Statistical approaches for highly correlated exposures

Andrea Bellavia

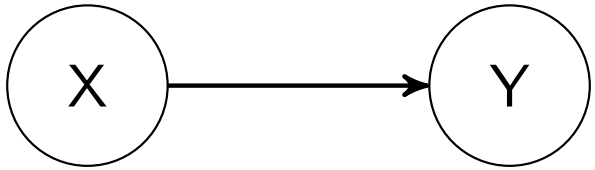
Harvard T.H. Chan School of Public Health
abellavi@hsph.harvard.edu

May 22, 2017

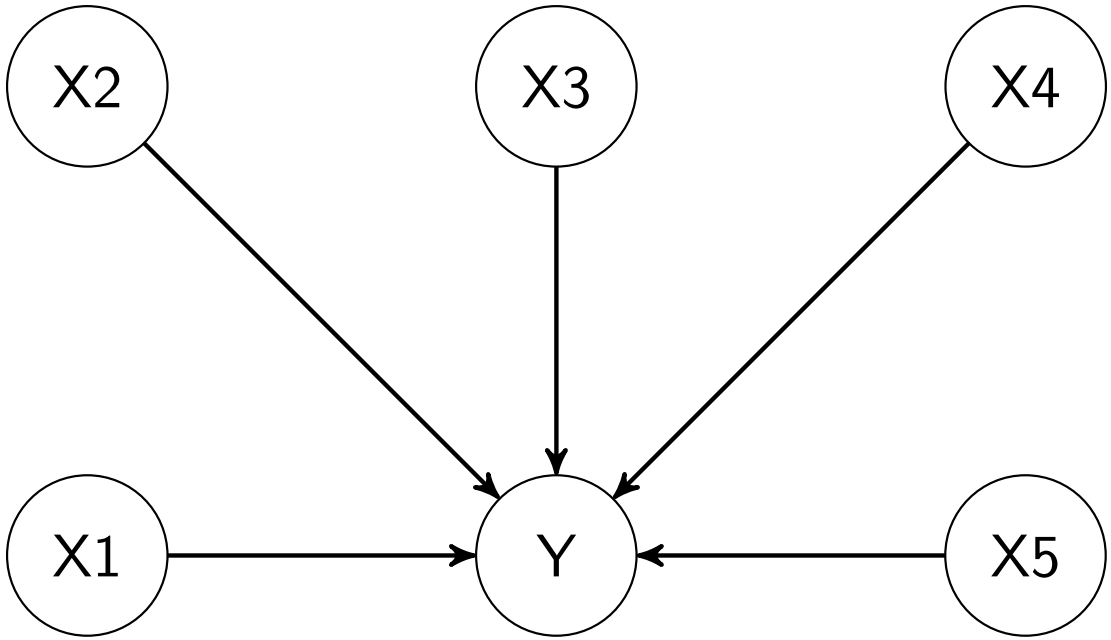
Motivating example - mixtures of phthalates and birthweight

- Phthalates are ubiquitous (e.g. in food, plasticizers, cosmetics, personal care products, toys) chemical disruptors that may be associated with the risk of several pregnancy complications
- 7-8 metabolites are generally present. They are highly correlated
- Each metabolite has different sources, with implications for interventions and recommendations
- Studies investigating phthalates and birth weight have provided inconsistent results

The main focus of epidemiologic and public health research is the identification of risk factors



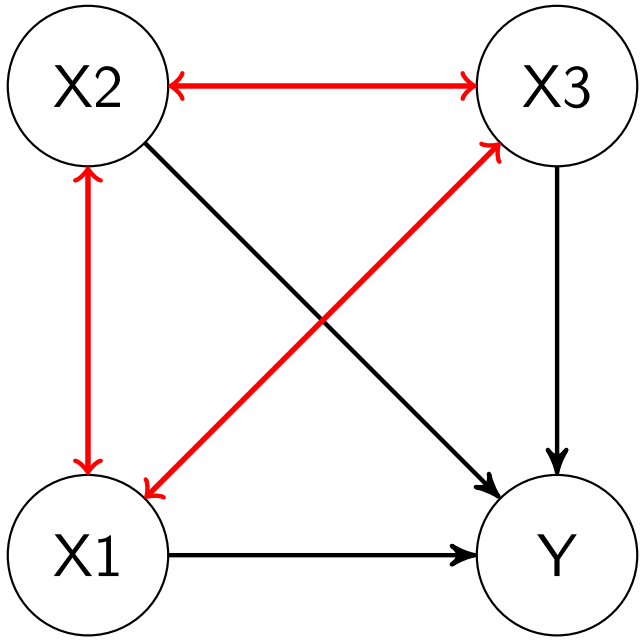
The common situation is that, for a given outcome, multiple risk factors can be identified (or hypothesized)



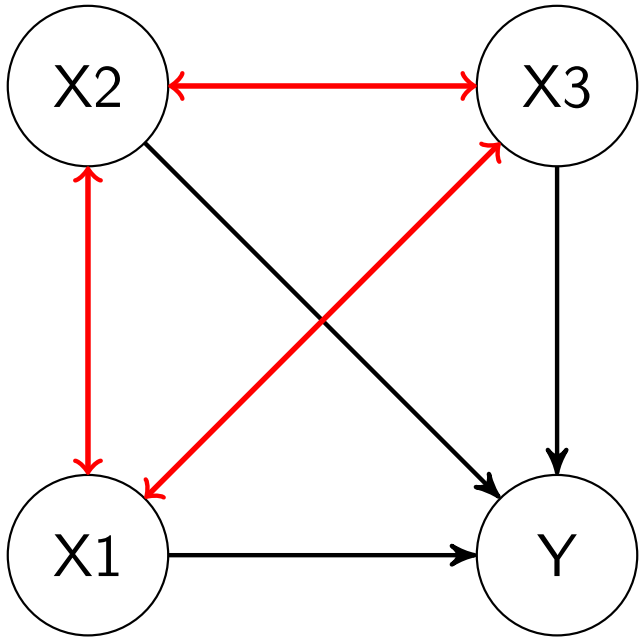
We generally address this kind of research question by using regression approaches

$$f(Y) = \beta_0 + \beta_1 \cdot X_1$$

Covariates, however, are generally related to each other



Covariates, however, are generally related to each other



In this situation, each covariate acts as a confounder of the relationship between every other covariate and the outcome

Standard approach: mutually adjust for all predictors in the same statistical model

$$f(Y) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$$

Standard approach: mutually adjust for all predictors in the same statistical model

$$f(Y) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$$

What if the covariates are correlated?

