Identifying potential significant factors impacting zero-inflated proportion data

—

Supporting Information

Mélina Ribaud[1,*], Edith Gabriel[1], Joseph Hughes[2], and Samuel Soubeyrand[1,*]

[1]INRAE, BioSP, 84914 Avignon, France
[2]MRC-University of Glasgow, Centre for Virus Research, Glasgow, Scotland, United Kingdom
[*]Corresponding authors: melina.ribaud@gmail.com and samuel.soubeyrand@inrae.fr

April 3, 2023

Appendix A: Why within-block permutations?

Here, we consider a degenerate case to show the huge loss of power when classical permutations are done instead of within-block permutations. The notations are the same as the ones presented in the main text. Consider $n_t n_c$ realizations of a factor X such that $x^{1,1} > x^{1,2} > \ldots > x^{n_t,n_c}$ and $n_t n_c$ realizations of the response variable Z such that for a fixed target (or receiver) i:

$$z_1^i \leq \ldots \leq z_{n_c}^i, \tag{A1}$$

$$\sum_{j=1}^{n_c} \mathbf{1}_{z_j^i > 0} = c,\ c \leq n_c, \tag{A2}$$

$$\sum_{j=1}^{n_c} z_j^i = \frac{i}{n_t}. \tag{A3}$$

This toy example can mimic real cases. For instance, consider a situation where the spread of a plant pathogen follows the wind flow, e.g. from West to East, and target and contributing

nodes are located along this gradient as illustrated in Figure A1. The response are the probabilities of transmission and the factor is the distance between hosts. The closer the contributor to the target, the higher the probability of transmission (Equation (A1)). Only a given number of hosts are potential contributors (Equation (A2)). Equation (A3) can result from an external contributor that transmits the virus from East to West by another path like underground river. See e.g.,[1] for the definition of penalization to favor transmissions at short (geographic or genetic) distances.

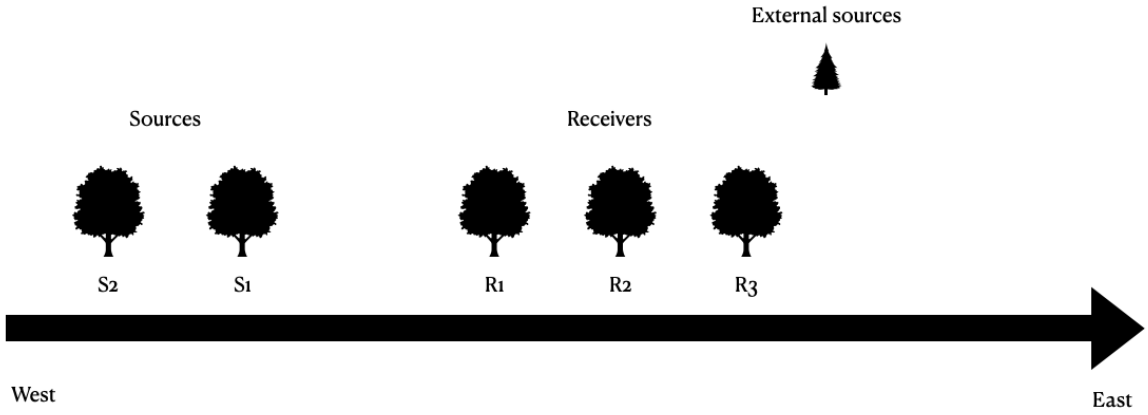


Figure A1: Schematic representation of plant locations.

In this context, the factor $\mathbf{x}$ highly impacts the response $\mathbf{z}$, so that the power of the test, $1-\beta = 1-\mathbb{P}(H_0|H_1)$ is equal to one when performing within-block permutations and much lower than one from classical permutations. Indeed, for b a permutation by block of receivers, $\mathbb{P}(H_0|H_1) = \mathbb{P}(T^b \geq T) = 0$ because no within-block permutation can give a Spearman correlation greater than the non-permuted sample (i.e. $T^b < T$, $\forall b$), while for π a permutation without block constraint, $\mathbb{P}(H_0|H_1) = \mathbb{P}(T^\pi \geq T) >> 0$ because some permutation can give a Spearman correlation greater than the non-permuted sample.

For the sake of illustration, we take $n_t = 10$, $n_c = 20$ and $c = 5$ with 1000 simulated responses. The response is computed as follows, $\forall i \in \{1, \ldots, n_t\}$:

1. Generate c realizations of the random variable $Y \sim \mathcal{U}([0;1])$ and order them so that $y_1 \leq \ldots \leq y_c$;
2. Compute the simulated response: $(z_1^i, \ldots, z_{n_c}^i) = \frac{i}{n_t \sum_{k=1}^{c} y_k}(0, \ldots, 0, y_1, \ldots, y_c)$.

The factor $\mathbf{x}$ is equal to $n_t n_c, n_t n_c - 1, \ldots, 2, 1$. The estimated powers are 1 with within-block-permutation and 0.05 with classical permutation.

Hence, within-block permutations may be crucial to identify factors that are correlated to the response variable.

Appendix B: $\Delta_{\mathbf{x},\mathbf{y}}$ calculation

The quantity $\Delta_{\mathbf{x},\mathbf{y}}$ comes from the optimal Spearman's correlation when the rank of two vectors $\mathbf{y^0} = (y_1^0, \ldots, y_n^0) \in \mathbb{R}_+^n$ and $\mathbf{x^0} = (x_1^0, \ldots, x_n^0) \in \mathbb{R}^n$ are equal except on a given set of indices. In our context, this set corresponds to the zeros of the response.[3] gives some formulas for the Spearman's correlation, and[5] details the calculation of the Spearman's correlation when the vectors $\mathbf{y^0}$ and $\mathbf{x^0}$ have consecutive ties. In the calculation below, we use the same reasoning as those presented in the two aforementioned articles to obtain the desired upper bound.

Let $y_i = R_{y_i^0}$ denote the rank of y_i^0 within $\mathbf{y^0}$, $x_i = R_{x_i^0}$ the rank of x_i^0 within $\mathbf{x^0}$, $I_0 = \{i | y_i^0 = 0\}$ the set of indices for which $y_i^0 = 0$ and $n_0 = \#\{I_0\}$ the number of such indices.

Assume that the ranks are equal $x_i = y_i$ for all $i \notin I_0$ and set $y_i = \frac{n_0+1}{2}$ for all $i \in I_0$, such that $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ (the assumption of equal ranks for all $i \notin I_0$ implies that the Spearman's correlation takes its maximum value given I_0). Then, the Spearman's correlation of $\mathbf{y^0}$ and $\mathbf{x^0}$, which is equal to the Pearson correlation of $\mathbf{y}$ and $\mathbf{x}$, satisfies:

$$\begin{aligned} r_s^2(\mathbf{x}, \mathbf{y}) &= r^2(\mathbf{x}, \mathbf{y}) \\ &= \frac{\widehat{Cov}^2(\mathbf{x}, \mathbf{y})}{\widehat{\sigma}_{\mathbf{x}}^2 \widehat{\sigma}_{\mathbf{y}}^2}, \end{aligned}$$

with

$$\begin{aligned} \widehat{Cov}(\mathbf{x}, \mathbf{y}) &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}}) \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - n \bar{\mathbf{x}} \bar{\mathbf{y}} \right] \\ &= \frac{1}{n-1} \left[y_0 \sum_{i=1}^{n_0} x_i + \sum_{i=n_0+1}^n y_i^2 - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right] \\ &= \frac{1}{n-1} \left[y_0 \sum_{i=1}^{n_0} y_i + \sum_{i=n_0+1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^{n_0} y_i^2 + \sum_{i=n_0+1}^n y_i^2 - n \bar{\mathbf{y}}^2 \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 - n \bar{\mathbf{y}}^2 \right] \\ &= \widehat{\sigma}_{\mathbf{y}}^2, \end{aligned}$$

and

$$\begin{aligned}\widehat{\sigma}_{\mathbf{x}}^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{\mathbf{x}}^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n_0} x_i^2 + \sum_{i=n_0+1}^{n} y_i^2 - n\bar{\mathbf{y}}^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=n_0+1}^{n} y_i^2 + \sum_{i=1}^{n_0} y_i^2 - n\bar{\mathbf{y}}^2 + \sum_{i=1}^{n_0} x_i^2 - \sum_{i=1}^{n_0} y_i^2\right] \\ &= \widehat{\sigma}_{\mathbf{y}}^2 + \frac{1}{n-1}\left[\sum_{i=1}^{n_0}(x_i^2 - y_i^2)\right].\end{aligned}$$

