Discriminability and uncertainty in principal component analysis (PCA) of temporal check-all-that-apply (TCATA) data

J.C. Castura[1*], D.N. Rutledge[2,3], C.F. Ross[4], T. Næs[5,6]

[1] Compusense Inc., 255 Speedvale Ave. W., Guelph, Ontario, N1H 1C5, Canada

[2] Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 75005 Paris, France

[3] National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, Australia

[4] School of Food Science, Washington State University, Pullman, WA 99164-6376, USA

[5] Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway

[6] Dept. of Food Science, Faculty of Sciences, University of Copenhagen, Rolighetsvej 30, 1958 Fredriksberg, Copenhagen, Denmark.

[*] jcastura@compusense.com

Abstract

Temporal check-all-that-apply (TCATA) data can be summarized and explored using principal component analysis (PCA). Here we analyze TCATA data on Syrah wines obtained from a trained sensory panel. We evaluate new and existing methods to explore the uncertainty in the PCA scores. To do so, we use the bootstrap procedure to obtain many virtual panels from the real panel's data. Virtual-panel PCA scores are obtained using two methods. The first method, called the partial bootstrap (PB), obtains virtual-panel scores from regression. The second method, called the truncated total bootstrap (TTB), applies PCA to the virtual-panel results to obtain scores, which are truncated and superimposed on the real-panel scores by Procrustes rotation. We use the virtual scores from each method to investigate uncertainty in the real-panel PCA scores visually and numerically. To understand the uncertainty of the scores, we obtain confidence ellipses (CEs) and their areas, as well as confidence intervals (CIs) and their widths. Next, to determine whether PCA scores for different samples are well separated, we propose a procedure for approximating the standard errors of sample differences and correcting for multiple comparisons. We propose a discriminability index, and show that it can enhance the interpretability of PCA results. We incorporate graphical features into our PCA biplots to visualize discriminability. We did

PREPRINT – see the published manuscript in *Food Quality and Preference* for all quotes and citations
https://doi.org/10.1016/j.foodqual.2021.104370

not find a large difference between the PB and TTB methods for understanding the uncertainty and discriminability in PCA scores. Although the TCATA data that we analyzed have a special structure, the methodological approaches presented here can be readily adapted to other applications of PCA.

Keywords: principal component analysis (PCA); temporal check-all-that-apply (TCATA); bootstrap; discriminability; uncertainty; wine.

1. Introduction

Multivariate sensory data are often summarized using principal component analysis (PCA). PCA is applied routinely to data from sensory descriptive studies (Lawless & Heymann, 2010), including data from temporal sensory studies based on time-intensity (Dijksterhuis, 1997), temporal dominance of sensations (Lenfant, Loret, Pineau, Hartmann, & Martin, 2009), and temporal check-all-that-apply (TCATA; Castura, Baker & Ross, 2016b). However, such PCA results contain uncertainty because the data are subject to natural variation. Potential sources of variability include individual differences of the assessors, variability of the samples, and other systematic or transitory factors. This paper proposes and compares methods for visualizing the uncertainty in PCA scores. We also propose an index to evaluate the discriminability of PCA scores (see Section 2.5). We apply and compare methods on the "Syrah data" (Baker et al., 2016), which are TCATA results from a trained sensory panel. We will call the original data set the "real panel". Communicating uncertainty along with PCA results could avoid the over-interpretation of spurious results and the generation of erroneous hypotheses.

Previously, uncertainty in PCA scores has been investigated using Hotelling's T^2 confidence regions (Johnson & Wichern, 2007, p. 459-465) or bootstrap procedures (Efron & Tibshirani, 1994), which in this context involves creating many virtual panels using only the data from the real panel (e.g., Josse, Wager & Husson, 2016; Babamoradi, van den Berg & Rinnan, 2013; Timmerman, Kiers & Smilde, 2007). One way to obtain virtual-panel scores is to use the "partial bootstrap" (henceforth abbreviated "PB"; Greenacre, 2007, p. 250-252; Lebart, 2007; Husson, Lê, & Pagès, 2005) method, which uses the real-panel loadings to map the virtual-panel results onto the real panel's principal components (PCs). Confidence ellipses (CEs) and confidence intervals (CIs) are then constructed from the virtual panels' scores.

Cadoret and Husson (2013) investigated uncertainty in various sensory data sets using multivariate analyses (not including PCA). They considered sensory quantitative-descriptive (QD) data, in which samples are evaluated by identifying and indicating the intensities of sensory attributes, and holistic sensory data, in which each sample is considered as a whole, not a collection of individual attributes. They showed under simulation that the PB method produced CEs that had acceptable coverage of true parameters for sensory QD results, but produced CEs that were too small and too well separated for holistic sensory data, including for unstructured data (product labels were permuted) whose CEs should be very large. For both QD and holistic sensory data, they recommend a different way to obtain scores from the virtual-panel results: the "truncated total bootstrap" (henceforth abbreviated "TTB"; Cadoret

& Husson, 2013) method, which was first used to investigate uncertainty in holistic sensometric data (Courcoux, Qannari, Taylor, Buck & Greenhoff, 2012). In the TTB method, each virtual panel's results are submitted to multivariate analysis, the virtual panel's scores matrix is truncated, then superimposed on the real panel's truncated scores matrix by Procrustes rotation (Schönemann, 1966). The TTB-derived scores are obtained through multivariate analyses and Procrustes rotations, not from a simple function of the real-panel loadings, as the PB-derived scores are. Cadoret and Husson (2013) recommend using the TTB method, but they found the PB method to be adequate for investigating uncertainty in most QD data sets.

TCATA is a type of temporal sensory method that characterizes products according to sensory attributes, thus it can be considered to yield a type of QD data. Such temporal sensory data sets are often organized to have sensory attributes in columns and a compound structure of products and times in rows. Rows are often highly correlated because the same products are evaluated repeatedly over time. Even though both TTB and PB methods have each been used and compared in sensory evaluation (Cadoret & Husson, 2013), they have never been applied and evaluated in the context of temporal sensory data, of which TCATA is a special case. This will be the topic of the present paper. We improve on a previously published procedure of resampling TCATA data (Castura et al., 2016b). In addition to presenting graphical ways of evaluating the methods, we propose an index to quantify discriminability for each component separately.

