Received: 20 March 2022 | Revised: 27 September 2022 | Accepted: 30 October 2022

DOI: 10.1002/cem.3455

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Automatic hierarchical model builder

Lorenzo Marchi[1] | Ivan Krylov[2] | Robert T. Roginski[3] | Barry Wise[3] | Francesca Di Donato[4] | Sonia Nieto-Ortega[5] | José Francielson Q. Pereira[6] | Rasmus Bro[7]

[1]Department of Chemical and Geological Sciences, University of Modena and Reggio Emila, Modena, Italy

[2]Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia

[3]Eigenvector Research, Inc., Manson, Wenatchee, Washington, USA

[4]Department of Physical and Chemical Sciences, University of L'Aquila, L'Aquila, Italy

[5]AZTI, Food Research, Basque Research and Technology Alliance (BRTA), Parque Tecnológico de Bizkaia, Derio, Spain

[6]Department of fundamental chemistry, Federal University of Pernambucano (UFPE), Recife, Brazil

[7]Department of Food Science, University of Copenhagen, Copenhagen, Denmark

Correspondence
Rasmus Bro, Department of Food Science, University of Copenhagen, Copenhagen, Denmark.
Email: rb@life.ku.dk

Funding information
Russian Foundation for Basic Research, Grant/Award Number: 20-33-90280

Abstract

When building classification models of complex systems with many classes, the traditional chemometric approaches such as discriminant analysis or soft independent modeling of class analogy often fail. Some people resort to advanced deep neural network, but this is only an option if there is access to very many samples. Another alternative often used is to build hierarchical models where subclasses are sort of peeled off one or a few at a time. Such approaches often outperform classical classification as well as deep neural network on small multi-class problems. The downside though is that it is very cumbersome to build such hierarchies of models. It requires substantial work of a skilled person. In this paper, we develop a fully automated approach for building hierarchical models and test the performance on a number of classification problems.

KEYWORDS

automation, classification, hierarchical

1 | INTRODUCTION

Building chemometric classification models has often been done with partial least squares regression discriminant analysis, partial least squares discriminant analysis (PLS-DA), soft independent modeling of class analogy, SIMCA,[1,2] or similar methods. These methods work very well even with highly collinear data, but it is commonly known that they do not perform well when there are very many classes to separate. An example of this is presented by Singh et al.[3] where 35 types of barley are analyzed using near-infrared hyperspectral imaging with the aim to discriminate varieties. The study showed that PLS-DA decreases abruptly in performance as the number of varieties grows. In particular, the sensitivity decreases from 100% to 49% when the number of varieties passed from two to 35. If there are many samples available, typically in the order of way above thousands, then deep learning may be a viable alternative, but when this is not the case, alternatives must be found. Several approaches have been suggested for handling the so-called many class classification problem. Important methods that can be mentioned are decision trees, k-nearest neighbor, and support vector machines.[4] Also, methods such as *one versus all* and *one versus one* are frequently used.[5] However, the most common approach in chemometrics is to develop hierarchical classification models typically but not exclusively based on PLS-DA models.[6,7]

It may require days of work to set up a hierarchical classification model, and typically, a lot of trial and error is needed for finding a suitable way to divide up the classification problem. The number of possible hierarchical models describing n_0 classes can be calculated as the number of possible semilabeled trees with n_0 leaves and no vertices of out-degree one[8]:

$$M(n_0) = \sum_{n=n_0}^{2n_0-2} \sum_{\{n_i\}} \frac{n!}{\prod_{i=1}^{n_0} n_i! i!^{n_i}}$$

$$\text{for all } \{n_i \in \mathbb{N} \cup \{0\}\} \text{ subject to } n_1 = 0, n = \sum\nolimits_{i=0}^{n_0} n_i, n_0 = 1 + \sum\nolimits_{i=2}^{n_0} (i-1) n_i.$$

This number grows faster than $O(\exp(n_0))$, which makes exhaustive evaluation of all possible models infeasible, and also means that only adequately skilled persons are able to benefit from the advantages of such models. In this paper, we seek to develop a fully automated hierarchical classification model.

2 | THEORY

We suggest an automated approach for building hierarchical models. The approach is based on a crucial observation. It is often difficult to qualify which groups can best be separated especially when classes are not well separated. On the other hand, it is often easy to ascertain which models are definitely not possible to separate. Rather than trying to find good candidate classes to separate, we will instead take a "negative" approach and look for classes that we can definitely not separate well. Such classes will initially be merged into a pseudo-class, and we will then investigate if the classification problem is now easier to handle.