Hence,

$$\begin{aligned}\frac{1}{r_s^2(\mathbf{x},\mathbf{y})} &= \frac{\left(\widehat{\sigma}_{\mathbf{y}}^2 + \frac{1}{n-1}\left[\sum_{i=1}^{n_0}(x_i^2 - y_i^2)\right]\right)\widehat{\sigma}_{\mathbf{y}}^2}{\widehat{\sigma}_{\mathbf{y}}^4} \\ &= \frac{\widehat{\sigma}_{\mathbf{y}}^2\widehat{\sigma}_{\mathbf{y}}^2}{\widehat{\sigma}_{\mathbf{y}}^4} + \frac{\left(\sum_{i=1}^{n_0}(x_i^2 - y_i^2)\right)\widehat{\sigma}_{\mathbf{y}}^2}{(n-1)\widehat{\sigma}_{\mathbf{y}}^4} \\ &= 1 + \frac{\sum_{i=1}^{n_0}(x_i^2 - y_i^2)}{(n-1)\widehat{\sigma}_{\mathbf{y}}^2},\end{aligned}$$

and

$$r_s^2(\mathbf{x},\mathbf{y}) = \frac{1}{1+\Delta_{\mathbf{x},\mathbf{y}}},$$

where $\Delta_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i=1}^{n_0}(x_i^2 - y_i^2)}{(n-1)\widehat{\sigma}_{\mathbf{y}}^2}$.

Consequently, under the same hypothesis for the vector $y \in \mathbb{R}_+^n$ we have:

$$r_s^2(\mathbf{x},\mathbf{y}) \leq \frac{1}{1+\Delta_{\mathbf{x},\mathbf{y}}} \Leftrightarrow r_s^2(\mathbf{x},\mathbf{y})(1+\Delta_{\mathbf{x},\mathbf{y}}) \leq 1,$$

for all vector $x \in \mathbb{R}^n$.

In addition, if $\mathbf{y}$ is such that $y_i \neq y_j$ for all $(i,j) \notin I_0^2$, $i \neq j$, and $\mathbf{x}$ is such that $x_i \neq x_j$ for all $(i,j) \in \{1,\ldots,n\}^2$, $i \neq j$, the quantity $\Delta_{\mathbf{x},\mathbf{y}}$ can be defined in a simple way:

$$\begin{aligned}\widehat{\sigma}_{\mathbf{y}}^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{\mathbf{y}}^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n_0}\left(\frac{n_0+1}{2}\right)^2 + \sum_{i=n_0+1}^{n} i^2 - n\left(\frac{n+1}{2}\right)^2\right] \\ &= \frac{1}{n-1}\left[\frac{n_0(n_0+1)^2}{4} + \frac{n(2n+1)(n+1)}{6} - \frac{n_0(2n_0+1)(n_0+1)}{6} - \frac{n(n+1)^2}{4}\right] \\ &= \frac{1}{12(n-1)}\left[n(n+1)(n-1) - n_0(n_0+1)(n_0-1)\right],\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{n_0}(x_i^2 - y_i^2) &= \sum_{i=1}^{n_0} x_i^2 - n_0 y_0^2 \\ &= \sum_{i=1}^{n_0} i^2 - \frac{n_0(n_0+1)^2}{4} \\ &= \frac{1}{12}\left[n_0(n_0+1)(n_0-1)\right],\end{aligned}$$

lead to the following expression,

$$\begin{aligned}\Delta_{\mathbf{x},\mathbf{y}} &= \frac{n_0(n_0+1)(n_0-1)}{n(n+1)(n-1) - n_0(n_0+1)(n_0-1)} \\ &= \frac{n_0(n_0^2-1)}{n(n^2-1) - n_0(n_0^2-1)}.\end{aligned}$$

Appendix C: Comparison of ranking methods with cross validation

The objective of the cross validation step is to compare our methodology, called multitest-based multivariate analysis (MMA), to linear regression (LM), rank-based linear regression (LMRank) and decision tree (Tree) in its ability to adequately rank the weights of the contributors for any target. For the linear regression, the least-squares error is minimized on each training data set. For the linear regression based on ranks, the L2 norm used in the linear regression is replaced by a pseudo-norm defined in[4], which is a function of the ranks of the residuals. For both LM and LMRank, the model is:

$$Z = \boldsymbol{\beta}'\mathbb{X} + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Note that these model assumptions are not satisfied by proportion data, which are not normally distributed and not independent, but here we only use the predictions of the response variables provided by the regression (i.e., the fitted values). The decision tree learns on the training set with the CART algorithm[2] and, like for the regressions, we use the predictions of the response variables provided by the tree in what follows.

We use two indicators to compare the four methodologies. The first one is the performance indicator introduced in Equation (3) of the article, that we rewrite here as follows $I(M_{\mathbb{X}}\hat{\boldsymbol{\beta}}, \mathbf{z}) = I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, \mathbf{z})$. Thus, the performance indicator is equal to $I(\hat{\mathbf{z}}, \mathbf{z})$ where $\hat{\mathbf{z}}$ is either $M_{\mathbb{X}}\boldsymbol{\beta}$ for the multitest-based multivariate analysis (MMA) or the prediction of the response variable for LM, LMRank and Tree.

The second indicator, say CR which stands for contributor ranking indicator, assesses whether the target-contributor pairs that have strictly positive probabilities are ranked as likely pairs by the method under consideration. For a fixed target i and any of the four methods, let $E_i = \{j \in \{1, \ldots, n_c\} : z_j^i > 0\}$ be the set of target-contributor pairs that have strictly positive probabilities, $N_i = \text{card}(E_i)$, $\hat{R}_j^i$ be the predicted rank of the pair (i, j) for $j \in \{1, \ldots, n_c\}$ (the larger the predicted value for the response variable, the larger the rank), $J_i = \{j \in \{1, \ldots, n_c\} : \hat{R}_j^i \in \{n_c - N_i, \ldots, n_c\}\}$ be the set of contributors ranked among the top N_i contributors by the predictor under consideration, and CR_i be the proportion of the N_i contributors with positive transmission probabilities for target i that are ranked among the top N_i contributors by the predictor under consideration:

$$\text{CR}_i = \frac{1}{N_i} \sum_{\ell \in J_i} \mathbf{1}(\ell \in E_i).$$

Table C1 gives examples of CR_i computations for a set of 10 contributors, among which only the first three have positive probabilities z_j^i. CR_i quantifies, for target i, the quality of identification of the contributors with the highest ranks. The order of the contributors among the group of contributors with the highest ranks does not matter (and the order of the contributors outside this group does not matter as well). The contributor ranking (CR) indicator is defined as the average over the n_t targets of CR_is:

$$\mathrm{CR} = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathrm{CR}_i.$$

Table C1: An illustrative example for the computation of the indicator CR_i with 10 possible contributors. First row: response variables (probabilities) for each target-contributor pair. Second row: true ranks based on the probabilities. Three last rows: three different predicted rankings. Last column: value of the indicator CR_i.

	1	2	3	4	5	6	7	8	9	10	CR_i
z_j^i	0.3	0.2	0.4	0	0	0	0	0	0	0	
$R_{z_j^i}$	9	8	10	4	4	4	4	4	4	4	1
$\hat{R}_j^i$	9	10	8	3	7	3	3	6	3	3	1
$\hat{R}_j^i$	7	10	8	3	9	3	3	6	3	3	2/3
$\hat{R}_j^i$	6	4	7	4	9	10	4	8	4	4	0

To sum up, the performance indicator I evaluates the method ability to globally order probabilities. The contributor ranking indicator CR focuses on the positive probabilities for each target. It is very useful in the context of our applications since, the higher the indicator, the more able the method to identify the main contributors for each target.

Appendix D: Exploration of the significance of factors related to sex

Here, we investigate eventual confounding effects related to the significant effects of "Same_Sex" and "Trans_Sex" factors on the transmission probability. Even if our permutation tests do not require balanced classes, we firstly explore whether the trend for higher probabilities of F→F transmissions coincides with an excess of female horses. Actually, the number of females is about the half of the number of males (Figure D1, left). Therefore, under complete (uniform) randomness, we would expect about two times more M→F transmissions than F→F transmissions (and two times more M→M than F→M). When we only consider the occurrences of "Trans_Sex" corresponding to positive probabilities (without accounting for null probabilities), we clearly see the excess of F→F and F→M transmissions compared to their expected values under complete randomness (Figure D1, right). To complete this observation, transmission probabilities were inferred to be positive for only 19% of all the possible M→F pairs (0.7 times less than expected under complete randomness), 53% for F→F (1.8 times more than expected under complete randomness); see Table D1. Hence, gender distribution is not likely to be involved in the significant effect of the factors related to sex.

