Supplementary material for ”Splitting on categorical predictors in random forests”

Marvin N. Wright and Inke R. König

Predictor type	Outcome type		Nominal predictor method		
		Order (split)	Order (once)	Dummy	Ignore
Digit	Regression	9.6×10^{-01}	**3.9×10^{-21}**	**9.7×10^{-55}**	**8.4×10^{-71}**
	Binary Class.	3.5×10^{-01}	**7.4×10^{-08}**	**6.1×10^{-17}**	**7.4×10^{-39}**
	Multiclass Class.	1.7×10^{-01}	**5.1×10^{-37}**	**6.8×10^{-77}**	**4.1×10^{-80}**
	Survival	**5.6×10^{-16}**	**4.2×10^{-02}**	**3.1×10^{-23}**	**2.9×10^{-08}**
SNP	Regression	6.8×10^{-01}	**2.2×10^{-17}**	**6.3×10^{-24}**	**4.1×10^{-40}**
	Binary Class.	1.2×10^{-01}	9.7×10^{-01}	3.8×10^{-01}	**1.3×10^{-06}**
	Multiclass Class.	9.9×10^{-01}	**9.8×10^{-15}**	**4.9×10^{-19}**	**3.6×10^{-40}**
	Survival	4.0×10^{-01}	3.5×10^{-01}	**7.2×10^{-03}**	**1.4×10^{-02}**
Group	Regression	5.1×10^{-01}	4.1×10^{-01}	**1.1×10^{-07}**	**4.0×10^{-23}**
	Binary Class.	6.6×10^{-01}	7.7×10^{-01}	6.5×10^{-01}	**4.9×10^{-06}**
	Multiclass Class.	1.6×10^{-01}	6.6×10^{-02}	**4.3×10^{-33}**	**1.3×10^{-29}**
	Survival	1.8×10^{-01}	**9.8×10^{-04}**	**1.1×10^{-28}**	**2.4×10^{-02}**

Table S1: p-values of simulations for $n = 100$. Each p-value corresponds to a two-sided paired t-test, compared to the Partition method. Significant results at $\alpha = 0.05$ are highlighted in bold.

Predictor type	Outcome type		Nominal predictor method		
		Order (split)	Order (once)	Dummy	Ignore
Digit	Regression	2.1×10^{-01}	**1.8×10^{-09}**	**6.0×10^{-20}**	**1.5×10^{-48}**
	Binary Class.	9.3×10^{-01}	3.8×10^{-01}	**3.0×10^{-09}**	**6.7×10^{-21}**
	Multiclass Class.	4.1×10^{-01}	**1.3×10^{-43}**	**1.9×10^{-34}**	**2.5×10^{-55}**
	Survival	**2.0×10^{-07}**	1.5×10^{-01}	**9.1×10^{-13}**	**5.7×10^{-04}**
SNP	Regression	9.4×10^{-01}	**1.6×10^{-09}**	**2.3×10^{-13}**	**3.0×10^{-21}**
	Binary Class.	8.0×10^{-01}	7.0×10^{-01}	7.1×10^{-01}	1.5×10^{-01}
	Multiclass Class.	8.5×10^{-01}	**1.1×10^{-17}**	**2.6×10^{-23}**	**6.8×10^{-16}**
	Survival	1.7×10^{-01}	3.8×10^{-01}	3.8×10^{-01}	**3.2×10^{-02}**
Group	Regression	2.4×10^{-01}	3.6×10^{-01}	**9.6×10^{-05}**	**3.4×10^{-10}**
	Binary Class.	1.3×10^{-01}	8.3×10^{-01}	2.7×10^{-01}	**1.8×10^{-04}**
	Multiclass Class.	6.8×10^{-01}	**1.0×10^{-02}**	**5.4×10^{-10}**	**4.8×10^{-10}**
	Survival	**4.7×10^{-02}**	4.6×10^{-01}	**7.3×10^{-16}**	**2.9×10^{-02}**

Table S2: p-values of simulations for $n = 50$. Each p-value corresponds to a two-sided paired t-test, compared to the Partition method. Significant results at $\alpha = 0.05$ are highlighted in bold.