It is not possible to determine, based on general principles, which of the two methods is to be preferred in this context. Therefore, an empirical approach is used. In Section 2, we describe the TCATA data set considered in this manuscript (Section 2.1). We describe how we compose virtual panels via multilevel resampling (Section 2.2). A description of PCA is given in Section 2.3. We describe the PB method and the TTB method, then outline how uncertainty in PCA scores is explored (Section 2.4). Next, we introduce an index for investigating discriminability (Section 2.5) and describe how statistical analyses are performed (Section 2.6). Then we present and discuss results (Sections 3 and 4) before making conclusions. Although TCATA data sets have a specific structure, the insights gained from our investigations have broader relevance.

2. Materials and Methods

Fig. 1 provides a visual overview of the methodological approach. Each subsection of Section 2 provides the rationale and relevant details on the methodology.

2.1. The "Syrah Data"

To empirically compare the performance of the PB and the TTB for investigating the uncertainty in PCA scores in the context of TCATA data, we used the "Syrah data" set. The Syrah data (Fig. 1) were first described by Baker et al. (2016) and have also been analyzed by Castura et al. (2016b) and Meyners and Castura (2018). The data arose from a trained panel ($n = 13$) that evaluated a high-ethanol wine ("H"), a low-ethanol wine ("L"), and a low-ethanol wine that was adjusted to have the same ethanol content as

the high-ethanol wine ("A"). Assessors were trained to characterize wine samples over time using the TCATA method (Castura, Antúnez, Giménez & Ares, 2016a). Their task was to interact continuously with a list of sensory attributes so that at each moment the attributes that did (not) describe the sensations perceived were (not) selected. Assessors were instructed to check any attribute that is unselected if it characterizes the sample at that moment, and to uncheck any attribute that is selected but no longer characterizes the sample. In this study, the list contained attributes related to taste (*bitter, sour*), flavour (*dark fruit*, *earthy*, *green, red fruit*, *spices*), and mouthfeel/texture (*astringent*, *heat / ethanol burn*), as well as a catch-all attribute (*other*). Attributes were organized into lists using a Williams Latin-square design then attribute lists were allocated to assessors for the study (Meyners & Castura, 2016). Data were captured using Compusense *at-hand* (since renamed to Compusense Cloud; Compusense Inc., Guelph, Canada).

A two-sip evaluation protocol was used (Fig. 2). Each combination of wine (H, L, A) and sip number (1, 2) is called a WineSip (H1, H2, L1, L2, A1, A2). Sip 1 was taken simultaneously with clicking *Start*. At 10 s, an onscreen prompt cued the assessor to expectorate and begin characterizing the wine finish. The evaluation of Sip 1 stopped if the assessor clicked *Stop* (indicating they no longer perceived any attributes) or if the timer reached 180 s. After a one-minute break (without palate cleansing), Sip 2 was evaluated using the same protocol as Sip 1. A two-minute break (with palate cleansing) was enforced between wine samples. Differences between Sip 1 and Sip 2 thus provide relevant information about potential carry-over effects related to astringency and other sensations (Ross, Hinken, & Weller, 2007; Colonna, Adams & Noble, 2004). For each wine treatment, there were four replicates of this procedure split across two testing occasions. Within each replicate, a Williams Latin-square design was used. Each assessor got a different sample presentation order in each replicate. Overall, each assessor evaluated wine treatment H four times, wine treatment A four times, and wine treatment L four times. Since each wine sample was evaluated using the two-sip evaluation protocol, each assessor contributed four evaluations of each of H1, H2, A1, A2, L1, and L2.

Only the nine sensory attributes were included in the subsequent analyses. The attribute *other*, which was endorsed by assessors only rarely, was dropped entirely. The reason for dropping the attribute *other* is that it was qualitatively different: it was not a sensory attribute but rather it served as a catch-all term to allow assessors to describe perceptions that were not included on the ballot. When evaluating the low-ethanol wine, there was one evaluation in which an assessor clicked *Start* for the second sip (of one replicate) but did not select any attributes. The result was recorded as being zero. To maintain focus on the methods for investigating uncertainty, data from this replicate were analyzed exactly as recorded.

Fig. 1. Overview of how results from the real panel (top left) and the virtual panels were aggregated and analyzed in the PB and TTB methods. Some assessors are included on the virtual panel (top right) more than once, and not always with the same replicates.