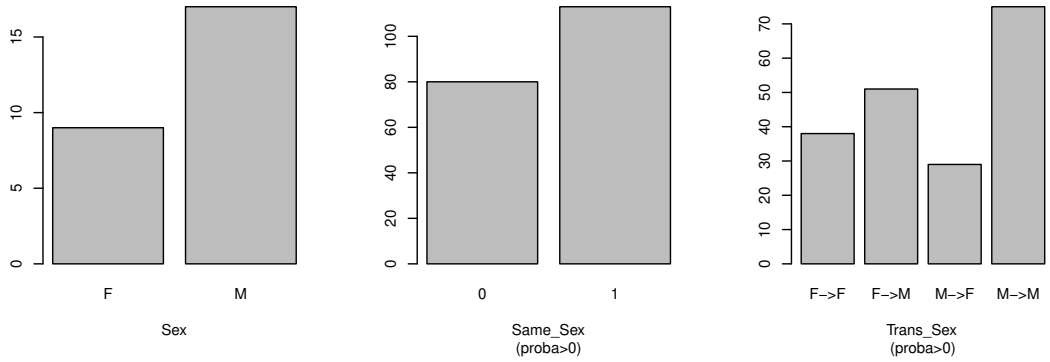


Figure D1: Left: gender distribution (female and male) in the equine influenza study. Center and right: distributions of the variables "Same_Sex" and "Trans_Sex" corresponding to pairs associated with positive transmission probabilities.

Secondly, we explore the eventual existence of confounding factors among those we have considered. As shown by Figure D2, "Same_Sex" and "Trans_Sex" are not correlated with "Same_Yard" and "Dist_Yard", and "Trans_Sex" is only slightly correlated with "Diff_Age". The absence of link between the two yard variables and the two gender variables is confirmed

Table D1: Statistics about "Trans_Sex" modalities for the equine influenza study. Line 1: Transmission probabilities inferred to be positive for all the possible source-receptor pairs with respect to each modality of the "Trans_Sex" variable. Lines 2 and 3: Risk ratio and odds ratio, respectively, for each "Trans_Sex" modality between the observed situation and the hypothetical case of completely random transmissions.

Statistic	F→F	F→M	M→F	M→M
% of positive proba.	53	33	19	28
Risk ratio	1.8	1.2	0.7	1.0
Odds ratio	2.9	1.3	0.4	0.9

by Figure D3 and Table D2. Hence, there seems to be no confounding factors in the data set, and the significance of gender-related factors has to be explained by external processes (e.g., the indirect contacts between hosts via groom, jockey or transport, the behavior of horses in herds, or different immune responses depending on the sex).

Table D2: Result of the independence chi-squared test applied to qualitative factors considered in the equine influenza study.

Factors	χ^2-test statistic	p-value
Same_Yard : Same_Sex	0.04	0.85
Same_Yard : Trans_Sex	1.76	0.62
Same_Yard : Diff_Age	4.3	0.12
Same_Sex : Trans_Sex	650	$< 2.2e^{-16}$

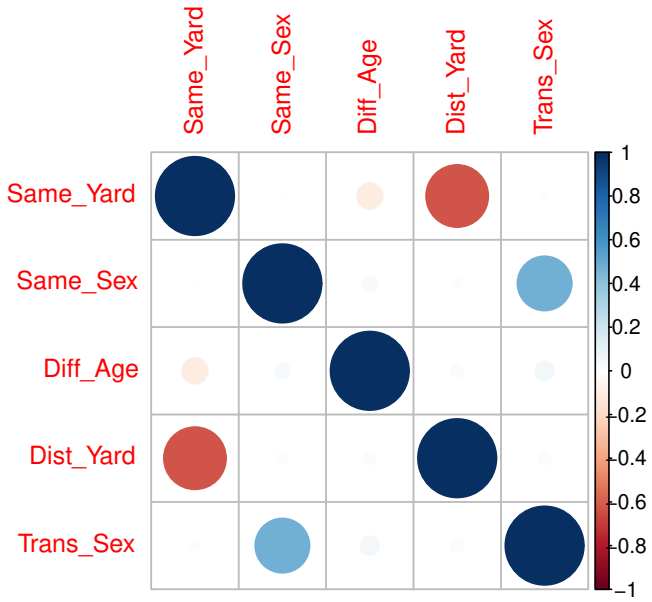


Figure D2: Matrix of Pearson's correlations between factors in the equine influenza study.

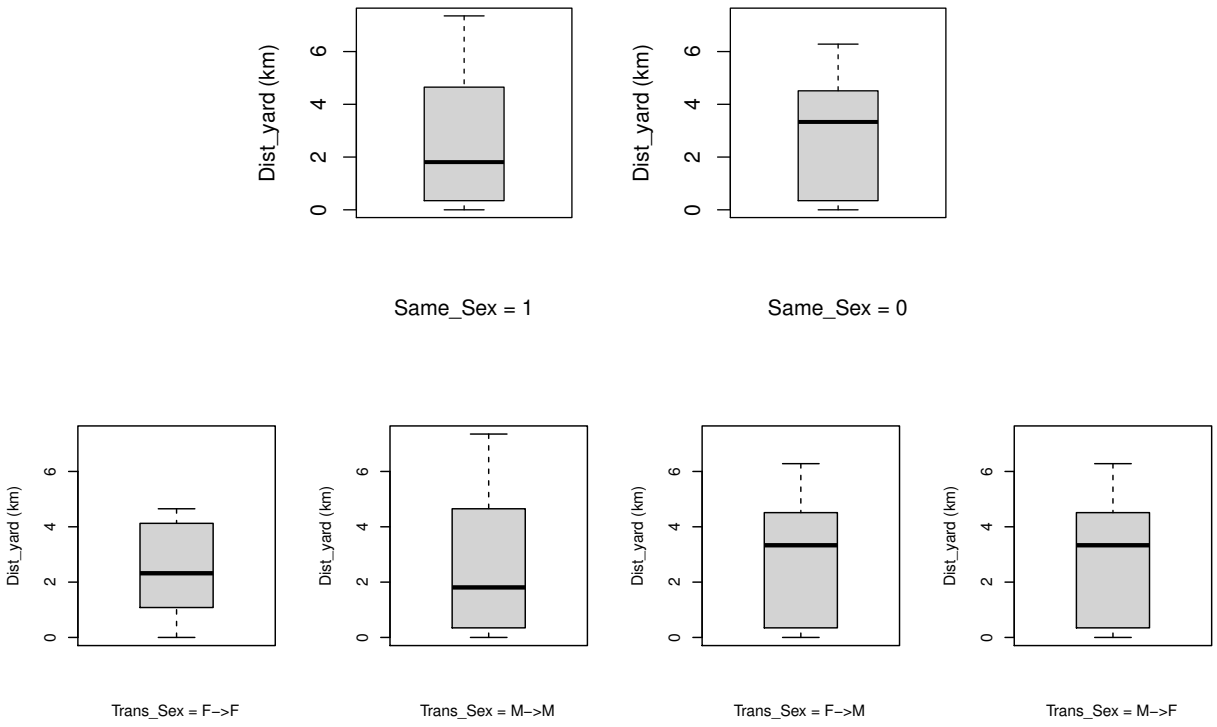


Figure D3: Distribution of the variable "Dist_Yard" by modality of the variable "Same_Sex" and "Trans_Sex" in the equine influenza study.

