Supplemental materials

Supplement A: Results on relevance by *loco*.

We use the notation

$$\hat{\sigma}^2_{n_1} = \widehat{\text{MSPE}}^{\text{Train}}(\hat{f}) = \frac{1}{n_1 - p - 1}(\mathbf{y}_1 - \hat{\mathbf{y}}_{1.X.z})^T(\mathbf{y}_1 - \hat{\mathbf{y}}_{1.X.z}).$$

Proposition 3 *Let*

$$RV^{Train}_{loco}(Z) = \frac{1}{\hat{\sigma}^2_{n_1}}\frac{1}{n_1}(\hat{\mathbf{y}}_{1.X.z} - \hat{\mathbf{y}}_{1.X})^T(\hat{\mathbf{y}}_{1.X.z} - \hat{\mathbf{y}}_{1.X})$$

be the relevance by loco *of the variable Z, evaluated in the training sample. It happens that*

$$n_1 RV^{Train}_{loco}(Z) = F_z.$$

Moreover,

$$RV^{Train}_{loco}(Z) = \frac{1}{\hat{\sigma}^2_{n_1}}\hat{\beta}^2_z\hat{\sigma}^2_{z.x,n_1},$$

where $\hat{\sigma}^2_{z.x,n_1}$ is a consistent estimator of $\sigma^2_{z.x}$, the residual variance in the model $Z = X^T\alpha + \varepsilon_z$, computed from the training sample.

Proof We are dealing with the estimation of models $f(x,z) = x^T\beta_x + z\beta_z$ and $f_p(x) = x^T\beta_0$ from the training set $(\mathbf{X}_1, \mathbf{z}_1, \mathbf{y}_1)$. The ordinary least squares (OLS) coefficient estimators are $\hat{\beta}_0 = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}_1$ and

$$\begin{pmatrix}\hat{\beta}_x \\ \hat{\beta}_z\end{pmatrix} = \begin{pmatrix}\mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{z}_1 \\ \mathbf{z}_1^T\mathbf{X}_1 & \mathbf{z}_1^T\mathbf{z}_1\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{X}_1^T \\ \mathbf{z}_1^T\end{pmatrix}\mathbf{y}_1.$$

The predicted values for the response are, respectively, $\hat{\mathbf{y}}_{1.X.z} = \mathbf{X}_1\hat{\beta}_x + \mathbf{z}_1\hat{\beta}_z$ and $\hat{\mathbf{y}}_{1.X} = \mathbf{X}_1\hat{\beta}_0$. Using the expression of the inverse of a partitioned matrix (see Appendix A.1), it is easy to obtain the well known result

$$\hat{\beta}_x = \hat{\beta}_0 - \hat{\alpha}_1\hat{\beta}_z \tag{6}$$

where $\hat{\alpha}_1 = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{z}_1$ is the vector of regression coefficients in the regression of the omitted variable on the other explanatory variables in the training set, and

$$\hat{\beta}_z = \frac{1}{k}(\mathbf{z}_1 - \hat{\mathbf{z}}_1)^T\mathbf{y}_1$$

where $\hat{\mathbf{z}}_1 = \mathbf{X}_1\hat{\alpha}_1$ and $k = (\mathbf{z}_1 - \hat{\mathbf{z}}_1)^T(\mathbf{z}_1 - \hat{\mathbf{z}}_1)$. These results (see, e.g., Seber and Lee 2003) show that the multiple regression coefficient of each variable, $\hat{\beta}_z$, is the slope in the simple regression of $\mathbf{y}$ on $\mathbf{z}_1 - \hat{\mathbf{z}}_1$, the part of $\mathbf{z}_1$ that is uncorrelated to the rest of explanatory variables. Also, $\text{Var}(\hat{\beta}_z) = \sigma^2/k$, the standard t-test statistic for the null hypothesis $H_0 : \beta_z = 0$ is $t_z = \hat{\beta}_z\sqrt{k}/\hat{\sigma}$, where $\hat{\sigma}^2_{n_1} = (\mathbf{y}_1 - \hat{\mathbf{y}}_{1.X.z})^T(\mathbf{y}_1 - \hat{\mathbf{y}}_{1.X.z})/(n_1 - p - 1)$, and the standard F-test statistic for the same null hypothesis is