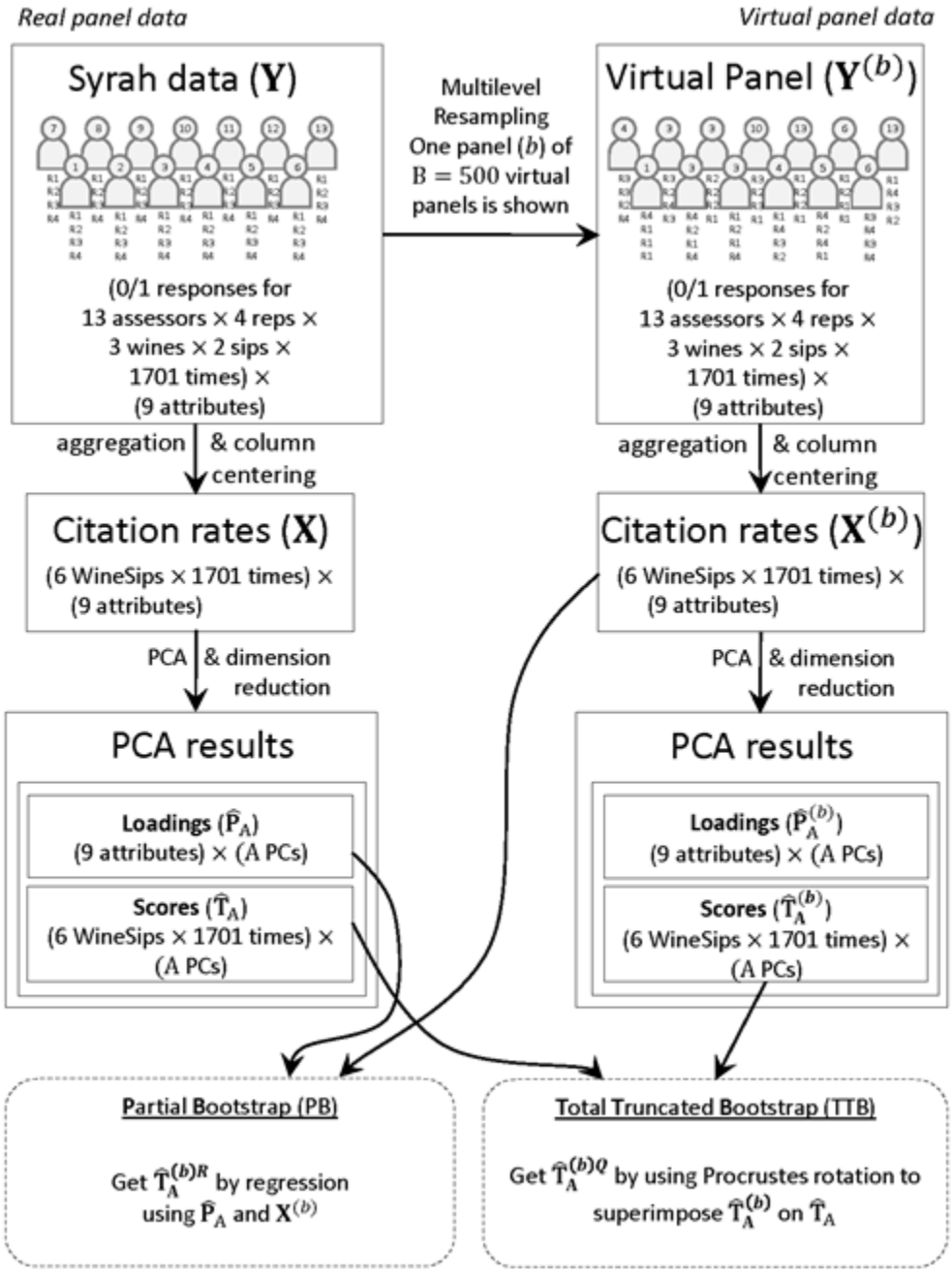
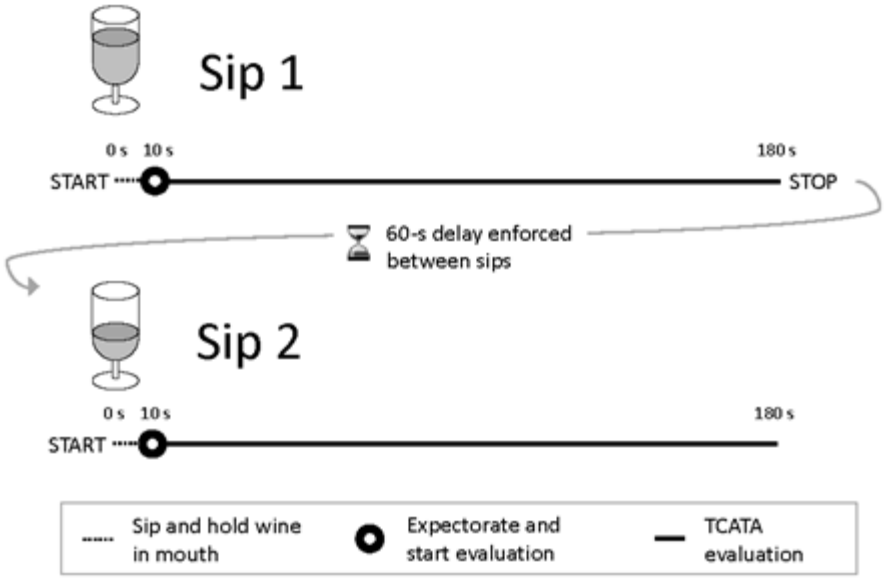


Fig. 2. Each TCATA evaluation of a wine sample consisted of two sips, as shown.



The raw Syrah data ($\mathbf{Y}$) is a multidimensional array of 3 wine treatments, 2 sips, 13 assessors, 4 replicates, 1701 time points (from 10.0 s to 180.0 s at 0.1-s increments, also called "time slices"), and 9 sensory attributes. Raw Syrah data ($\mathbf{Y}$) were aggregated over assessors and replicates. Attributes were mean centered. This produced a matrix of column-centered citation rates ($\mathbf{X}$) with 10206 rows (6 WineSips × 1701 time slices) and 9 columns (sensory attributes). The data set `syrah` in the R package `tempR` (Castura, 2020) gives an aggregation of the Syrah data; complete data are publicly available (Baker et al., 2019).

2.2. Multilevel (assessor-replicate) resampling

Bootstrap procedures were used to investigate the uncertainty in PCA results by means of $\mathrm{B} = 500$ virtual panels. First, assessors were sampled with replacement. At each instance that an assessor was selected, that assessor's four replicates were resampled with replacement such that both the real and virtual assessor had four replicates. This multilevel resampling from $\mathbf{Y}$ maintains the structure of the raw data set while capturing the variability that exists among the assessors and the assessors' replicates. The raw data for virtual panel b, denoted $\mathbf{Y}^{(b)}$, was then aggregated across assessors and replicates to get citation rates, which were column centered to obtain $\mathbf{X}^{(b)}$, which has the identical dimensions as $\mathbf{X}$.

A particular virtual panel is presented in Fig. 1 (top right). Note that some assessors were added to this virtual panel one or more times, whereas other assessors are absent from this virtual panel. For example, one instance of assessor 4 is accompanied by replicates R3, R3, R2, and R4, whereas another instance of assessor 4 is accompanied by replicates R1, R4, R3, and R2. To match the constraints of the

original experimental design, further resampling was not done within replicates: when replicate R2 is selected, then data from replicate R2 are used for all six WineSips (H1, H2, L1, L2, A1, and A2).