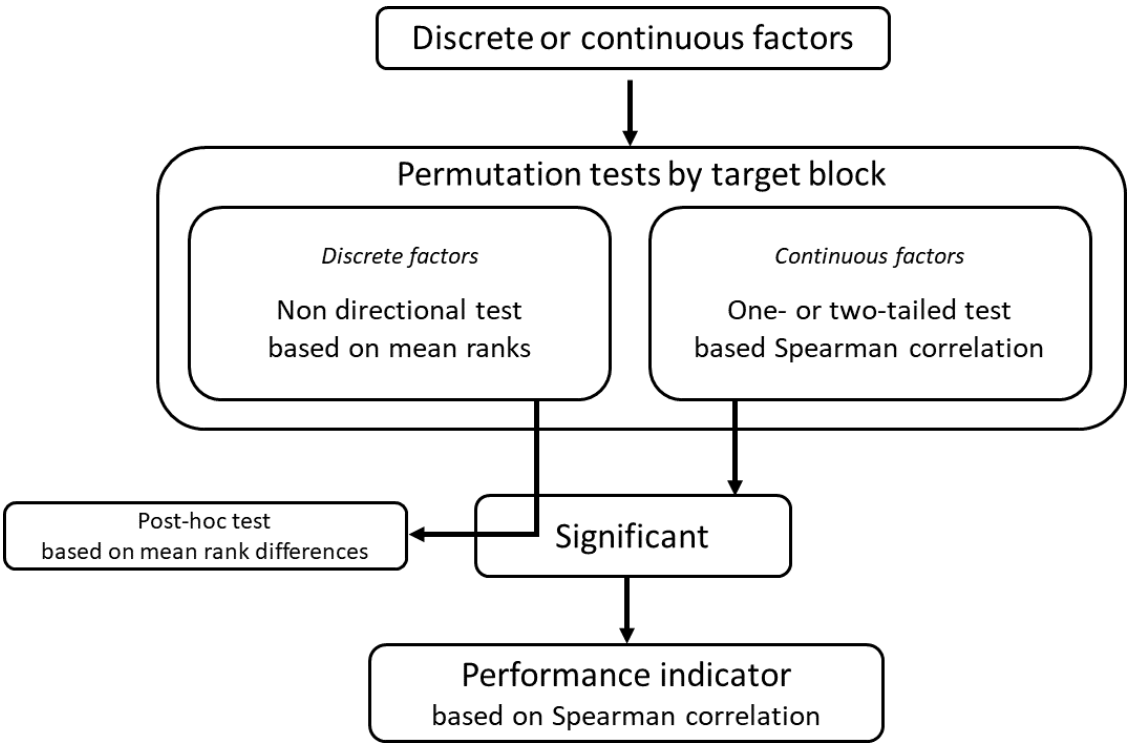
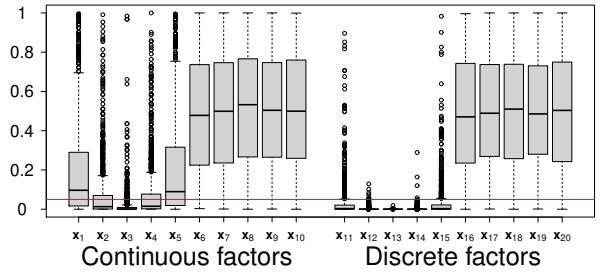
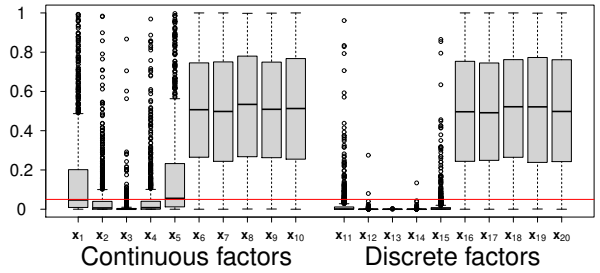


Figure S1: Workflow of the within-block permutation-based methodology.

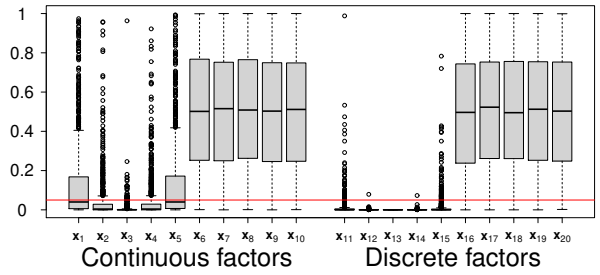
(a) P-value, $m = 0.1$



(b) P-value, $m = 0.15$



(c) P-value, $m = 0.2$



(d) P-value, $m = 0.25$

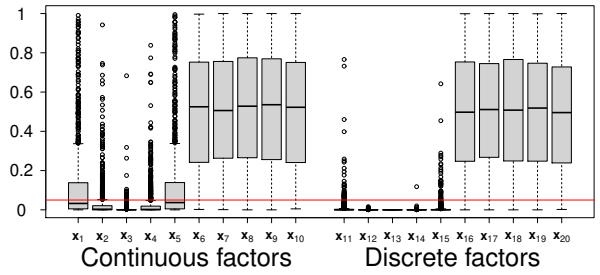
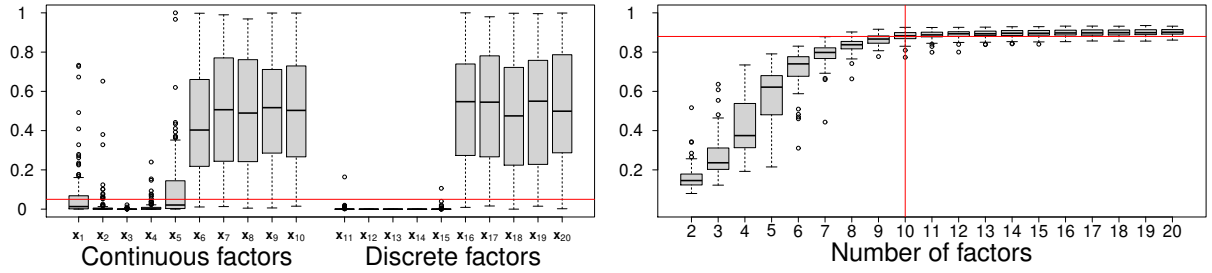


Figure S2: P-values of two-tailed permutation tests for each factor with $m \in \{0.1, 0.15, 0.2, 0.25\}$. The factors $\mathbf{x}_k$ are continuous for $k = \{1, \ldots, 10\}$ and discrete for $k = \{11, \ldots, 20\}$. The data are simulated 1000 times for each value of m.

(a) P-value

(b) Performance indicator



(c) Estimated coefficient

(d) Relative importance

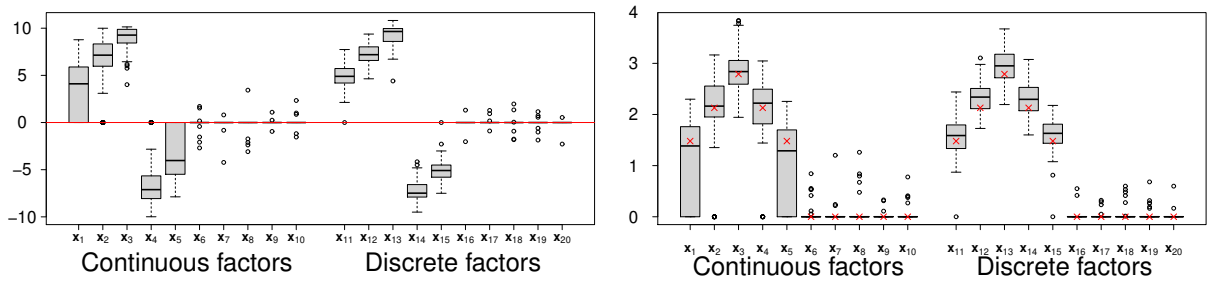


Figure S3: Factor significance and importance in the simulation study when the standard deviation of the noise is set to $\eta = 5$. a) P-values of two-tailed permutation tests for each factor (the red line indicates the 0.05 value). b) Distribution of the performance indicator for varying number of included factors (factors are successively incorporated by first including those with lowest p-values). The horizontal red line gives the median value of the performance indicator computed with the true value of $\boldsymbol{\beta}$ (0.88). c) Distribution of estimated coefficients (i.e., the components of $\hat{\boldsymbol{\beta}}$) for each factor. d) Distribution of the relative importance $\tilde{e}_k$ of each factor; the red cross gives the expected relative importance given the simulation scheme described in Section 2.4. The distributions are drawn with $m = 0.25$ and from 1000 repetitions for a) and 100 repetitions for b), c) and d).