- Including multiple correlated covariates in the same statistical model can lead to the statistical issue of multicollinearity (or simply collinearity).
- If the correlation between two covariates (say X_1 and X_2) is very high, then one is a pretty accurate linear predictor of the other

- Including multiple correlated covariates in the same statistical model can lead to the statistical issue of multicollinearity (or simply collinearity).
- If the correlation between two covariates (say X_1 and X_2) is very high, then one is a pretty accurate linear predictor of the other
- The associated coefficients may be (even severely) biased
- Of important note, collinearity does not influence the overall performance of the model, but has an important impact on individual predictors. Unfortunately, this is exactly what we are interested in

- Including multiple correlated covariates in the same statistical model can lead to the statistical issue of multicollinearity (or simply collinearity).
- If the correlation between two covariates (say X_1 and X_2) is very high, then one is a pretty accurate linear predictor of the other
- The associated coefficients may be (even severely) biased
- Of important note, collinearity does not influence the overall performance of the model, but has an important impact on individual predictors. Unfortunately, this is exactly what we are interested in
- The focus of today's workshop is on how to deal with those situations

Collinearity scenarios are very common, much more than expected:

- Environmental exposures
- Nutrients
- Dietary factors
- Psychology
- Anthropometric factors (e.g. height and BMI)
- . . . others

Collinearity scenarios are very common, much more than expected:

- Environmental exposures
- Nutrients
- Dietary factors
- Psychology
- Anthropometric factors (e.g. height and BMI)
- . . . others

Multiple correlated exposures are often referred to as mixture of exposures, and related methods as mixture models. We will use the terminology interchangeably

Classification of statistical approaches for mixtures

At least 25-30 approaches have been proposed in the last decades (listed in Taylor et al., 2016). They can be grouped in 4 major categories:

- Classification and prediction
- Variable selection
- Variable shrinkage
- Exposure-response surface estimation

Illustrative examples

Two simulated datasets (dataset1.xls, dataset2.xls) are provided in the workshop material. All results presented are based on Dataset2 (R code also provided). You can use the other for individual practice

Standard approach

Simulated data: 1 continuous outcome Y; 14 predictors $X_1 - X_{14}$; 3 Confounders Z_1, Z_2, Z_3.

Standard approach

Simulated data: 1 continuous outcome Y; 14 predictors $X_1 - X_{14}$; 3 Confounders Z_1, Z_2, Z_3.

Single regressions (for each X)

$$Y = \beta_0 + \beta_1 \cdot X + \sum_{i=1}^{3} \beta_{i+1} \cdot Z_i$$

Standard approach

Simulated data: 1 continuous outcome Y; 14 predictors $X_1 - X_{14}$; 3 Confounders Z_1, Z_2, Z_3.

Single regressions (for each X)

$$Y = \beta_0 + \beta_1 \cdot X + \sum_{i=1}^{3} \beta_{i+1} \cdot Z_i$$

Mutually adjusted regression

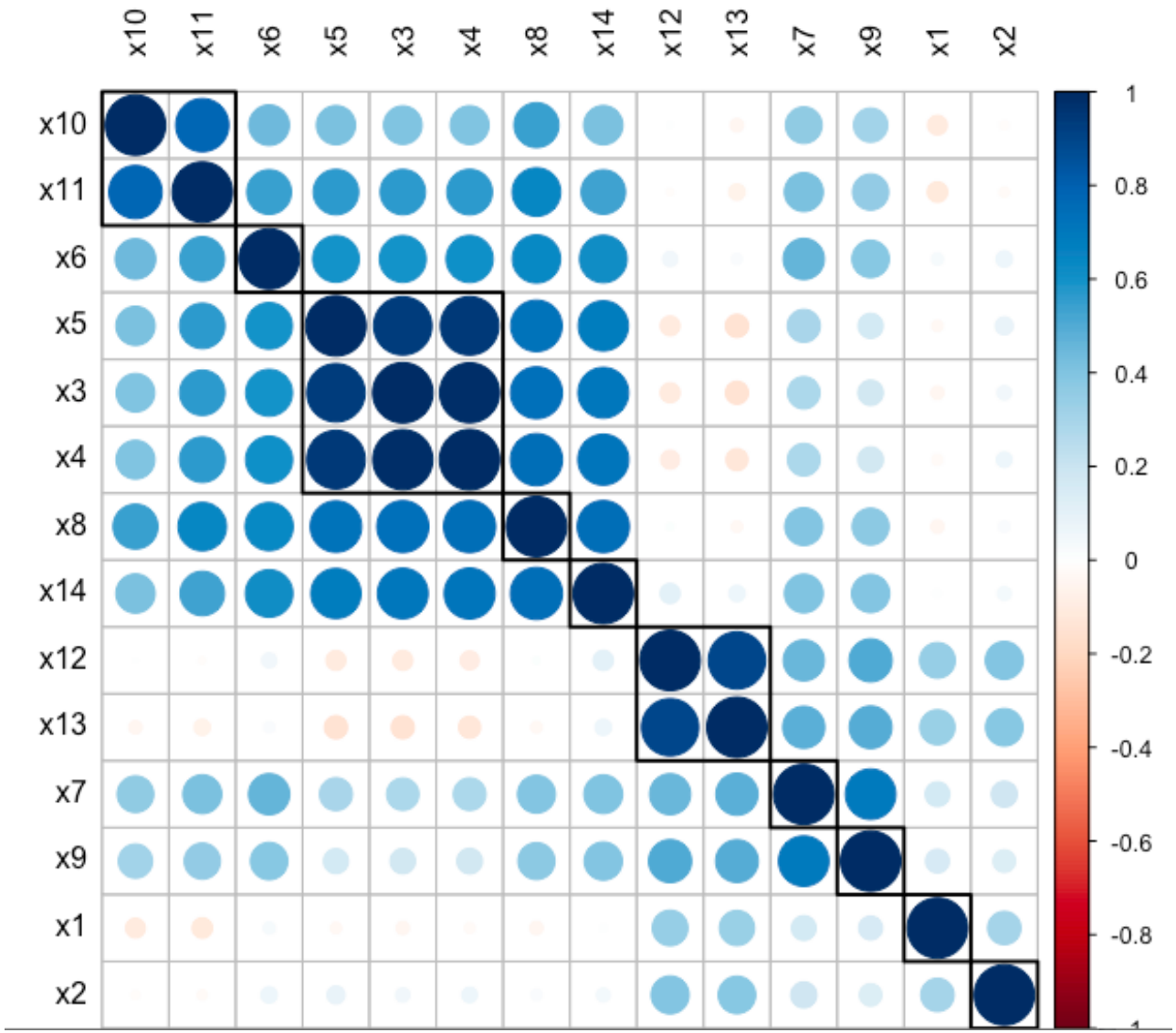
$$Y = \beta_0 + \sum_{i=1}^{14} \beta_i \cdot X_i + \sum_{i=1}^{3} \beta_{i+14} \cdot Z_i$$