2.3. Principal Component Analysis & dimension reduction

Singular value decomposition (SVD; see Mardia et al., 1979, p. 473-474; Johnson & Wichern, 2007, p. 100-102) of the real panel's attribute mean-centered (i.e., column-centered) citation rates matrix can be written as

$$\mathbf{X} = \hat{\mathbf{S}}\hat{\mathbf{D}}\hat{\mathbf{P}}^{\mathrm{T}}, \qquad \text{Eq. (1)}$$

where $\hat{\mathbf{S}}$ and $\hat{\mathbf{P}}$ are the left and right singular vectors ($\hat{\mathbf{S}}^{\mathrm{T}}\hat{\mathbf{S}} = \mathbf{I}$ and $\hat{\mathbf{P}}^{\mathrm{T}}\hat{\mathbf{P}} = \mathbf{I}$) and $\hat{\mathbf{D}}$ is the diagonal matrix of singular values. Columns of $\hat{\mathbf{S}}$ are the eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and columns of $\hat{\mathbf{P}}$ are the eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. The SVD algorithm used in PCA (Næs, Brockhoff, & Tomic, 2010; Mardia et al., 1979; Johnson & Wichern, 2007; Legendre & Legendre, 2012) transforms data to a new coordinate system in which the first few PCs extract most of the variability in $\mathbf{X}$. In the field of sensory evaluation, it is conventional to take $\hat{\mathbf{S}}$ and $\hat{\mathbf{D}}$ together as component scores, which allows Eq. (1) to be rewritten

$$\mathbf{X} = \hat{\mathbf{T}}\hat{\mathbf{P}}^{\mathrm{T}}. \qquad \text{Eq. (2)}$$

$\hat{\mathbf{T}}$ is the scores matrix; the PCs are in columns. $\hat{\mathbf{P}}$ is the loadings matrix; its columns correspond to the PCs. The eigenvalues can be obtained from the diagonal of the matrix $\hat{\mathbf{T}}^{\mathrm{T}}\hat{\mathbf{T}} = \hat{\mathbf{D}}^2$. An eigenvalue corresponds to the sum of squares extracted in its PC. The variance accounted for (VAF) in a PC is obtained by dividing the PC's eigenvalue by $n - 1$, where n is the number of rows in $\mathbf{X}$. The sum of the eigenvalues (the trace of $\hat{\mathbf{D}}^2$) equals the total sum of squares in $\mathbf{X}$ (the trace of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$). The total variance is the sum of the variances for all columns in $\mathbf{X}$; it equals the sum of VAFs across all the PCs. The percentage of VAF (%VAF) by a PC is obtained by dividing its variance by the total variance, or by dividing the PC's eigenvalue by the sum of the eigenvalues.

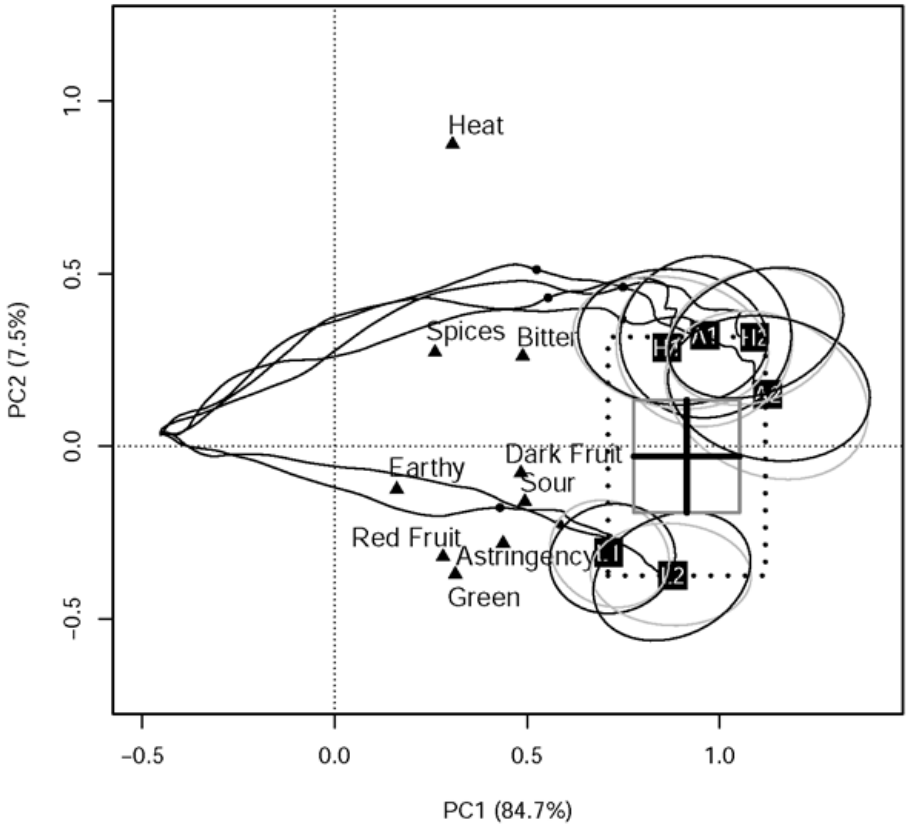
After PCA, we selected four PCs, which account for more than 98% of the variance in $\mathbf{X}$ (Table 1). Using four PCs allows us to investigate uncertainty in three PCs that account for relatively large proportions of variability as well PC4, which accounts for a relatively low proportion of variability. After truncation to $\mathrm{A} = 4$ PCs, the scores and loadings matrices were denoted $\hat{\mathbf{T}}_{\mathrm{A}}$ and $\hat{\mathbf{P}}_{\mathrm{A}}$, respectively.

2.4. Graphical/visual presentation of uncertainty of PCA scores in a biplot

The real panel's PCA results were truncated to four PCs then visualized in two PCA distance biplots (Legendre & Legendre, 2012), which is sometimes called a JK biplot. Scores from adjacent times from each of the six WineSip were adjoined and smoothed (Castura et al., 2016a). The resulting curves are called WineSip trajectories. WineSip trajectories can be seen in Suppl. Video 1. Fig. 3 shows the WineSip trajectories at 30.0 s. To understand how the six WineSips are perceived over time with respect to the nine sensory attributes, the reader can imagine projections of each of the six WineSip scores onto the

nine attribute loading vectors (emanating from the origin) in each biplot in Suppl. Video 1. The uncertainty of the scores were explored by constructing confidence ellipses and confidence intervals on the basis of the PB and TTB methods. Suppl. Video 1 and Fig. 3 contain additional features related to WineSip discriminability that will be discussed in Section 3.2.1.

Fig. 3. Smoothed trajectories for the six WineSips (H1, H2, A1, A2, L1, L2) in the PC1 vs. PC2 plane are shown up to 30.0 s. Each WineSip is overlaid with its PB-derived 95% CE (grey line) and its TTB-derived 95% CE (black line). The evaluation starts at 10 s (far left). The rate of sensory change is indicated by dots that appear along the trajectories at 20.0 s. The Range box (dotted line) and h-cross (thick solid lines in a thin grey box) show that WineSips are discriminated in both PCs.



