SciMedCentral

Journal of Human Nutrition & Food Science

Mini Review

Variable Selection in Chemometrics for Food and Nutrition: From Predictive Features to Defensible Biomarkers

Christophe BY Cordella[1,2]*, Fabrice Mulotti[1,2], Carl Duchesne[3], and Eric Guilbert[4]

[1]*LARTIC - Laboratoire de Recherche et de Traitement de l'Information Chimiosensorielle, Department Sciences des Aliments, Université Laval, Canada*

[2]*INAF - Institute of Nutrition and Functional Foods, Université Laval, Canada*

[3]*Département de génie chimique, Faculté des sciences et de génie, Université Laval, Canada*

[4]*CRDIG - Centre de recherche en données et intelligence géospatiales, Université Laval, Canada*

***Corresponding author**

Christophe BY Cordella, LARTIC - Laboratoire de Recherche et de Traitement de l'Information Chimiosensorielle, Université Laval, Québec, QC, Canada

Submitted: 03 September 2026

Accepted: 30 October 2026

Published: 31 October 2026

ISSN: 2333-6706

Copyright

© 2026 Cordella CBY, et al.

OPEN ACCESS

Keywords

- Chemometrics
- Variable selection
- Food analysis
- Nutritional metabolomics
- Biomarkers
- Validation

Abstract

High-dimensional analytical data are increasingly used in food science and human nutrition to predict composition, authenticate foods, characterize dietary exposure, and discover candidate biomarkers. Variable selection is often used to simplify these data and improve model performance, but the scientific meaning of a selected variable is frequently overstated. In spectra, chromatograms, hyperspectral images, or metabolomic profiles, a selected coordinate may be one member of a redundant information structure, a proxy for a latent process, or a stable correlate of batch, origin, or acquisition conditions. Its predictive usefulness therefore does not establish chemical specificity or biological causality. This mini-review proposes an evidence-based interpretation of variable selection for food and nutrition chemometrics. We distinguish four common objectives-prediction, analytical or sensor simplification, marker or biomarker discovery, and mechanistic interpretation-and show that each requires a different level of validation. Selection and tuning should remain inside the validation loop, stability should be evaluated beyond exact-coordinate frequency, and candidate food or dietary biomarkers should be supported by independent chemical identification and external reproducibility. Variable selection is most informative when the strength of the scientific claim matches the strength of the evidence.

INTRODUCTION

Food and nutrition research increasingly relies on analytical platforms that generate hundreds to millions of descriptors per sample. Near- and mid-infrared spectroscopy, Raman spectroscopy, nuclear magnetic resonance, chromatography–mass spectrometry, fluorescence excitation–emission matrices, hyperspectral imaging, and metabolomics can all produce data in which the number of measured variables greatly exceeds the number of independent samples. Variable selection is therefore attractive because it can remove uninformative or redundant signals, reduce computation, simplify sensors, and sometimes improve predictive robustness [1,2]. In food analysis, variable-selection methods have been used for calibration, authenticity testing, quality assessment, and classification [1,3]. In human nutrition, related strategies are central to metabolomic studies seeking signatures of food intake, dietary patterns, or metabolic response [4,5]. Yet these applications create the same interpretative difficulty: a variable selected by an algorithm is first a statistical object. It is not automatically a unique chemical marker, a validated biomarker, or evidence of a biological mechanism.

This distinction matters because the intended conclusions are very different. A wavelength can be useful for predicting protein content without identifying one molecular species. A metabolite can discriminate consumers from non-consumers of a food while remaining insufficiently specific for habitual intake. A mass-spectral feature can be reproducible yet reflect sample handling or batch structure. The purpose of this mini-review is therefore not to compare variable-selection algorithms, but to clarify what selected variables can legitimately support in food and nutrition chemometrics.

What does a selected variable represent?

Analytical variables are rarely independent. Neighboring wavelengths can describe the same broad absorption band, adjacent retention-time or mass-to-charge coordinates can belong to one chromatographic or mass-spectral feature, and nearby pixels or wavelengths in hyperspectral images can reflect the same physical structure. Two selection methods may consequently retain different coordinates while exploiting essentially the same chemical information [2-6]. Exact coordinate agreement is therefore not always the most meaningful measure of concordance. For food spectroscopy, the scientifically relevant object may be a spectral interval, a peak, or a molecular contribution rather than one wavelength. For fluorescence or hyperspectral data, it may be a two-dimensional region rather than a single channel. Multiway approaches such as N-CovSel make this issue explicit by allowing multidimensional features to be selected [7]. The selectable unit should therefore be defined before modelling according to the analytical question.