Multiple testing corrections (e.g. Bonferroni, FDR) should be considered when the latter is used

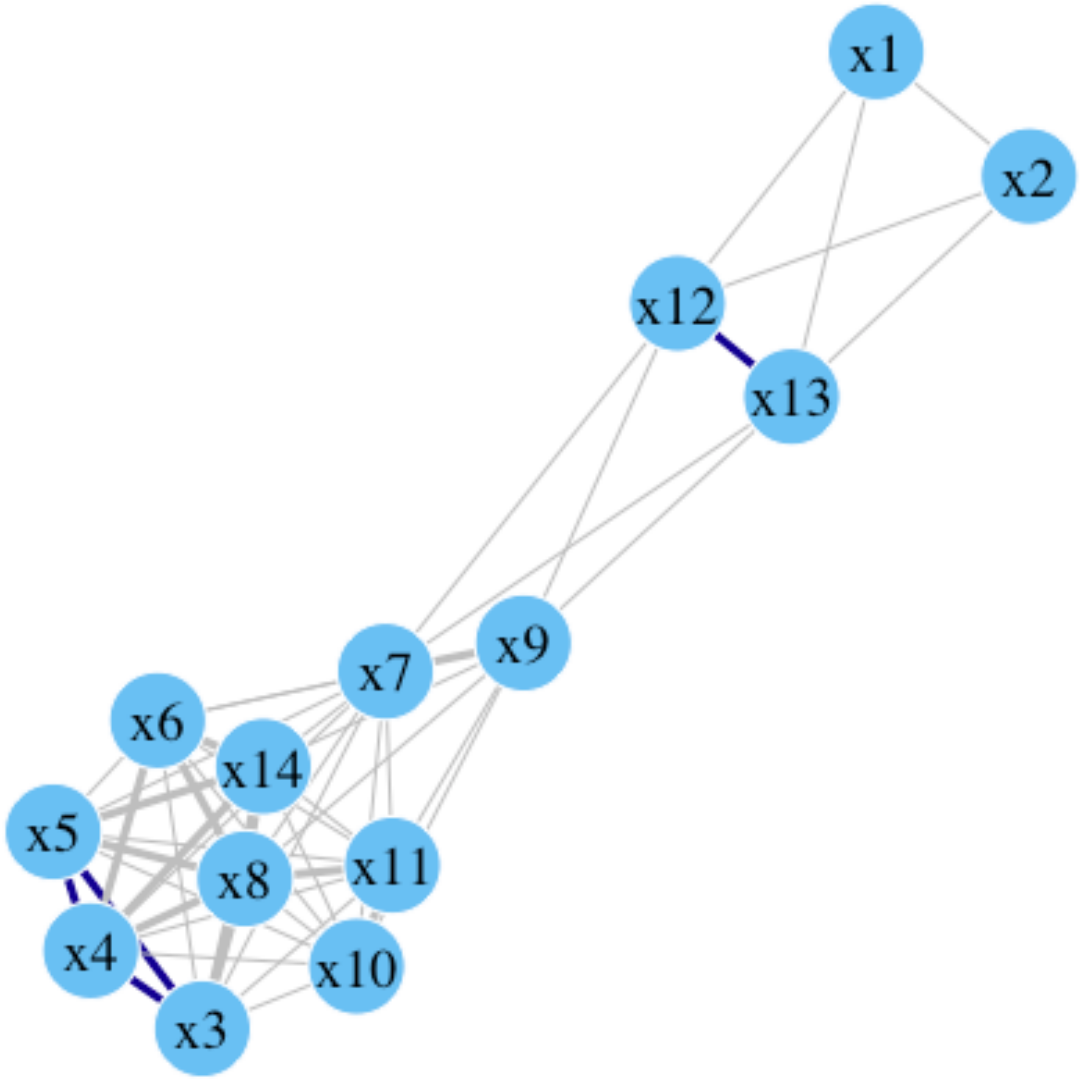
Multiple regression - Results

	β (one at the time)	p-value	β (mutually adjusted)	p-value
X_1	0.11	< 0.001	0.06	0.08
X_2	0.07	0.01	0.02	0.55
X_3	0.08	0.007	-0.03	0.77
X_4	0.09	0.003	0.05	0.64
X_5	0.07	0.005	0	0.92
X_6	0.12	<0.001	0.06	0.47
X_7	0.15	<0.001	-0.03	0.62
X_8	0.14	<0.001	0.02	0.68
X_9	0.16	<0.001	0.02	0.67
X_{10}	0.12	<0.001	0.05	0.26
X_{11}	0.15	<0.001	0.05	0.34
X_{12}	0.29	<0.001	0.22	0.14
X_{13}	0.24	<0.001	-0.08	0.59
X_{14}	0.18	<0.001	0.05	0.29

What about the correlation?