Predictor type	Outcome type	Nominal predictor method			
		Order (split)	Order (once)	Dummy	Ignore
Digit	Regression	9.5×10^{-01}	**2.0×10^{-42}**	**1.8×10^{-74}**	**5.8×10^{-91}**
	Binary Class.	3.6×10^{-01}	**3.2×10^{-20}**	**3.4×10^{-40}**	**5.2×10^{-62}**
	Multiclass Class.	4.5×10^{-01}	**1.8×10^{-19}**	**2.6×10^{-86}**	**5.0×10^{-108}**
	Survival	**4.2×10^{-14}**	7.6×10^{-01}	**1.3×10^{-45}**	**1.5×10^{-06}**
SNP	Regression	4.5×10^{-01}	**7.2×10^{-16}**	**2.4×10^{-21}**	**1.0×10^{-37}**
	Binary Class.	2.9×10^{-01}	8.1×10^{-01}	5.4×10^{-01}	**5.3×10^{-04}**
	Multiclass Class.	1.8×10^{-01}	**4.9×10^{-05}**	**9.0×10^{-08}**	**2.2×10^{-50}**
	Survival	4.8×10^{-01}	9.9×10^{-01}	**2.6×10^{-07}**	**6.2×10^{-04}**
Group	Regression	5.5×10^{-01}	**2.1×10^{-02}**	**4.5×10^{-11}**	**1.8×10^{-32}**
	Binary Class.	2.3×10^{-01}	**3.1×10^{-04}**	**3.7×10^{-05}**	**1.1×10^{-17}**
	Multiclass Class.	3.3×10^{-01}	**1.4×10^{-17}**	**1.4×10^{-62}**	**1.3×10^{-50}**
	Survival	**6.5×10^{-05}**	**7.4×10^{-04}**	**4.9×10^{-29}**	**3.0×10^{-06}**

Table S3: p-values of simulations for $n = 200$. Each p-value corresponds to a two-sided paired t-test, compared to the Partition method. Significant results at $\alpha = 0.05$ are highlighted in bold.

Dataset	Nominal predictor method			
	Order (split)	Order (once)	Dummy	Ignore
Tic-tac-toe	9.5×10^{-01}	8.1×10^{-01}	9.4×10^{-01}	5.1×10^{-02}
Splice	8.5×10^{-01}	7.8×10^{-01}	1.8×10^{-01}	**1.5×10^{-02}**
MPG	NA*	7.8×10^{-01}	9.3×10^{-02}	**4.4×10^{-02}**
Servo	9.5×10^{-01}	8.1×10^{-01}	**3.2×10^{-02}**	**2.5×10^{-02}**
RA SNPs	9.0×10^{-01}	8.6×10^{-01}	8.3×10^{-01}	2.9×10^{-01}
AIDS	8.8×10^{-01}	8.2×10^{-01}	5.6×10^{-01}	7.3×10^{-01}

Table S4: p-values of real data analyses. Each p-value corresponds to a two-sided corrected paired t-test (Nadeau and Bengio, 2003), compared to the Partition method. Significant results at $\alpha = 0.05$ are highlighted in bold. *No results could be obtained for the Partition method on the MPG dataset because the category limit was reached. The other methods are compared to the Order (split) method instead.

