ChemHouse: A research and development centre for chemometrics

Jean-Claude Boulet, Marion Brandolini-Bunlon, Gilles Chaix, Benoit Jaillais, Eric Latrille, Matthieu Lesnoff, Alexandre Mallet, Sílvia Mas Garcia, Maxime Metz, Jean-Michel Roger, et al.

ChemHouse: A research and development centre for chemometrics

JC Boulet[1,2], M Brandolini-Bunlon[2,3], G Chaix[2,4,5], B Jaillais[2,6], E Latrille[2,7], M Lesnoff[2,8], A Mallet[2,7,9], S Mas Garcia[2,9], M Metz[2,9], JM Roger[2,9], V Rossard[2,7], DN Rutledge[2,10,11] and R Servien[2,7]

[1]SPO, PROBE, INRAE, Montpellier SupAgro, University of Montpellier, Montpellier, France
[2]ChemHouse Research Group, Montpellier, France
[3]Université de Clermont Auvergne, INRAE, UNH, MetaboHub Clermont, Clermont-Ferrand, France
[4]CIRAD, AGAP Institut, Montpellier, France
[5]AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France
[6]INRAE-ONIRIS Unit_e Statistiques, Sensométrie, Chimiométrie (Stat SC), Nantes, France
[7]INRAE, Univ Montpellier, LBE, Narbonne, France
[8]SELMET, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France
[9]ITAP, INRAE, Institut Agro, University Montpellier, Montpellier, France
[10]Universit_e Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, Paris, France
[11]National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, Australia

Corresponding author: JM Roger, INRAE, 361 rue JF Breton, BP 5095, Montpellier 34000, France.
Email: jean-michel.roger@inrae.fr

The aim of this article is to present the different activities of ChemHouse and to give an overview of the first years of operation. Detailed information can be found on the ChemHouse website (http://chemproject.org/chemhouse).

A few years ago, some researchers from Montpellier, France, created ChemHouse, a multiinstitute research cluster (INRAE, CIRAD, Irstea, University of Montpellier). This creation was guided by the development of three tools, mainly applied to near infrared spectrometry (NIRS): CheMoocs, a MOOC dedicated to chemometrics for NIRS; ChemFlow, a free and open software tool, allowing anyone to implement the techniques learned in CheMoocs and ChemData, an open database.

ChemHouse aims at ensuring an open and shared scientific animation to encourage national and international collaborations in chemometrics, in particular in the form of hosting researchers, and to allow the collaborative development of own research. ChemHouse also hosts the forges of the three tools: CheMoocs, ChemFlow and ChemData. Today, ChemHouse has 48 members (Cf https://chemproject.org/chemHouse/team).

Every fortnight, ChemHouse members are invited to meet to discuss the operational and research issues of the cluster, without any restrictions. At each session, a member (or an outsider, if invited by a member) leads a scientific seminar around a presentation on a topic of their choice. More than 40 scientific presentations have been held in ChemHouse over the two years: 2019 and 2020. Some specific sessions are organised in the form of collective work on data and processing methodology, with for example participation in scientific conference shootouts. A list of these presentations and their content is available on the ChemHouse website at http://chemproject.org/chemhouse/ressources.

Many ChemHouse seminars have been devoted to topical research issues:

• The calibration of NIRS models is certainly the research question that has attracted the most contributions. The use of NIRS in areas of massive data acquisition, such as precision agriculture, poses problems of model calibration. The adaptation of multivariate calibration to large databases is therefore a current topic, which has been the subject of seven ChemHouse seminars. A first solution consists of performing linear models (PLS) on a neighbourhood of the point to be predicted. This approach, known as Local-PLS, makes it possible to manage the non-linearities inherent in large databases. Three seminars were dedicated to local PLS: weighted PLS, ParSketch-PLS and RoBoost-PLS.[1–3] Some seminars were held on other non-linear methods, such as artificial neural networks.