2.4.1. Obtaining virtual-panel PCA scores

2.4.1.1. Partial bootstrap ("PB") – The attribute-centered citation rates from the virtual panels were mapped onto the PCs from the real panel using the real panel's truncated loadings matrix, which contain coefficients which are fixed. Specifically, virtual-panel scores in the first four PCs were obtained as estimates from an ordinary least squares regression:

$$\widehat{\mathbf{T}}_{\mathrm{A}}^{(b)R} = \mathbf{X}^{(b)}\widehat{\mathbf{P}}_{\mathrm{A}}. \qquad \text{Eq. (3)}$$

Since these virtual-panel scores are obtained from a linear function that is based on the real-panel loadings, no further reordering or reflection of axes (PCs) is required to enable direct comparison between the virtual-panel scores and the real-panel scores.

2.4.1.2. Truncated total bootstrap ("TTB") – Each virtual panel's attribute-centered citation rates ($\mathbf{X}^{(b)}$) were submitted to PCA (Section 2.3). The scores and loadings matrices were truncated to $A = 4$ PCs. Since PCA results are arbitrary with respect to direction of axes/components and since eigenvalues of two axes may be quite similar (switching their importance in the two panels), one has to make a choice on how to compare the virtual-panel scores with the true-panel scores. Some sort of adjustment is needed to make the comparison relevant. In this paper, we align the configuration of virtual-panel scores to the real-panel scores using Procrustes rotation (see Lebart, 2007, Total Bootstrap, Type 3), in the way described by Cadoret and Husson (2013). Specifically, each virtual panel's truncated PCA scores ($\widehat{\mathbf{T}}_{\mathrm{A}}^{(b)}$) were superimposed onto the real panel's truncated PCA scores by finding the Procrustes rotation matrix ($\widehat{\mathbf{Q}}_{\mathrm{A}}^{(b)}$). We refer to the virtual panel's four-component rotated scores matrix as $\widehat{\mathbf{T}}_{\mathrm{A}}^{(b)Q}$. Since we retain four PCs, Procrustes rotation can resolve mixing of virtual-panel scores in the first four PCs that correspond to the real-panel scores in the first four PCs. Any virtual-panel scores in subsequent PCs that correspond to the real-panel scores in PC1 to PC4 are lost to truncation, so cannot be resolved by Procrustes rotation. The reason that scores are truncated prior to Procrustes rotation is to avoid superimpositions that fit overly well (Cadoret & Husson, 2013).

2.4.2. Determining confidence intervals and confidence ellipses

We used the PB-derived scores ($\widehat{\mathbf{T}}_{\mathrm{A}}^{(b)R}$) and the TTB-derived scores ($\widehat{\mathbf{T}}_{\mathrm{A}}^{(b)Q}$) to investigate uncertainty in the real panel's PCA scores.

2.4.2.1. Confidence intervals for WineSip scores in a PC

We obtained PB- and TTB-derived 95% CIs for each WineSip and time slice combination in each PC. Each CI is constructed using the percentile method: the 95% confidence limits at the 2.5th and 97.5th quantiles enclose the "middle 95%" of the virtual scores. The mid-point of the CI is determined by the virtual-panel scores, not the real-panel scores.

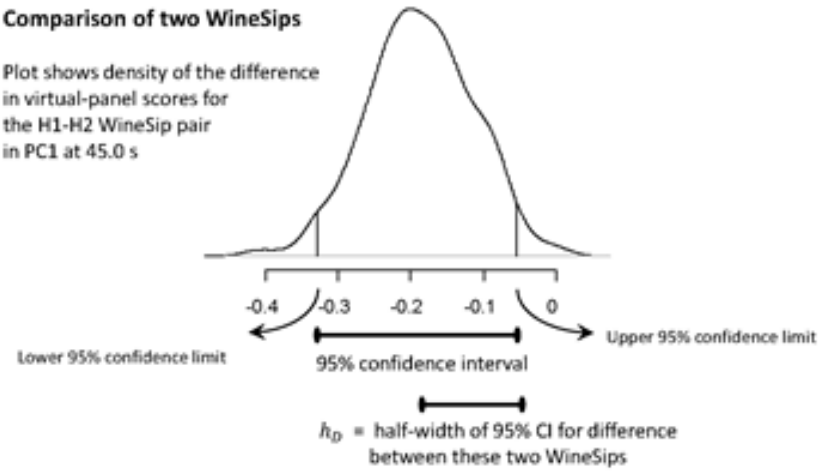
2.4.2.2. Confidence ellipses for WineSip scores in a plane of two PCs

In sensory evaluation, a normal practice when reviewing PCA results is to focus on PC1 vs. PC2. We obtained PB- and TTB-derived 95% CEs for each WineSip and time slice combination in the PC1 vs. PC2 plane and in the PC3 vs. PC4 plane. For each method, we obtained six WineSip CEs per time slice in each plane. Each CE follows the 5% probability contour of the multivariate normal distribution, such that 95% of the virtual-panel scores are contained within. In each case, the mean and the covariance matrix of the multivariate normal distribution were calculated from the virtual-panel scores.

2.4.2.3. Confidence intervals and standard errors for pairwise differences between two WineSip scores

The uncertainty of the differences between two WineSips is investigated using the approach illustrated in Fig. 4. In each PC and time slice, we obtain the differences between the virtual-panel scores from the two WineSips in a manner analogous to a paired t-test. Next, we construct a 95% CI for each paired difference distribution using the percentile method (Section 2.4.2.1). The half-width of the 95% CI, which is denoted h_D, approximates twice the standard error (SE) for the paired WineSip difference at the time slice and PC. If the difference between the real-panel WineSip pair is d, then the 95% CI is $d \pm h_D$. If this 95% CI excludes zero, then we conclude that the difference between the two WineSips is significant. We compute h_D using both the PB- and TTB-derived scores to enable comparison. The h_D values are used to investigate the discriminability among a limited number of planned WineSip comparisons, specifically the sip-to-sip comparisons described in Section 2.5.1.

