Approaches to Sample Size Determination for Multivariate Data

Saccenti, Edoardo; Timmerman, Marieke E.

Published in:
Journal of Proteome Research

DOI:
10.1021/acs.jproteome.5b01029

Publication date:
2016

Journal of proteome research

Article

pubs.acs.org/jpr

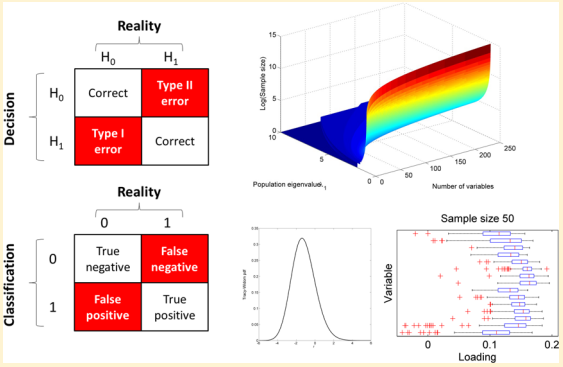
Approaches to Sample Size Determination for Multivariate Data: Applications to PCA and PLS-DA of Omics Data

Edoardo Saccenti*,† and Marieke E. Timmerman‡

†Laboratory of Systems and Synthetic Biology, Wageningen University and Research Center, Dreijenplein 10, 6703 HB, Wageningen, The Netherlands

‡Department Psychometrics & Statistics, University of Groningen, Grote Kruissstraat 2/1, 9712 TS, Groningen, The Netherlands

ABSTRACT: Sample size determination is a fundamental step in the design of experiments. Methods for sample size determination are abundant for univariate analysis methods, but scarce in the multivariate case. Omics data are multivariate in nature and are commonly investigated using multivariate statistical methods, such as principal component analysis (PCA) and partial least-squares discriminant analysis (PLS-DA). No simple approaches to sample size determination exist for PCA and PLS-DA. In this paper we will introduce important concepts and offer strategies for (minimally) required sample size estimation when planning experiments to be analyzed using PCA and/or PLS-DA.



KEYWORDS: *loading estimation, covariance estimation, eigenvalue distribution, random matrix theory, hypothesis testing, dimensionality, multivariate analysis, power analysis*

INTRODUCTION

Data originating from high-throughput experimental techniques are usually multivariate in nature and being analyzed using multivariate statistical methods. In metabolomics and proteomics, principal component analysis (PCA) and partial least-squares discriminant analysis (PLS-DA) have been so far the most commonly used types of analysis, and they occupy, for historical reasons, a special place among the tools available to the practitioners in these fields.[1] Great efforts have been made in the past to make the omics community aware of the merits and demerits of these techniques and to provide extensions and new approaches to overcome bottlenecks and limitations.[1−4]

In the context of metabolomics and proteomics data analysis, the issue of sample size determination has seldom been addressed. In contrast, in other omics disciplines, such as genomics and genetics, there is abundant literature on this topic. The reason for this is obvious. In most cases (but not always), genomics data are treated in a univariate fashion; that is, each dependent variable is considered separately (for instance using a *t*-test or analysis of variance (ANOVA)). In this setting, theoretical tools are readily available and new ones have been developed.[5−11] When data are analyzed using a multivariate approach, such as PCA, the problem complicates considerably. Tools for sample size determination in the multivariate case are available in the form of power analysis, but only for the multivariate extensions of the classical *t* test (i.e., Hotelling T^2) and ANOVA (i.e., Multivariate ANOVA (MANOVA)). This means that no simple approach to power analysis exists for PCA and PLS-DA and, as a consequence, it may happen that many studies are underpowered.

There is mounting evidence that a vast majority of published clinical research suffers from low statistical power, owing to limited sample size and other design issues, yielding serious concerns about the generalizability of results.[12−17] Because meta-analysis and retrospective studies are common in the biomedical, behavioral, and social disciplines, but less commonly employed among the metabolomics and proteomics disciplines, it is unknown to what extent studies are underpowered in this area. Notably, the importance is being acknowledged, as indicated by the fact that the last release of the popular Metaboanalyst (www.metaboanalyst.ca) server for metabolomics analysis now offers a module for power analysis and sample size estimation in the conventional ANOVA/MANOVA setting.[18] Therefore, some reflections on sample size estimation in PCA and PLS-DA seem timely.

Sample size determination and power analysis are intertwined. For this reason we set the scene by recalling some basic concepts and terminology of hypothesis testing and power analysis. We proceed by discussing two major topics for PCA. The first is the problem of determining the (minimal) sample size to obtain stable and reproducible component loading estimations. The importance of this issue for the generalizability of results to the population is widely acknowledged in the social sciences, especially in the PCA related common factor analysis context,[19−21] but it has received little attention in the *omics* field. The second topic addressed is determining the minimal sample size required to assess the dimensionality of a data set, which is also relevant in

Received: November 9, 2015
Published: June 20, 2016

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

derived applications of PCA, such as in deconvolution/curve resolution. The first topic is investigated empirically by making use of simulated and real life *omics* data, and the second is addressed from a more formal theoretical point of view, building upon both recent results from advanced multivariate statistical inference theory and a vast array of simulations. A Mathematical Appendix provides the reader with more details and references about the technical details.

In the PLS-DA framework we first show analogies between the discrimination problem in the classical case-control setting and hypothesis testing. Then we propose a simulation strategy for sample size estimation, incorporating the definition of multivariate effect as derived from the MANOVA practice, when PLS is used for discrimination. We conclude with some final considerations and new directions for further investigation.

MATERIALS AND METHODS

Experimental data

We make use of four experimental data sets obtained with different analytical platforms on different biofluids.

Data set D.1. Serum blood metabolites were measured during the large scale metabolomics project on twins TwinGene.[22] We considered here 2139 spectra and an array of 133 metabolites. Data were downloaded from the Metabolights public repository[23] (www.ebi.ac.uk/metabolights) with accession number MTBLS93. This is a designed case-cohort of incident coronary heart disease, diabetes, dementia, and ischemic stroke events and a matched subcohort (controls) stratified on age and sex which resulted in the inclusion, for metabolomics profiling, of 2139 individuals out of 12591 from the TwinGene project. For full details on the study protocol, sample collection, chromatography and GC-MS experiments and metabolites identification and quantification see the original publication[24] and the Metabolights accession page.

Data set D.2. Serum blood metabolites (29 quantified) that were measured on 864 adult healthy blood donor volunteers. The study investigates a highly homogeneous cohort where no obvious clustering of subjects could be observed in the quantified serum metabolites. For full details on the study protocol, sample collection, NMR experiments and metabolites identification and quantification see.[25,26]

Data set D.3. Urine NMR spectra (bucketed) were collected for 22 subjects over a period of three months. Two different subdata sets were considered; D.3A: containinig the NMR bucketed spectra (0.004 ppm bin width) for 22 subjects (data size: 733 × 1225 after removal of water and urea resonances and empty regions) and D.3B: containing data for 31 subjects (0.02 bucketing, data size: 1604 × 490).

This was a normality study aiming to investigate subject-specific metabolic urinary profiles in healthy subjects who provided ∼40 urine samples each. For full details on the study protocol, sample collection, NMR experiments see the original publications.[27−29] Data set D.3B is used only in the analysis shown in Figure 8.

Data set D.4. Quantified urine NMR metabolites (62 quantified) measured on 79 urine samples from pigs (*Sus scrofa domesticus*). Data were downloaded from the Metabolights public repository with accession number MTBLS123.

The study investigated carbohydrate prefeed in pigs undergoing simulated polytrauma and hemorrhagic shock with resuscitation with the aim of determining whether the metabolic response to shock is dependent on fed state. The experimental setup was as follows: 64 Yorkshire pigs were divided into two experimental groups: fasted and prefed in addition to two control groups. Experimental animals were subjected to a standardized hemorrhagic shock protocol, including pulmonary contusion and liver crush injury. Multiple urine samples were collected at different time points during the study. For full details on the study protocol, sample collection, chromatography and MS experiments and metabolites identification and quantification see the original publication[27] and the Metabolights accession page.

Data preprocessing

Urine data were normalized to correct for differences in urine volume. NMR spectra of data set D.3 were normalized to creatine peak before a bucketing of 0.004 and 0.02 ppm. Quantified NMR metabolite concentrations of data set D.4 were normalized to total urine output as provided by the original publication.[27]

Data were centered but not scaled before analysis. Scaling affects the data structure and the variance in the data. As a consequence, loading estimations will be affected. For a discussion of the impact of scaling on loading estimation, we refer to.[30]

Software

Power calculations for the Hotelling T^2 test has been performed with the G*power 3 software[33,34] available at www.gpower.hhu.de/en.html. All other calculations have been performed in the Matlab environment (MATLAB 8.5, The MathWorks, Inc., Natick, MA, US).

Matlab code for determining the sample size for the cases in PCA and PLS-DA considered, is available at semantics.systemsbiology.nl.

The PLS-DA models has been estimated using in house-scripted routines. Model optimization has been carried out by means of a double cross-validation scheme.[31,32]

SAMPLE SIZE DETERMINATION, POWER ANALYSIS AND RELATED CONCEPTS

Before moving to PCA and PLS-DA applications, we recall some basic concepts of power analysis.

Univariate case

The power of a statistical test is defined as the probability of rejecting the null hypothesis H_0 when it is actually false. The power of a statistical test is determined by the interplay of three parameters: sample size, Type I error, which is the probability of rejecting H_0 when actually true, and effect size, which is the quantity that indexes the degree of deviation from H_0 in the underlying population.[33]

The Type I error, also known as the significance level or α, is to be specified by the investigator; typical values specified are 0.01 and 0.05. The concept of effect size is crucial in power analysis. It relates to the population parameter that one is interested in assessing by performing the experiments. The specific effect size to consider has to be tailored to the statistical test procedure involved. Power analysis can be performed, for example, for tests involving means, proportions, regression coefficients and correlations. Usually for those scenarios more than a single definition of effect size is available.[33,35,36]

We focus on *a priori* power analysis, in which the experimenter is interested in determining the minimal number of samples to attain a given power, say 0.80, for a specific statistical procedure, given the specified Type I error and effect size.

To introduce the core ideas of power analysis we will consider a case-control scenario. The interest may then be to assess the equivalence of the means of the groups, making the difference in

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

population means a natural parameter of interest. When testing the difference with a two-sample *t*-test, the effect size *d* can be defined as

$$d = \frac{\mu_1 - \mu_2}{\sigma} \quad (1)$$

where μ_1 and μ_2 are the population means of the two groups and σ^2 is the pooled variance. To compute *d*, one needs to know the population means and their variances. This may seem counterintuitive because these are the quantities of interest, which are to be estimated on the basis of the study conducted. Arriving at a reasonable effect size for the power analysis is an educated guess which should be based on the researcher's expectations.