• Another important topic in ChemHouse is the preprocessing of spectra. Several contributions have been made on orthogonal projections, which can be used to avoid unwanted spectral variations.[4–8] However, completely new methods have also been developed. The Variable Sorting for Normalization (VSN)[9] offers an interesting alternative to the widely used but also much criticised Standard Normal Variate (SNV) method. The combination of several preprocessings via multiblock methods is also a new idea. Two new methods, Sequential preprocessing through ORThogonalization (SPORT)[8] and Parallel pre-processing through orthogonalization (PORTO),[7] have recently been developed in ChemHouse, in collaboration with external partners.

• Another topic of interest for ChemHouse researchers is variable selection. Thus, Successive Orthogonalized Covariate Selection (SO-CovSel),[10] a method for selecting variables in multi-block data sets, was developed. A seminar was also devoted to the selection of variables in functional data.[11]

• In addition, many other seminars were devoted to open questions that did not lead to publication, such as the selection of the dimension of a PLS model, the estimation of prediction uncertainties, the estimation of the Net Analyse Signal (NAS), robust principal component analysis, the statistical properties of standard prediction errors, cross-validation, oblique projections or the boosting/bagging/stacking of models.

• Furthermore, ChemHouse regularly organises workshops on key research issues. These workshops, lasting from half a day to a full day, bring together ChemHouse researchers and external partners. Among others, the topics of local PLS, discrimination, hyperspectral imaging, soil NIRS, Python development, prediction

uncertainties, software versioning and Git, use of the RNIRS package (https://github.com/mlesnoff/rnirs), etc. have been studied and discussed.

ChemHouse has produced papers and seminars dedicated to NIRS applications in the fields of phenotyping, wood quality,[12] waste characterisation,[13] process monitoring[14] or calibration transfer.[15]
ChemHouse is also a place for hosting researchers. In total, 57 weeks of hosting have been carried out since the cluster was created. These stays are of very diverse natures. Doctoral students are hosted for short periods to assist them in processing their data. More experienced researchers are hosted for longer periods to carry out methodological developments, such as SO-CovSel,[10] SPORT[8] or VSN.[9] International partnerships are established with the University of Rome "La Sapienza" (IT), the University of Aquilla (IT), the University of Barcelona (ES), the University of Wageningen (NL), the University of Dublin (IRL), the University of Bilbao (ES), the University of Tarragona (ES) and the Polytechnic University of Madrid (ES). In addition to these direct hosts, ChemHouse has a wider impact due to the privileged interactions between its members and their scientific partners. This is the case, for example, of CIRAD's ChemHouse members who work in partnership in tropical countries and transfer the chemometrics skills and know-how produced and exchanged in ChemHouse, in Africa (Madagascar, Côte d'Ivoire) and South America (Brazil, Argentina). ChemHouse also maintains a private partnership. The company Ondalys is actively involved in the scientific activities and in the organisation of events. The company Pellenc Selective Technologies participates as a donor.
The CheMoocs MOOC (chemproject.org/chemoocs) was developed by a broad collective of French-speaking chemometricians, many of whom have joined ChemHouse to continue its development. Since 2016, this MOOC offers a complete teaching in chemometrics, requiring very few prerequisites. Each year, about 1500 people register for CheMoocs and about a hundred people successfully complete the course.
ChemHouse is also the forge of ChemFlow[16] (chemproject.org/chemflow), a free and open source chemometrics software. Based on a Galaxy-project platform, this software allows any user, after registration, to access a large number of chemometrics tools such as spectral preprocessing, unsupervised analysis, multivariate regression, discriminant analysis, multiblock analysis, variable selection methods. ChemFlow can host code written in different languages (R, Python, Scilab, Octave), which allows the collection of scripts from the ChemHouse community and its partners. ChemFlow can generate processing workflows, making process automation and collaboration between users possible. ChemFlow is also used as an educational tool, even internationally (Cf https://chemproject.org/ressources/trainings). It is also the support of CheMoocs, for the realization of exercises and challenges. ChemFlow training courses are for an audience of 10 to 50 people. These training sessions aim to make ChemFlow users autonomous and to teach good practices.
ChemHouse is part of an open science approach. With this in mind, ChemFlow and ChemData (https://chemproject.org/ChemData) have been developed. This project proposes to share the data of ChemHouse members in a three-sided scheme. The data are described in a data paper,[17] they are the purpose of CheMoocs exercises and they are made available to the community in a dataverse.[18]

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: ChemHouse is supported by the sponsorship of the company PELLENC Selective Technologies (www.pellencst.com), via a SupAgro Foundation project and grants from the INRAe MathNum department. (https://www.inrae.fr/departements/mathnum).