There are multiple ways for investigating the correlation structure. Networks are another useful tool



Principal Component Analysis

- Classification approaches reduce collinearity by identifying *orthogonal factors* that summarize a set of covariates of interest

Principal Component Analysis

- Classification approaches reduce collinearity by identifying *orthogonal factors* that summarize a set of covariates of interest
- Simplifying the definition, from a set of correlated covariates to a smaller set of uncorrelated principal components

Principal Component Analysis

- Classification approaches reduce collinearity by identifying *orthogonal factors* that summarize a set of covariates of interest
- Simplifying the definition, from a set of correlated covariates to a smaller set of uncorrelated principal components
- In practice, we look for a first component that maximises the variance between covariates. We then look for a second component to maximize the residual variance, under the constraint of this component being orthogonal (i.e. uncorrelated) to the first. And so on until all variance is explained (or as decided by the user)

Principal Component Analysis

- Classification approaches reduce collinearity by identifying *orthogonal factors* that summarize a set of covariates of interest
- Simplifying the definition, from a set of correlated covariates to a smaller set of uncorrelated principal components
- In practice, we look for a first component that maximises the variance between covariates. We then look for a second component to maximize the residual variance, under the constraint of this component being orthogonal (i.e. uncorrelated) to the first. And so on until all variance is explained (or as decided by the user)
- A common practical advice is to select a number of components that jointly explain around 80% of the original variance

Principal Component Analysis

- Classification approaches reduce collinearity by identifying *orthogonal factors* that summarize a set of covariates of interest
- Simplifying the definition, from a set of correlated covariates to a smaller set of uncorrelated principal components
- In practice, we look for a first component that maximises the variance between covariates. We then look for a second component to maximize the residual variance, under the constraint of this component being orthogonal (i.e. uncorrelated) to the first. And so on until all variance is explained (or as decided by the user)
- A common practical advice is to select a number of components that jointly explain around 80% of the original variance
- Scores called loading factors are used to understand the relationship between the components and the original covariates

Principal Component Analysis

- Classification approaches reduce collinearity by identifying *orthogonal factors* that summarize a set of covariates of interest
- Simplifying the definition, from a set of correlated covariates to a smaller set of uncorrelated principal components
- In practice, we look for a first component that maximises the variance between covariates. We then look for a second component to maximize the residual variance, under the constraint of this component being orthogonal (i.e. uncorrelated) to the first. And so on until all variance is explained (or as decided by the user)
- A common practical advice is to select a number of components that jointly explain around 80% of the original variance
- Scores called loading factors are used to understand the relationship between the components and the original covariates
- Accessible and brief introduction: `http://www.lauradhamilton.com/introduction-to-principal-component-analysis-pca`

PCA - example

- Can be run with *principal* command within the *psych* library in R
- We run multiple PCA models by changing the number of components, and then focus on the explained variance to select the best one
- It is always recommended to rescale the covariates (mean 0, unit sd) before running a PCA model

PCA - 3 components

Explained variance=74% (39 %, 19%, 16%, respectively)

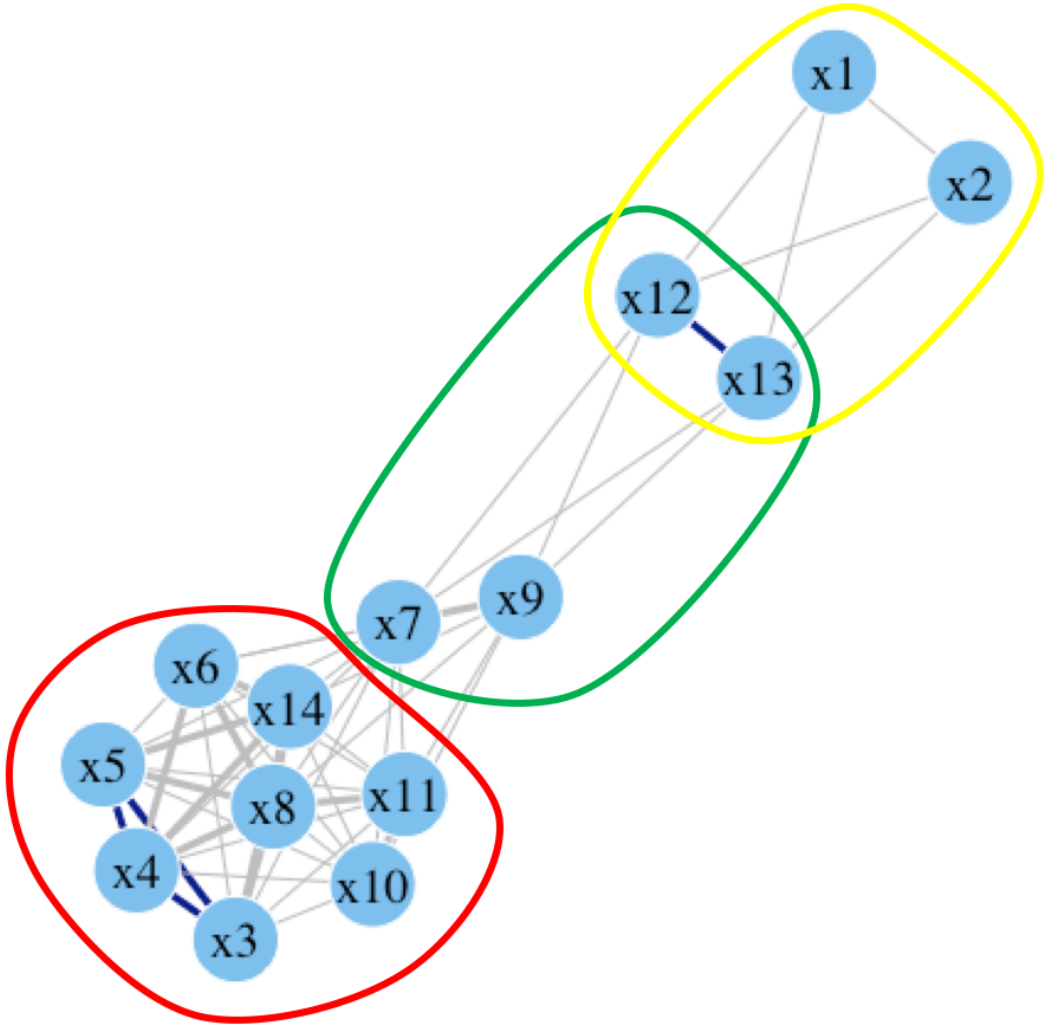
PCA - 3 components

Explained variance=74% (39 %, 19%, 16%, respectively)

	Loading factors		
	1st	2nd	3rd
X_1	0.00	0.05	**0.66**
X_2	0.12	-0.01	**0.71**
X_3	**0.96**	-0.01	0.01
X_4	**0.97**	0.00	0.02
...	...	...	...
X_{12}	-0.16	**0.56**	**0.70**
...	...	...	...