As the effect size *d* is a standardized measure, its value can be directly interpreted. In the behavioral sciences, Cohen's definitions for effect size[37] are usually utilized and there is general consensus on which effects can be considered large ($d \sim 0.8$), medium ($d \sim 0.5$), small ($d \sim 0.5$), and trivial ($d < 0.2$) effects;[38] trivial effects are effects that can be detected statistically but may be not relevant from a content point of view (see reference[39]). How this nomenclature transfers to the biological sciences is open to discussion: we are not aware of such studies concerning metabolomics/proteomics, thus it is difficult to deliberate on what value of *d* should be considered, for instance, for a medium effect. However, Cohen's definition has found some application in metabolomics studies[40] (for that matter, this was the only paper we could find with an embryonic power analysis in the metabolomics field). Some indications of what can be considered a trivial effect (*i.e. biologically non relevant*) in the case of metabolomics will be provided in the section dedicated to PLS-DA.

Multivariate case

When hundreds to thousands of variables are measured simultaneously on complex biological systems, one is often interested in extracting information that relies not only on the mean level of the variables (e.g., metabolites, peptides) but also in the mutual relationships among these variables.

Many models for the univariate case can be extended to the multivariate case, when more than one variable (metabolite) is measured in one or more groups. For the two-group case the Hotelling T^2 test (i.e., the multivariate extension of the classical *t*-test) can be used; multigroup cases and complex experimental designs involving repeated measures or different experimental factors are addressed with MANOVA (Multivariate Analysis of Variance), or variants thereof.

Estimating effects in the multivariate case is less straightforward than in the univariate counterpart. For example, for the T^2 test, one needs knowledge of the population variance-covariance matrix $\mathbf{\Sigma}$, rather than only the variance as in the *t*-test. Under the assumption of equality of covariance matrices across the groups, set $\mathbf{w} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and the multivariate effect size is given by the formula

$$\Delta = \sqrt{\mathbf{w}^T \Sigma^{-1} \mathbf{w}} \quad (2)$$

where $\mathbf{\Sigma}$ is a $p \times p$ symmetrical matrix.[33] In eq 2, the matrix $\mathbf{\Sigma}$ plays the role of σ in eq 1. Multivariate effects are more difficult to understand than their univariate counterparts, because they are a function of the population means and standard deviations, as well as the strength of the relationships between the variables, described by the off-diagonal terms of the variance-covariance matrix $\mathbf{\Sigma}$.

Sample size determination for PCA and PLS

When data are going to be analyzed using PCA or PLS, the matter complicates considerably. PCA and PLS are typically viewed as summarizing and exploratory, rather than as testing procedures. This implies that it is uncommon to interpret the model parameters as statistical effects, i.e. as quantities that index the degree of deviation from H_0 in the underlying population. Although defining the null hypotheses and alternative hypotheses in this context may be complicated, it is, by any means, possible. A further complication arises because no obvious single definition for the effect size is available, as many parameters defining the multivariate model can be considered as parameters of interest: one can be interested in the loadings, the number of components that best fit the data, or a given amount of variance explained by the model. The fact that for most parameters the theoretical null distribution is unknown, is not prohibitive, because they could be empirically estimated by a resampling procedure, as permutation or bootstrapping. However, even when the power analysis problem cannot be explicitly stated, the problem remains of determing the (minimal) number of samples needed to obtain an accurate estimation (according to some rule that will be defined) of the parameters of interest. For PCA we will consider the stability of the loadings and the number of significant components (i.e., the significant variance explained by the model). The latter problem is addressed in the context of hypothesis testing.

For PLS-DA we will consider the discrimination accuracy of the model and we will show how the discrimination problem can be rephrased in terms of hypothesis testing language and classical power analysis. Working in this setting we will show how the concept of trivial effect (i.e., non relevant biological effects) may become of importance when designing experiments to be analyzed with PLS-DA.

PRINCIPAL COMPONENT ANALYSIS

The reason for focusing on principal component analysis is 2-fold. First, the use of PCA is ubiquitous in the analysis of *omics* data where the aim is summarizing the information as well as possible using a limited number of variables.[1] Second, recent theoretical developments open the way to power analysis in a PCA context.

Given a $n \times p$ (objects × variables) data matrix $\mathbf{X}$ the PCA model is

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (3)$$

where $\mathbf{T}$ is the score matrix of size $n \times k$, containing the projections of the observations onto the *k*-dimensional PCA subspace, $\mathbf{P}$ the loading matrix of size $p \times k$ and $\mathbf{E}$ the residual matrix of size $n \times p$. The principal components and loadings can be found using a singular value decomposition of $\mathbf{X}$, but the PCA problem can be solved also by considering the eigendecomposition of the covariance matrix of $\mathbf{X}$.

In *omics* literature it is customary to show score plots resulting from PCA with the aim of showing patterns in the data (such as, clustering of different groups). Although this may be instructive, it should be recalled that the interpretation of a PCA model is dependent upon the loadings. The loadings describe the relative importance of each variable to the model and are pivotal to understand the patterns seen in the score plot. In this light it is clear that it is crucial to obtain a reliable estimation of the loadings. Moreover it should be noted that the objects (i.e., subjects providing samples in a study) are usually con-

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

sidered to be randomly sampled from a certain population, with the aim to form a sample that is representative of the population under study. In this case it is of interest to consider how close the estimated loadings typically are to those in the population, and thus to what extent the results obtained on the sample can be generalized to the overall population. This leads to the question of determining the number of samples that are needed to obtain a stable (and accurate) estimation of the population loadings. This problem will be addressed in the section Sample size estimation for the loadings in PCA.

Another critical aspect in PCA is the choice for the number of components k. This is often based on either interpretational issues or criteria that are directly related to the eigenvalues of the covariance matrix associated with the principal components. In the section Sample size estimation for the determination of the number of components in PCA, we address the related problem of estimating the number of samples needed to reliably estimate the number of components in a PCA model. These two aspects are intertwined and it can be argued that no stability can be reached for nonsignificant components (i.e., components accounting only for noise). However, since the two problems are addressed using different numerical and theoretical approaches they are presented here as two different subjects.

Sample size estimation for the loadings in PCA

We start with a simulated example, where data matrices $\mathbf{X}$ are generated with two leading components by considering a diagonal $p \times p$ population covariance matrix $\mathbf{\Sigma}$ of the form

$$\Sigma = \begin{bmatrix} \lambda_1 & & & & 0 \\ & \lambda_2 & & & \\ & & 1 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{bmatrix} \quad (4)$$

with $\lambda_1 > \lambda_2 > 1$. In this case, the $p - 2$ variables associated with population eigenvalues equal to 1 ($\lambda_i = 1$ with $i > 2$) describe noise. This particular model is known as the spiked covariance model[41] and describes a situation in which all variables are uncorrelated and the information is concentrated within a few variables (two, in this case). This simple framework allows the development of extremely powerful and widely applicable statistical methods, which will be introduced and used in the next sections. Under this model, the population loading matrix $\mathbf{\Pi}$ (that can be obtained by an eigendecomposition of $\mathbf{\Sigma}$) is a $p \times p$ diagonal matrix with standard deviations on the diagonal.

To illustrate the effect of sample size on the estimation of PCA loadings we implemented a simulation schemes as follows. We fixed the population eigenvalues λ_1 and λ_2 to 10 and 8, respectively. We generated sets of sample data matrices $\mathbf{X}$ of different size n, with n in the range 50 to 2000, and different number of variables p, taking p with values 50, 100, and 500. Each generated data matrix $\mathbf{X}$ is decomposed via PCA.

The resulting sample loadings $\mathbf{P}$ on the principal components are Procrustes rotated toward the known population loadings (to account for the arbitrary rotational ambiguity of the loadings). The rotated loadings are then compared with the known population loadings using a measure of congruence, Tucker's ϕ.[42] Tucker's ϕ (see Mathematical Appendix for a definition) takes values from -1 to 1, with 1 (and -1) indicating perfect congruence (though opposite in sign), and values closer to 0 indicating smaller degrees of congruence. The exercise is repeated 100 times to provide an indication of the variability introduced by the sampling procedure. The sample loadings should converge to the population loadings as the sample size n increases, and hence the congruence should converge to 1. Figure 1 shows the results of this exercise for

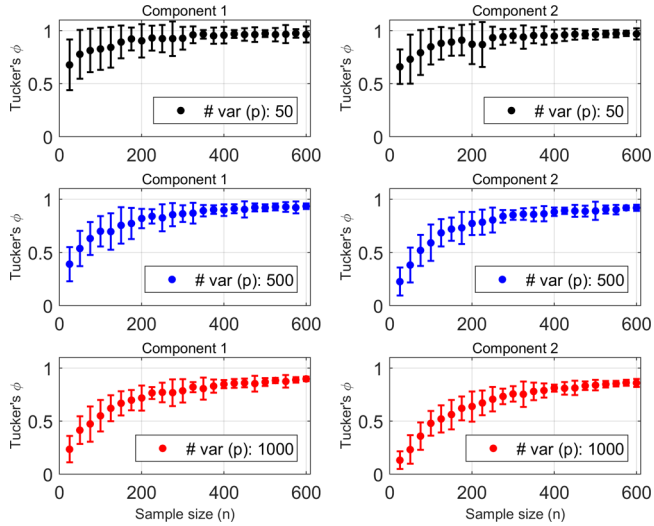


Figure 1. Estimation of the loadings of the first two principal components as a function of the sample size. Data are randomly generated under a PCA model with 2 components using different numbers of variables (from top to bottom: 50, 500, and 1000). The mean and the error bar (given as ±1 standard deviation) are calculated over 100 replications for each value of the sample size. Loadings for each model are Procrustes rotated to a target (the known population loadings) before calculations of Tucker's ϕ.

data generated under a multivariate normal distribution $N(\mathbf{0},\mathbf{\Sigma})$. Results are the same using a nonsymmetrical distribution (not shown), indicating that they may hold generalizability to other distributions.

As expected, the sample loadings $\mathbf{P}$ approximate the population loadings $\mathbf{\Pi}$ as the sample size increases: around 300 objects are needed to obtain a congruence >0.9, which is an accepted value in the behavioral sciences to establish equivalence between two sets of loadings.[43]

This implies that, based on this simulation scheme, ~ 300 objects should be considered to build a PCA model when the focus of the analysis is to generalize the results obtained from a sample to the population level. Results in Figure 1 further suggest, in line with earlier findings,[44] that the ratio between the numbers of variables and observations influences the loading estimation. It appears that for a given sample size the congruence is lower when the number of variables is higher, and that this effect is larger for smaller sample sizes. It should be also considered that increasing the number of variables will increase the variance in the data. The first two components explain a fixed amount of variance (i.e., 10 + 8), while the remaining components explain increasing amounts of variance (32, 482, and 982 for the three cases). Adding variables, which are non informative in this case, also increases the difficulty of the problem as a result of the curse of dimensionality.

Analytical formulas exist for the estimation of confidence intervals for loadings in specific situations,[45−47] while the bootstrap approach can be used for all types of loadings. This problem is discussed in reference[48] for a PCA setting and in reference[49] in the closely related multilevel simultaneous component analysis.[50−52]

The pattern illustrated by simulation also holds in the case of real experimental data. We consider the large metabolomics data set D.1 (2139 × 133) (see Materials and Methods section) and take as population loadings those obtained by fitting a PCA

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

model on the *full* data set. Subsets of different size are obtained by random sampling from the full data set. Again we consider samples sizes in the range 50 to 2000; loadings are Procrustes rotated toward the loadings obtained fitting a PCA model on the full experimental data set(s). Results are shown in Figure 2A.