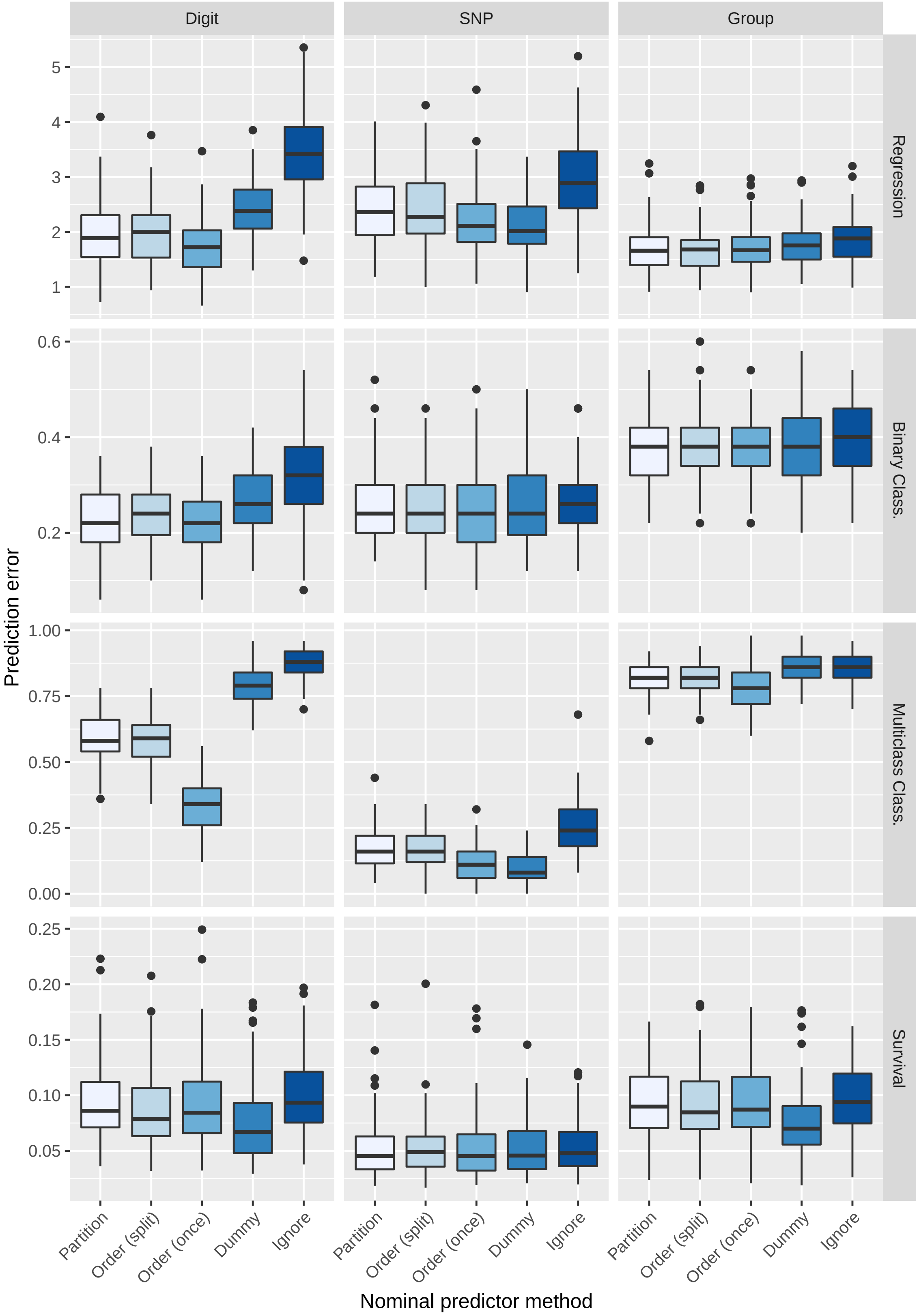

Figure S1: Simulation results for $n = 50$. The vertical panels correspond to the predictor type (Digit, SNP or Group), the horizontal panels to the outcome type (regression, binary classification, multiclass classification or survival). Each boxplot represents the prediction error for one method to handle nominal predictor variables, applied to one simulation scenario. The prediction error is measured by the mean squared error for regression, by the proportion of misclassifications for classification and by the integrated Brier score for survival.