$$F_z = t_z^2 = \frac{\hat{\beta}_z^2 k}{\hat{\sigma}^2_{n_1}} = \frac{1}{\hat{\sigma}^2_{n_1}}\hat{\beta}_z(\mathbf{z}_1 - \hat{\mathbf{z}}_1)^T(\mathbf{z}_1 - \hat{\mathbf{z}}_1)\hat{\beta}_z =$$

$$\frac{1}{\hat{\sigma}^2_{n_1}}(\hat{\mathbf{y}}_{1.X.z} - \hat{\mathbf{y}}_{1.X})^T(\hat{\mathbf{y}}_{1.X.z} - \hat{\mathbf{y}}_{1.X}) = n_1\text{RV}^{\text{Train}}_{\text{loco}}(Z).$$

The proof concludes when defining $\hat{\sigma}^2_{z.x,n_1} = (\mathbf{z}_1 - \hat{\mathbf{z}}_1)^T(\mathbf{z}_1 - \hat{\mathbf{z}}_1)/n_1$. □

Let us use the notation

$$\hat{\sigma}^2_{n_1,n_2} = \widehat{\mathrm{MSPE}}(\hat{f}) = \frac{1}{n_2}(\mathbf{y} - \hat{\mathbf{y}}_{2.X.z})^T(\mathbf{y} - \hat{\mathbf{y}}_{2.X.z}).$$

Proposition 4 *Assume that the regression function of Y over (X, Z) is linear and that it is estimated by OLS. Then*

$$n_1 RV_{loco}(Z) = F_z \frac{\hat{\sigma}^2_{z.x,n_1,n_2}}{\hat{\sigma}^2_{z.x,n_1}} \frac{\hat{\sigma}^2_{n_1,n_2}}{\hat{\sigma}^2_{n_1}} = F_z \left(1 + O_p\left(\min\{n_1, n_2\}^{-1/2}\right)\right),$$

and

$$RV_{loco}(Z) = \frac{1}{\hat{\sigma}^2_{n_1,n_2}} \hat{\beta}^2_z \hat{\sigma}^2_{z.x,n_1,n_2},$$

where $\hat{\sigma}^2_{z.x,n_1,n_2}$ and $\hat{\sigma}^2_{z.x,n_1}$ are consistent estimators of the same parameter $\sigma^2_{z.x}$ (the residual variance in the linear regression model $Z = X^T\alpha + \varepsilon_z$), the first one depending on both, the training sample and the test sample, and the second one (also appearing in Proposition 3) only on the training sample.

Proof The vectors of predicted values in the test sample are $\hat{\mathbf{y}}_{2.X.z} = \mathbf{X}_2\hat{\beta}_x + \mathbf{z}_2\hat{\beta}_z$ and, using equation (6) in the proof of Proposition 3,

$$\hat{\mathbf{y}}_{2.X} = \mathbf{X}_2\hat{\beta}_0 = \mathbf{X}_2\left(\hat{\beta}_x + \hat{\alpha}_1\hat{\beta}_z\right) = \mathbf{X}_2\hat{\beta}_x + \hat{\mathbf{z}}_{2.1}\hat{\beta}_z,$$

where $\hat{\mathbf{z}}_{2.1} = \mathbf{X}_2\hat{\alpha}_1$ is the prediction of $\mathbf{z}_2$ using the linear model fitted in the training sample to predict Z from X. Therefore, using the same notation as in Proposition 3, the *relevance by* loco of the variable Z is

$$\mathrm{RV}_{\mathrm{loco}}(Z) = \frac{1}{\hat{\sigma}^2_{n_1,n_2}} \frac{1}{n_2}(\hat{\mathbf{y}}_{2.X.z} - \hat{\mathbf{y}}_{2.X})^T(\hat{\mathbf{y}}_{2.X.z} - \hat{\mathbf{y}}_{2.X}) =$$

$$\frac{1}{\hat{\sigma}^2_{n_1,n_2}} \frac{1}{n_2}\hat{\beta}_z(\mathbf{z}_2 - \hat{\mathbf{z}}_{2.1})^T(\mathbf{z}_2 - \hat{\mathbf{z}}_{2.1})\hat{\beta}_z =$$