In short, our algorithm proceeds as follows assuming we have C classes, and for each class, we have a dataset $\mathbf{X}_c$ and assuming that one global model is not already giving perfect classifications.

3 | AUTOMATIC HIERARCHICAL CLASSIFICATION MODEL BUILDER (AHIMBU)

1. Run binary classification models of each class versus each class.
2. Select the two classes that have the lowest success rate (classification error) and merge them into one class. The minimum cross-validated nonerror rate is used to estimate the classification error.
3. If all classes can now be perfectly separated in one total classification model *or* if there are only two classes remaining, then stop and build a classification model to separate these. If not, then go to step one.
4. Repeat the procedure for classified groups that consisted of combined classes.

4 | DATA

The algorithm has been tested on different datasets. They present increasing levels of complexity: from a 3-class problem to an 18-class dataset. Most of the data sets have been described in detail in other publications and will only be shortly described here. In order to test the goodness of the algorithm, each dataset has been split into a calibration and a validation set, in a percentage 75%–25% using the Kennard–Stone algorithm. The classification results will be given in terms of true positive ratio (TPR), the ratio between the number of samples rightly categorized as true and the total number of samples of a class.

All chemometric treatment was performed with Matlab software (MATLAB® R2019b MathWorks) and using PLS_Toolbox 8.9.1 (Eigenvector Research Inc., USA). The AHIMBU algorithm can be downloaded together with the data but is also integrated in PLS_Toolbox from version 9.2 and onwards. The publicly available data and the algorithm can be found at https://sid.erda.dk/share_redirect/eliS6vjC43 [Aug 29, 2022]. The following data sets can be found in this location: Sugar, Chickpeas, Saffron, Chimiométrie, and Blood. Furthermore, the White Fish data set is available for download as specified in the description of this data set.

4.1 | Sugar

This dataset[9] contains 41 ordinary white sugar samples from three different sugar factories (coded by C, D, F). Ten different quality parameters were analyzed: ash content, color, turbidity, two types of grain sizes, SO_2, invert sugar, flocculation, residue, and amino-nitrogen. They can be used as chemical fingerprints to identify the factory site origin. Autoscaling was used as preprocessing, and the target is to classify which factory a sample comes from.

4.2 | Chickpeas

This dataset contains 70 samples of chickpeas (*Cicer arietinum* L.) harvested in the Italian territories of Cicerale (Campania), Valentano (Lazio), and Navelli (Abruzzo) in 2019, and the task is to see if we can classify which region a sample is coming from.[10] The samples were characterized by determination of the content of 10 elements (Ca, K, P, Mg, Mo, Cu, Fe, Mn, Zn, and Sr) with inductively coupled plasma-optical emission spectrometry (ICP-OES). The data set was preprocessed using mean centering.

4.3 | Saffron

Unpublished data, courtesy of Francesca Di Donato, PhD student. Three hundred and fifty-three saffron samples coming from the main production areas of Italy (Abruzzo, Campania, Friuli-Venezia Giulia, Sardinia, Sicily, Umbria, Emilia-Romagna, Tuscany, Basilicata, Liguria, Veneto, Apulia, Latium, and Molise) and belonging to six production years (from 2015 to 2020) were directly provided by the producers of the related Consortia. The saffron stigmas were gently ground in a mortar, and the multispectral imaging of the ground sample was performed using a VideometerLab Instrument (https://videometer.com, Accessed March 7, 2022) measuring 18 specific wavelengths from ultraviolet (430, 450, 470, 505, 565, 590, 630, 645, 660, and 700 nm) to near-infrared[11](850, 870, 890, 910, 920, 940, 950, and 970 nm). The mean spectrum was calculated on the region of interest of each multispectral image. The preprocessing used was mean centering. The classification task is to classify which of the six production years a sample is coming from.

4.4 | Potatoes

This dataset contains information about eight different potatoes species (*Solanum tuberosum* L.): three of them grown in the area of Majella National Park (Abruzzo, Italy) and five commercial varieties cultivated in the same area.[11] A total of 279 attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectra (in the range of 4000–940 cm^{-1}) was used to acquire a fingerprint of the tuber flesh composition. Preprocessing was SNV followed by mean centering, and the classification problem was to predict which of the eight production sites a sample came from.