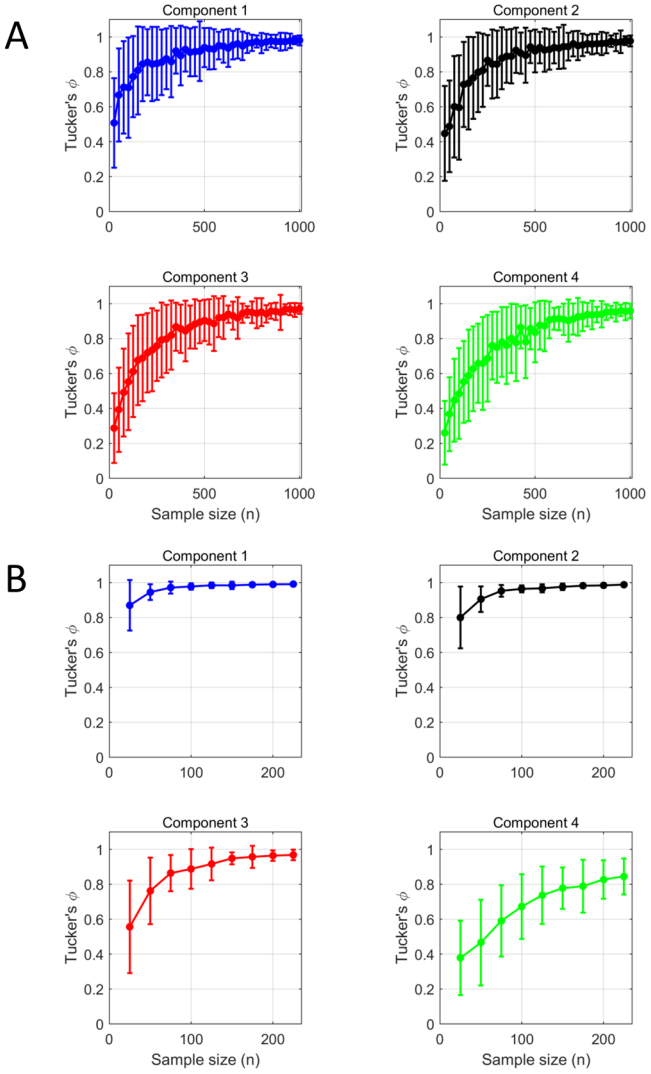


Figure 2. Estimation of the loadings of the first four principal components as a function of the sample size (n) for real data. Data are sampled without replacement from two experimental data sets (Panel A: Data set D.1, of size 2139 × 133; Panel B: Data set D.3, of size 733 × 1225). The mean and the error bar (given as ±1 standard deviation) are calculated over 100 replications for each value of the sample size.

For the first four components it can be seen that loadings converge toward the population loadings with increasing sample size, and that as many as 300−400 samples are needed to obtain a congruence larger than 0.9. We have repeated the same exercise with data set D.3A for which the number of variables far exceeds the number of observations (1225 versus 733); results are presented in Figure 2B: in this case as many as ∼75 samples are needed to obtain a congruence larger than 0.9 for the first component.

Tables 1 and 2 summarize the results of similar exercises on the four experimental data sets (D.1 and D.2, D.3A, D.4) for the first and second component, respectively. It appears that the level of congruence attainable depends both on the number of samples and on other factors, including the nature of the data, the order of components considered and the variance explained. For instance, for data set D.2 as few as 25 samples are needed to obtain a stable estimation of the loadings for the first component, while this increases to 300 for data sets D.1. Herewith, the inherent structure of data seems to play a crucial role, and also the way the data are presented. For instance when NMR data are bucketed, higher covariances between bucketed variables than metabolites can be expected, since many bucketed variables may represent the same compound increasing the degree of covariation.

It is interesting to note that the congruence level attainable with a given sample size n seems to diminish with increasing order of components. For instance in the case of data set D.4, n = 300 samples are needed to obtain an average congruence ∼0.9 for the loadings of the first component; to obtain a similar level of congruence for the second component, n = 400 are needed. It should be noted that we have chosen the 0.9 threshold as a reference for discussion and that such a value should be considered, given the application, the data at hand, and the resources available. For instance, considering data set D.1, it would probably be a waste of resources to acquire and measure 300 samples to increase the congruence from 0.8, which can already be obtained with 150 samples.

Figure 3 shows the estimated loadings of the first 15 variables of the experimental data set D.1 over 100 replications for four different samples sizes. For a small sample size ($n \sim 50$), the sampling variability of the loadings is huge. This implies that a PCA model obtained on such a small data set holds limited generalizability. For some variables the loading estimate ranges from 0 (no importance to the model) to 0.2. This can have a dramatic effect on the biological interpretation of the data, since variables usually have some biological meaning (such as being metabolites or protein concentrations).

As loadings can be derived from the covariance matrix via an eigendecomposition, the problem of loading estimation is related to the problem of estimating the population covariance matrix starting from the experimental data: an inaccurate estimation of the covariance matrix will result in nonaccurate estimation of the loadings and thus in a lack of generalizability. This ubiquitous problem in data analysis[1,53−55] and bioinformatics (especially in the network inference context[56]) can be stated by asking how many samples N are needed to guarantee an estimation of the covariance matrix with a fixed accuracy. Using arguments from advanced probability theory, it has been shown that $N \sim O(p)$; that is, N is smaller than or equal to $A \times p$, where A is some positive constant[57,58] or even $p \times \log_2(p)$,[59] depending on distributional properties and on the structure of the covariance matrices. It is noteworthy that these limiting values are consistent with what was observed in the simulations for loading estimates, and this is in line with some older results:[60] It should be noted that $N \geq p$ unless some structure is imposed on the covariance matrix or some shrinkage or thresholding is applied: in such cases the sample size required can be on the order of $\log_2(p) < p$.[53,57,61,62] However, as our simulation shows, a good agreement for metabolomics applications can also be derived with a smaller number of samples.

We provide a Matlab function (TuckerLoadingEstim.m) to evaluate the accuracy and the stability of loadings through numerical simulation. The function takes as input the population covariance matrix or the eigenvalues of the population covariance matrix. Random data (normally distributed by default, but any distribution could be used) with the prespecified covariance/loading structure is generated with different sample sizes, and the agreement between the k-th sample and population loadings

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

Table 1. Summary of Results for the Congruence of Loadings for the *First* Component of Four Different Experimental Data Sets[a]

	Data set D.1		Data set D.2		Data set D.3A		Data set D.4	
	Serum metabolites MS (2139 × 133)		Serum metabolites qNMR (864 × 29)		Urine bucketed (733 × 1225)		Urine qNMR (343 × 62)	
Sample size n	ϕ	%var	ϕ	%var	ϕ	%var	ϕ	%var
5	0.331	44.5	0.825	77.5	0.651	62.9	0.783	86.1
25	0.507	20.8	0.977	71.9	0.893	41.8	0.744	68.8
50	0.657	19.2	0.99	70.1	0.941	39.3	0.798	61.8
100	0.759	17.7	0.996	70.3	0.979	37.8	0.855	57.7
150	0.795	16.5	0.998	70.2	0.988	37.9	0.95	57.8
200	0.876	16.1	0.998	70.4	0.989	37.3	0.967	56.8
250	0.871	15.5	0.999	70.6	0.994	37.2	0.989	56.3
300	0.919	15.6	0.999	70.6	0.995	37.3	0.996	56.2
350	0.915	15.1	0.999	70.7	0.996	37.2		
400	0.955	15.7	0.999	70.7	0.997	37.1		
450	0.933	14.9	0.999	70.5	0.998	37.1		
500	0.961	15.4	1	70.7	0.998	37.2		
550	0.97	15.3	1	70.6	0.999	37.1		
600	0.948	15.1	1	70.5	0.999	37		
650	0.974	15.1	1	70.6	1	37.1		
700	0.98	15	1	70.5	1	37		
750	0.978	15	1	70.6				
800	0.976	14.8	1	70.7				
850	0.983	15	1	70.6				
900	0.986	15.1	1	70.5				
950	0.989	15.1	1	70.6				
1000	0.988	14.9						
1050	0.987	15						

[a] n indicates the total number of samples used to build the model; *%var* indicated the variance explained by the 1st component; ϕ is the Tucker's congruence coefficient.[43] *%var* and ϕ values are averaged over 100 repetitions for each n. For more details on the procedure, see the text.

(expressed in terms of the Tucker's ϕ) is plotted/tabulated as a function of the sample size. Tucker's $\phi \geq 0.9$ is taken as a threshold for good agreement.

Sample size for the determination of the number of components in PCA

The variation observed in the data can be split into the informative variation, which is related to the biology of the problem, and the noninformative variation, denoted as noise. The complete set of (p) principal components of a given data set captures both types. To disentangle both types of variation, one relies on the crucial assumption that the informative variation is linked to a relatively large variation and the noninformative variation to a relatively little and/or about equal variation. This implies that the first few components (i.e., having large variance) capture the relevant information while higher order components (i.e., having small and/or about equal variance) describe noise. A fundamental problem in PCA is thus how to determine the number of relevant components. The literature on this topic is abundant, and many solutions, either statistical or numerical, have been proposed (see, for instance, refs 63−66 for an overview). Here we consider a statistical procedure to determine the number of relevant components based on results and approaches from modern Random Matrix Theory. This statistical framework, which we will introduce with a minimal formulation, allows for sample size determination in PCA when the problem is the following: How many samples are needed to build a PCA model for which the number of relevant components can be reliably inferred? Stated otherwise: what is the minimal sample size required to correctly assess how many components describe meaningful information with a specific probability?