$$\frac{1}{\hat{\sigma}^2_{n_1,n_2}} \frac{1}{n_2}\hat{\beta}^2_z k_{\mathrm{loco}} = \frac{1}{\hat{\sigma}^2_{n_1,n_2}} \frac{\hat{\sigma}^2_{n_1} F_z}{n_1} \frac{k_{\mathrm{loco}}/n_2}{k/n_1},$$

where $k_{\mathrm{loco}} = (\mathbf{z}_2 - \hat{\mathbf{z}}_{2.1})^T(\mathbf{z}_2 - \hat{\mathbf{z}}_{2.1})$ and k has been defined in the Appendix A.1 of the paper. Observe that both, $\hat{\sigma}^2_{n_1}$ and $\hat{\sigma}^2_{n_1,n_2}$, are consistent estimators of the residual variance in the linear regression of Y over (X, Z). In a similar way, both k_{loco}/n_2 and $k/(n_1 - p)$, are consistent estimators of the residual variance in the linear regression model $Z = X^T\alpha + \varepsilon_z$. The proof concludes when defining $\hat{\sigma}^2_{z.x,n_1,n_2} = k_{\mathrm{loco}}/n_2$ and using $\hat{\sigma}^2_{z.x,n_1}$ defined in Proposition 3. The expression involving the O_p notation is derived by standard arguments for the limit of a quotient. □

In the linear regression model, Proposition 4 and Theorem 1 establish a parallelism between deleting the variable Z and replacing it by a ghost variable. This relationship goes even farther. Consider the linear model $f_z(x) = x^T\alpha$. Let $\hat{\alpha}_1$ and $\hat{\alpha}_2$ be the OLS estimators of α in the training and test samples, respectively. They are expected to be close to each other, because both are OLS

estimators of the same parameter. This expected proximity and the results stated in Appendix A.1, lead us to write

$$\hat{\mathbf{y}}_{2.X.\hat{z}} = \mathbf{X}_2\hat{\beta}_x + \hat{\mathbf{z}}_{2.2}\hat{\beta}_z = \mathbf{X}_2\hat{\beta}_x + \mathbf{X}_2\hat{\alpha}_2\hat{\beta}_z = \mathbf{X}_2\left(\hat{\beta}_x + \hat{\alpha}_2\hat{\beta}_z\right) \approx$$

$$\mathbf{X}_2\left(\hat{\beta}_x + \hat{\alpha}_1\hat{\beta}_z\right) = \mathbf{X}_2\hat{\beta}_0 = \hat{\mathbf{y}}_{2.X}.$$

That is, using the ghost variable $\hat{\mathbf{z}}_{2.2}$ leads to similar predictions of Y in the test sample than removing the variable $\mathbf{z}_1$ when fitting the model in the training sample.

Supplement B: Relevance matrix by random permutations in linear regression

We analyze now the structure of the relevance matrix $\mathbf{V}$ when random permutations are used instead of ghost variables. We focus in the case of multiple linear regression. Define

$$\tilde{\mathbf{A}} = (\mathbf{X}_2 - \mathbf{X}_2')\text{diag}(\hat{\beta}),$$

where the j-th column of matrix $\mathbf{X}_2'$ is $\mathbf{x}_{2.j}'$, a random permutation of $\mathbf{x}_j$. Therefore,

$$\tilde{\mathbf{V}} = \frac{1}{\hat{\sigma}^2_{n_1,n_2}}\frac{1}{n_2}\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}} = \frac{1}{\hat{\sigma}^2_{n_1,n_2}}\frac{1}{n_2}\text{diag}(\hat{\beta})(\mathbf{X}_2 - \mathbf{X}_2')^{\mathrm{T}}(\mathbf{X}_2 - \mathbf{X}_2')\text{diag}(\hat{\beta}) \approx$$

$$\frac{1}{\hat{\sigma}^2_{n_1,n_2}} 2\,\text{diag}(\hat{\beta})\,\mathbf{S}_2\,\text{diag}(\hat{\beta}) =$$