4.5 | Chimiométrie

This dataset was originally used as part of a chemometric challenge (Data challenge of the Chimiométrie 2018 conference). It contains NIR absorbance data of about 10 different kinds of animal feed and additives. The provided 3903 spectra are raw absorbances directly extracted from several standardized FOSS instruments.

The data were preprocessed with mean centering [chemom2018.sciencesconf.org/resource/page/id/5.html, Mar 5, 2022].

4.6 | Blood

This dataset contains different bloodstain (human and animal) and common false positive (lipstick, soy sauce, jam, ketchup, pepper sauce, balsamic vinegar, and red wine). Those samples were deposited on 10 different fabrics of two types (five synthetic and five cotton-based), dried for 3 days, and hyperspectral imagens were acquired in a spectral range of 928 to 2524 nm.[7] Noise and scattering effects were minimized by reducing the working range to 1187–2265 nm. The preprocessing adopted was smoothing through a Savitzky–Golay filter (11-point window widths, second-order polynomial), SNV, generalized least squared weighting (GLSW) where the source of clutter is the x-block classes which is used to suppress the effect of the differences between the matrices of different classes and mean centering. The class target is whether the sample contains human or animal blood or whether it is a "common false positive".

4.7 | White fish

This dataset contains information of seven species of white fish: turbot (*Psetta maxima*), panga (*Pangasius hypophthalmus*), alaskan pollock (*Theragra chalcograma*), tilapia (*Oreochromis Niloticus*), sole (*Solea solea*, wild and *Solea senegalensis*, farmed), seabass (*Dicentrarchus labrax*), and cod (*Gadus morhua*), collected during 3 years. For each sample, the NIR spectra (from 900 to 1650 nm) were collected with a hand-held device (MicroNIR OnSite from Viavi) in fresh, frozen, and thawed state. The data set was mean centered. The class target is a combination of the seven species in the three different states, for a total of 18 classes (three species are not present in the fresh state). The data can be downloaded at https://www.azti.es/en/withefish-database/ (June 2, 2022).

5 | RESULTS

Rather than providing details on all data sets, we will show some example results and then present an overview of the performance on all data sets. We will start with a simple problem of three classes. We have quality parameters of sugar samples produced at three different sugar factories. As can be seen from a principal component analysis (PCA) score plot (Figure 1), one class is easily separated but two classes overlap quite a lot.

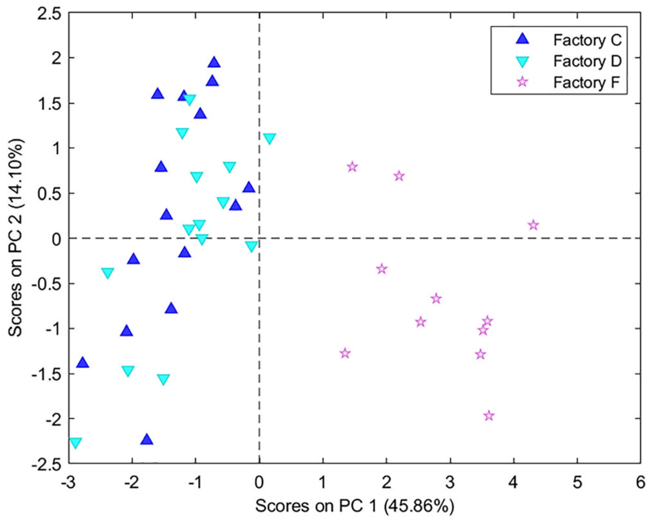

FIGURE 1 A simple three-class problem

Hence, according to our suggested algorithm, we will initially merge factory C and D into on class and then make a classification model separating those two from factory F. Subsequently, we will make a classification model separating C from D. The resulting hierarchical model as derived using AHIMBU is shown in Figure 2. In the supporting information, we have provided a more detailed description of the process.

In this case, the dataset is very simple, yet a PLS-DA will give back an average of 85% as TPR (or sensitivity) when tested with a validation set. The application of the hierarchical model allows us to achieve a perfect classification (100% as TPR).

Moving up to a more challenging classification problem, we can see a score plot of the Chimiométrie dataset in Figure 3. We can see that every class is overlapped with at least one other class.