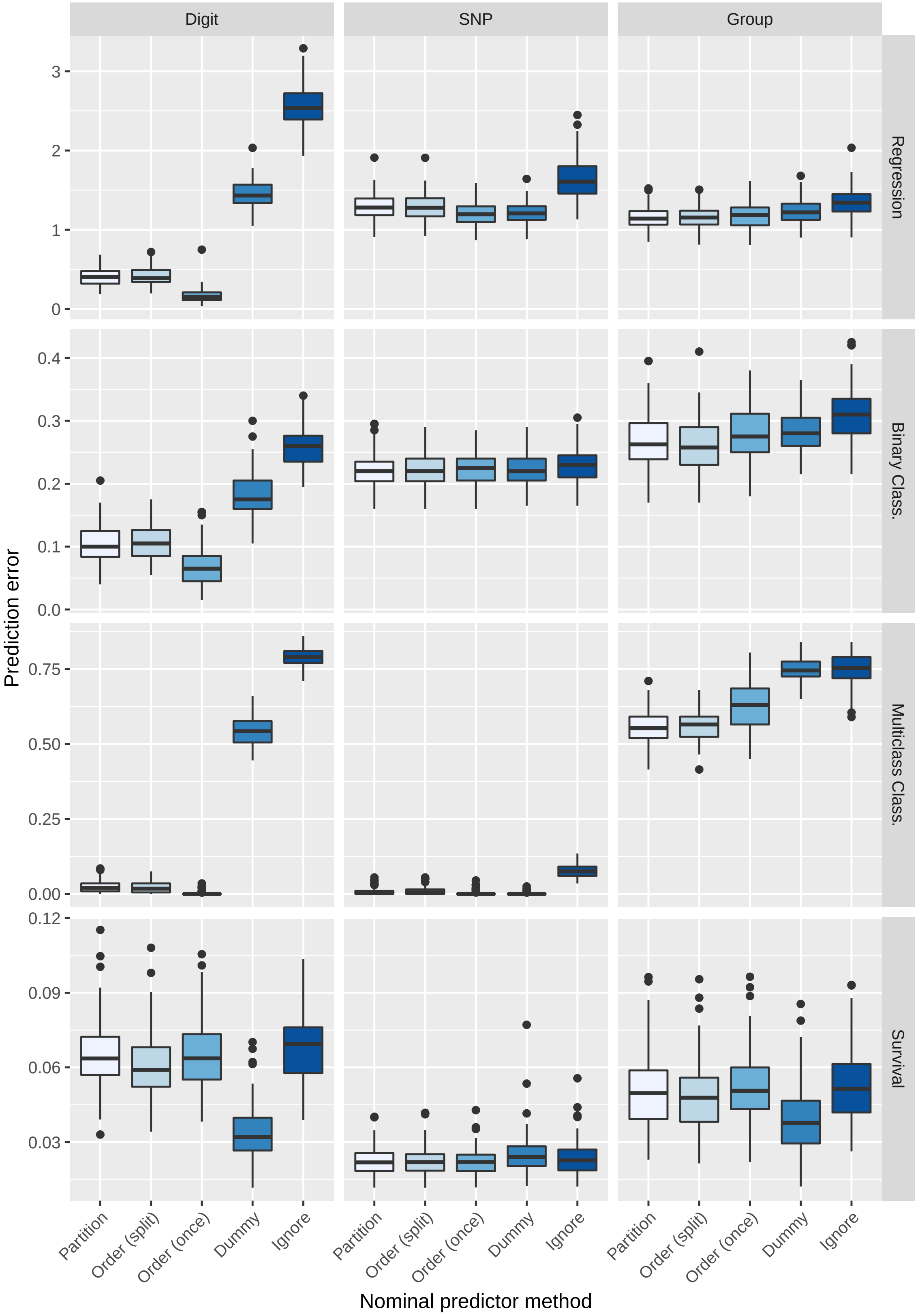

Figure S2: Simulation results for $n = 200$. The vertical panels correspond to the predictor type (Digit, SNP or Group), the horizontal panels to the outcome type (regression, binary classification, multiclass classification or survival). Each boxplot represents the prediction error for one method to handle nominal predictor variables, applied to one simulation scenario. The prediction error is measured by the mean squared error for regression, by the proportion of misclassifications for classification and by the integrated Brier score for survival.