$$\frac{1}{\hat{\sigma}^2_{n_1,n_2}} 2\,\text{diag}(\hat{\beta})\,\text{diag}(S_1,\ldots,S_p)\,\mathbf{R}\,\text{diag}(S_1,\ldots,S_p)\,\text{diag}(\hat{\beta}),$$

where S_j^2 is the sample variance (computed dividing by n_2) of $\mathbf{x}_j$, and $\mathbf{R}$ is the correlation matrix of the test sample $\mathbf{X}_2$. The *"approximately equal to"* sign ($\approx$) indicates that $\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2'$ is a matrix with elements approximately equal to 0, because 0 is their expected value under random permutations. We conclude that the correlation structure of $\tilde{\mathbf{V}}$ coincides with that of the sample correlation matrix $\mathbf{R}$, and it has diagonal elements $2\hat{\beta}_j^{\,2}S_j^2/\hat{\sigma}^2_{n_1,n_2} = \text{RV}_{\text{rp}}(X_j)$.

We have found analogous expressions for $\mathbf{V}$ and $\tilde{\mathbf{V}}$ that allow us to identify the main differences between both relevance matrices. First, $\mathbf{V}$ is related with partial correlations, while $\tilde{\mathbf{V}}$ is associated with standard correlations. Second, the expression of $\mathbf{V}$ includes the estimated residual variances in the regressions of each variable over the rest, while the usual sample variances appear in the expression of $\tilde{\mathbf{V}}$. These findings suggest that the eigen-structure of $\tilde{\mathbf{V}}$ will probably be related with the principal component analysis of the test sample explanatory matrix $\mathbf{X}_2$, but hopefully new knowledge can be found when analyzing the eigen-structure of $\mathbf{V}$.

Table 5 Rent housing prices: Standard output of the linear model.

```
## lm(formula = log.price ~ ., data = rhBM.price[Itr, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73746 -0.17693 -0.02142  0.15657  1.46787
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.8173767  0.0346767 110.085  < 2e-16 ***
## Barcelona        0.1136419  0.0052738  21.548  < 2e-16 ***
## categ.distr      0.1180689  0.0034057  34.668  < 2e-16 ***
## type.chalet     -0.0936153  0.0201548  -4.645 3.44e-06 ***
## type.duplex     -0.0298433  0.0152426  -1.958 0.050267 .
## type.penthouse   0.0519714  0.0101414   5.125 3.03e-07 ***
## type.studio     -0.0959335  0.0137110  -6.997 2.76e-12 ***
## floor            0.0129268  0.0009929  13.020  < 2e-16 ***
## hasLift          0.0426585  0.0119147   3.580 0.000345 ***
## floorLift       -0.0045339  0.0044637  -1.016 0.309772
## log.size         0.6203055  0.0090442  68.586  < 2e-16 ***
## exterior        -0.0325094  0.0068722  -4.731 2.27e-06 ***
## rooms           -0.0504532  0.0033378 -15.116  < 2e-16 ***
## bathrooms        0.1442336  0.0046214  31.210  < 2e-16 ***
## hasParkingSpace -0.0016900  0.0129244  -0.131 0.895968
## ParkingInPrice  -0.0571662  0.0138070  -4.140 3.49e-05 ***
## log_activation   0.0397800  0.0018478  21.528  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2647 on 11519 degrees of freedom
## Multiple R-squared:  0.7621,Adjusted R-squared:  0.7618
## F-statistic:  2306 on 16 and 11519 DF,  p-value: < 2.2e-16
```

Supplement C: Linear and additive models for the real data example (rent housing prices).

Linear model

Table 5 shows the standard output of the fitted linear model. Thirteen variables and the intercept are significant at level 0.001. There are 7 variables with t-value larger than 10 in absolute value: `Barcelona`, `categ.distr`, `floor`, `log.size` (this one being the most significant variable), `rooms`, `bathrooms`, and `log_activation`. The adjusted coefficient of determination R^2 is 0.7618.