A classical classification model such as PLS-DA fails when trying to build one global model to separate all classes. Yet, it remains a challenge to manually set up a hierarchical model and it is also quite time-consuming and frustrating. With the AHIMBU approach though, we automatically get a hierarchical model as shown in Figure 4.

This hierarchical model is made by nine different PLS-DA models, represented as nodes or rules. In each node, the class or classes with the lowest misclassification error are peeled off and separated from the others.

As shown in Table 1, with using one global PLS-DA model, a good classification can be achieved. However, in a few classes, the TPR is far from satisfying (e.g., for grass silage and milk powder and whey). Thus, there is a significant

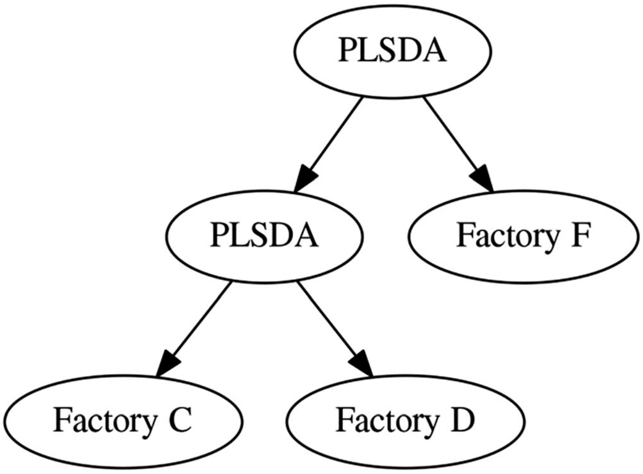

FIGURE 2 The hierarchical tree obtained with our automatic hierarchical classification model builder (AHIMBU) algorithm

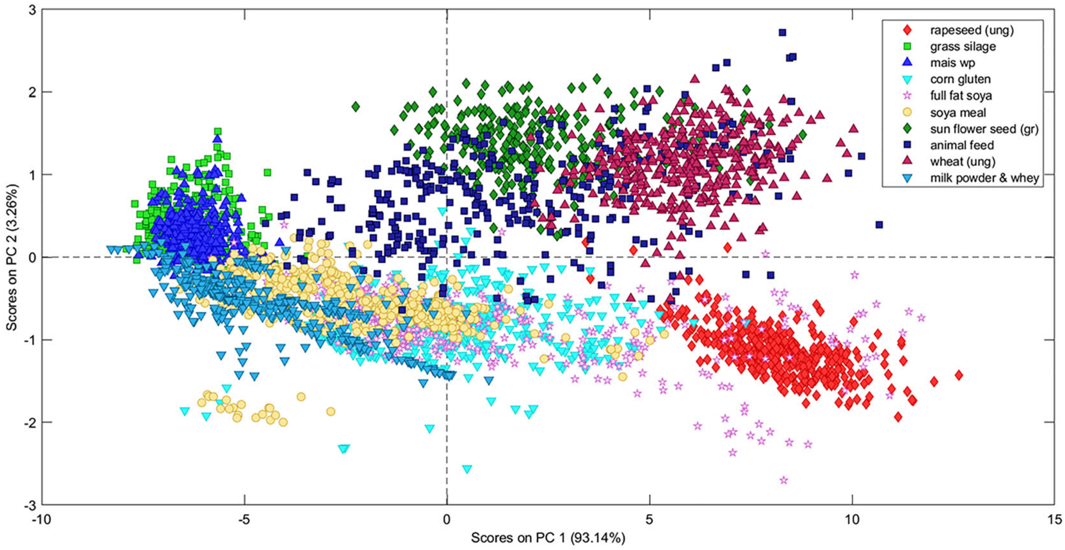

FIGURE 3 Principal component analysis (PCA) score plot of a data set of Chimiométrie preprocessed with mean centering

