How to determine the amount of filtering

TERMINOLOGY

We use **samples** to denote units such as patients, persons, study participants, animals, plants, cells,...

We use **variables** to denote quantities that have been measured for each sample, such as gene expression levels, miRNA expression levels, protein concentrations, antibody concentrations, methylation levels, answers to questions in a questionnaire, ...

Normally a **data set** is described by a matrix where the columns represent samples and the rows represent variables.

HOW TO DETERMINE THE AMOUNT OF FILTERING

The amount of filtering will be very dependent on your data and your application and it is hence difficult to give a precise quantitative answer on how much you should filter. First of all, you have to take into account if your data has already been filtered or not. In many cases the available datasets are already filtered in some way. At least doubtful measurements (outliers) have usually been taken out.

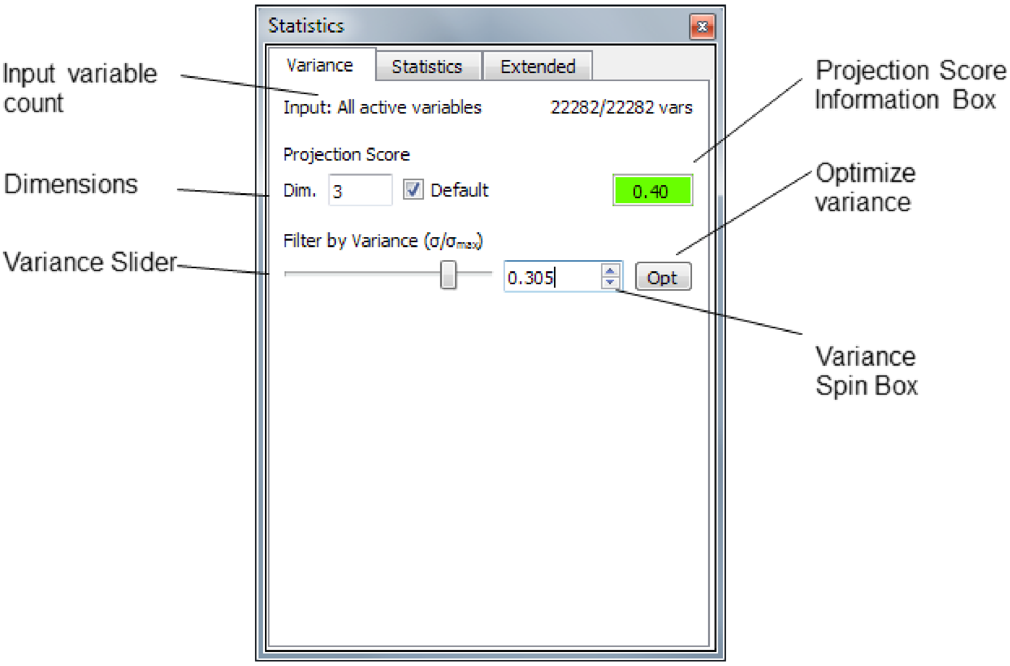
If your dataset has not been pre-filtered, then it is almost always best to start to filter by variance to get rid of some of the effects of disturbing noise. What you regard as noise will differ between data sets and the question you are trying to answer. As an example, if you are looking for genes that are discriminating groups and has gene expression data you can regard the genes that do not vary between the groups as noise since they cannot explain your question. Normally you can take away the variables that have a standard deviation that is less than 2-5 % of the maximal standard deviation. i.e. move the variance slider to somewhere in between 0.02-0.05. If your dataset has already been filtered, then this filtering by variance should only remove a small portion of the variables, whereas if your data set is unfiltered, then it can be acceptable to lose up to 90% or more of your variables in this step.

PROJECTION SCORE

The projection score is a tool that is unique to Qlucore Omics Explorer. It measures the informativeness of a low-dimensional representation obtained by PCA.

The goal of exploratory visualization is to find a representation from which we can extract interpretable and potentially interesting information, that is, one that contains structures and patterns that are likely to be non-random.

By following the evolution of the projection score in real time during variance filtering, the user can easily find the variable subset (and thus implicitly the variance cutoff) giving the most informative representation.



Input: All active variables 22282/22282 vars

Projection Score

Dim. 3 ☑ Default 0.40

Filter by Variance (σ/σ_{max}) 0.305

By monitoring the evolution of the projection score during variance filtering, the informativeness of the variable subsets thus obtained can be followed and the optimal one (the one with the highest projection score) can be found. By selecting the Opt button the program will find the maximum projection score by adjusting the variance cut-off automatically (see figure above). Red color in the Projection score box indicates a low projection score, yellow color indicates a medium-high score and green color corresponds to a high projection score.

In practice, almost all real data sets contain some non-random structure, and therefore it is very uncommon to get a projection score very close to zero. The colors, and thus the boundaries between what is considered to be a ”good” or a ”bad” projection score, are based on our experience from applying the projection score to high-throughput (mainly gene expression microarray) data sets and should be interpreted mainly as rough guidelines suggesting the quality of the representations. For other types of data sets, other quality cutoffs may be more suitable.

The Projection score is calculated based on the number of dimensions provided in the Dimensions box.

STATISTICAL FILTERS

After getting rid of some of the noise by filtering by variance you might try to filter using any of the statistical tests offered (DESeq2, ANOVA,…). What you chose will depend on your data and they question you want to answer. T

After selecting the appropriate test, use the p-value slider or the spin-boxes to enter values. Traditionally a p-value of 0.05 is viewed as the threshold for statistical significance. However, it is difficult to interpret p-values when testing multiple hypotheses and in these situations, the q-value is a better measurement. The q-value is more or less equal to the False Discovery Rate (FDR), see the Reference Manual for more details.

You select a statistical test, the active samples and an input list of variables, possibly all variables in a data frame, the program then applies the test to all active samples and variables and calculates a p-value and a q-value for each variable.

When you have obtained potentially interesting findings, you need to validate your results. The p- and q-values will give a good indication of the statistical significance of your results. In Qlucore Omics Explorer you also have other methods such as cross-validation and permutation of annotations to test your results. More information about this can be found in the reference manual.

DISCLAIMER

The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing.

Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document.

Qlucore Omics Explorer is only intended for research purposes.