The relevance by ghost variables results are shown in Figure 9. We can see (first row, first column plot) that the 7 most relevant variables are the seven we cited before (those with largest t-values in absolute value). This is a consequence of the existing relation (Theorem 1) between the relevance by ghost variables and F-values (the squares of t-values) in the linear model. This plot shows that `log.size` is the most relevant variable, followed by `categ.distr`,

`bathrooms`, `log_activation`, and `Barcelona`. The relevance of `rooms` and `floor` is much lower. The second graphic in the first row of Figure 9 compares the values of the variables relevance with the corresponding F-statistics (divided by n_1, the training sample size). It can be seen that, for every explanatory variable, both values are almost equal. Blue dashed lines in these two first graphics indicate the critical value beyond which an observed relevance can not be considered null, at significance level $\alpha = 0.01$. According to Theorem 1, this critical value is computed as

$$F_{1,n-p-1,1-\alpha}/n_1,$$

where $F_{1,n-p-1,1-\alpha}$ is the $(1-\alpha)$ quantile of a $F_{1,n-p-1}$ distribution ($p = 16$ in this example).

Regarding the analysis of the relevance matrix $\mathbf{V}$, only the eigenvectors explaining more than 1% of the total relevance are plotted. The first eigenvector accounts for the 60% of total relevance, and it is associated with the size of houses (mainly with `log.size`, and to a lesser extent with `bathrooms` and `rooms`). The second eigenvector (16% of total relevance) is mostly related to the district price level (`categ.distr`) and least to `bathrooms` and `rooms`. The 3rd, 4th and 5th eigenvectors are combination of the six most relevant variables, with `bathrooms` having the largest weight in the 3rd one, while `log_activation` and `Barcelona` are dominant in the 4th and the 5th. Given that the eigenvalues corresponding to these three eigenvectors (the first and second could be included here as well) have different values, it follows that the six most relevant variables have related effects on the model predictions, and that they hardly admit isolated interpretations. Finally, the eigenvector 6th is related to `floor`, and the eigenvector 7th to `rooms`, `bathrooms`, and `type.studio`.

Comparing the relevance results for the linear model with those for the neural network, it can be said that they present small differences. The most important one is that less variables are considered relevant in the linear model, but the order of their relevance does not change with respect to the results for the neural network.

Additive model

The standard output of the additive model is shown in Table 6. Three variables are not significant at level size 0.001 (one of them, `floorLift`, enters in the model in a non-parametric way). The adjusted R^2 is 0.784 (larger than in the linear model, and lower than in the neural network).

Let us examine the relevance by ghost variables results (Figure 10). Now, in the additive model, there is no theoretical support for a direct relation between the t and F-values shown in Table 6, and the relevance values plotted in the first plot of Figure 10. In fact we can see that this direct relation does not happen in this case (for instance, `log.size` appears in Figure 10 as much more relevant than `categ.distr`, but the F-value of the term `s(categ.distr)` is larger than that of `s(log.size)` in Table 6). The 7 most relevant variables

Table 6 Rent housing prices: Standard output of the additive model.

```
## log.price ~ Barcelona + s(categ.distr, k = 3) + type.chalet +
##   type.duplex + type.penthouse + type.studio + s(floor) + hasLift +
##   s(floorLift, k = 6) + s(log.size) + exterior + s(rooms) +
##   s(bathrooms, k = 6) + hasParkingSpace + ParkingInPrice +
##   s(log_activation)
##
## Parametric coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.280683   0.012121 600.690  < 2e-16 ***
## Barcelona        0.102610   0.005251  19.540  < 2e-16 ***
## type.chalet     -0.103370   0.020701  -4.993 6.02e-07 ***
## type.duplex     -0.051914   0.014772  -3.514 0.000442 ***
## type.penthouse   0.062322   0.009977   6.247 4.35e-10 ***
## type.studio     -0.022915   0.024865  -0.922 0.356773
## hasLift          0.053318   0.012274   4.344 1.41e-05 ***
## exterior        -0.023472   0.006580  -3.567 0.000363 ***
## hasParkingSpace  0.002794   0.012322   0.227 0.820600
## ParkingInPrice  -0.055495   0.013192  -4.207 2.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df       F p-value
## s(categ.distr)    1.996  2.000 682.908  <2e-16 ***
## s(floor)          8.470  8.867  46.036  <2e-16 ***
## s(floorLift)      1.000  1.000   6.589  0.0103 *
## s(log.size)       8.760  8.977 567.263  <2e-16 ***
## s(rooms)          7.275  8.048  48.518  <2e-16 ***
## s(bathrooms)      4.843  4.979 122.147  <2e-16 ***
## s(log_activation) 3.510  4.363 114.468  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.784   Deviance explained = 78.5%
## GCV = 0.063817  Scale est. = 0.063564  n = 11536
```

coincide with the 7 most statistically significant in the additive model, which were also detected as the most relevant in the linear model.