The same caution is essential in nutritional metabolomics. A selected metabolite may be associated with a food exposure because it originates from the food, from host metabolism, from gut microbial transformation, or from a correlated dietary behavior. Reviews of dietary biomarker research show that many reported candidates remain insufficiently validated across populations and study designs [4,5]. Statistical importance should thus be described using vocabulary proportional to the evidence: “discriminant feature” or “associated metabolite” requires less evidence than “candidate biomarker,” and a validated biomarker requires substantially more.

Prediction, simplification, biomarker discovery, and mechanism are different goals

The central question in variable selection is not only which variables are retained, but why they are being retained. Four common objectives can be distinguished. For prediction, the requirement is reliable generalization to genuinely new samples. A selected subset is acceptable if the complete modelling procedure performs robustly under appropriate validation, even when the retained variables are not chemically unique. For analytical or sensor simplification, predictive performance is necessary but not sufficient. The selected wavelengths or features must also be transferable to the intended instrument, robust to calibration transfer, and compatible with the physical measurement process. A small wavelength subset chosen from a benchtop spectrum may not retain the same value on a portable multispectral device. For marker or biomarker discovery, the evidentiary threshold is higher. Food-authenticity markers should be reproducible across independent batches, origins, seasons, and analytical runs and should be supported by chemical assignment using an appropriate orthogonal technique. In nutrition, proposed biomarkers of food intake require evidence for plausibility, dose and time response, robustness, reliability, stability, analytical performance, and inter-laboratory reproducibility [8]. Discovery from a multivariate model is therefore only the beginning of biomarker validation.

Finally, mechanistic interpretation requires evidence beyond association. A selected variable may covary with an outcome because of confounding or correlated processes. Establishing a mechanism generally requires intervention, experimental manipulation, temporal evidence, or another design capable of testing causality. Predictive importance alone cannot provide this level of inference (Table 1).

Stability and validation must follow the experimental structure

Variable selection is itself part of model fitting and must remain inside the validation loop. Selecting variables once from the complete dataset and then cross-validating only the reduced model allows information from validation samples to influence the feature set, producing optimistic estimates. The entire sequence—preprocessing, selection, hyperparameter tuning, and model fitting—should be repeated within each resampling iteration [9]. This principle is particularly important in spectroscopy and metabolomics, where many candidate variables can be explored. Stability should also be evaluated. A variable may disappear when the calibration set, preprocessing method, or selection depth changes. Stability selection provides a formal framework for resampling-based reproducibility [10], but exact-coordinate frequency can be misleading in strongly correlated analytical data. Low stability at the wavelength level may coexist with high stability at the spectral-region level. Conversely, a highly stable feature

Table 1: Evidence required for common uses of variable selection in food and nutrition chemometrics

Intended use	What the selected variables must demonstrate	Appropriate interpretation
Prediction	Generalization to genuinely new samples after nested selection and tuning	Predictive feature/ subset
Analytical or sensor simplification	Prediction plus robustness to instrument implementation and calibration transfer	Operational wavelength/ feature set
Food marker or dietary biomarker discovery	Stability, external replication, control of confounding, chemical identification; for dietary biomarkers, biological validation criteria	Candidate marker/ biomarker
Mechanistic interpretation	Evidence capable of distinguishing causal or mechanistic explanations from association	Mechanistically supported feature

may be a reproducible artifact. The resampling unit must reflect the population to which the model is intended to generalize. Technical replicates from the same food sample are not independent foods; repeated spectra from the same biological specimen are not independent participants. Likewise, production batch, site, season, acquisition day, instrument, dietary intervention arm, and repeated measurements can create hierarchical dependence. Randomly splitting closely related observations between training and validation sets can inflate both predictive performance and apparent selection stability. Structured confounding is especially dangerous because it can be stable. Baseline drift, scattering, matrix effects, sample preparation, run order, batch effects, or storage conditions may align with the response. Appropriate quality-control samples, randomized acquisition, balanced experimental design, batch-aware validation, and sensitivity analyses across preprocessing strategies are therefore part of the interpretation of variable selection, not merely technical details.

From selected features to defensible food and nutrition biomarkers

Food science and nutrition provide useful examples of why the interpretation of selected variables must be staged. In food authentication, a selected spectral or chromatographic feature may be sufficient for classification but may still represent a family of correlated compounds or a geographical or processing covariate rather than a unique authenticity marker [3-11]. In nutritional metabolomics, metabolic profiling can identify features associated with specific foods or dietary patterns, but independent replication and targeted quantification are needed before those features can be used as objective measures of intake [4-8].