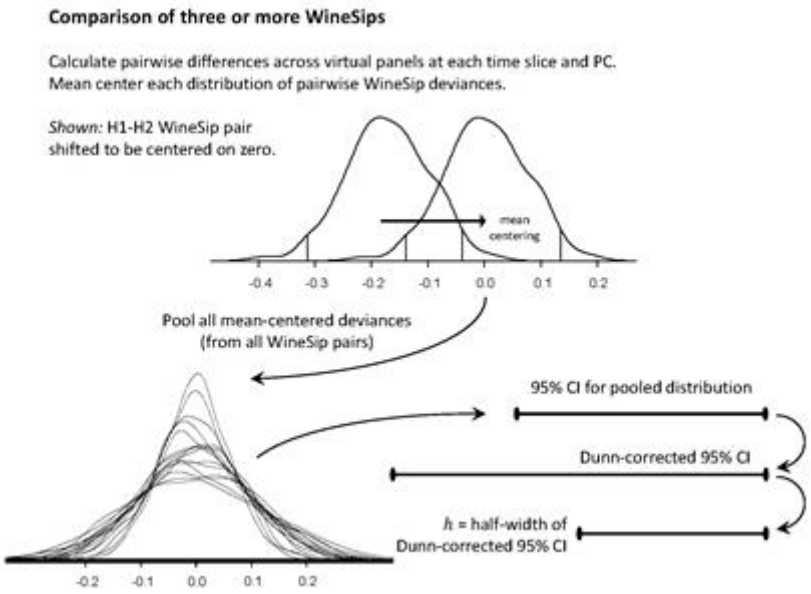
Fig. 4. Method of obtaining 95% CIs for paired differences between WineSips.



2.4.2.4. Confidence intervals and pooled standard errors for differences among three or more WineSips scores

Next, we extend the approach in the previous section to investigate the uncertainty of the differences among three or more WineSips at each time slice and PC. If we investigated multiple comparisons between every pair of WineSips using the method in Section 2.4.2.3, then it would increase the overall Type I error rate unacceptably. To address this problem, we use the approach illustrated in Fig. 5 to construct a set of 95% CIs for all pairs of WineSips that controls the overall Type I error rate at the nominal level α.

Fig. 5. Method of obtaining Dunn-corrected 95% CIs for paired differences for all WineSips.



The first step is to get the 95% CIs for the pooled WineSip paired difference distribution in each PC and time slice. We start by obtaining the differences between the virtual-panel scores from each pair of WineSips in a manner analogous to a paired t-test. Each paired difference distribution is mean-centered. This shifts the mean of each distribution to zero without changing its shape. Results from all of these mean-centered difference distributions are then pooled into one large distribution. We obtain the unadjusted 95% CI for the pooled differences using the percentile method. The half-width of the 95% CI approximates twice the SE of the pooled distribution of paired differences.

The next step is to construct a set of 95% CIs that accounts for multiple comparisons. To do this, we use Bonferroni's inequality (Dunn, 1961). We calculate the critical value for the standard normal distribution

$$c = \Phi^{-1}(1 - \alpha/2m), \qquad \text{Eq. (4)}$$

where Φ^{-1} is the quantile function of the standard normal distribution; α is the Type I error rate and there are m paired comparisons. For six WineSip pairs and the conventional level $\alpha = 0.05$, $m = 15$ and $c \approx 3$, which is larger (by a factor of 1.5) than the critical value for one paired comparison. To determine the width of the Dunn-corrected 95% CI, we multiply the width of the unadjusted 95% CI for the pooled differences by a factor of 1.5. The half-width of the Dunn-corrected 95% CI is h. A pair of WineSips whose real-panel scores are separated by d has a Dunn-corrected 95% CI of $d \pm h$. If this interval excludes zero, then we conclude that the difference between these two WineSips is significant. To enable comparison, we calculate all values of h using the PB- and TTB-derived scores. Later, in Section 2.5.2, these h values are used to investigate the discriminability among the six WineSip scores.

2.5. Reciprocal of discriminability

Temporal sensory studies are nearly always conducted to understand how sensory characterizations of products evolve over time and perceptual differences that occur at matching times during consumption.

2.5.1. Discriminability of two WineSips

We start with a planned comparison of two WineSips. For example, in this study, the two-sip experimental protocol is designed to test for sip-to-sip differences within a wine treatment, such as H1 vs. H2. There are only three such comparisons. For a particular WineSip pair, PC, and time slice, we obtain d and h_D as described in Section 2.4.2.3. The absolute deviance d between two WineSips that are truly the same would by chance alone be separated by more than h_D in only 5% of cases. To quantify how well the panel discriminates this pair of WineSips, we calculate the "reciprocal of discriminability",

$$R_d = h_D/d, \qquad \text{Eq. (5)}$$

at each time slice and PC. The threshold of discriminability ($R_d = 1$) for these two WineSips occurs at $h_D = d$. The reciprocal index R_d is somewhat analogous to an inverted pseudo-t statistic: a lower value of R_d indicates higher discriminability, and $R_d = 0$ indicates perfect discriminability.

2.5.2. Discriminability of three or more WineSips

Next, we extend the approach to investigate discriminability for three or more WineSips. In each PC and time slice, we test whether the panel discriminates the two WineSips that are the most different. The most different WineSips have the largest and smallest real-panel WineSip scores; the distance between these two WineSips have the largest value of d, which we call $Range$. If the panel does not discriminate this WineSip pair, then it does not discriminate any WineSip pair. We obtain h as described in Section 2.4.2.4. By chance, if all WineSips are truly the same, then the most-different pair of WineSips will be separated by more than h in only 5% of cases. We calculate the following reciprocal of discriminability:

$$R_d = h/Range. \qquad \text{Eq. (6)}$$

$R_d < 1$ indicates that at least one WineSip pair is discriminated. Eq. (6) generalizes Eq. (5): if there are only two WineSips, Eq. (6) reduces to Eq. (5) because $h = h_D$ and $Range = d$. Later, in Section 3.2.2, we visualize the discriminability for the six WineSips by incorporating h and $Range$ into our PCA biplots.

2.6. Statistical analysis

Analyses were conducted and visualizations rendered in R version 4.1.0 (R Core Team, 2021). The functions `sample` and `prcomp` were used to perform resampling and PCA, respectively. The function `Procrustes` from the R package `psych` (Revelle, 2020) was used to conduct Procrustes rotation. The R package `tempR` (Castura, 2020) was used to visualize the six WineSip trajectories. The `quantile` function was used to construct CIs as described in Section 2.4. We constructed 95% CEs using the

function `dataEllipse` in the R package `car` (Fox & Weisberg, 2019). The R package `sp` (Bivand, Pebesma & Gomez-Rubio, 2013) was used to calculate ellipse areas. We used the R package `av` (Ooms, 2021) to render Suppl. Video 1.