PCA - 3 components

1st - red; 2nd - green; 3rd - yellow



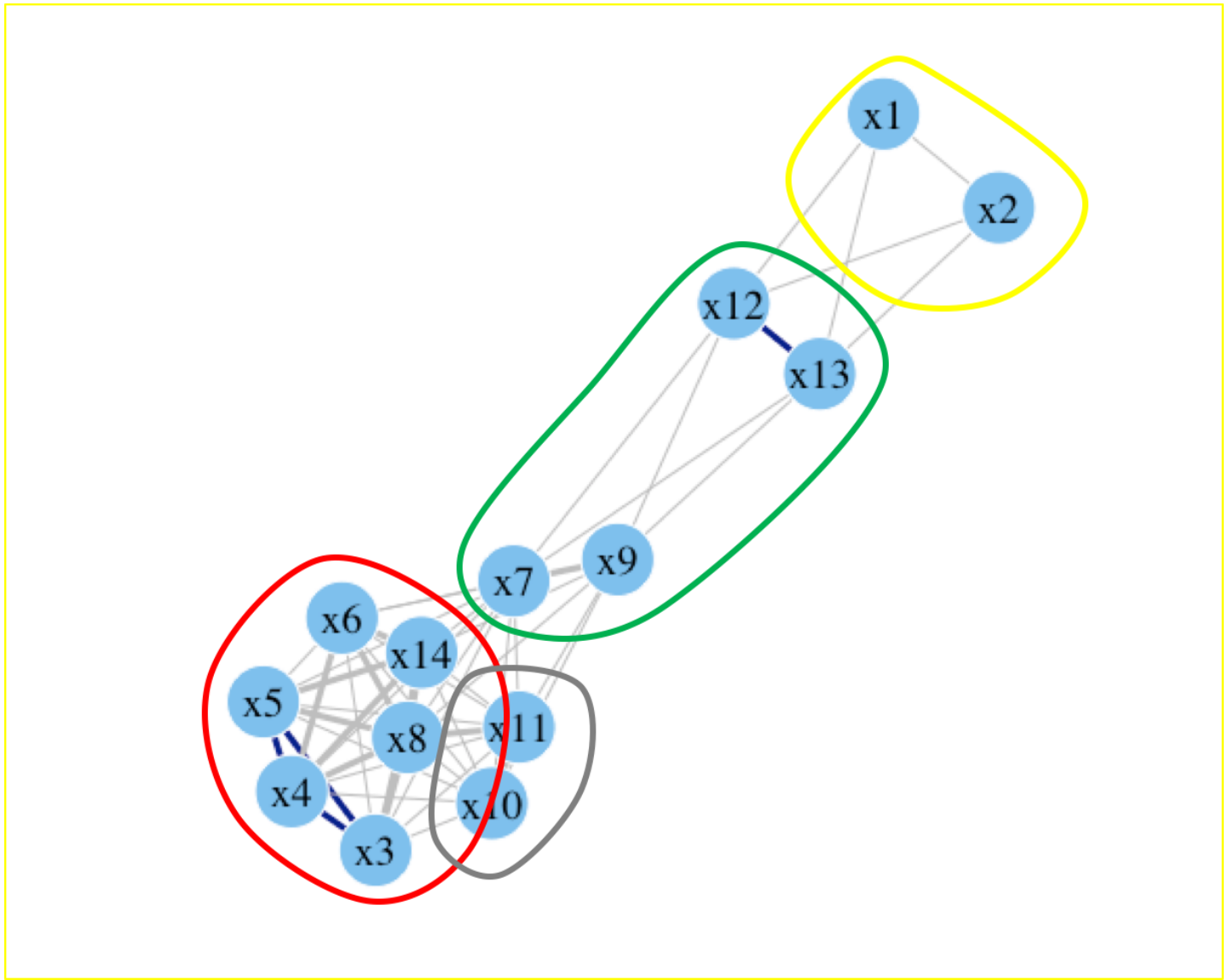
PCA - 4 components

Explained variance=80% (35 %, 20%, 14%, 11%, respectively)

	Loading factors			
	1st	2nd	3rd	4th
X_1	-0.05	0.19	-0.01	**0.74**
X_2	0.07	0.14	-0.03	**0.81**
X_3	**0.96**	-0.05	0.14	0.02
X_4	**0.96**	-0.04	0.15	0.03
...	...	...	...	
X_{11}	**0.48**	0.08	**0.79**	-0.05
X_{12}	-0.09	**0.82**	-0.12	**0.39**
...	...	...	...	

PCA - 4 components

1st - red; 2nd - green; 3rd - yellow, 4th - grey



PCA - Regression

3 components

$$Y = \beta_0 + \beta_1 \cdot PC_1 + \beta_2 \cdot PC_2 + \beta_3 \cdot PC_3 + \sum_{i=1}^{3} \beta_{i+3} \cdot Z_i$$

	β	p-value
PC_1	0.18	< 0.001
PC_2	0.11	< 0.001
PC_3	0.04	0.06

4 components

$$Y = \beta_0 + \beta_1 \cdot PC_1 + \beta_2 \cdot PC_2 + \beta_3 \cdot PC_3 + \beta_4 \cdot PC_4 + \sum_{i=1}^{3} \beta_{i+4} \cdot Z_i$$

	β	p-value
PC_1	0.16	< 0.001
PC_2	0.08	< 0.001
PC_3	0.05	0.01
PC_4	0.10	< 0.001

Structural Equation Model

- PCA could be integrated within a SEM framework.
- The main contribution of SEM in a PCA environment is a better account for measurement error

Structural Equation Model

- PCA could be integrated within a SEM framework.
- The main contribution of SEM in a PCA environment is a better account for measurement error
- In this case (you can check the code and results), we get exactly the same coefficients estimated with PCA

Summary of classification approaches

- + They **reduce data dimension**
- + A certain amount of flexibility is allowed

Summary of classification approaches

- + They **reduce data dimension**
- + A certain amount of flexibility is allowed
- - Subjective choice of components
- - Difficult interpretation if groups with a specific biological meaning are not identified
- - Difficult to retrieve the individual contribution of the predictors
- - Difficult, sometimes impossible, to account for non-linear effects