The Tracy−Widom test

Recall that the variance of a component is given by the corresponding eigenvalue of the covariance matrix: with reference to eq 4 the variance for the first component is given by λ_1, that for the second by λ_2, and so on. In PCA it is customary to give the proportion of variance explained by each component (ve_i for the i-th component), which is simply defined as

$$ve_i = \frac{\lambda_i}{\sum_q^p \lambda_q} \tag{5}$$

In every day practice the population covariance matrix $\mathbf{\Sigma}$ is unknown, and we can only access the sample covariance matrix $\mathbf{S}$, which is defined as

$$\mathbf{S} = \mathbf{X}^T\mathbf{X} \tag{6}$$

Based on the sample covariance matrix, we can compute the (ordered) sample eigenvalues $l_1 > l_2 > ... > l_p$. The matrix $\mathbf{S}$ is usually fed to the PCA algorithm. A beautiful result from Johnstone[41] shows (see Mathematical Appendix for a full statement) that if data is randomly generated under a multivariate model with $\mathbf{\Sigma} = \mathbf{I}$, the sample eigenvalues, when properly normalized to L_1 (see eq 15 in the Appendix), are distributed like the Tracy−Widom distribution;[67,68] this is illustrated in Figure 4, panel A. The situation in which $\mathbf{\Sigma} = \mathbf{I}$ implies that (*i*) all population eigenvales are equal to 1; (*ii*) each component explains the same amount of variance $1/p$ (cf., eq 5); and (*iii*) all variables are uncorrelated. Thus, $\mathbf{\Sigma} = \mathbf{I}$ represents the situation in which all components describe pure noise; thus, the Tracy−Widom distribution describes the distribution of the largest eigenvalue associated with noise. Stated otherwise, the

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

Table 2. Summary of Results for the Congruence of Loadings for the *Second* Component of Four Different Experimental Data Sets[a]

	Data set D.1		Data set D.2		Data set D.3A		Data set D.4	
	Serum metabolites MS (2139 × 133)		Serum metabolites qNMR (864 × 29)		Urine bucketed (733 × 1225)		Urine qNMR (343 × 62)	
Sample size n	ϕ	*%var*	ϕ	*%var*	ϕ	*%var*	ϕ	*%var*
5	0.258	25.8	0.7	18.3	0.467	23.4	0.404	12.7
25	0.421	13.8	0.959	21.1	0.806	22.7	0.737	24.4
50	0.464	11.8	0.985	20.7	0.93	21.3	0.789	30.7
100	0.625	11.1	0.993	21.7	0.96	21.6	0.814	35.4
150	0.757	10.8	0.996	21.2	0.975	21.4	0.929	37
200	0.839	10.8	0.997	21.2	0.983	21.6	0.978	37.9
250	0.863	10.7	0.998	21.7	0.989	21.3	0.989	37.8
300	0.872	10.6	0.998	21.4	0.991	21.2	0.997	38.9
350	0.888	10.8	0.998	21	0.993	21.3		
400	0.932	10.8	0.999	21.2	0.996	21.4		
450	0.95	10.8	0.999	21.1	0.996	21.4		
500	0.946	10.7	0.999	21	0.997	21.3		
550	0.954	10.7	0.999	21.2	0.998	21.3		
600	0.959	10.7	1	21	0.999	21.3		
650	0.973	10.7	1	21.1	0.999	21.3		
700	0.963	10.7	1	21.2	1	21.3		
750	0.976	10.7	1	21.2				
800	0.976	10.7	1	21.1				
850	0.977	10.7	1	21.1				
900	0.982	10.6	1	21.2				
950	0.985	10.6	1	21.2				
1000	0.985	10.6						
1050	0.986	10.7						

[a] n indicates the total number of samples used to build the model; *%var* indicates the variance explained by the 2nd component; ϕ is the Tucker's congruence coefficient.[43] *%var* and ϕ values are averaged over 100 repetitions for each n. For more details on the procedure, see the text.

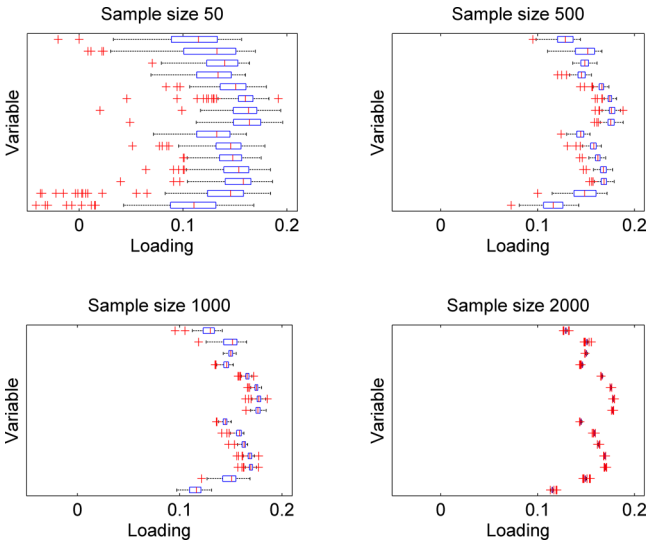


Figure 3. Box plot representation of the loading estimation (first principal component) for the first 15 variables of the data set D.1 as a function of sample size. The variability observed for a small sample size (n = 50) is remarkably large. Red crosses indicate outliers; outliers are larger than $q_3 + 1.5(q_3 - q_1)$ or smaller than $q_1 - 1.5(q_3 - q_1)$, where q_1 and q_3 are the 25th and 75th percentiles, respectively.

Tracy–Widom distribution gives a null model against which to test the sample eigenvalues. The *P*-value is the proportion under the Tracy–Widom distribution associated with the normalized (see eq 15 in the Appendix) sample eigenvalue L_1 or larger. If this *P*-value is smaller than the significance level α, the first component is deemed significant. Here we limit ourselves to the first eigenvalue, but results hold true also for higher order eigenvalues. The complete testing procedure for higher order eigenvalues, first presented by Johnstone[41] and reviewed and adapted by several other authors,[69−71] is outlined in the Mathematical Appendix. The L_1 for some real data is shown in Figure 4 panel B. For the first eigenvalue the difference between what is expected in the absence of information in the data (i.e., the distribution in panel A) and what is observed in the real data is large; this is associated with a low *P*-value, and thus a significant first component for a PCA model for this data set. The testing procedure has been derived working under the spiked model in eq 4. Good performances of the test have been found on a large number of different simulated data sets including covariance structures other than the spiked model.[66]

Toward a power analysis in PCA

When considering a Tracy–Widom testing procedure, the population effect of interest is the population eigenvalue associated with the first component. To introduce a power analysis for the first eigenvalue with a Tracy–Widom test, we first need to present the so-called Baik–Ben Arous–Péché (BBP) conjecture, which we state in a simplified form. The BBP conjecture[72] concerns the behavior of the largest sample eigenvalue and has been found to have implications in data analysis.[70] Its importance here lies in the fact that it has implications for power analysis in a PCA setting.

Let us consider the situation in which the first population eigenvalue $\lambda_1 > 1$ and all other eigenvalues are equal to 1 and $(n,p) \to \infty$, with a finite limit of the ratio p/n.

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

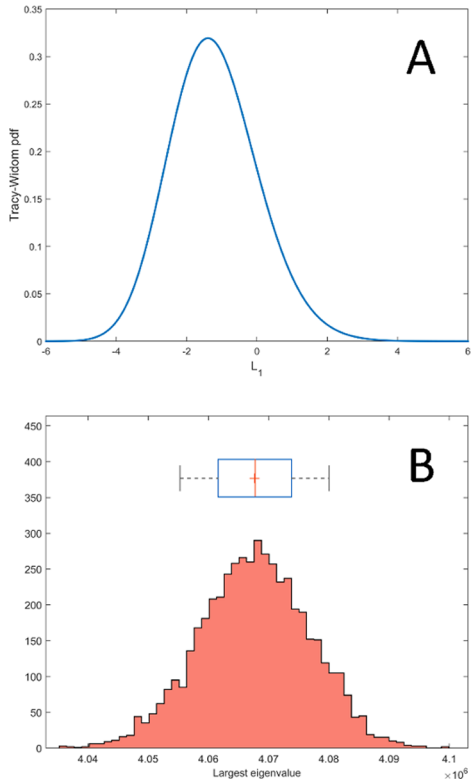


Figure 4. (A) Plot of the Tracy–Widom probability density function (pdf, see eq 20 in the Appendix), describing the distribution of the largest eigenvalue (normalized using eq 16) of the covariance matrix of random $N(0,I)$ data of size $n \times p$. (B) Distribution of the largest eigenvalue (normalized using eq 16) of sample covariance matrices obtained by random sampling from the full data set D.1.

1. If

$$\lambda_1 < 1 + \sqrt{\frac{p}{n}} \tag{7}$$

then l_1, when properly normalized to L_1, will have the same distribution as when $\lambda_1 = 1$; that is, it will be Tracy–Widom distributed.

2. If

$$\lambda_1 \geq 1 + \sqrt{\frac{p}{n}} \tag{8}$$

then L_1 will be almost surely detectable; that is, it will be well separated from the eigenvalues describing the noise.

We refer to the quantity $1 + \sqrt{p/n}$ as the BBP threshold. The BBP conjecture provides the following limits of convergence for the sample eigenvalues under conditions (1) and (2):

$$l_1 \to \left(1 + \sqrt{\frac{p}{n}}\right)^2 \text{ if } \lambda_1 \leq 1 + \sqrt{\frac{p}{n}} \tag{9}$$

$$l_1 \to \lambda_1\left(1 + \sqrt{\frac{p}{n}}\frac{1}{\lambda_1 - 1}\right) \text{ if } \lambda_1 > 1 + \sqrt{\frac{p}{n}} \tag{10}$$

This indicates that the behavior of the sample eigenvalues depends on whether the corresponding population eigenvalues are below or above the BBP threshold. This effect is shown graphically in Figure 5, panel A, where the sample eigenvalues are plotted together with the population eigenvalues. Statement

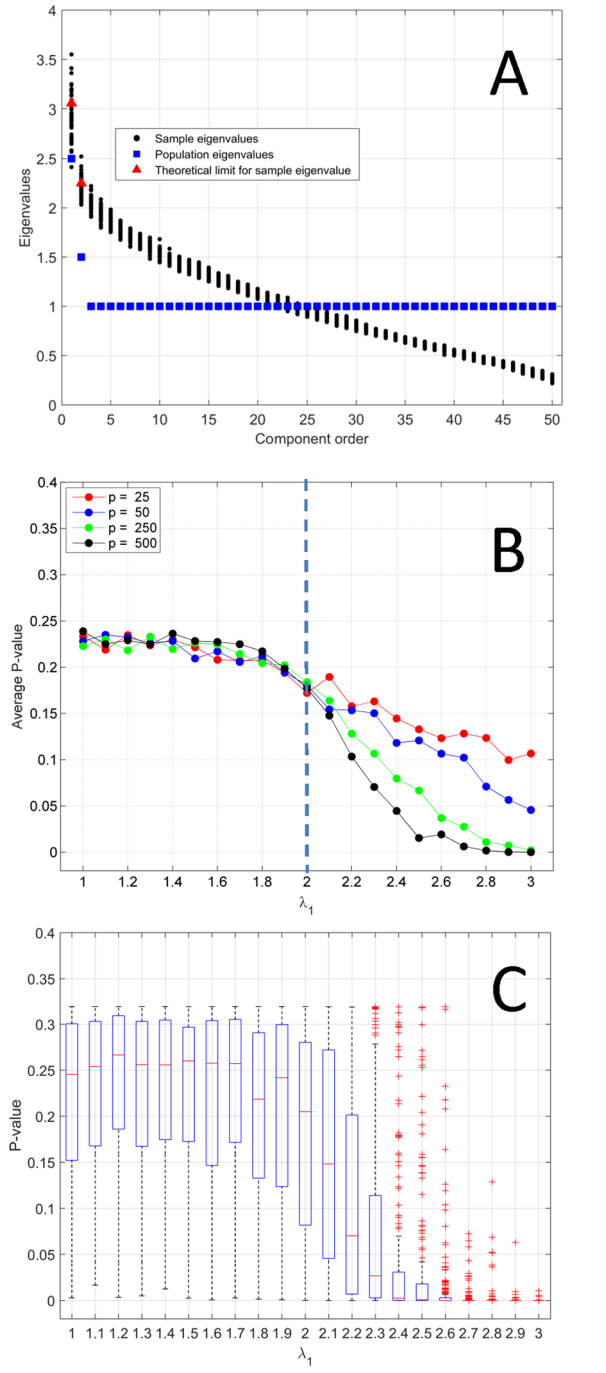


Figure 5. (A) Illustration of the BBP phase transition: here $p = 50$ and $n = 200$. The population eigenvalues are $\lambda_1 = 2.5$, $\lambda_2 = 1.5$, and $\lambda_j = 1$ for $j > 2$; the BBP threshold is 1.5. Blue boxes indicate the population eigenvalues; black boxes indicate the sample eigenvalues (over 50 replicates), and red triangles indicate the theoretical limits for the sample eigenvalues as given by eq 8. This figure is an adaptation of Figure 3 in the review paper by Paul.[91] (B) Average P-value of the Tracy–Widom test for the first principal component (computed across 200 replicates) as a function of λ_1, for four data sizes, with $p = 25$, 50, 250, 500, and $n = p$. The vertical line indicates the phase transition. The change of behavior is evident at the BBP threshold (2 for all four data matrices, see eq 8): if the population eigenvalue is above the BBP threshold, the power of the test increases dramatically, as indicated by the sharp reduction of the P-values associated with the test. (C) Box plots of the P-values for the same Tracy–Widom tests as depicted in part B, for $p = 500$, to show the variability across the 200 replicates, as a function of λ_1.