3. Results

3.1. *Uncertainty of WineSip perception dynamics*

Suppl. Video 1 shows the temporal evolution of the six smoothed WineSip trajectories in the planes of PC1 vs. PC2 (left panel) and PC3 vs. PC4 (right panel). Attribute labels are centered at the endpoints of the real panel's loading vectors. In each plane, the six WineSips are overlaid with their PB- and TTB-derived 95% CEs. The CEs produced from these two methods are similar in size. The high-ethanol wine treatments (H, A) appear to be characterized differently than the low-ethanol wine treatment (L), especially in the first minute of the evaluation. Specifically, the high-ethanol wine treatments are characterized by the attributes *heat*, *spices*, and *bitter* more often than the low-ethanol wine treatments, which are characterized more often with attributes such as *green* and *red fruit*. The video shows how the sensory characterizations of the WineSips evolve over time. It can be paused at any time to allow inspection. A still image that is adapted from the 30.0-s time slice of Suppl. Video 1 is shown in Fig. 3.

3.1.1. Interpretation of axes/components

Eigenvalues for the first four PCs are (in order) 2372.6, 210.7, 154.4, and 23.5 and account for 84.7%, 7.5%, 5.5%, and 0.8% of variance in $\mathbf{X}$, respectively (totaling 98.6%). Briefly, PC1 is related to the citation rates and separates the (nearly) zero attribute citation rates that occur at the start (and end) of the evaluation period from the peak citation rates that occur around 30.0 s. PC2 and PC3 separate the high-ethanol wine treatments (A and H) from the low-ethanol wine treatment (L) especially well. PC2 is an ethanol impact dimension that opposes *bitter*, *heat/ethanol burn*, and *spices*, which are characteristic of higher-ethanol wines in early evaluation, against the other attributes. PC3 contrasts *sour* and *red fruit* with *astringent* and *dark fruit*. PC4 opposes *bitter* and *sour* against the other attributes and seems to capture a later-onset bitterness. For a more detailed description of the PCs, we refer the reader to a three-component PCA solution discussed by Castura et al. (2016b).

3.1.2. %VAF by the virtual panels

The total variance in $\mathbf{X}$ is 0.2745. The total variance in $\mathbf{X}^{(b)}$ is on average 0.3067, but ranges from 0.1711 to 0.5411 (95% CI: 0.2207, 0.4372). This indicates that the multilevel bootstrap procedure tends to compose virtual panels with more variable results compared to the real panel. A potential reason is that assessor-replicate combinations are often repeated in the virtual panels, whereas this never occurs in the real panel.

Table 1 provides details on the %VAF in the virtual-panel results ($\mathbf{X}^{(b)}$) that is accounted for in each of the first four PCs by the PB scores and by the TTB scores (before rotation). Collectively, the first four real-panel PCs account for a significantly higher proportion of the variance in $\mathbf{X}$ than the first four PB- or TTB-derived components account for in $\mathbf{X}^{(b)}$ (only by a few percent). Table 1 also shows %VAF in PC5 and PC6. Overall, the PB scores and TTB scores (before rotation) account for a lower percentage of variance than the real panel's scores in PC1, and a higher percentage of variance than the real panel's scores in subsequent PCs. However the real-panel %VAF mostly falls within the 95% CIs for proportion of variance accounted for by the PB- or TTB-derived virtual-panel scores. Given the resemblances in %VAF in the first four PCs for real panel and for PB- and TTB-derived results from the virtual panels, we find it reasonable to proceed with investigating the uncertainty in the real panel's PCA results using both methods.

3.1.3. Uncertainty of WineSip scores over time

The WineSip 95% CEs are shown over time in the PC1 vs. PC2 plane and the PC3 vs. PC4 plane in Suppl. Video 1. Visual review of virtual scores for each WineSip did not reveal obvious departures from bivariate normality. The CE areas for each of the WineSips were calculated and plotted in Suppl. Fig. 1. There, we indicate the proportion of time slices that the TTB-derived CEs were larger than the PB-derived CEs and by how much, in terms of their cumulative areas. Generally, the PB-derived ellipses were slightly larger than the TTB-derived ellipses for H1, H2, A1 and A2 and smaller than the PB-derived ellipses for L1 and L2 in the PC1 vs. PC2 plane. In the PC3 vs. PC4 plane, neither method consistently produced larger or smaller CEs than the other method.

*Table 1. %VAF by PB- and TTB-derived PCA scores (before rotation). The two largest PCs that are lost to truncation (italics) and the cumulative %VAF by the first 4 PCs (**bold**) are also shown.*

	Real	PB method		TTB method	
%VAF in…	Panel	Mean	95 CI	Mean	95 CI
PC1	84.7	78.9	(70.0, 85.3)	80.4	(73.0, 86.3)
PC2	7.5	7.7	(4.9, 11.5)	9.1	(5.9, 13.1)
PC3	5.5	6.4	(3.1, 10.8)	5.5	(2.9, 9.2)
PC4	0.8	1.8	(0.8, 3.5)	1.7	(1.0, 2.6)
First 4 PCs	**98.6**	**94.7**	**(91.6, 96.7)**	**96.7**	**(95.3, 97.7)**
PC5	*0.4*	*1.5*	*(0.8, 3.0)*	*1.1*	*(0.7, 1.6)*
PC6	*0.3*	*1.1*	*(0.6, 2.0)*	*0.8*	*(0.5, 1.2)*

Instead of investigating the ellipse areas in other planes of PCs (PC1 vs. PC3, etc.), we found it more useful to investigate WineSip 95% CI widths in each component. The real panel's WineSip scores were near the mid-points of the 95% CIs obtained from each of the methods. Results are shown for the six WineSips in Suppl. Fig. 2. There, we describe the proportion of time slices that the TTB-derived CIs were larger than the PB-derived CIs, and by how much. Overall, we find that the PB-derived 95% CIs are roughly the same width or wider than the TTB-derived 95% CIs (an exception is L2 in PC2). In PC4, the TTB-derived 95% CIs are generally wider than the PB-derived 95% CIs (an exception is H2). It is worth noting that the magnitude of these differences is modest: the average PB-derived 95% CI width is 0.6% larger than the average TTB-derived 95% CI width in PC1, 5.4% larger in PC2, 11.8% larger in PC3, and 3.3% smaller in PC4. Overall, the average CI width from the full timeline in each PC is similar in magnitude for the PB and TTB methods. For both methods, the average CI width is narrowest in PC2. Compared with PC2, the average CI width is about 7% wider in PC4, 33% wider in PC3, and 78% wider in PC1.