Bayesian Kernel Machine Regression

- BKMR (Bobb et al., 2015) was (mainly) motivated by the last two final points (i.e. predictor-specific contribution, non-linear effects)

Bayesian Kernel Machine Regression

- BKMR (Bobb et al., 2015) was (mainly) motivated by the last two final points (i.e. predictor-specific contribution, non-linear effects)
- Kernel machine regression are a popular tool in machine learning
- The main idea is that we model the relationship between a high-dimensional set of M predictors and the outcome by using a flexible function h of the covariates

$$g(E(Y_i)) = h(x_{i1}, ..., x_{iM}) + \beta \cdot Z_i$$

- h is called *exposure-response* function, and is estimated with a Gaussian Kernel, which flexibly captures a wide range of underlying functional forms for h
- BKMR implements Markov Chain Monte Carlo (MCMC, an iterative estimation algorithm) for the estimation of the function (n.b. it takes long)

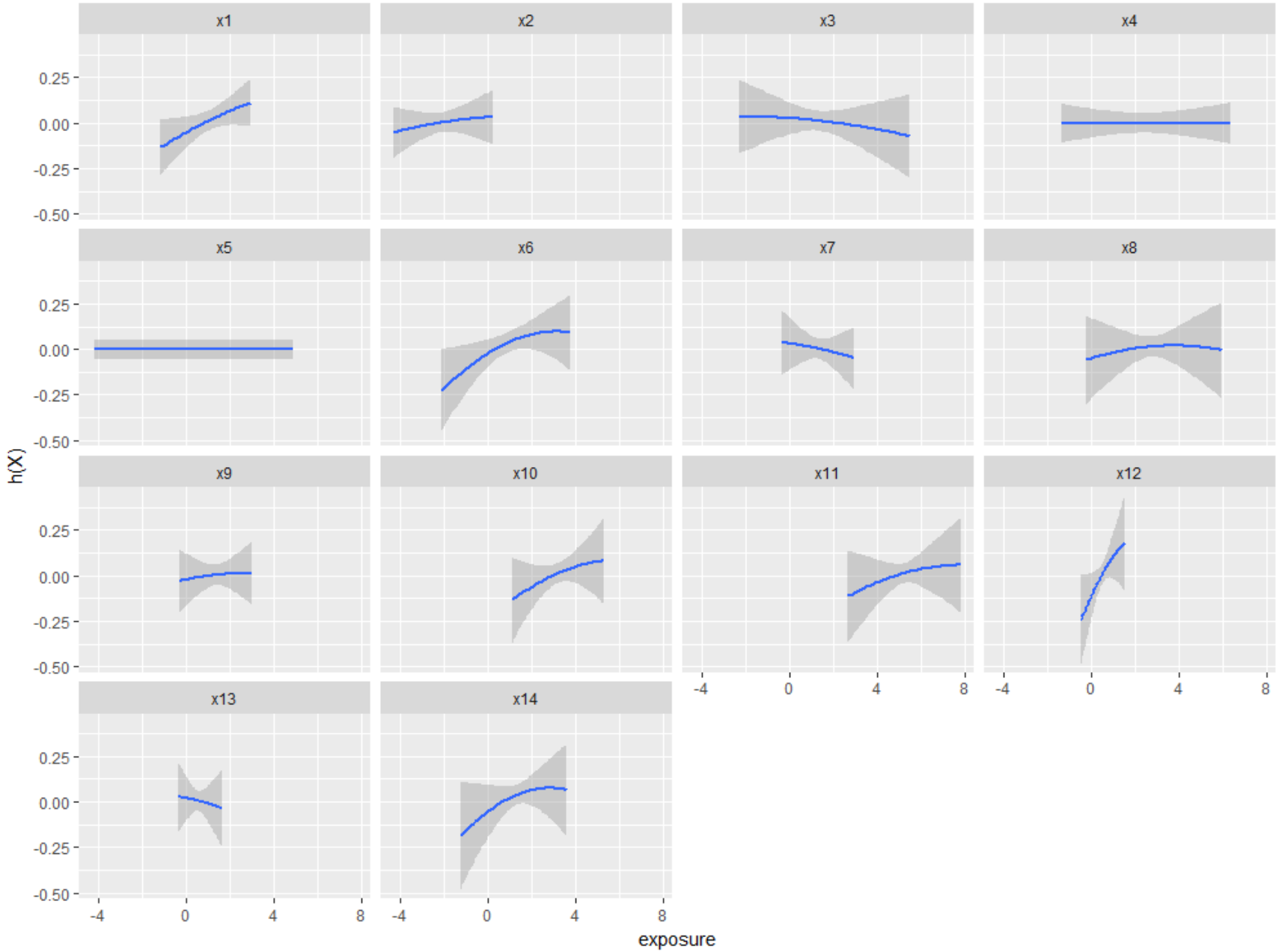
- h is called *exposure-response* function, and is estimated with a Gaussian Kernel, which flexibly captures a wide range of underlying functional forms for h
- BKMR implements Markov Chain Monte Carlo (MCMC, an iterative estimation algorithm) for the estimation of the function (n.b. it takes long)
- One can also provide preliminary information to the model (such as the grouping of the predictors that we have already seen).
- In the following slide we will however proceed in a non-informative environment.

BKMR - predictor-response function

- Estimation is straightforward. The package *bkmr* does everything, just write one line and wait few minutes
- Different tools are then available to summarize results
- Introduction to the package:
  `https://jenfb.github.io/bkmr/overview.html`