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379–2393

(2) was proved for real data,[73] while Paul[74] showed that, in the real case, L_1 will be normally distributed. This result is stated for λ_1, but it holds true also for higher order eigenvalues. It should be noted that λ_1 here plays the role of the effective estimate of interest. Assuming the k-th sample eigenvalue to be a good representation of the corresponding population eigenvalue, eq 8 could, in principle, also be used as a posteriori test to avoid the inclusion of nonsignificant components.

Power of the Tracy–Widom test

From eq 8 it follows that the possibility of detecting structure in the data is completely determined by the size of the eigenvalue(s) of the population covariance matrix. This has repercussions on the ability of a Tracy–Widom test to detect structure in the data. If λ_1 is below the BBP threshold, l_1 (and thus L_1) is Tracy–Widom distributed and it is indistinguishable from noise. To illustrate this phenomenon, we simulated square data matrices (i.e., $n = p$, thus associated with a BBP threshold equal to 2 (see eq 7) of four sizes (p = 25, 50, 250, 500) and different population eigenvalues λ_1 (from 1 to 3, with steps of 0.1), and 200 replicates per condition. For each simulated data set, we performed the Tracy–Widom test for the first principal component. In Figure 5, panel B, the average P-value across the 200 replicates is plotted against the population eigenvalue λ_1, for the four sizes. It appears that, below the BBP threshold, there is little chance to have a significant component, while this chance increases sharply above the BBP threshold. Figure 5, panel C shows the variability in the P-values when performing such an exercise.

If λ_1 is above the BBP threshold, the detection becomes feasible. This means that the power of the test depends on the magnitude of the population eigenvalue and it increases with λ_1, as in any statistical test. This is shown in Figure 6, panel A, where the empirical power of the Tracy–Widom test is plotted against λ_1, the effective size. As expected, the power increases with the effective size itself, but it also strongly depends on the number of variables considered. Given the same effective size, the power becomes larger with increasing size of the data matrices, that is the ratio p/n. Note that, for $\lambda_1 = 1$, the H_0 is true, rendering the term power—strictly speaking—to be incorrect; the represented probability refers to the actual Type I error.

Sample size estimation for PCA

Equation 8 provides a direct way to estimate the absolute minimally required sample size in a PCA application for the detection of λ_1. Suppose that λ_1 is known a priori (or some sort of educated guess is available, as it always must be when performing a power analysis). Then solving eq 8 for n gives the absolute minimal sample size N_t required for detecting λ_1

$$N_t \geq \frac{p}{(\lambda_1 - 1)^2} \tag{11}$$

Thus, the possibility of detecting λ_1 depends on the ratio between the sample size and the number of variables. The latter usually depends on the measurement platform and is typically determined by the experimental setting: in targeted metabolomics and metabonomics, p is usually in the range of 10 to 500 (although binning techniques allow manipulation of the actual number of variables (bin) to be analyzed); in genomics, it is 10000–50000; and in genetics, it can easily reach 10^6, while the experimenter can decide on how many samples to consider in the study.

Above the absolute minimal sample size N_t, increasing the sample size yields an increase in power of the test. From the BBP conjecture, it is inferred that increasing the sample size is useful only if λ_1 is above the detection limit because if λ_1 is below the detection limit it will be distributed like noise. If λ_1 is above the BBP threshold, increasing the sample size enhances the convergence of the sample eigenvalue(s) to the population eigenvalues.

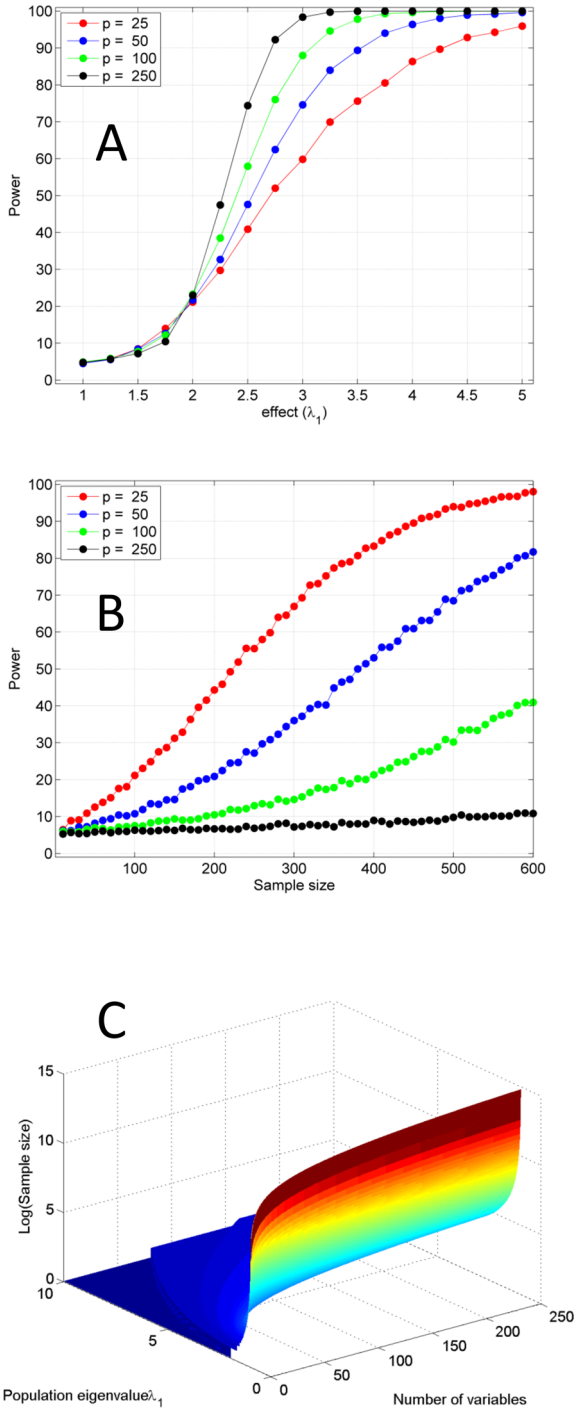


Figure 6. (A) Power of the Tracy–Widom test for the first sample eigenvalues, as a function of the magnitude of the first largest population eigenvalue (λ_1) for four data matrices, with p = 25, 50, 100, 250, and $n = p$. (B) Power of the Tracy–Widom test for the first sample eigenvalue as a function of the sample size, for the number of variables p = 25, 50, 100, 250, and $n = p$, with λ_1 fixed at 1.5. (C) (Log of) minimal sample size as a function of the first population eigenvalue (λ_1) and the number of variables p. These three quantities are interrelated through eq 9.

Figure 6, panel B illustrates the power of a TW test, as a function of sample size for a given value of λ_1. As expected, the

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379–2393

power increases with the sample size, and strongly depends on the ratio p/n (here we consider square matrices). This indicates that if the number of samples cannot be increased, then reducing the number of variables measured can increase the power of the test. Figure 6, panel C gives the minimal sample size N_i as a function of both λ_1 and p; N_i diverges rapidly as $\lambda_1 \to 1$ and p increases.

To make use of the BBP conjecture and the power of the TW test to estimate the sample size needed to assess the significance of a component, one needs to have an estimation of the population eigenvalues of the empirical data at hand. This should preferably be done on the basis of existing data. Alternatively, it could be based on the collected data, which would make the power analysis *a posteriori*, with the associated risk of unrealistic results because of sampling fluctuations.

We provide a Matlab script (empowerTW.m) to evaluate the empirical power of a Tracy−Widom test for the first component: the function takes as input the data matrix dimensions and the value of the largest population eigenvalue and calculates whether the signal is below or above the BBP threshold; in the latter case, the empirical power of the test (at both the 0.01 and 0.05 levels) is tabulated.

PARTIAL LEAST-SQUARES DISCRIMINANT ANALYSIS

In many metabolomics applications, the interest is in discriminating between two or more groups (like in a classical case-control setting) with the aim of selecting variables (i.e., metabolites) important to the biological problem under study. This is mostly done in a multivariate context by making use of discriminating techniques, such as partial least-squares discriminant analysis (PLS-DA)[75] or principal component discriminant analysis (PCA-DA).[76] PLS-DA (and its extensions as kernel PLS-DA,[77] orthogonal PLS-DA,[78] and multilevel PLS-DA[79]) is ubiquitous in the metabolomics literature, but we restrict ourselves to PLS-DA, for the sake of simplicity. Our considerations apply generally to the other methods mentioned.

Relationships between classification and hypothesis testing

In PLS-DA, discrimination turns into a classification problem: if a difference exists between two groups (labeled 0 for controls and 1 for cases), is it possible to build a model that can correctly classify unknown samples. The problem of how to build and optimize such a PLS-DA model has been widely reviewed;[31,32,80−82] here we assume that the model has been properly defined (see Material and Methods).

The quality of a classification model can be assessed by considering the sensitivity attained by the model, which is defined as

$$sensitivity = \frac{TP}{TP + FN} \tag{12}$$

where TP and FN are the number of true positives and false negatives, respectively. Sensitivity can be interpreted as the probability of a correct classification of case samples. This bears some resemblance to the power of a statistical test,[83] which is the probability of rejecting H_0 when actually false. As the power of a statistical test increases with the sample size, legitimate questions are whether also the sensitivity of a PLS-DA model increases with the sample size, and whether it is possible to provide indications for sample size determination in the PLS-DA setting.

A strategy for sample size determination in PLS-DA

We will illustrate that the sensitivity of a PLS-DA model, given an effect size, increases with the number of samples considered. We simulate a case-control study, where the data for the two groups are generated under the multivariate normal model with $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ is the 133×133 population covariance matrix set to be equal to the covariance matrix of the experimental data set D.1; $\boldsymbol{\mu}_1$ is the multivariate population mean of the Control group (133×1) set equal to the mean of the data set D.1 data; $\boldsymbol{\mu}_2$ is mean of the Case group and is constructed as

$$\mu_2 = \mu_1 + \mathbf{d} \tag{13}$$

where $\mathbf{d}$ is a $p \times 1$ vector whose elements are set equal to d, considering values of d equal to 0, 0.1, 0.2, and 0.4 to simulate different situations; $\mathbf{d}$ describes the magnitude of the separation between the two groups. However, it cannot be considered as expressing the multivariate effect fully, because then the covariance structure should be taken into account as well.