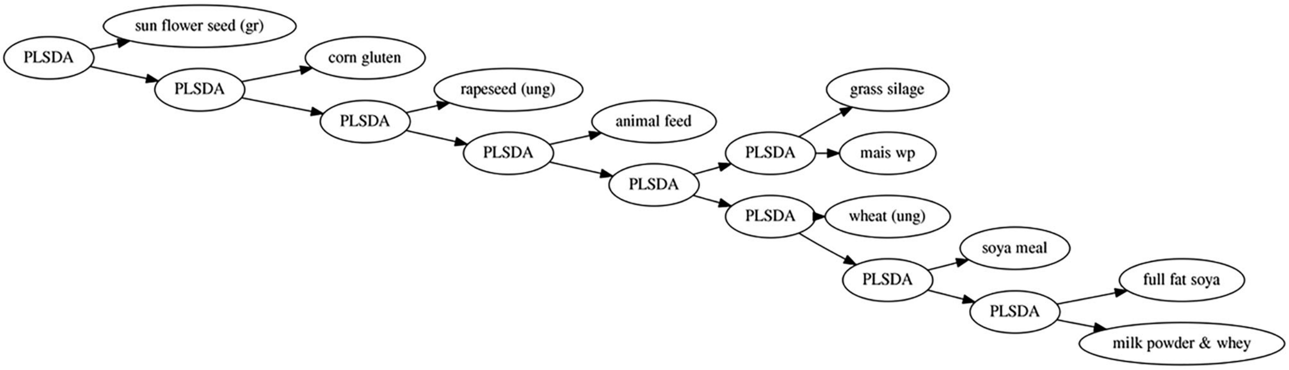

FIGURE 4 The hierarchical model obtained from the Chimiométrie dataset

TABLE 1 Classification results (TPR) comparing one global PLS-DA model with a hierarchical approach

		PLS-DA	AHIMBU	Manual
Rapeseed (ung)	1	99%	98%	98%
Grass silage	2	44%	98%	98%
mais wp	3	62%	100%	100%
Corn gluten	4	75%	100%	100%
Full fat soya	5	76%	96%	96%
Soya meal	6	91%	98%	98%
Sunflower seed (gr)	7	100%	100%	100%
Animal feed	8	86%	98%	98%
Wheat (ung)	9	100%	100%	100%
Milk powder and whey	10	55%	70%	70%
	Tot	79%	96%	96%

Note: Results are from applying the developed models on a test set.
Abbreviations: AHIMBU, automatic hierarchical classification model builder; TPR, true positive ratio.

improvement in the sensitivity after modeling the data with AHIMBU. The TPR goes from an average of 79% to an average of 96%, and hence, the classification is almost perfect. Every class is predicted with at least a better TPR. Moreover, this process of getting the hierarchical model is very fast taking minutes which is orders of magnitude faster than a manual approach.

It is worth noticing that at the fifth node in Figure 4, AHIMBU peeled off two classes at a time instead of one as at the other nodes. Taking a glimpse at the PCA score plot in Figure 3, it can be seen that those two classes, namely, “grass silage” and “mais wp,” are almost completely overlapped. Hence, in the binary classification of the AHIMBU workflow, those classes returned the highest error, and the algorithm put them together in a new combined class, as explained in the algorithm. Moreover, this new class is now easily separable from all the other classes.

The results for all data sets are shown in Table 2. The column Manual shows the classification results for a manually derived hierarchical classification model. For the simpler models that approach is often the same as the automatic models but for more complex models, there can be differences.

Comparing the performance of the classical PLS-DA with the hierarchical model (both the automatic and the manual), we can see that we always obtain worse or the same results using just one PLS-DA model. The only time in which we obtain the same result as the hierarchical model is in the Chickpeas dataset. In that case, we obtain a perfect separation (100% in sensitivity) because the classification problem is extremely simple. In fact, with just a simple PCA, we are able to clearly separate each class visually.

On the other hand, comparing the performance of the automatic model builder with the manual model builder, we can see that they both give pretty similar, and almost comparable, TPR. From this, we can conclude that a hierarchical model is the right approach when we face a classification problem with many classes. Furthermore, and more

TABLE 2 Classification results (TPR) of the eight different datasets

				Hierarchical model	
Dataset	**No. of classes**	**No. of nodes**	**PLS-DA**	**Automatic**	**Manual**
Chickpeas	3	2	100%	100%	100%
Sugar	3	2	85%	100%	100%
Saffron	6	5	56%	65%	71%
Potatoes	8	7	29%	51%	74%
Chimiométrie	10	9	79%	96%	96%
Blood	11	10	62%	72%	74%
White fish	18	17	54%	74%	74%

Note: The complexity of each increase from the top to the bottom.
Abbreviation: TPR, true positive ratio.

importantly, we can conclude that AHIMBU is a convenient tool, since it is able to return a hierarchical model with performances comparable with a manual approach. For the very complex data sets, a manual approach can be extremely time-consuming.