BKMR - predictor-response function

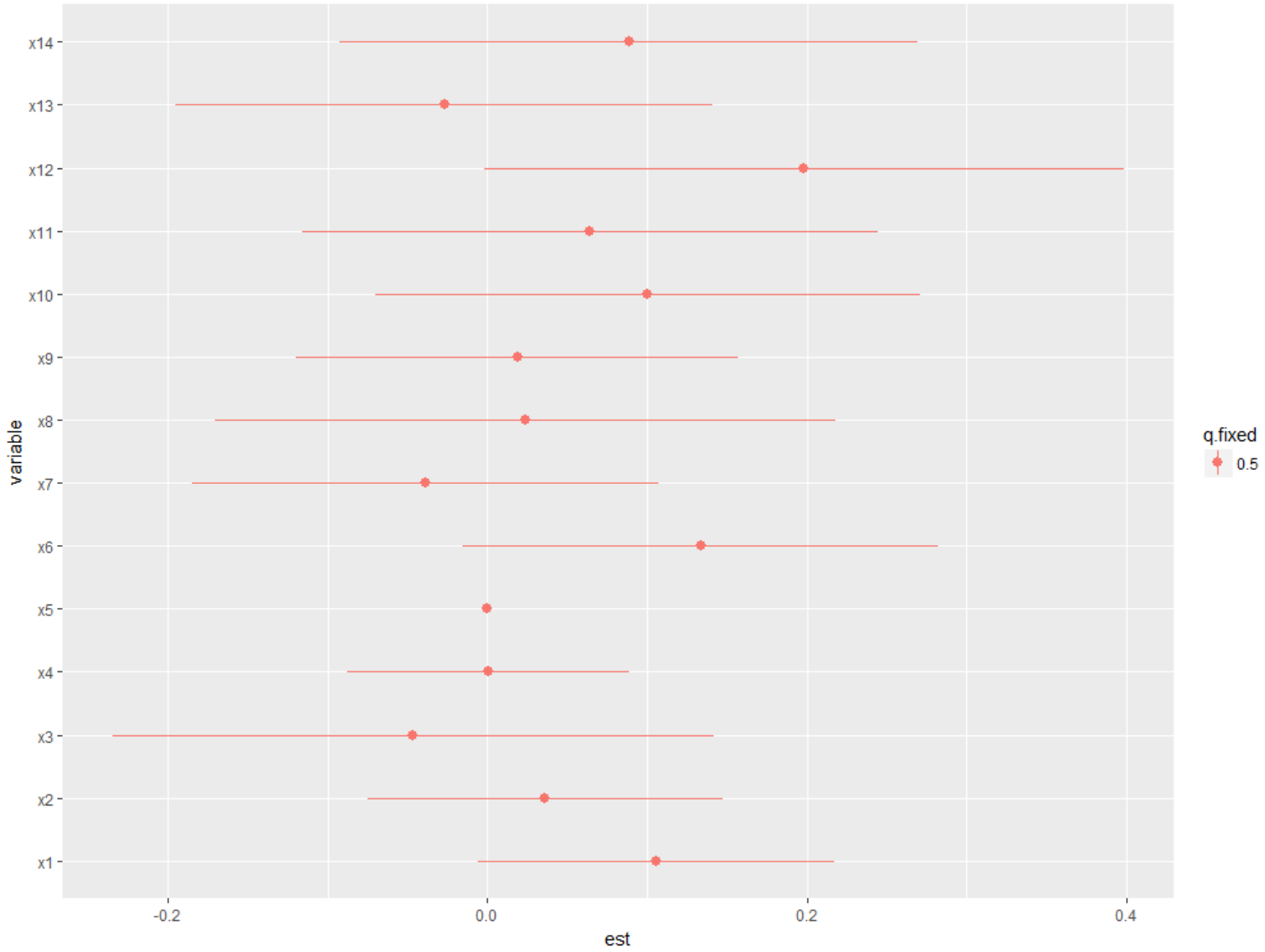
- Estimation is straightforward. The package *bkmr* does everything, just write one line and wait few minutes
- Different tools are then available to summarize results
- Introduction to the package: `https://jenfb.github.io/bkmr/overview.html`
- Clearly, it is not possible to visualize the entire function h (in our example, this is a 14-dimension function)
- The best way of presenting results is to visualize the relationship between 1 predictor and the outcome while fixing the other to a specific value (e.g. the median)


x1
x2
x3
x4
x5
x6
x7
x8
x9
x10
x11
x12
x13
x14
h(X)
0.25
0.00
-0.25
-0.50
-4
0
4
8
exposure

(These were the original results)

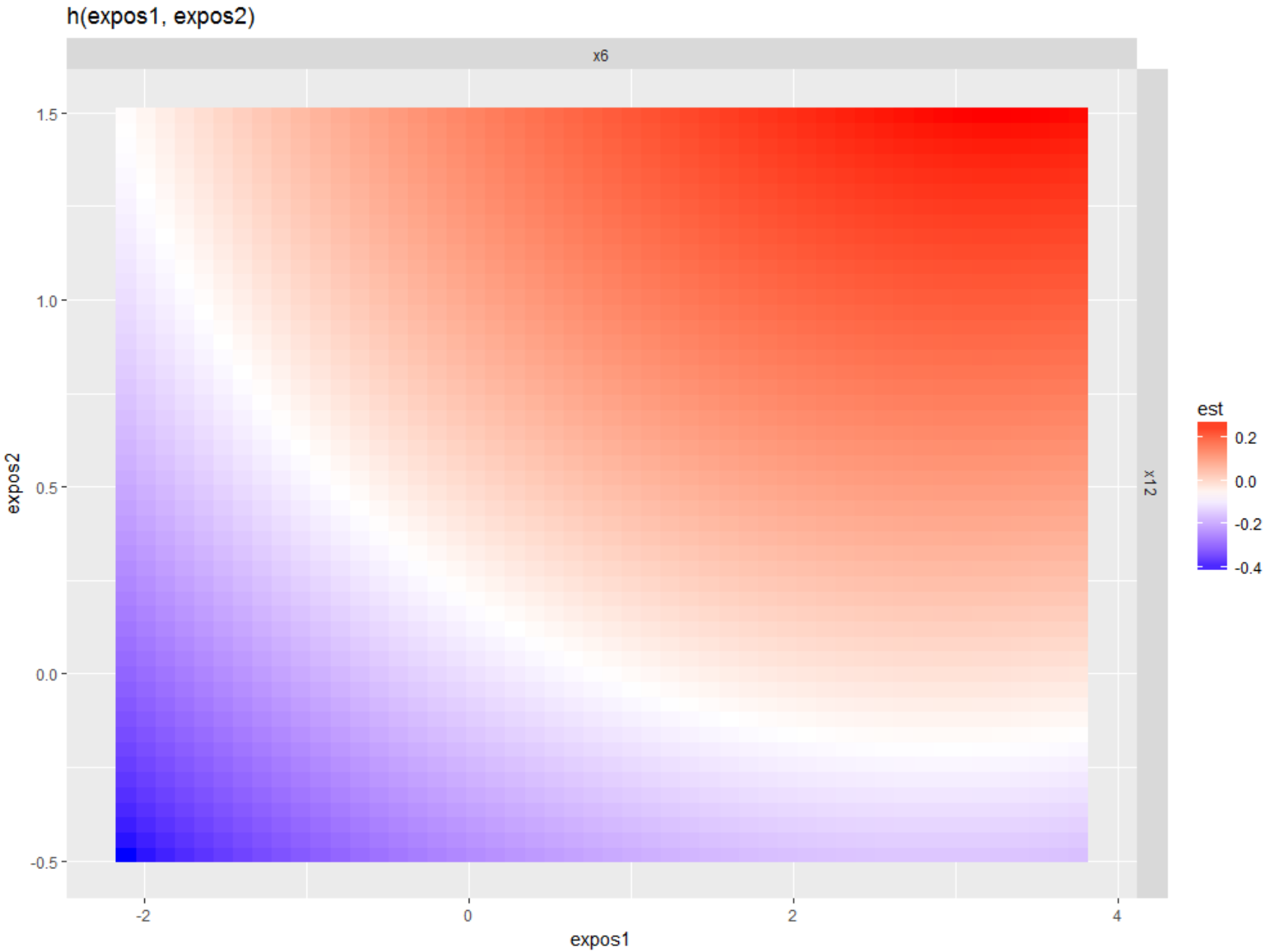
	β (one at the time)	p-value	β (mutually adjusted)	p-value
X_1	0.11	< 0.001	0.06	0.08
X_2	0.07	0.01	0.02	0.55
X_3	0.08	0.007	-0.03	0.77
X_4	0.09	0.003	0.05	0.64
X_5	0.07	0.005	0	0.92
X_6	0.12	<0.001	0.06	0.47
X_7	0.15	<0.001	-0.03	0.62
X_8	0.14	<0.001	0.02	0.68
X_9	0.16	<0.001	0.02	0.67
X_{10}	0.12	<0.001	0.05	0.26
X_{11}	0.15	<0.001	0.05	0.34
X_{12}	0.29	<0.001	0.22	0.14
X_{13}	0.24	<0.001	-0.08	0.59
X_{14}	0.18	<0.001	0.05	0.29