Several subsets may also be equally defensible. LASSO, elastic net, covariance-based selection, wrappers, filters, and evolutionary methods optimize different criteria and may choose different members of a correlated information structure [2-12]. When several subsets perform within the uncertainty of the validation estimate, presenting one "optimal" list can create false chemical specificity. Region-level summaries, co-selection patterns, selection paths, and the performance distribution across plausible subsets provide a more faithful representation of uncertainty. The number of retained variables is likewise a modelling decision rather than an intrinsic property of the data. Thresholds, regularization parameters, the number of selected CovSel features, or wrapper stopping rules should be tuned within validation. Conclusions about chemistry or nutrition should then be checked across plausible selection depths. If five and twenty variables yield essentially equivalent prediction but imply different chemical narratives, the interpretation should emphasize the information common to both solutions rather than the smallest list.

Minimum reporting recommendations

For food and nutrition chemometric studies using variable selection, we recommend reporting at least: [1] the intended purpose of selection; [2] the selectable analytical unit; [3] the complete preprocessing–selection–tuning–validation sequence; [4] predictive uncertainty on independent or appropriately resampled data; [5] stability at both coordinate and information-structure levels; [6] plausible alternative subsets and redundancy; [7] sensitivity to preprocessing and selection depth; and [8] controls for batch effects, artifacts, and confounding.

When a chemical or nutritional biomarker claim is made, the analytical identification procedure and the evidence for external reproducibility should be stated explicitly. For dietary biomarkers, established validation dimensions such as dose response, time response, robustness, reliability, and inter-laboratory reproducibility provide a useful benchmark [8]. These practices do not require every predictive study to demonstrate causality; they require only that the terminology and scientific claim match the evidence actually obtained.

CONCLUSION

Variable selection can make high-dimensional food and nutrition data more useful, but it does not automatically make them more interpretable. A selected wavelength, peak, metabolite, or multidimensional feature may be predictive without being unique, chemically specific, biologically specific, or causal. The meaning of selection depends on the analytical unit, the validation design, the stability of the information structure, and the intended use of the result.

For prediction, honest generalization may be sufficient. For sensor simplification, transferability is additionally required. For food or dietary biomarker discovery, independent replication and chemical confirmation are necessary. For mechanistic claims, stronger experimental evidence is essential. Treating variable selection as an evidentiary step rather than a purely algorithmic operation can improve reproducibility and help transform high-dimensional analytical signatures into defensible information for food science and human nutrition.

CRediT Authorship Contribution Statement

CC: Conceptualization, Investigation, Writing - original draft, Writing - review & editing.

FM: Conceptualization, Investigation, Writing - review & editing.

CD: Conceptualization, Writing - review & editing.

EG: Conceptualization, Writing - review & editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Gomes AA, Azcarate SM, Diniz PHGD, Fernandes DDS, Veras G. Variable selection in the chemometric treatment of food data: A tutorial review. Food Chem. 2022; 370: 131072.
2. Andersen CM, Bro R. Variable selection in regression-a tutorial. J Chemom. 2010; 24: 728-737.
3. Jiménez-Carvelo AM, González-Casado A, Bagur-González MG, Cuadros-Rodríguez L. An overview on the application of chemometrics tools in food authenticity and traceability. Foods. 2022; 11: 3940.
4. Rafiq T, Azab SM, Teo KK, Thabane L, Anand SS, Morrison KM, et al. Nutritional metabolomics and the classification of dietary biomarker candidates: A critical review. Adv Nutr. 2021; 12: 2333-2357.
5. Brennan L. Biomarkers of food intake: current status and future opportunities. Proc Nutr Soc. 2026; 85: 121-125.
6. Westad F, Marini F. Variable selection and redundancy in multivariate regression models. Front Anal Sci. 2022; 2: 897605.
7. Biancolillo A, Roger JM, Marini F. N-CovSel, a new strategy for feature selection in N-way data. Anal Chim Acta. 2022; 1231: 340433.
8. Dragsted LO, Gao Q, Scalbert A, Vergères G, Kolehmainen M, Manach C, et al. Validation of biomarkers of food intake-critical assessment of candidate biomarkers. Genes Nutr. 2018; 13: 14.
9. Ezenarro J, Schorn-García D. How are chemometric models validated? A systematic review of linear regression models for NIRS data in food analysis. J Chemom. 2025; 39: e70036.
10. Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Series B Stat Methodol. 2010; 72: 417-473.
11. Biancolillo A, Marini F, Ruckebusch C, Vitale R. Chemometric strategies for spectroscopy-based food authentication. Appl Sci. 2020; 10: 6544.
12. Sorochan Armstrong MD, de la Mata AP, Harynuk JJ. Review of variable selection methods for discriminant-type problems in chemometrics. Front Anal Sci. 2022; 2: 867938.