From the study of the relevance matrix $\mathbf{V}$ we observe that its first 3 eigenvectors are very similar to the corresponding ones in the linear model. The first eigenvector is mainly associated with `log.size`, while the second one is mainly related with `categ.distr`. The relationship between `bathrooms` and `rooms` is reflected in eigenvectors 3 and 7. The 4th eigenvector is related with `floor`, the 5th with `log_activation`, and the 6th with `Barcelona`.

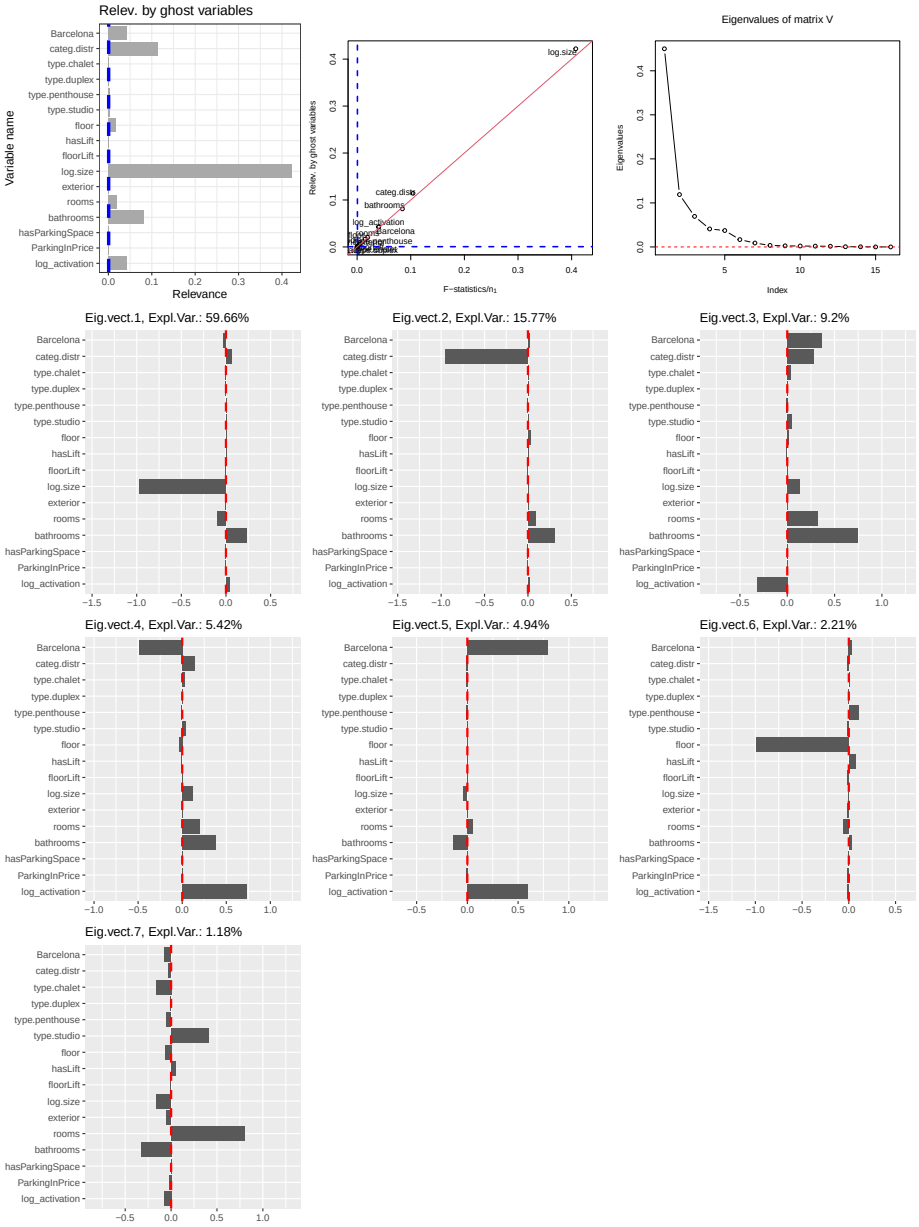

Fig. 9 Rent housing prices: Relevance by ghost variables for the linear model. Only the eigenvectors explaining more than 1% of the total relevance are plotted.