Given $\mathbf{d}$, we built a series of PLS-DA models using an increasing number of samples per group (from 25 controls + 25 cases to 500 controls + 500 cases). Here PLS-DA is used to discriminate between two groups: this is a common problem in multivariate statistics that can be addressed using a standard Hotelling T^2 test for which the power can be analytically calculated.

Nonetheless, the T^2 and PLS-DA tests test different hypotheses, although the aim is the same (assessing difference between groups): by using the Hotelling's test, evidence can only be provided that some linear combination of the population means (of the measured variables) exists for which a nonzero difference between the groups exists. In contrast, in PLS-DA, sensitivity relates to the correctness of classification of individuals, embedding the predictive power of the model. Herewith, also the specificity of the model needs to be taken into account:

$$specificity = \frac{TN}{TN + FP}$$

because the sensitivity alone is not enough to judge the overall quality of the model. This bears resemblance to a statistical test, where the power attained can be judged only by taking into account the significance level considered. In the classification context, combined measures of sensitivity and specificity (i.e., power and significance) could be considered, such as ROC curves (*sensitivity* versus $1 - specificity$) or the AUROC, which is the area under the ROC curve. Indeed, the latter measure has been suggested to optimize the PLS-DA model.[31]

Figure 7 shows the sensitivity (panel A) and specificity (panel B) of a PLS-DA model as a function of sample size for different values of d (0, 0.1, 0.2, 0.4). As expected, the sensitivity and specificity increase both with the sample size and with the magnitude of the separation. Further, the variability of the results decreases with the sample size and magnitude of the separation. This indicates that increasing the sample size not only increases the power, but also increases the stability of the classification solution. Panel C shows the power of the T^2 Hotelling test, which also increases with sample size and the effect size, as expected.

One could be tempted to compare directly the two methods as a function of effect size and sample size. However, this cannot be done, as the two methods depart from different principles. In classification, sensitivity and specificity measure the correctness of individual classifications. In a statistical test, alpha and beta indicate the error probability for a null-hypothesis, for example concerning the difference in means in the population. By increasing the sample size, the power for the Hotelling T^2 would come arbitrarily close to 1, whereas the maximal achievable sensitivity

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

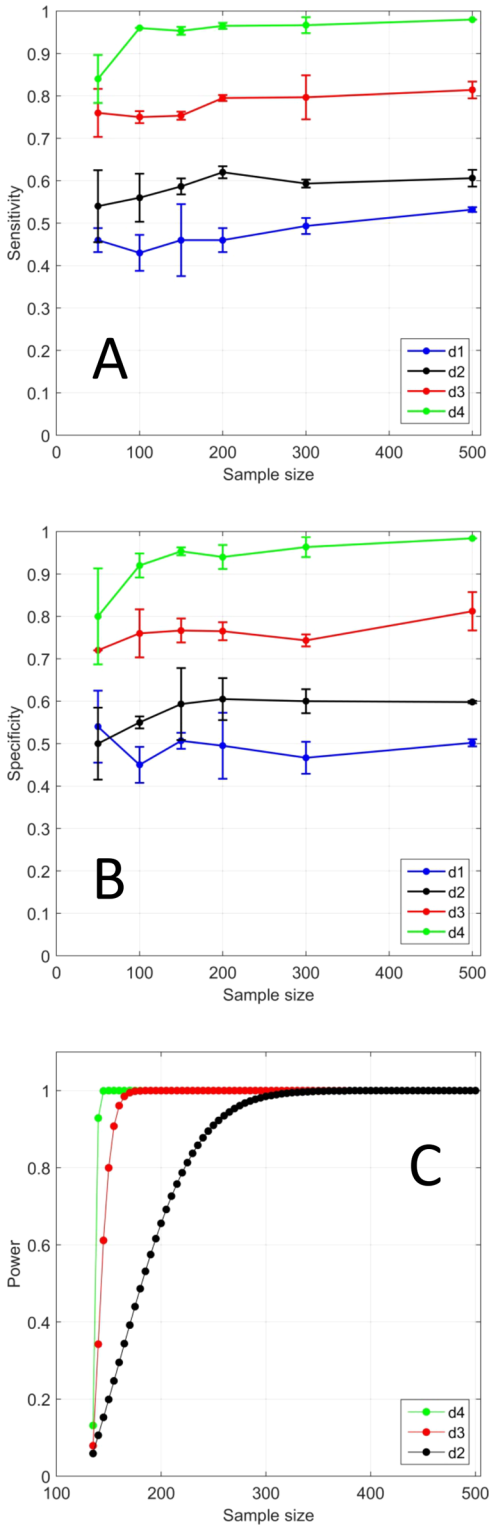


Figure 7. Sensitivity (A) and specificity (B) of a PLS-DA model as a function of the total sample size for the discrimination between two group in a case-control design; d1 = 0, d2 = 0.1, d3 = 0.2, and d4 = 0.4. (Equal size groups; thus each group size is half of the total sample size.) (C) Power of T^2 Hotelling test (calculated) as a function of the total sample size (same as PLS-DA case) with a significance level $\alpha = 0.05$. The *d* values correspond to multivariate effects Δ (as defined by eq 2) of 0, 1.2, 2.4, and 4.9, respectively.

obtained by a PLS-DA model is limited by the Bayes overlap between the case/control distributions.

Determining effect size in a PLS-DA context

The overall quality of a PLS-DA model is influenced by many factors, such as the number of components selected, the cross-validation strategy used when building the model, and the optimization criterion.[31] Thus, sample size is only one of the factors that affects sensitivity. Nonetheless, if all other parameters have been carefully chosen, simulations can be performed to obtain at least a rough estimation of the sample size required to attain a certain sensitivity given an effect size, as shown in Figure 8. This would require the knowledge (or at least an educated guess) of the population means and the variance-covariance matrix $\mathbf{\Sigma}$, just as in a power analysis for Hotelling T^2 and other multivariate tests. Thus, the problem here is the *a priori* knowledge of $\mathbf{\Sigma}$ underlying a given (biological) phenomon: the number of different covariance patterns between *p* variables (say metabolites) that can occur in reality is virtually infinite. Population means also need to be estimated *a priori*. This appears to be feasible in empirical practice, because the range of concentrations that can be expected for metabolites is usually bounded by physiological constraints that can be easily obtained.

If estimating a multivariate effect may seem unfeasible, it is certainly possible to provide a lower bound, by making use of simulations and real experimental data. By this we mean that it is possible to obtain an estimation of what can be considered to be a trivial effect, i.e. biologically nonrelevant.

To set the scene, we start considering a case-control setting where data from both groups is from a multinormal distribution $N(\mathbf{0},\mathbf{I})$ and thus the two groups are completely equivalent. The multivariate effect can be calculated using eq 2; note that, in this case, we set $\mathbf{\Sigma}$ equal to the identity matrix (thus, all variables are uncorrelated and are just the sum of the squares of $\mathbf{w}$). One could be tempted to infer that, given $\mathbf{w}$, the effect of Δ is mimimum for $\mathbf{\Sigma} = \mathbf{I}$ (uncorrelated variables) but it is easy to generate random covariance matrices (with $\mathbf{\Sigma} \neq \mathbf{I}$) for which the corresponding Δ is smaller than the one obtained for $\mathbf{\Sigma} = \mathbf{I}$. By taking different realizations of the two groups, a distribution of effects under the null hypothesis can be created ($\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$). This is shown in the first panel of Figure 8.

The situation $\mathbf{\Sigma} = \mathbf{I}$ may appear rather unrealistic, as biological and *omics* data, in particular, usually show complex correlation patterns; for this reason, it is useful to repeat the same exercise using real data to estimate Δ; $\mathbf{\Sigma}$ is not known, but it can be approximated with the covariance matrix calculated from the data. Figure 8 shows also the dynamic ranges of effect obtained from NMR and MS metabolomics experiments on plasma, serum, and urine, using both quantified metabolite concentrations and buckets. Here the two groups are obtained by random sampling from larger experimental data sets (more details are provided in the figure caption; see the Materials and Methods for the data description). It can be noted that the effective size under the condition of equivalence of the two groups (i.e., the distribution of biologically relevant effects) seems to depend on the platform used (here, MS or NMR), and this should be considered when setting up simulations. It is interesting to note that, on average, trivial effects obtained from real data are not very dissimilar from those obtained from simulated data, indicating that using $\mathbf{\Sigma} = \mathbf{I}$ may be a reasonable choice. However, it should be noted that Cohen's definition of trivial effects (which are defined, we recall, for the univariate case) does not transfer to the multivariate case, at least for what concerns the data explored here (average trivial effects range here from 0.4 to 1.5, which in a behaviorial setting would correspond to medium to very large effects).

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

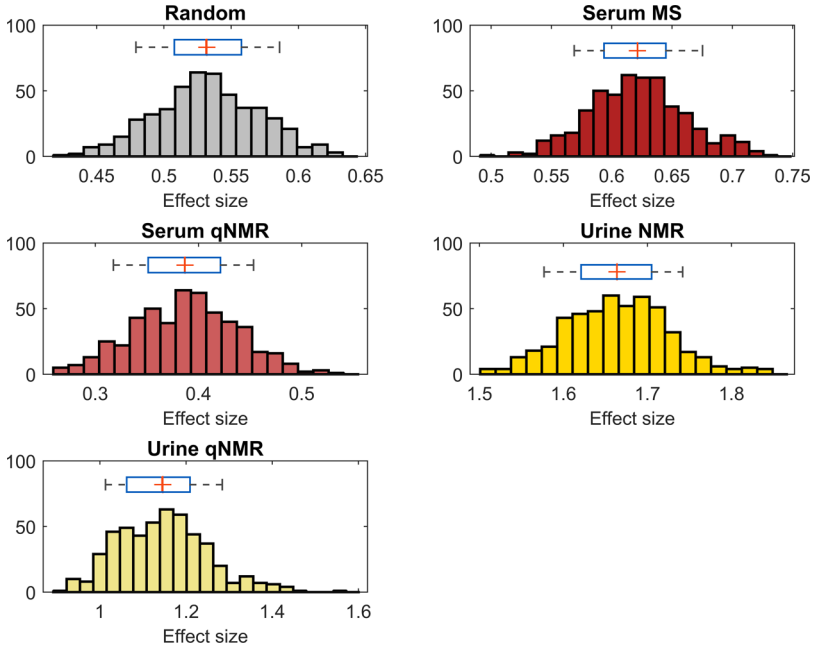


Figure 8. Distributions of biologically nonrelevant effects for simulated (random) and real data. Serum MS (data set D.1), serum NMR quantified (qNMR, data set D.2), urine NMR (data set D.3B, bucketed), and urine qNMR (data set D.4). See Materials and Methods for more details about the data sets. To arrive at a distribution of the nonrelevant effects, the data sets are randomly split into two groups and the multivariate effects are calculated using eq 2. As the two groups are biologically equivalent, the observed difference should be considered biologically nonrelevant.