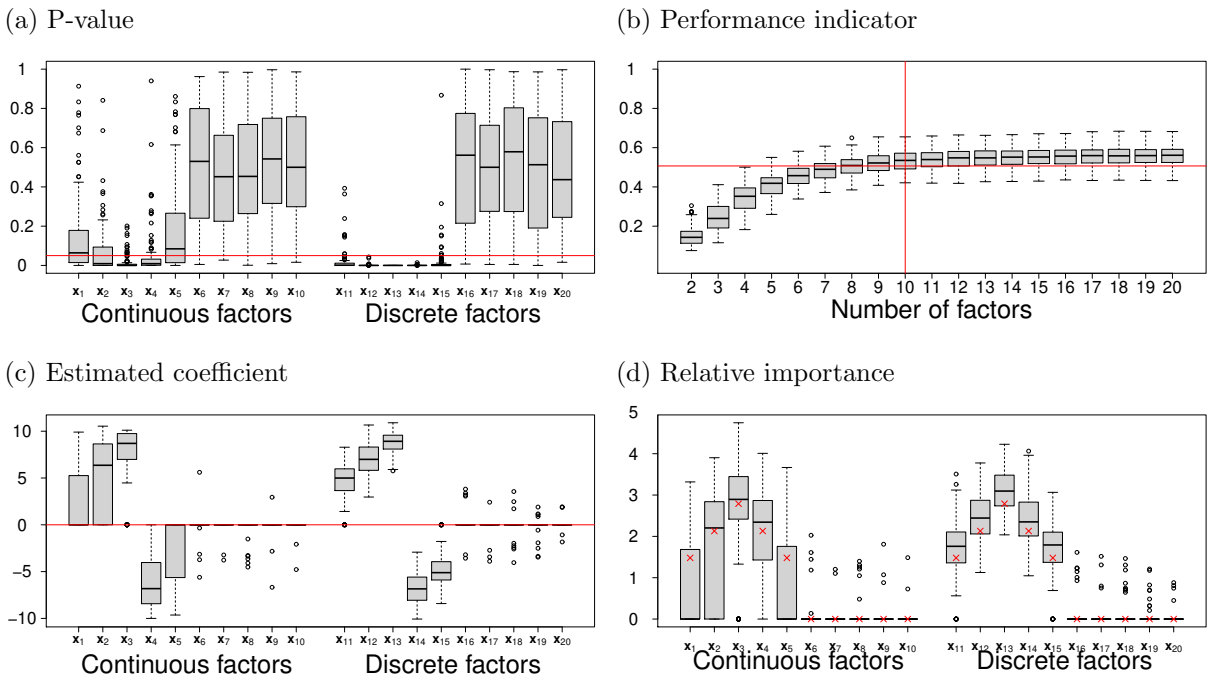
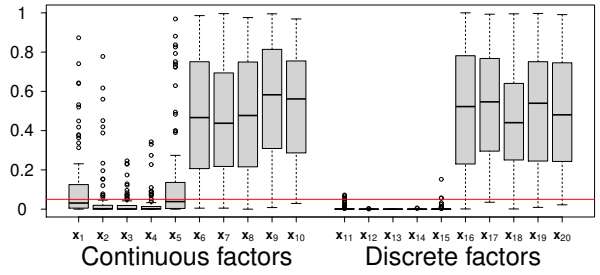
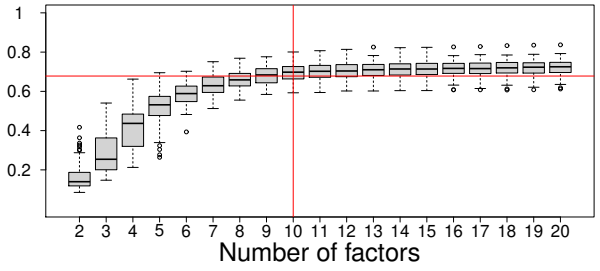


Figure S4: Factor significance and importance in the simulation study when the standard deviation of the noise is set to $\eta = 15$. a) P-values of two-tailed permutation tests for each factor (the red line indicates the 0.05 value). b) Distribution of the performance indicator for varying number of included factors (factors are successively incorporated by first including those with lowest p-values). The horizontal red line gives the median value of the performance indicator computed with the true value of $\boldsymbol{\beta}$ (0.51). c) Distribution of estimated coefficients (i.e., the components of $\hat{\boldsymbol{\beta}}$) for each factor. d) Distribution of the relative importance $\tilde{e}_k$ of each factor; the red cross gives the expected relative importance given the simulation scheme described in Section 2.4. The distributions are drawn with $m = 0.25$ and from 1000 repetitions for a) and 100 repetitions for b), c) and d).

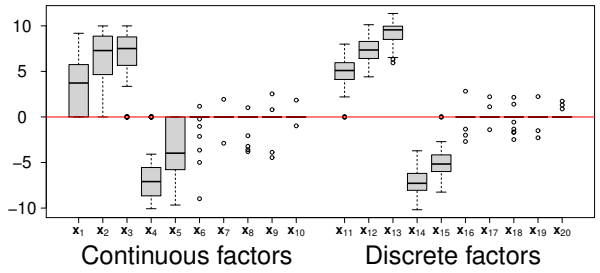
(a) P-value



(b) Performance indicator



(c) Estimated coefficient



(d) Relative importance

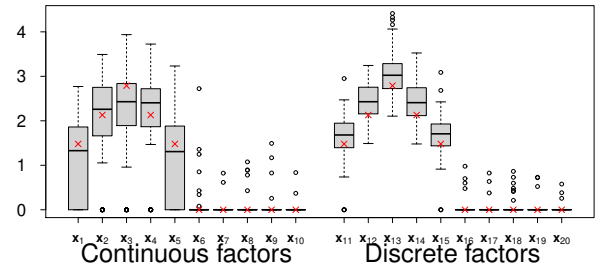
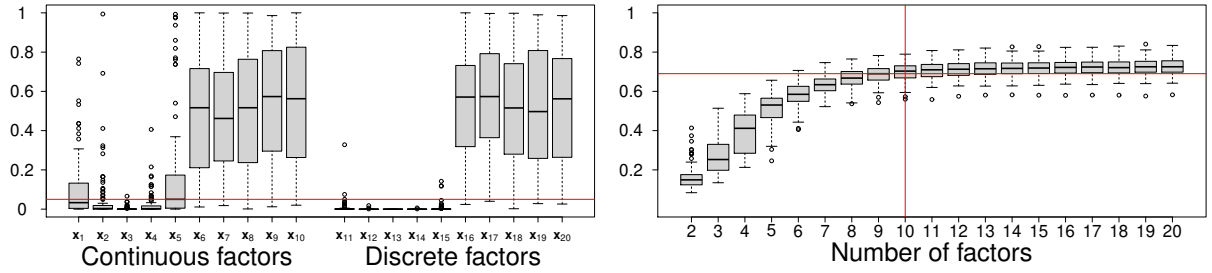


Figure S5: Factor significance and importance in the simulation study when the standard deviation of the noise is set to $\eta = 10$ and factors are transformed with the non-linear function f. a) P-values of two-tailed permutation tests for each factor (the red line indicates the 0.05 value). b) Distribution of the performance indicator for varying number of included factors (factors are successively incorporated by first including those with lowest p-values). The horizontal red line gives the median value of the performance indicator computed with the true value of $\boldsymbol{\beta}$ (0.69). c) Distribution of estimated coefficients (i.e., the components of $\hat{\boldsymbol{\beta}}$) for each factor. d) Distribution of the relative importance $\tilde{e}_k$ of each factor; the red cross gives the expected relative importance given the simulation scheme described in Section 2.4. The distributions are drawn with $m = 0.25$ and from 1000 repetitions for a) and 100 repetitions for b), c) and d).

(a) P-value

(b) Performance indicator



(c) Estimated coefficient

(d) Relative importance

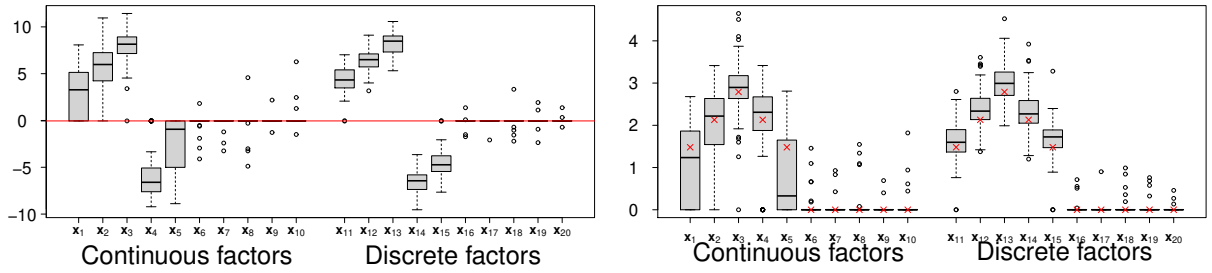
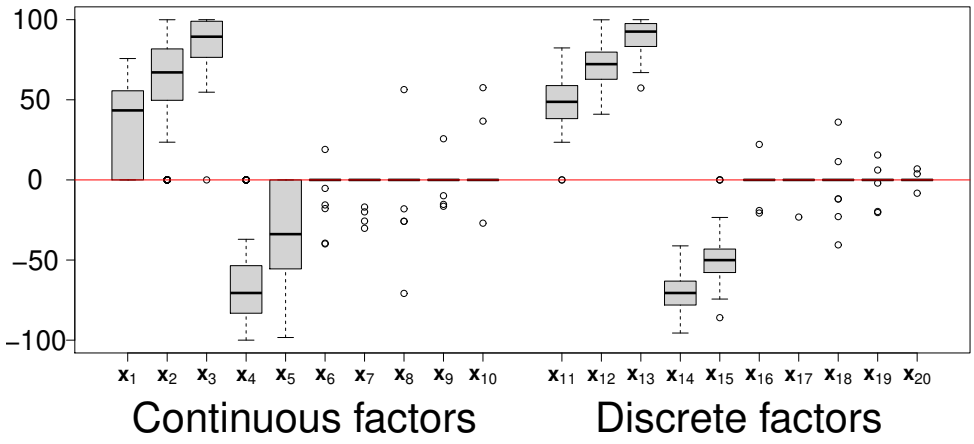