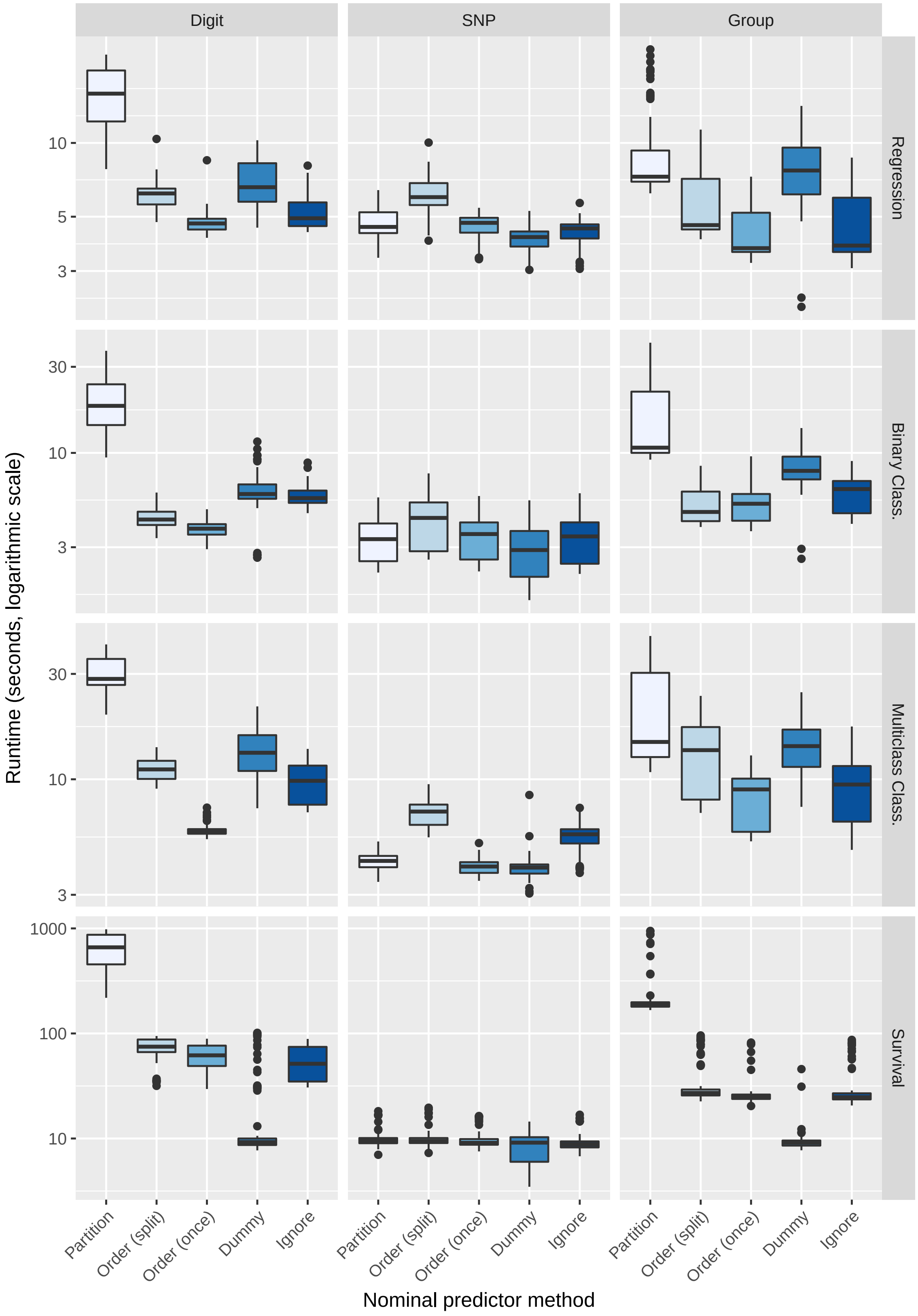

Figure S3: Runtimes of simulations. The vertical panels correspond to the predictor type (Digit, SNP or Group), the horizontal panels to the outcome type (regression, binary classification, multiclass classification or survival). Each boxplot represents the runtime (in seconds) for one method to handle nominal predictor variables, applied to one simulation scenario. The vertical axis is displayed on a logarithmic scale.