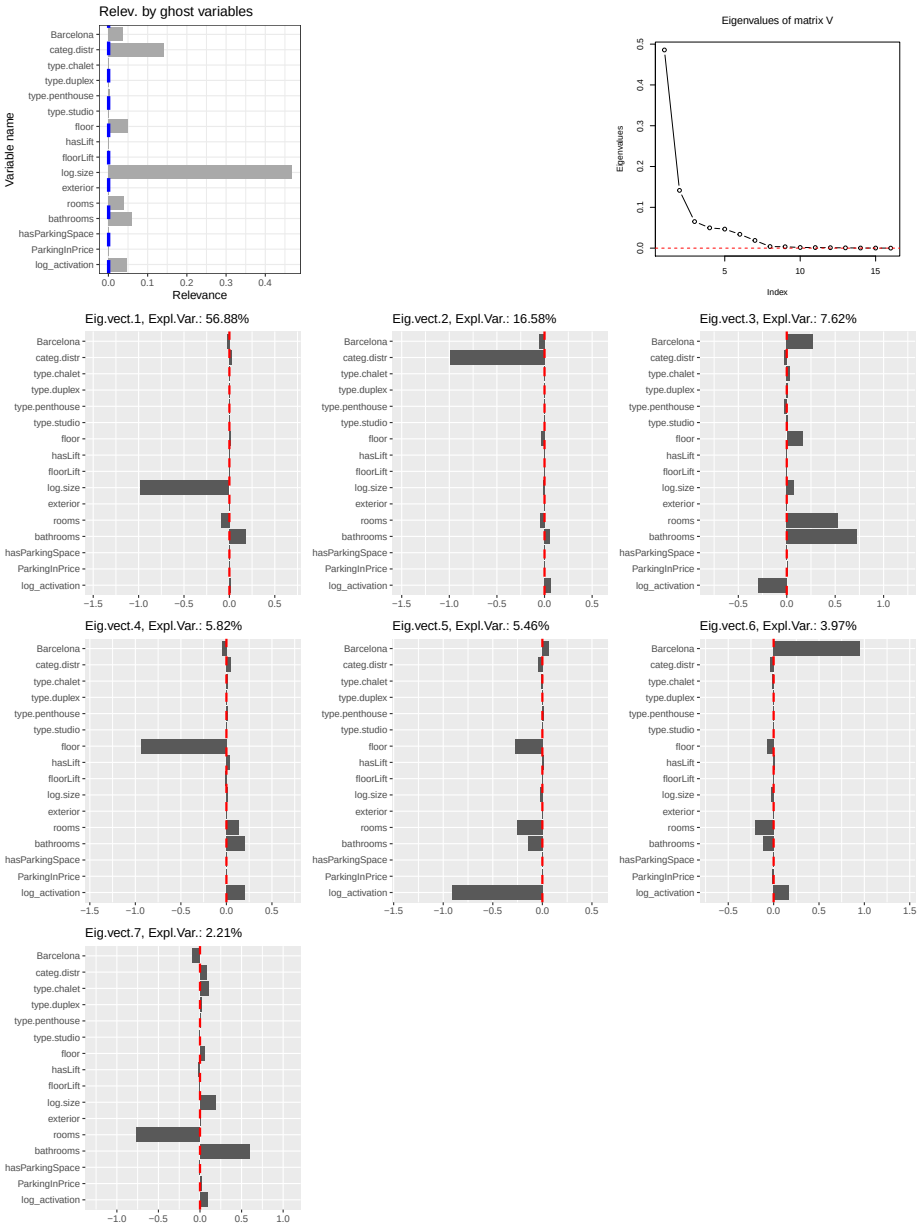

Fig. 10 Rent housing prices: Relevance by ghost variables for the additive model. Only the eigenvectors explaining more than 1% of the total relevance are plotted.