Figure S6: Factor significance and importance in the simulation study when the standard deviation of the noise is set to $\eta = 10$, and each sum $\sum_{j=1}^{n_c} z_j^i$ is drawn from a uniform distribution between 0.5 and 1 (which mimics the non observation of some contributors), instead of fixing $\sum_{j=1}^{n_c} z_j^i$ to the value 1 as described in stage 4 of the simulation algorithm detailed in Section 3.1. a) P-values of two-tailed permutation tests for each factor (the red line indicates the 0.05 value). b) Distribution of the performance indicator for varying number of included factors (factors are successively incorporated by first including those with lowest p-values). The horizontal red line gives the median value of the performance indicator computed with the true value of $\boldsymbol{\beta}$. c) Distribution of estimated coefficients (i.e., the components of $\hat{\boldsymbol{\beta}}$) for each factor. d) Distribution of the relative importance $\tilde{e}_k$ of each factor; the red cross gives the expected relative importance given the simulation scheme described in Section 2.4. The distributions are drawn with $m = 0.25$ and from 1000 repetitions for a) and 100 repetitions for b), c) and d).

(a) Estimated coefficient, without additional non-linearity



(b) Estimated coefficient, with additional non-linearity

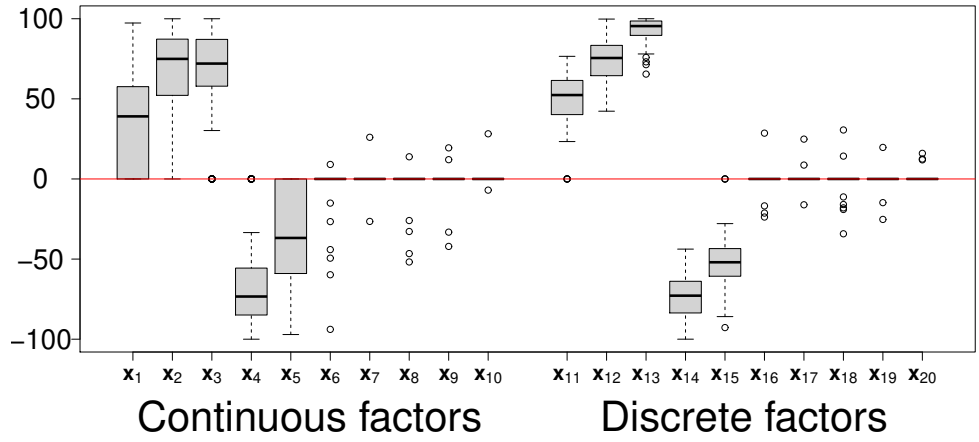


Figure S7: Distributions of estimated coefficients (appearing in β) for each factor when the coefficients are optimized within the interval [-100;100] instead of [-10,10] that was used to produce Figure 3 in the article. The distributions are drawn from 100 repetitions for $m = 0.25$. The top panel corresponds to the case without additional non-linearity, the bottom one to the case with additional non-linearity; the non-linear case refers to the use of the transformation function f specified in the main text, Section 3.1.

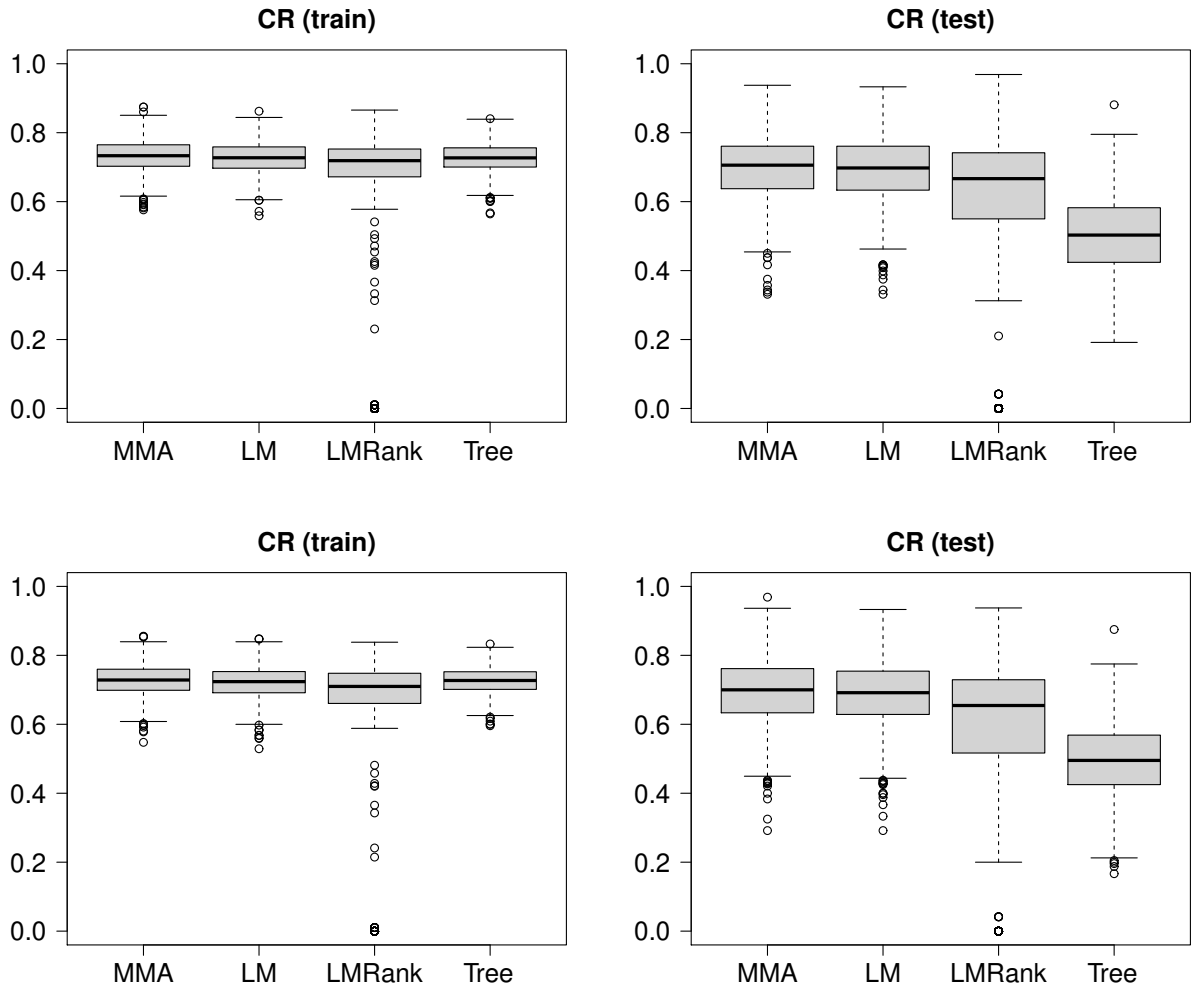


Figure S8: Boxplots of the contributor ranking indicator calculated from the training and test samples for MMA, LM, LMRank and Tree in the simulation study without additional non-linearity (top panels) and with additional non-linearity (bottom panels); the non-linear case refers to the use of the transformation function f specified in the main text, Section 3.1.