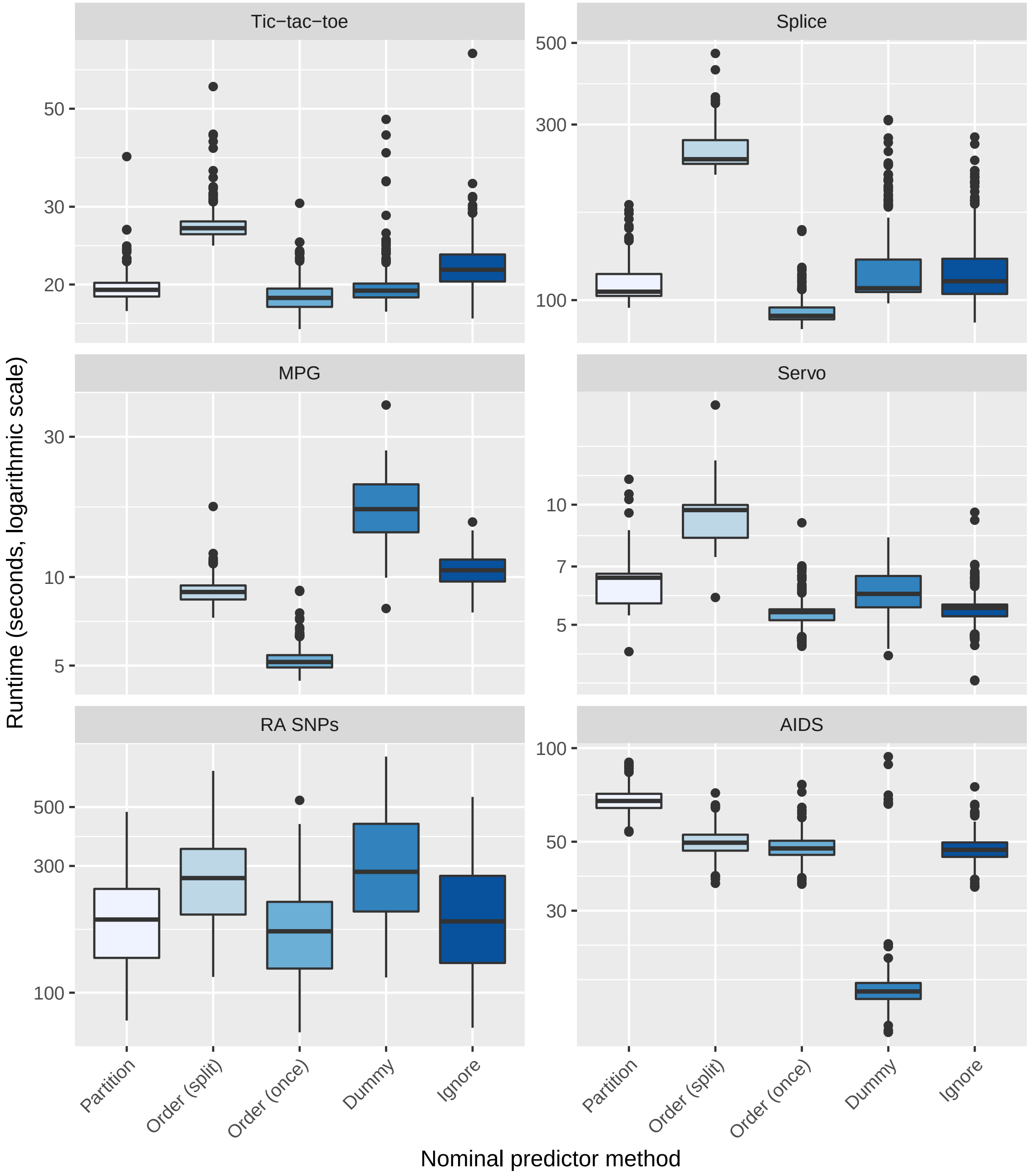

Figure S4: Runtimes of real data analyses. The panels correspond to the datasets. Each boxplot represents the runtime (in seconds) of one method to handle nominal predictor variables, applied to one dataset. No results could be obtained for the Partition method on the MPG dataset because the category limit was reached. The vertical axis is displayed on a logarithmic scale.