References

1. Lesnoff M, Metz M and Roger JM. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. J Chemometrics 2020; 34: e3209.
2. Metz M, Lesnoff M and Roger JM. Un algorithme big-data pour la local-PLS. France: H_elioSPIR, 2019.
3. Metz M, Abdelghafour F, Roger JM, et al. RoBoost-PLSR: a novel robust PLS regression method inspired from boosting principles. Anal Chim Acta (under review) 2021; 1179: 338823.
4. Herrero-Langreo A, Gorretta N, Beghin A, et al. Orthogonal projection as a spectral pre-treatment method to reduce the interference of polystyrene signal in NIR imaging of agar on petri-dishes. In: 2019 10th workshop on hyperspectral imaging and signal processing: evolution in remote sensing (WHISPERS). Piscataway: IEEE, 2019, pp.1–4.
5. Boulet JC and Sabatier R. Further investigations on the relationship between the OPLS preprocessing and the NAS. Chemometrics Intell Lab Syst 2020; 206: 104159.
6. Ryckewaert M, Gorretta N, Henriot F, et al. Reduction of repeatability error for analysis of variance-Simultaneous Component Analysis (REP-ASCA): application to NIR spectroscopy on coffee sample. Anal Chimica Acta 2020; 1101: 23–31.
7. Mishra P, Roger JM, Marini F, et al. Parallel preprocessing through orthogonalization (PORTO) and its application to near-infrared spectroscopy. Chemometrics Intell Lab Syst 2020; 212: 104190.
8. Roger JM, Biancolillo A and Marini F. Sequential preprocessing through ORThogonalization (SPORT) and its application to near-infrared spectroscopy. Chemometrics Intell Lab Syst 2020; 199: 103975.

9. Rabatel G, Marini F, Walczak B, et al. VSN: variable sorting for normalization. J Chemometrics 2020; 34: e3164.
10. Biancolillo A, Marini F and Roger JM. SO-CovSel: a novel method for variable selection in a multiblock framework. J Chemometrics 2020; 34: e3120.
11. Picheny V, Servien R and Villa-Vialaneix N. Interpretable sparse SIR for functional data. Stat Comput 2019; 29: 255–267.
12. Chaix G, Pires Franco M, Chambi Legoas R, et al. Impact of drought on eucalyptus wood chemistry by near-infrared hyperspectral imaging. Pesq.f/01: bras .. Colombo. 2019; 39: e2019020–13. Special issue. p. 1–768.
13. Mallet A, Charnier C, Latrille E , et al. Unveiling nonlinear water effects in near-infrared spectroscopy: a study on organic wastes during drying using chemometrics. Waste Manage 2021; 122: 36–48.
14. Rey-Bayle M, Bendoula R, Caillol N, et al. Multiangle near-infrared spectroscopy associated with common components and specific weights analysis for in-line monitoring. J Near Infrared Spectrosc 2019; 27: 134–146.
15. Mishra P, Roger JM, Rutledge DN, et al. Two standardfree approaches to correct for external influences on near-infrared spectra to make models widely applicable. Postharvest Biol Technol 2020; 170: 111326.
16. Rossard V, Boulet JC, Gogé F, et al. ChemFlow, chemometrics using Galaxy. In: Galaxy Community Conference– GCC2016, Bloomington, USA, (24 June 2016–29 July 2016).
17. Zgouz A, H_eran D, Barth_es B, et al. Dataset of visible-near infrared handheld and micro-spectrometers – comparison of the prediction accuracy of sugarcane properties. Data Brief 2020; 31: 106013.
18. Chauvergne C, Latrille E, Bonnal L, et al. Dataset of organic sample near infrared spectra acquired on different spectrometers. Data Brief 2020; 32: 106264.