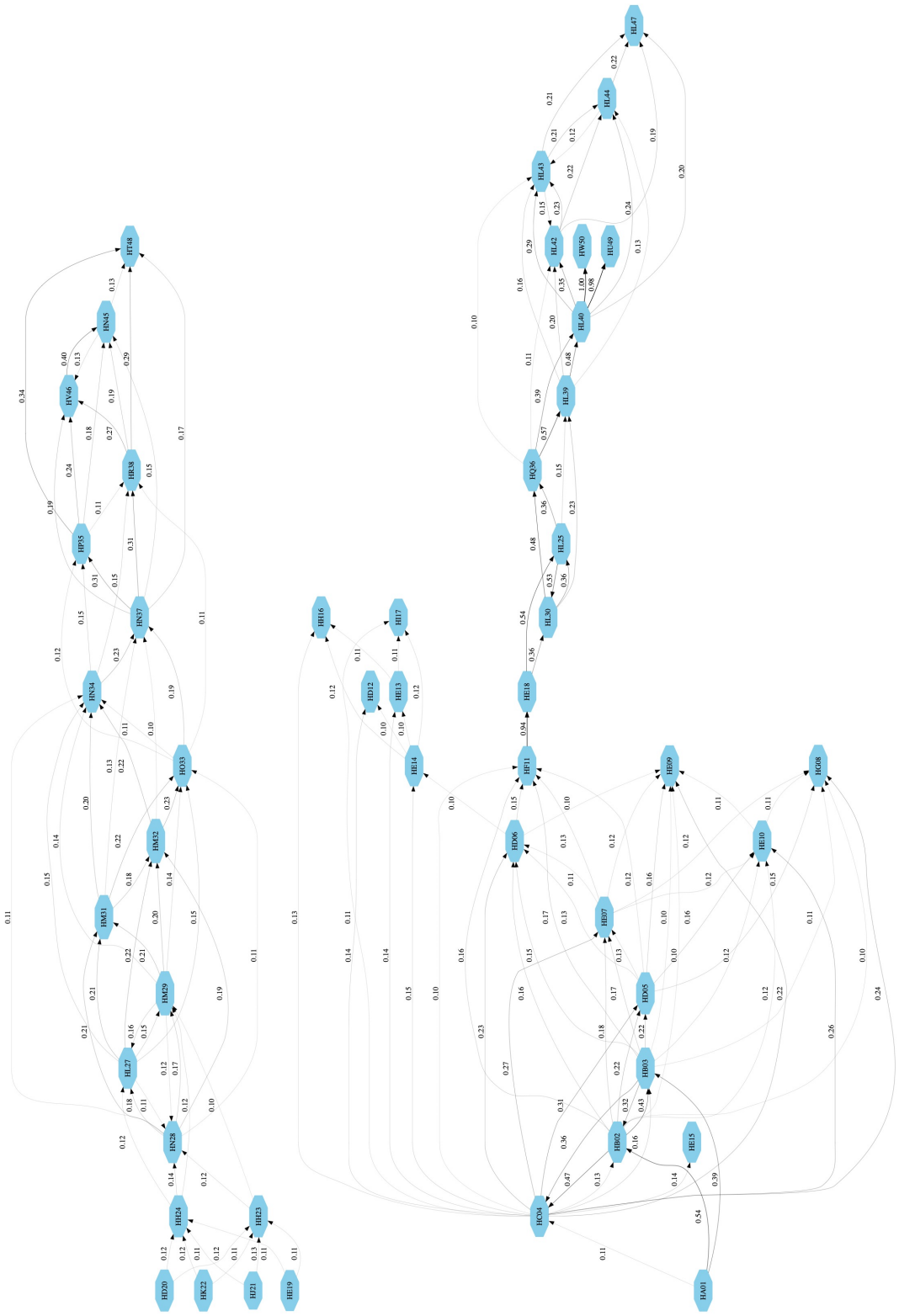


Figure S9: Transmission tree for the Equine Influenza outbreak inferred with BadTrIP. Blue ellipses: hosts; arrows: transmission links; values accompanying arrows: transmission probabilities.

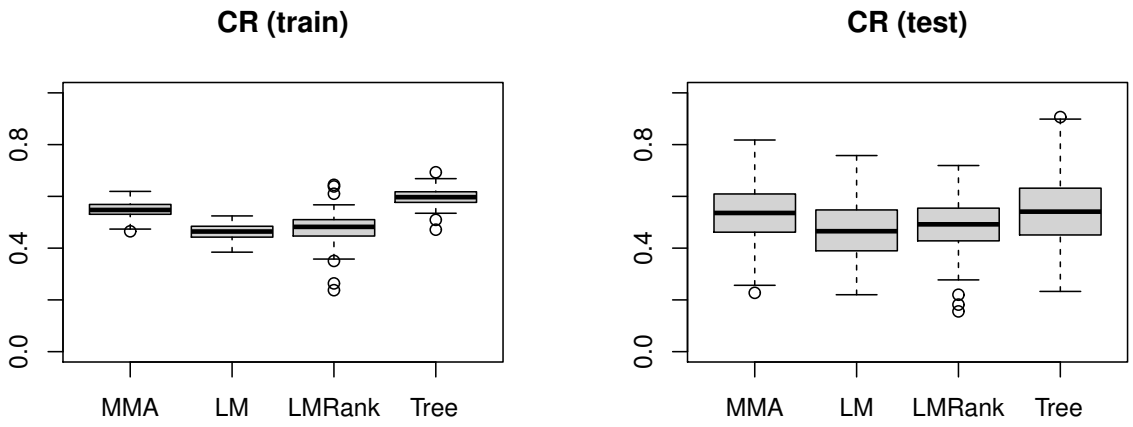


Figure S10: Boxplots of the contributor ranking indicator calculated from the training and test samples for MMA, LM, LMRank and Tree in the equine influenza study.

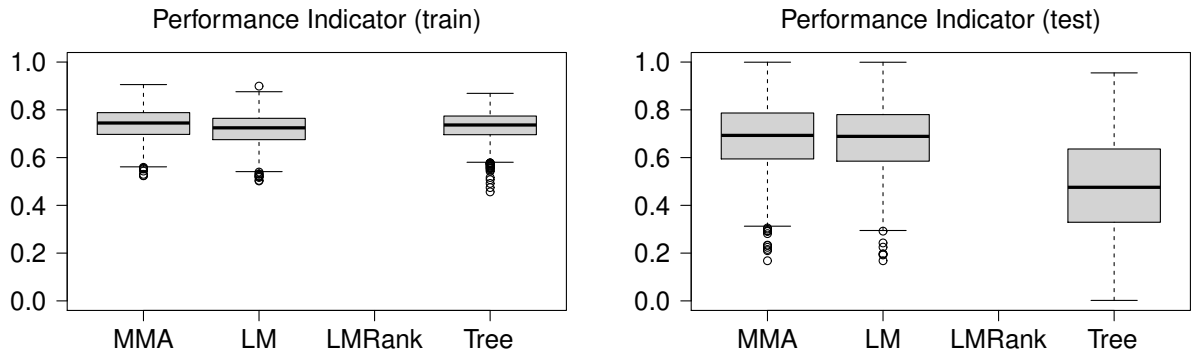


Figure S11: Boxplots of the performance indicator calculated from the training and test samples for MMA, LM, LMRank and Tree in the simulation study without the non-linear transformation f and where the coefficients of all continuous factors have been fixed at 0 (the performance indicator could not be computed for LMRank because this method only predicts null probabilities, which result on a null variance of the ranks and, therefore, an undefined value for the performance indicator).