In fact, for all the complex data sets (Potatoes, Chimiométrie, Blood), we were not able to develop a hierarchical model within a reasonable time. Instead, we took the tree from the AHIMBU model and implemented it manually. The differences in errors lay only in a different number of components chosen in some of the PLS-DA models.

6 | CONCLUSIONS

In this paper, a new tool for automatic classification model building was presented. In classification problems with large number of classes, a PLS-DA model will eventually fail, and a hierarchical model approach is needed. This new approach gives the opportunity to obtain nice results comparable with a manual approach and without substantial work of a skilled data scientist needed. AHIMBU was tested on seven different datasets, with increasing level of complexity. The performances have shown that a classic PLS-DA model performs worse compared with AHIMBU especially with complex data sets. Also, the algorithm returns equal or at least similar performances when compared with a manual approach.

ACKNOWLEDGMENTS

We thank Dr Vincent Baeten, director of the quality and authentication of agricultural products Unit, Walloon Agricultural Research Centre, Gembloux, Belgium, for providing the Chimiométrie dataset.

The contribution of Ivan Krylov to the study was partially funded by Russian Foundation for Basic Research (RFBR), project number 20-33-90280, and Danish Government Scholarship under the Cultural Agreements.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/cem.3455.

DATA AVAILABILITY STATEMENT

Most data sets are available as reported in the paper.

ORCID

Rasmus Bro https://orcid.org/0000-0002-7641-4854

REFERENCES

1. Frank IE, Lanteri S. Classification models: discriminant analysis, SIMCA, CART. *Chemom Intel Lab Syst*. 1989;5(3):247-256. doi:10.1016/0169-7439(89)80052-8

2. Lee LC, Liong C-Y, Jemain AA. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst*. 2018;143(15):3526-3539. doi:10.1039/C8AN00599K
3. Singh T, Garg NM, Iyengar SRS. Nondestructive identification of barley seeds variety using near-infrared hyperspectral imaging coupled with convolutional neural network. *J Food Process Eng*. 2021;44(10):e13821.
4. Mohamed A. Survey on multiclass classification methods. *Neural Netw*. 2005;19:1-9.
5. Mathew RM, Gunasundari R. *A Review on Handling Multiclass Imbalanced Data Classification In Education Domain*. In 2021 *International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. 2021.
6. Ríos-Reina R, Azcarate SM, Camiña J, Callejón RM, Amigo JM. Application of hierarchical classification models and reliability estimation by bootstrapping, for authentication and discrimination of wine vinegars by UV–vis spectroscopy. *Chemom Intel Lab Syst*. 2019;191:42-53. doi:10.1016/j.chemolab.2019.06.001
7. Pereira JFQ, Pimentel MF, Honorato RS, Bro R. Hierarchical method and hyperspectral images for classification of blood stains on colored and printed fabrics. *Chemom Intel Lab Syst*. 2021;210:104253. doi:10.1016/j.chemolab.2021.104253
8. Erdős PL, Székely LA. Applications of antilexicographic order. I. An enumerative theory of trees. *Adv Appl Math*. 1989;10(4):488-496. doi:10.1016/0196-8858(89)90026-2
9. Nørgaard L. *Classification and prediction of quality and process parameters of thick juice and beet sugar by fluorescence spectroscopy and chemometrics*. Zuckerindustrie 1995 [cited 120; 970-981].
10. Di Donato F, Squeo F, Biancolillo A, Rossi L, D'Archivio AA. Characterization of high value Italian chickpeas (Cicer arietinum L.) by means of ICP-OES multi-elemental analysis coupled with chemometrics. *Food Control*. 2022;131:108451. doi:10.1016/j.foodcont.2021.108451
11. Di Donato F, Di Cecco V, Torricelli R, et al. Discrimination of potato (Solanum tuberosum L.) accessions collected in Majella National Park (Abruzzo, Italy) using mid-infrared spectroscopy and chemometrics combined with morphological and molecular analysis. *Appl Sci*. 2020;10(5):1630. doi:10.3390/app10051630

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Marchi L, Krylov I, Roginski RT, et al. Automatic hierarchical model builder. *Journal of Chemometrics*. 2022;36(12):e3455. doi:10.1002/cem.3455