CONCLUSIONS

In this paper we offered some ideas and suggestions for sample size estimation in a multivariate setting considering principal component analysis and partial least-squares discriminant analysis. The results offered for PCA are grounded by solid statistical characterization of the distributional properties of the PCA solution. It is possible, at a certain point, to treat PCA as an inferential method deriving formulas for sample size estimation or numerical recipes for simulation.

We presented here results obtained on experimental data stemming from metabolomics experiments. However, the methodologies proposed are general and can be applied to a large variety of data. Random matrix theory based methods and the Tracy–Widom limits have been successfully applied in genetics[70] and econometrics[84] and are being discussed in the chemometrics field.[40,66,69,71]

The distributional properties of the PLS-DA solutions are not statistically characterized. Nonetheless, the discrimination problem bears an analogy with classical hypothesis testing. This renders it possible to define power analysis in the PLS-DA context by borrowing the concept of effect from multivariate mean testing. The ideas introduced here regarding PLS-DA warrant further investigation. For instance, it would be interesting to investigate how the sample size and variable to sample ratio may influence the PLS loadings, the regression coefficients, or other measures derived from the PLS models, such as VIPs or sensitivity ratios.

MATHEMATICAL APPENDIX

Definition of Tucker's ϕ

The Tucker's congruence coefficient ϕ[42,85] for two loading vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$\phi = \frac{\sum_i y_i \cdot x_i}{\sqrt{\sum_i y_i^2 \cdot \sum_i x_i^2}} \tag{14}$$

A congruence $\phi > 0.9$ is an accepted value to establish equivalence between two sets of loadings.[43]

Johnstone's theorem

In a ground-breaking paper with the programmatic title *"On the distribution of the largest eigenvalue in principal component analysis"*,[41] Johnstone established a link between RMT and inferential multivariate statistics, demonstrating that the limiting distribution of the largest eigenvalues of large random covariance matrices, when properly centered and scaled, is the Tracy–Widom distribution (see Figure 4, bottom panel for a graphical illustration). This result is summarized by the following fundamental theorem:

Theorem 1 (Johnstone)[41]

Let the entries of the columns X_j of a $n \times p$ matrix $\mathbf{X}$ be identical and independently distributed random variables with Gaussian distribution $N(0,\mathbf{I})$, and let l_1 be the largest eigenvalue of the $n \times p$ sample covariance matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$. Define the centering and scaling parameters

$$\mu_{np} = (\sqrt{n-1} + \sqrt{p})^2$$

$$\sigma_{np} = (\sqrt{n-1} + \sqrt{p})\left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}}\right)^{1/3} \tag{15}$$

If

$$\lim_{(n,p)\to\infty} \frac{n}{p} = \gamma$$

with γ in $(0,1)$, then the statistic

$$L_1 = \frac{l_1 - \mu_{np}}{\sigma_{np}} \tag{16}$$

is distributed like the Tracy–Widom distribution $F_1(s,1)$.

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

This theorem holds true with dimensions going to infinity. However, the Tracy–Widom limit holds also for the finite case[41] (n and p as small as 5), providing an excellent approximation of the distribution of the largest eigenvalue of random covariance matrices. It has also been generalized to the case in which the data matrix entries have an arbitrary symmetric distribution. Further generalizations concern rectangular matrices,[86] different asymptotic behaviour depending on the ratio n/p (and p/n),[87] and non-identity of the population covariance matrix.[72,73,88]

The Tracy–Widom testing procedure

The procedure to determine the number of components in a PCA model introduced by Johnstone[41] consists of a series of nested tests for the null hypothesis

H_0: at least k components

against the alternative

H_1: at most $k - 1$ components

The procedure for assessing the signficance of the k-th component (*i.e.* the k-th sample eigenvalue l_k) is given by

$$l_k > \tau^2(k)[\mu_{n,p-k} + x_{1-\alpha}\sigma_{n,p-k}] \tag{17}$$

where $\mu_{n,p-k}$ and $\sigma_{n,p-k}$ are given by eq 13 and $x_{1-\alpha}$ is the Tracy–Widom percentile value corresponding to an α confidence threshold: common values are $x_{90} = 0.4501$, $x_{95} = 0.9793$, and $x_{99} = 2.0234$. The parameter $\tau^2(k)$ is obtained from the sample eigenvalues of the sample covariance matrix $\mathbf{S}$ (see eq 6)

$$\tau^2(k) = \frac{1}{n(p-k-1)}\sum_{r=k}^{p} l_r \tag{18}$$

Other approaches have been proposed to estimate $\tau^2(k)$; see, for instance, the *KN* methods.[69] If eq 17 is satisfied, the H_0 is not rejected: the procedure is repeated until H_0 is rejected and the estimated number of components K at a significant level α is given by

$$K = \underset{k}{\operatorname{argmin}}\{l_k > \tau^2(k)[\mu_{n,p-k} + x_{1-\alpha}\sigma_{n,p-k}]\} - 1 \tag{19}$$

The Tracy–Widom distribution

In two key papers, Tracy and Widom[67,68] demonstrated that the function

$$F_1(s, 1) = \exp\left(-\frac{1}{2}\int_s^\infty q(x) + (x-s)q^2(x)dx\right) \tag{20}$$

is the limiting distribution of the largest eigenvalue of a certain class of $p \times p$ random matrices (the so called Gaussian orthogonal ensemble, GOE). The function $q(x)$ appearing in eq 20 is the unique Hastings–McLeod solution[89] of the nonlinear Painlevé differential equation

$$\frac{d^2}{dt^2}q(x) = tq(x) + 2q^3(x) \tag{21}$$

satisfying the boundary condition $q(x) \sim Ai(x)$ when $x \to \infty$ and $Ai(x)$ is the Airy function.[90] The distribution $F_1(s,1)$ is the now-called Tracy–Widom distribution (see Figure 1, panel A). $F_1(s,1)$ was found by Johnstone (see Theorem 1) to be the limiting distribution of the largest eigenvalues of random covariance matrices.

AUTHOR INFORMATION

Corresponding Author

*E-mail: esaccenti@gmail.com; Tel: +31 (0)317 482018; Fax: +31 (0) 3174 83829.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was partly supported by the European Commission-funded FP7 project INFECT (Contract No. 305340). We thank Leonardo Tenori at University of Florence and Megan E. Romano at Brown University for the fruitful comments on the manuscript.

NOTATION

A, matrix (**bold** uppercase); **a**, column vector (**bold** lowercase); *a*, scalar (*italic* lower case); **Σ**, population covariance matrix; **S**, sample covariance matrix; **Π**, population loading matrix; **I**, identity matrix (1's on the diagonal, 0's otherwise); λ_i, population eigenvalue associated with the i-th component; l_i, sample eigenvalue associated with the i-th component; L_i, Tracy–Widom statistic; n, number of samples (observations, objects); p, number of variables

REFERENCES

(1) Saccenti, E.; Hoefsloot, H. C.; Smilde, A. K.; Westerhuis, J. A.; Hendriks, M. M. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* **2014**, *10*, 361–374.

(2) Trygg, J.; Gullberg, J.; Johansson, A.; Jonsson, P.; Moritz, T. Chemometrics in metabolomics—an introduction. In *Plant metabolomics*; Springer: 2006; pp 117–128.

(3) Kjeldahl, K.; Bro, R. Some common misunderstandings in chemometrics. *J. Chemom.* **2010**, *24*, 558–564.

(4) Worley, B.; Powers, R. Multivariate analysis in metabolomics. *Curr. Metabolomics* **2013**, *1*, 92–107.

(5) Hwang, D.; Schmitt, W.; Stephanopoulos, G.; Stephanopoulos, G. Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics* **2002**, *18*, 1184–1193.

(6) Pawitan, Y.; Michiels, S.; Koscielny, A.; Gusnanto, S.; Ploner, A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* **2005**, *21*, 3017–24.

(7) Li, S.; Bigler, J.; Lampe, J.; Potter, J.; Feng, Z. Fdr-controlling testing procedures and sample size determination for microarrays. *Stat. Med.* **2005**, *24*, 2267–2280.

(8) Muller, P.; Parmigiani, G.; Robert, C.; Rousseau, J. Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Am. Stat. Assoc.* **2004**, *99*, 990–1001.

(9) Jung, S.-H.; Young, S. S. Power and sample size calculation for microarray studies. *Journal of Biopharmaceutical Statistics* **2012**, *22*, 30–42.

(10) Lin, W.; Hsueh, H.; Chen, J. Power and sample size estimation in microarray studies. *BMC Bioinf.* **2010**, *11*, 48.

(11) Lee, M.-L.; Whitmore, G. Power and sample size for microarray studies. *Stat. Med.* **2002**, *21*, 3543–3570.

(12) Ioannidis, J. P. Why most published research findings are false. *PLoS medicine* **2005**, *2*, e124.

(13) Eng, J. Sample Size Estimation: How Many Individuals Should Be Studied? 1. *Radiology* **2003**, *227*, 309–313.

(14) Moher, D.; Dulberg, C. S.; Wells, G. A. Statistical power, sample size, and their reporting in randomized controlled trials. *Jama* **1994**, *272*, 122–124.

(15) Freiman, J. A.; Chalmers, T. C.; Smith, H., Jr; Kuebler, R. R. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71″ negative″ trials. *N. Engl. J. Med.* **1978**, *299*, 690–694.