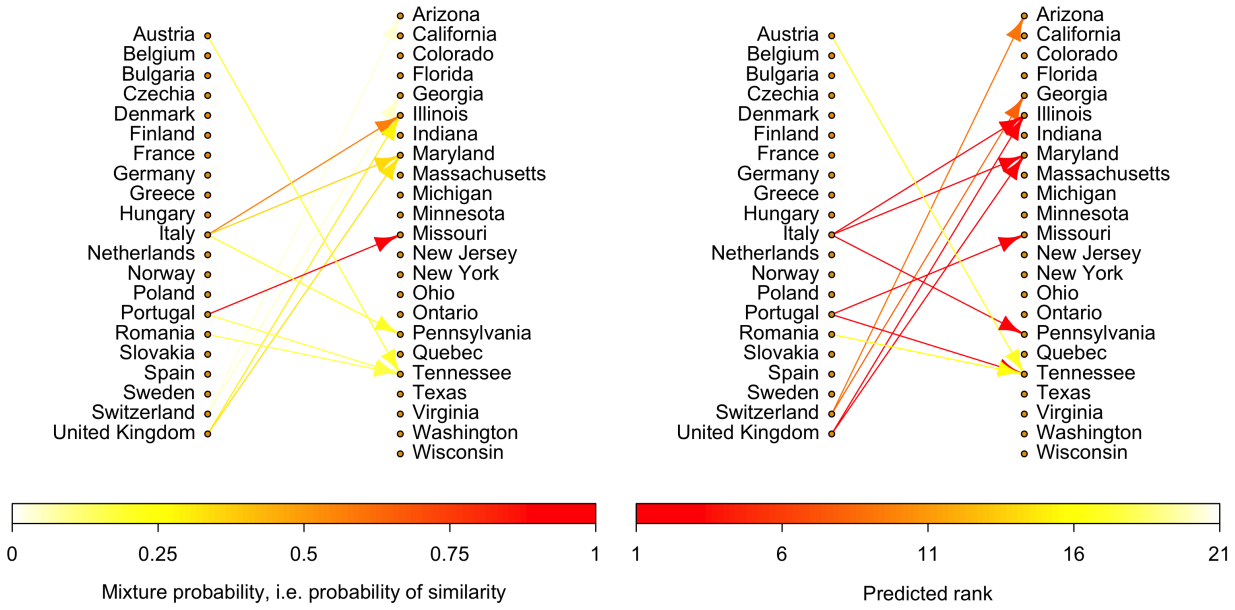


Figure S12: Mixture probabilities z_j^i interpreted as probabilities of similarity (left) and predicted ranks coinciding with the ranks of the linear combination of selected factors $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ (right). In both panels, only arrows corresponding to $z_j^i > 0.01$ are plotted. Most of the contributor-target links on the left are well ranked based on the factor combination $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$. In addition, the four links with a bad rank on the right panel, namely Austria→Tennessee, Romania→Tennessee, Switzerland→Arizona and Switzerland→Georgia correspond to low probabilities of similarity.

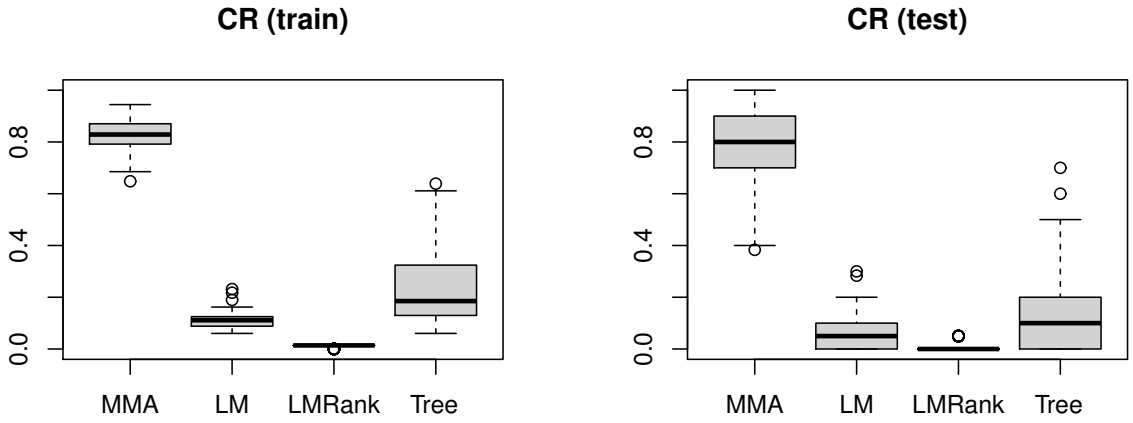


Figure S13: Boxplots of the contributor ranking indicator calculated from the training and test samples for MMA, LM, LMRank and Tree in the COVID-19 study.

Table S1: Estimated type I errors of the one-tailed (ii) permutation tests with 1000 repetitions.

	Continuous factors					Discrete factors				
m	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{16}$	$\mathbf{x}_{17}$	$\mathbf{x}_{18}$	$\mathbf{x}_{19}$	$\mathbf{x}_{20}$
0.1	0.052	0.056	0.050	0.038	0.045	0.048	0.058	0.047	0.065	0.047
0.15	0.043	0.046	0.048	0.052	0.049	0.050	0.054	0.058	0.051	0.053
0.2	0.056	0.063	0.045	0.059	0.048	0.046	0.044	0.058	0.054	0.040
0.25	0.051	0.062	0.048	0.056	0.056	0.038	0.052	0.055	0.055	0.048

Table S2: Estimated type I errors of the one-tailed (iii) permutation tests with 1000 repetitions.

	Continuous factors					Discrete factors				
m	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{16}$	$\mathbf{x}_{17}$	$\mathbf{x}_{18}$	$\mathbf{x}_{19}$	$\mathbf{x}_{20}$
0.1	0.050	0.047	0.048	0.037	0.053	0.048	0.058	0.047	0.065	0.047
0.15	0.051	0.046	0.044	0.042	0.036	0.050	0.054	0.058	0.051	0.053
0.2	0.048	0.053	0.047	0.054	0.038	0.046	0.044	0.058	0.054	0.040
0.25	0.068	0.044	0.044	0.044	0.051	0.038	0.052	0.055	0.055	0.048

Table S3: Estimated type II errors of the one-tailed (ii) permutation tests with 1000 repetitions.

	Continuous factors		Discrete factors	
m	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_{14}$	$\mathbf{x}_{15}$
0.1	0.219	0.480	0.003	0.094
0.15	0.147	0.387	0.001	0.048
0.2	0.113	0.352	0.000	0.037
0.25	0.095	0.320	0.001	0.027

Table S4: Estimated type II errors of the one-tailed (iii) permutation tests with 1000 repetitions.

	Continuous factors			Discrete factors		
m	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$
0.1	0.490	0.194	0.043	0.096	0.001	0.000
0.15	0.380	0.129	0.019	0.056	0.001	0.000
0.2	0.336	0.097	0.016	0.029	0.000	0.000
0.25	0.304	0.088	0.007	0.017	0.000	0.000

Table S5: Description of equine influenza data. Left: Counts of hosts with non-missing information for each observed variable. Right: counts of potential source-receptor pairs without non-missing information for pairwise factors.

	#
All	48
Yard	48
Age	27
Sex	26 (9 females, 17 males)

	#pairs
All	2256
Same_Yard	2256
Dist_Yard	2256
Diff_Age	702
Same_Sex	650
Trans_Sex	650

Table S6: Explanatory factors for the Covid-19 application.

Category	Variable	Description	Unit
Economy	gdp2019	Gross domestic product in 2019	M$
	gdp_capita	Gross domestic product per capita in 2019	$
	healthexp	Health expenditure	M$
Demography	pop	Total population	units
	density	Population density	units per km^2
	urbanpop	Percentage of population living in urban areas	%
	popmale	Percentage of male	%
	pop_tot_0_14	Percentage of population in the age group 0-14 (male, female, total)	%
	pop_tot_15_64	Percentage of population in the age group 15-64 (male, female, total)	%
	pop_tot_65_up	Percentage of population in the age group 65 or more (male, female, total)	%
	mediange	Median age	years
	life_expectancy	Life expectancy at birth	years
Health	lung	Death rate for lung diseases per 100,000 people	units
	fertility	Average number of children per woman	units
	obesity	Percentage of obese people within the popoulation	%
	smokers	Percentage of smokers within the population	%
Healthcare system	hospibed	Number of hospital beds per 1,000 people	units
	physicians_per_1K	Number of physicians per 1,000 people	units
	nurses_per_1K	Number of nurses per 1,000 people	units
Climate	tmin	Average minimum temperature in the first semester	°C
	tmax	Average maximum temperature in the first semester	°C
	prec	Average precipitation in the first semester	mm
	avghumidity	Average relative humidity	%

References

[1] Alamil, M., Hughes, J., Berthier, K., Desbiez, C., Thébaud, G., and Soubeyrand, S. (2019). Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases. Philosophical Transactions of the Royal Society B, 374(1775):20180258.

[2] Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). Classification and Regression Trees. The Wadsworth and Brooks-Cole Statistics-Probability Series. Taylor & Francis.

[3] Du Bois, P. (1939). Formulas and tables for rank correlation. The Psychological Record, 3:46.

[4] Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. The Annals of Mathematical Statistics, 43(5):1449–1458.

[5] Kendall, M. G. (1945). The treatment of ties in ranking problems. Biometrika, 33:239–251.