We could also plot the effects for a specific change in the predictors (e.g. from the 10th to the 90th percentile). This is the best summary plot if all relationships are linear



BKMR - bivariate predictor-response function

- We may be interested in the joint effect of two predictors, while fixing the other to a predefined value (eg. the median)
- We could focus, for example, on those predictors where we observed the strongest individual effects (X_6 and X_{12})


h(expos1, expos2)
x6
x12
expos2
expos1
1.5
1.0
0.5
0.0
-0.5
-2
0
2
4
est
0.2
0.0
-0.2
-0.4

Discussion

Discussion

- Correlation between multiple exposures is an important threat. It is more common and less addressed than expected
- This issue, if not addressed, may have severe impact on the results. In our example we found that only 2/3 exposures had a significant effect. The other individual effects were a results of the high correlation with these 2/3 main contributors

Discussion

- Correlation between multiple exposures is an important threat. It is more common and less addressed than expected
- This issue, if not addressed, may have severe impact on the results. In our example we found that only 2/3 exposures had a significant effect. The other individual effects were a results of the high correlation with these 2/3 main contributors
- Several methods have been presented, all of them with strengths and limitations (that is, we don't have an optimal method)
- Classification approaches (e.g. PCA) reduce dimensionality and are useful if biological groups (i.e. exposures that are expected to have the same behaviour with respect to the outcome) are clearly identified
- However, they fail to identify the specific contribution of each predictor and cannot take into account non-linear effects

Discussion (2)

- BKMR, while taking into account the grouping of the exposure, addresses these two limitations, thus proposing itself as a complete and flexible method
- This is potentially true, as much work remains to be done (e.g. binary outcomes, multiple measurements)

Practical tips

- Try to focus on the big picture (i.e. go beyond risk factors epidemiology)

Practical tips

- Try to focus on the big picture (i.e. go beyond risk factors epidemiology)
- Always check the correlation matrix

Practical tips

- Try to focus on the big picture (i.e. go beyond risk factors epidemiology)
- Always check the correlation matrix
- Identify if you have biologically similar exposures, and in that case use PCA

Practical tips

- Try to focus on the big picture (i.e. go beyond risk factors epidemiology)
- Always check the correlation matrix
- Identify if you have biologically similar exposures, and in that case use PCA
- BKMR, give it a try!

Practical tips

- Try to focus on the big picture (i.e. go beyond risk factors epidemiology)
- Always check the correlation matrix
- Identify if you have biologically similar exposures, and in that case use PCA
- BKMR, give it a try!
- All other methods are worthy consideration and were not covered here only because of time

Are all studies I published so far wrong?

Are all studies I published so far wrong?

- Definitely not, but they may not have taken into account the full picture
- Use your previous findings as a starting point for your first mixture model

Main References

- Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the health effects of exposure to multi-pollutant mixture. Annals of epidemiology. 2012 Feb 29;22(2):126-41.
- Bobb JF, Valeri L, Henn BC, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2014 Dec 22:kxu058.
- Braun JM, Gennings C, Hauser R, Webster TF. What can epidemiological studies tell us about the impact of chemical mixtures on human health?. Environmental health perspectives. 2016 Jan;124(1):A6.
- Schisterman EF, Perkins NJ, Mumford SL, Ahrens KA, Mitchell EM. Collinearity and causal diagramsa lesson on the importance of model specification. Epidemiology. 2017 Jan;28(1):47.
- Taylor KW, Joubert BR, Braun JM, Dilworth C, Gennings C, Hauser R, Heindel JJ, Rider CV, Webster TF, Carlin DJ. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. Environmental Health Perspectives. 2016 Dec;124(12):A227.
- Thomas DC, Witte JS, Greenland S. Dissecting effects of complex mixtures: who's afraid of informative priors? Epidemiology. 2007 Mar 1;18(2):186-90.
- VanderWeele TJ. Invited commentary: structural equation models and epidemiologic analysis. American journal of epidemiology. 2012 Sep 6:kws213.