-0.34	-0.32	0.32	-0.33	-0.35	-0.39	0.35	0.31	-0.35	-0.07
-0.80	-0.26	0.65	1.00	-0.67	0.37	0.31	-0.31	-0.28	-0.32	0.32	-0.29	-0.35	-0.30	0.31	0.33	-0.32	-0.08
0.80	0.33	-0.64	-0.67	1.00	-0.33	-0.33	0.34	0.31	0.33	-0.33	0.26	0.34	0.36	-0.35	-0.29	0.31	0.10
-0.40	-0.17	0.31	0.37	-0.33	1.00	0.25	-0.16	-0.20	-0.21	0.22	-0.15	-0.22	0.00	0.13	0.24	-0.17	-0.05
-0.40	-0.14	0.29	0.31	-0.33	0.25	1.00	-0.15	-0.30	-0.19	0.05	-0.20	-0.26	-0.22	0.22	0.02	-0.17	-0.04
0.40	0.15	-0.27	-0.31	0.34	-0.16	-0.15	1.00	0.15	0.23	-0.20	0.03	0.07	0.12	-0.15	-0.15	0.17	0.04
0.40	0.18	-0.34	-0.28	0.31	-0.20	-0.30	0.15	1.00	0.18	-0.01	0.20	0.32	0.18	-0.16	-0.15	0.32	0.05
0.40	0.18	-0.32	-0.32	0.33	-0.21	-0.19	0.23	0.18	1.00	-0.17	-0.03	0.25	0.14	-0.23	-0.13	0.06	0.05
-0.40	-0.24	0.32	0.32	-0.33	0.22	0.05	-0.20	-0.01	-0.17	1.00	-0.12	-0.13	-0.15	0.09	0.17	-0.12	-0.07
0.40	0.12	-0.33	-0.29	0.26	-0.15	-0.20	0.03	0.20	-0.03	-0.12	1.00	0.16	0.17	-0.26	-0.17	0.24	0.04
0.40	0.14	-0.35	-0.35	0.34	-0.22	-0.26	0.07	0.32	0.25	-0.13	0.16	1.00	0.18	0.02	-0.22	0.18	0.04
0.40	0.19	-0.39	-0.30	0.36	0.00	-0.22	0.12	0.18	0.14	-0.15	0.17	0.18	1.00	-0.29	-0.04	0.21	0.06
-0.40	-0.23	0.35	0.31	-0.35	0.13	0.22	-0.15	-0.16	-0.23	0.09	-0.26	0.02	-0.29	1.00	0.12	-0.15	-0.07
-0.40	-0.15	0.31	0.33	-0.29	0.24	0.02	-0.15	-0.15	-0.13	0.17	-0.17	-0.22	-0.04	0.12	1.00	-0.20	-0.05
0.40	0.17	-0.35	-0.32	0.31	-0.17	-0.17	0.17	0.32	0.06	-0.12	0.24	0.18	0.21	-0.15	-0.20	1.00	0.05
0.09	0.30	-0.07	-0.08	0.10	-0.05	-0.04	0.04	0.05	0.05	-0.07	0.04	0.04	0.06	-0.07	-0.05	0.05	0.18

TABLE B3  
The Marginal Covariance Matrix for the Data of the Third Illustrative Simulation

$X$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$	$M_6$	$M_7$	$M_8$	$M_9$	$M_{10}$	$M_{11}$	$M_{12}$	$M_{13}$
1.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.30	1.00	-0.01	-0.05	0.26	-0.06	-0.02	0.03	0.07	0.07	-0.15	0.00	0.03	0.08
0.00	-0.01	1.00	0.08	-0.03	-0.02	-0.09	0.15	-0.06	0.00	-0.01	-0.05	-0.09	-0.24
0.00	-0.05	0.08	1.00	-0.27	0.17	-0.05	0.04	0.12	0.00	-0.01	0.09	-0.08	0.08
0.00	0.26	-0.03	-0.27	1.00	-0.04	-0.03	0.06	-0.02	0.03	-0.04	-0.20	0.07	0.13
0.00	-0.06	-0.02	0.17	-0.04	1.00	0.13	0.00	-0.05	-0.07	0.08	0.01	-0.08	0.22
0.00	-0.02	-0.09	-0.05	-0.03	0.13	1.00	0.02	-0.20	-0.05	-0.16	-0.06	-0.14	-0.08
0.00	0.03	0.15	0.04	0.06	0.00	0.02	1.00	-0.01	0.09	-0.06	-0.19	-0.13	-0.06
0.00	0.07	-0.06	0.12	-0.02	-0.05	-0.20	-0.01	1.00	0.03	0.21	0.06	0.23	0.03
0.00	0.07	0.00	0.00	0.03	-0.07	-0.05	0.09	0.03	1.00	-0.02	-0.26	0.13	-0.03
0.00	-0.15	-0.01	-0.01	-0.04	0.08	-0.16	-0.06	0.21	-0.02	1.00	0.06	0.05	0.02
0.00	0.00	-0.05	0.09	-0.20	0.01	-0.06	-0.19	0.06	-0.26	0.06	1.00	0.00	0.01
0.00	0.03	-0.09	-0.08	0.07	-0.08	-0.14	-0.13	0.23	0.13	0.05	0.00	1.00	0.03
0.00	0.08	-0.24	0.08	0.13	0.22	-0.08	-0.06	0.03	-0.03	0.02	0.01	0.03	1.00
0.00	-0.13	0.10	-0.03	-0.08	-0.04	0.09	0.01	0.01	-0.09	-0.10	-0.14	0.26	-0.19
0.00	-0.04	-0.02	0.03	0.09	0.12	-0.20	0.01	0.02	0.04	0.02	-0.01	-0.09	0.18
0.00	0.06	-0.11	0.01	-0.03	-0.01	-0.01	0.01	0.22	-0.14	0.05	0.12	0.03	0.07
0.09	0.85	-1.03	-1.04	1.17	-0.68	-0.47	0.20	0.59	0.44	-0.27	0.16	0.72	0.62