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379–2393

(16) Button, K. S.; Ioannidis, J. P.; Mokrysz, C.; Nosek, B. A.; Flint, J.; Robinson, E. S.; Munafò, M. R. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **2013**, *14*, 365−376.
(17) Ransohoff, D. F.; Gourlay, M. L. Sources of Bias in Specimens for Research About Molecular Markers for Cancer. *J. Clin. Oncol.* **2010**, *28*, 698−704.
(18) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S. MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Res.* **2015**, *43*, W251.
(19) MacCallum, R. C.; Widaman, K. F.; Zhang, S.; Hong, S. Sample size in factor analysis. *Psychological methods* **1999**, *4*, 84.
(20) Tanaka, J. S. ″ How big is big enough?″: Sample size and goodness of fit in structural equation models with latent variables. *Child Dev.* **1987**, *58*, 134−146.
(21) Jung, S.; Lee, S. Exploratory factor analysis for small samples. *Behavior research methods* **2011**, *43*, 701−709.
(22) Magnusson, P. K.; Almqvist, C.; Rahman, I.; Ganna, A.; Viktorin, A.; Walum, H.; Halldner, L.; Lundström, S.; Ullén, F.; Långström, N. The Swedish Twin Registry: establishment of a biobank and other recent developments. *Twin Res. Hum. Genet.* **2013**, *16*, 317−329.
(23) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; Maguire, E.; González-Beltrán, A.; Sansone, S.-A.; Griffin, J. L.; Steinbeck, C. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **2013**, *41*, D781−D786.
(24) Fall, T.; Lee, W.; Hagg, S.; Magnusson, P. K.; Prenni, J.; Lind, L.; Pawitan, Y.; Ingelsson, E. A workflow for UPLC-MS non-targeted metabolomic profiling in large human population-based studies. *bioRxiv* **2014**, 002782.
(25) Saccenti, E.; Suarez-Diez, M.; Luchinat, C.; Santucci, C.; Tenori, L. Probabilistic Networks of Blood Metabolites in Healthy Subjects As Indicators of Latent Cardiovascular Risk. *J. Proteome Res.* **2015**, *14*, 1101−1111.
(26) Bernini, P.; Bertini, I.; Luchinat, C.; Tenori, L.; Tognaccini, A. The Cardiovascular Risk of Healthy Individuals Studied by NMR Metabonomics of Plasma Samples. *J. Proteome Res.* **2011**, *10*, 4983−4992.
(27) Lusczek, E.; Lexcen, D.; Witowski, N.; Mulier, K.; Beilman, G. *Metabolomics* **2013**, *9*, 223−235.
(28) Bernini, P.; Bertini, I.; Luchinat, C.; Nepi, S.; Saccenti, E.; Scha fer, H.; Schu tz, B.; Spraul, M.; Tenori, L. Individual human phenotypes in metabolic space and time. *J. Proteome Res.* **2009**, *8*, 4264−4271.
(29) Assfalg, M.; Bertini, I.; Colangiuli, D.; Luchinat, C.; Schafer, H.; Schutz, B.; Spraul, M. Evidence of different metabolic phenotypes in humans. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 1420−4.
(30) Van Den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; Van Der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006**, *7*, 142.
(31) Szymańska, E.; Saccenti, E.; Smilde, A. K.; Westerhuis, J. A. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **2012**, *8*, 3−16.
(32) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. Assessment of PLSDA cross validation. *Metabolomics* **2008**, *4*, 81−89.
(33) Faul, F.; Erdfelder, E.; Lang, A.-G.; Buchner, A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* **2007**, *39*, 175−191.
(34) Faul, F.; Erdfelder, E.; Buchner, A.; Lang, A.-G. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* **2009**, *41*, 1149−1160.
(35) Kirk, R. E. Practical significance: A concept whose time has come. *Educ. Psychol. Meas.* **1996**, *56*, 746−759.
(36) Nakagawa, S.; Cuthill, I. C. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews* **2007**, *82*, 591−605.
(37) Cohen, J. *Statistical power analysis for the behavioral sciences*; Academic Press: 2013.
(38) Lipsey, M. W.; Wilson, D. B. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *Am. Psychol.* **1993**, *48*, 1181.
(39) Friston, K. Ten ironic rules for non-statistical reviewers. *NeuroImage* **2012**, *61*, 1300−1310.
(40) Bertini, I.; Luchinat, C.; Miniati, M.; Monti, S.; Tenori, L. Phenotyping COPD by 1H NMR metabolomics of exhaled breath condensate. *Metabolomics* **2014**, *10*, 302−311.
(41) Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics* **2001**, *29*, 295−327.
(42) Tucker, L. R. *A method for synthesis of factor analysis studies*; Department of the Army: 1951.
(43) Lorenzo-Seva, U.; ten Berge, J. M. F. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* **2006**, *2*, 57−64.
(44) Osborne, J. W.; Costello, A. B. Sample size and subject to item ratio in principal components analysis. *Practical assessment, research & evaluation* **2004**, *9*, 8.
(45) Ogasawara, H. Standard errors of the principal component loadings for unstandardized and standardized variables. *British Journal of Mathematical and Statistical Psychology* **2000**, *53*, 155−174.
(46) Ogasawara, H. Concise formulas for the standard errors of component loading estimates. *Psychometrika* **2002**, *67*, 289−297.
(47) Ogasawara, H. Asymptotic biases of the unrotated/rotated solutions in principal component analysis. *British Journal of Mathematical and Statistical Psychology* **2004**, *57*, 353−376.
(48) Timmerman, M. E.; Kiers, H. A. L.; Smilde, A. K. Estimating confidence intervals for principal component loadings: A comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology* **2007**, *60*, 295−314.
(49) Timmerman, M. E.; Kiers, H. A. L.; Smilde, A. K.; Ceulemans, E.; Stouten, J. Bootstrap confidence intervals in multi-level simultaneous component analysis. *British Journal of Mathematical and Statistical Psychology* **2009**, *62*, 299−318.
(50) Saccenti, E.; Tenori, L.; Verbruggen, P.; Timmerman, M. E.; Bouwman, J.; van der Greef, J.; Luchinat, C.; Smilde, A. K. Of Monkeys and Men: A Metabolomic Analysis of Static and Dynamic Urinary Metabolic Phenotypes in Two Species. *PLoS One* **2014**, *9*, e106077.
(51) Timmerman, M. E. Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology* **2006**, *59*, 301−320.
(52) Jansen, J. J.; Hoefsloot, H. C. J.; van der Greef, J.; Timmerman, M. E.; Smilde, A. K. Multilevel component analysis of time-resolved metabolic fingerprinting data. *Anal. Chim. Acta* **2005**, *530*, 173−183.
(53) Schafer, J.; Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*.10.2202/1544-6115.1175
(54) Dahmen, J.; Keysers, D.; Pitz, M.; Ney, H., Structured covariance matrices for statistical image object recognition. In *Mustererkennung 2000*; Springer: 2000; pp 99−106.
(55) MacCallum, R. C.; Browne, M. W.; Sugawara, H. M. Power analysis and determination of sample size for covariance structure modeling. *Psychological methods* **1996**, *1*, 130.
(56) Suarez-Diez, M.; Saccenti, E. Effects of sample size and dimensionality on the performance of four algorithms for inference of association networks in metabonomics. *J. Proteome Res.* **2015**, *14*, 5119.
(57) Vershynin, R. How Close is the Sample Covariance Matrix to the Actual Covariance Matrix? *Journal of Theoretical Probability* **2012**, *25*, 655−686.
(58) Adamczak, R.; Litvak, A.; Pajor, A.; Tomczak-Jaegermann, N. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society* **2010**, *23*, 535−561.
(59) Rudelson, M. Random vectors in the isotropic position. *Journal of Functional Analysis* **1999**, *164*, 60−72.
(60) Gupta, P. L.; Gupta, R. D. Sample size determination in estimating a covariance matrix. *Comput. Stat. Data Anal.* **1987**, *5*, 185−192.

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393

(61) Rothman, A. J.; Levina, E.; Zhu, J. Generalized thresholding of large covariance matrices. *J. Am. Stat. Assoc.* **2009**, *104*, 177−186.
(62) Bien, J.; Tibshirani, R. J. Sparse estimation of a covariance matrix. *Biometrika* **2011**, *98*, 807−820.
(63) Peres-Neto, P. R.; Jackson, D. A.; Somers, K. M. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.* **2005**, *49*, 974−997.
(64) Bro, R.; Kjeldahl, K.; Smilde, A.; Kiers, H. Cross-validation of component models: A critical look at current methods. *Anal. Bioanal. Chem.* **2008**, *390*, 1241−1251.
(65) Jackson, D. A. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **1993**, *74*, 2204−2214.
(66) Saccenti, E.; Camacho, J. Determining the number of components in principal components analysis: A comparison of statistical, cross-validation and approximated methods. *Chemom. Intell. Lab. Syst.* **2015**, *149* (Part A), 99−116.
(67) Tracy, C. A.; Widom, H. On orthogonal and symplectic matrix ensembles. *Commun. Math. Phys.* **1996**, *177*, 727−754.
(68) Tracy, C. A.; Widom, H. Level-spacing distributions and the Airy kernel. *Commun. Math. Phys.* **1994**, *159*, 151−174.
(69) Kritchman, S.; Nadler, B. Determining the number of components in a factor model from limited noisy data. *Chemom. Intell. Lab. Syst.* **2008**, *94*, 19−32.
(70) Patterson, N.; Price, A. L.; Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2006**, *2*, e190.
(71) Saccenti, E.; Smilde, A. K.; Westerhuis, J. A.; Hendriks, M. M. W. B. Tracy−Widom statistic for the largest eigenvalue of autoscaled real matrices. *J. Chemom.* **2011**, *25*, 644−652.
(72) Baik, J.; Ben Arous, G.; Péché, S. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability* **2005**, *33*, 1643−1697.
(73) Baik, J.; Silverstein, J. W. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis* **2006**, *97*, 1382−1408.
(74) Paul, D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* **2007**, *17*, 1617.
(75) Wold, S.; Sjöström, M.; Eriksson, L. Partial least squares projections to latent structures (PLS) in chemistry. *Encyclopedia of computational chemistry* **1998**, DOI: 10.1002/0470845015.cpa012.
(76) Smit, S.; van Breemen, M. J.; Hoefsloot, H. C. J.; Smilde, A. K.; Aerts, J.; de Koster, C. G. Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta* **2007**, *592*, 210−217.
(77) Rosipal, R.; Trejo, L. J. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research* **2002**, *2*, 97−123.
(78) Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemom.* **2006**, *20*, 341−351.
(79) van Velzen, E. J. J.; Westerhuis, J. A.; van Duynhoven, J. P. M.; van Dorsten, F. A.; Hoefsloot, H. C. J.; Jacobs, D. M.; Smit, S.; Draijer, R.; Kroner, C. I.; Smilde, A. K. Multilevel data analysis of a crossover designed human nutritional intervention study. *J. Proteome Res.* **2008**, *7*, 4483−4491.
(80) Westerhuis, J. A.; van Velzen, E. J. J.; Hoefsloot, H. C. J.; Smilde, A. K. Discriminant Q 2 (DQ 2) for improved discrimination in PLSDA models. *Metabolomics* **2008**, *4*, 293−296.
(81) Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119−128.
(82) Bro, R. Multiway calibration. Multilinear PLS. *J. Chemom.* **1996**, *10*, 47−61.
(83) Browner, W. S.; Newman, T. B. Are all significant p values created equal?: The analogy between diagnostic tests and clinical research. *JAMA* **1987**, *257*, 2459−2463.
(84) Harding, M. C. Explaining the single factor bias of arbitrage pricing models in finite samples. *Economics Letters* **2008**, *99*, 85−88.
(85) Burt, C. Factor analysis and canonical correlations. *British Journal of Statistical Psychology* **1948**, *1*, 95−106.
(86) Soshnikov, A. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J. Stat. Phys.* **2002**, *108*, 1033−1056.
(87) Karoui, N. E. On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. *arXiv preprint math/0309355*; 2003.
(88) El Karoui, N. Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Annals of Probability* **2007**, *35*, 663−714.
(89) Hastings, S.; McLeod, J. A boundary value problem associated with the second Painlevé transcendent and the Korteweg-de Vries equation. *Arch. Ration. Mech. Anal.* **1980**, *73*, 31−51.
(90) Airy, G. B. On the intensity of light in the neighbourhood of a caustic. *Transactions of the Cambridge Philosophical Society* **1838**, *6*, 379.
(91) Paul, D.; Aue, A. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference* **2014**, *150*, 1.

DOI: 10.1021/acs.jproteome.5b01029
J. Proteome Res. 2016, 15, 2379−2393