Checklist for Multivariate Analysis Best Practice

A concise checklist of the most important issues to consider and steps to follow when applying multivariate analysis methods to your data

Produce results and interpret your data with confidence

Includes tips and advice from world-leading experts

camo.com

Bring data to life

Below is our 9-point checklist with practical tips on **Multivariate Analysis Best Practice** based on over 25 years experience.

☐ **Model iteratively**

You are unlikely to create the final model in your first attempt. Use the power of multivariate analysis to select interesting samples based on the instrumental measurements before you put them through the lab. This will save you a lot of time and money by identifying those samples which are the 'needles in the haystack' i.e. the important ones.

Add important samples to an initial model in order to create the most robust model.

☐ **Visualize before you analyze**

This is a key step. Use various plotting tools to isolate bad situations before you analyze the data, otherwise your analysis will be poor. Depending on the type of data use line- 2D-, 3D- or matrix-plots. If you have multiple responses be sure to visualize the correlation structure among them. On the other hand, by looking at raw data, the human eye can not decide if there is enough information to model your response variables.

Biased models or non-representative models are often a result of poor sampling or poor scanning. Check for this using Principal Component Analysis (PCA) and histograms to check the quality of reference data. You should also look for transcription errors using histograms.

☐ **Check Spectral and Reference data quality**

Both spectral data and reference data must be robust and reliable otherwise the final model won't be reliable. Sampling is critical. You can have excellent lab results but if the spectra are not representative the model is not representative. This also applies to reference data for which an estimate of the uncertainty is crucial. Also be sure that your instrument is actually suited for observing the chemistry or biology you are trying to model.

Bad spectral data can be picked up using PCA. Use Scores and Influence plots to look for extreme/interesting samples. Scores plots check if samples come from one homogenous class or if the expected variability lies in the samples. Look for the presence of clusters. Remember that different sample groups may require the use of more than one model.

☐ **Pre-processing**

Only pre-process data only when you understand its action on the data and why you are doing it. Apply the pre-processing method that matches the measurement principle (transmittance, reflectance etc.) The blind application of pre-processing or over pre-processing just because the options are there can lead to non-robust models.

Start simple, using smoothing of raw data. If purely offset effects exist, use baseline correction or derivatives. If scatter effects dominate use Multiplicative Scatter Correction (MSC) or Standard Normal Variate (SNV) etc

☐ **Validate your models**

Models can be validated using Cross Validation or Test Set Validation. However, don't use random or leave one out (full) cross validation just for the sake of it. Think carefully about the cross validation method you use and why you are using it. Sample grouping is a visual way of assessing the model stability and as a visual Analysis of Variance. Do you see the underlying structure you expect in your system in the various plots of important variables? Is your model based on indirect correlations that may not hold for future samples?

Test set is preferred in all cases and forces you to better understand your data.

There's many ways to cross validate a model. For example, if you're measuring replicates, cross validate by removing one replicate and modeling the other replicates or conceptually validate based on known classes or subgroups of samples within the data set. Wherever possible use a test set to validate your data.

☐ **Build for simplicity**

A simple model is often the best model. Don't add excessive numbers of components, as this will make the model specific to the calibration set but less robust for new samples. The tradeoff is simplicity versus slightly improved accuracy.

As a rule, you should normally opt for simplicity, or parsimony. Variable selection may improve the model but then a test-set is highly recommended for confirmation

☐ **Check for easy fit**

Don't try to fit a square peg in a round hole. Avoid trying to find one out of many combinations of pre-processing that make your models look better. Usually, if a simple model cannot be developed it's likely there is no real relationship or the data are nor suitably sampled for robust model development.

If, after modeling, the fit isn't as good as you would like or there is curvature in the model, re-evaluate the pre-processing and model again. However, don't try to make the model fit if it clearly doesn't Remember parsimony

☐ **Implement your models**

If your objective is to build models for on- or at-line prediction, models that sit on a computer without being used are worthless. Don't be afraid to use the model in real applications. Multivariate methods provide extensive diagnostics that give you an indication of the quality of the prediction, so you are not left in the dark.

Start using the model alongside the reference method in the early stages to gain confidence, then migrate fully over to the model for predictions.

☐ **Check for quality**

Use Inlier and outlier statistics to check if the prediction quality is good. Many outliers indicate the model is not as robust as you thought or the samples collected come from a new population, which can easily be verified with the analysis of projected scores plots.

Multivariate analysis is highly visual in nature, use plots of prediction uncertainty intervals, projected scores and influence plots to make the best, most well informed decisions.

We hope this checklist has been helpful. You can find other useful resources at www.camo.com or contact us directly.
Good luck with your multivariate modeling!

camo.com

Bring data to life

NORWAY

Nedre Vollgate 8,
N-0158
Oslo
Tel: (+47) 223 963 00
Fax: (+47) 223 963 22

USA

One Woodbridge Center
Suite 319, Woodbridge
NJ 07095
Tel: (+1) 732 726 9200
Fax: (+1) 973 556 1229

INDIA

14 & 15, Krishna Reddy
Colony, Domlur Layout
Bangalore - 560 071
Tel: (+91) 80 4125 4242
Fax: (+91) 80 4125 4181

AUSTRALIA

PO Box 97
St Peters
NSW, 2044
Tel: (+61) 4 0888 2007

JAPAN

Shibuya 3-chome Square Bldg 2F
3-5-16 Shibuya Shibuya-ku
Tokyo, 150-0002
Tel: (+81) 3 6868 7669
Fax: (+81) 3 6730 9539