3.2. Discriminability of WineSips

3.2.1. Sip-to-sip differences within each wine treatment

The rationale for the two-sip evaluation protocol was to determine whether wines are perceived differently in Sip 1 vs. Sip 2. Now we investigate the discriminability of sips within wine treatments by evaluating whether differences exist in A1 vs. A2, in H1 vs. H2, and L1 vs. L2. We reach similar conclusions whether discriminability is derived from the PB or from the TTB method. In both cases, we identify sip-to-sip differences, indicated by $R_d < 1$, between 10 s and 50 s (Suppl. Fig. 3). The two sips of the same wine are best discriminated in PC1, which captures much of the variability in overall citation rates. During this 40-s interval, H1 vs. H2 and A1 vs. A2 are discriminable just more than 50% of the time, and L1 and L2 just more than 20% of the time, in at least one PC. Sip-to-sip differences within each of these wine treatments are discriminable at 30.0 s in PC1, but not in PC2 or any other PC. The most prominent sip-to-sip differences outside of PC1 are detected only by the discriminability derived from the PB method in PC3, where A1 and A2 are discriminated intermittently between 15 and 20 s, when A2 is characterized by *sour* and *red fruit* more often than A1. Suppl. Fig. 3 shows moments when A1 and A2 are well separated during this time interval.

3.2.2. Six WineSips over time

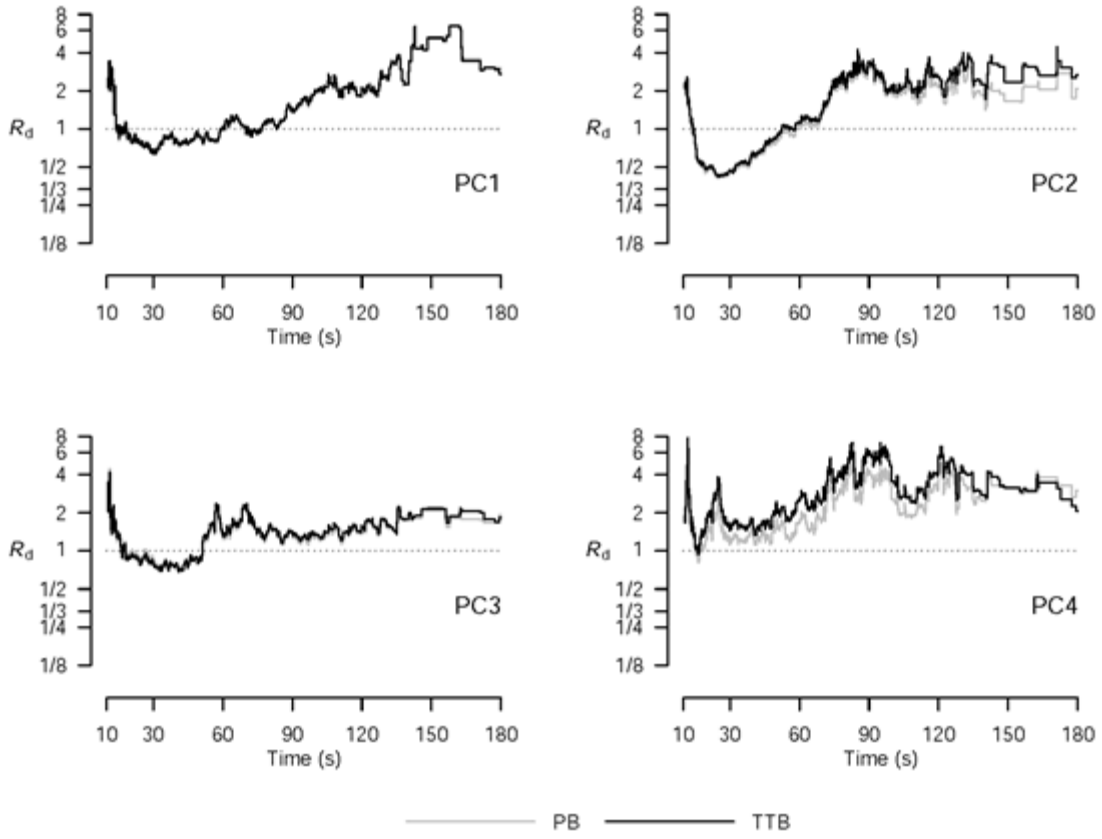
Now we investigate the discriminability of the six WineSips over time using the method described in Section 2.5.2. The purpose is to determine whether the panel discriminates any of the WineSips. In PC1, the h values obtained from the PB and TTB methods are nearly the same, on average. In the other PCs, the TTB-derived h values are larger than PB-derived h values by 8.4% in PC2, 1.5% in PC3, and 29.7% in PC4. This indicates a reversal of the trend that we reported earlier, because when 95% CIs are constructed for one WineSip at a time (Section 2.4.2.1), the CIs from the PB method were wider (Section 3.1.3).

Suppl. Video 1 shows how the six WineSips are perceived over time, along with information related to discriminability. (A black-and-white still from this video is shown in Fig. 3.) The $Range$ box is in blue (dotted black line in Fig. 3); it is the smallest rectangle that contains all real-panel WineSips scores. The h-cross appears in a grey rectangle. The horizontal and vertical lines represent the size of h in the PCs on the horizontal and vertical axes. If the $Range$ is larger than h in a PC ($R_d < 1$), then the corresponding line of the h-cross is emphasized in blue (solid black in Fig. 3); otherwise it is shown in black (grey in Fig. 3). At moments when the panel shows good discriminability in both PCs, the appearance of the inner rectangle resemblances the flag of Finland.

To further visualize discriminability, we show the R_d value for each PC over time in the right margin of each PCA biplot in Suppl. Video 1. The markers are shown in blue when $R_d < 1$ and in black otherwise, which matches the line colours of the h-cross. The PB-derived R_d values are not shown in Suppl. Video 1, but they are similar in size and interpretation.

Fig. 6 shows that the discriminability of the six WineSips based on the PB and TTB methods coincide approximately. WineSips are discriminated best in PC1, and less so in each successive PC. The panel discriminates WineSips for approximately 30% of the 170-s evaluation time in PC1, about 25% in PC1, and about 20% in PC3. WineSips are mostly discriminated between 14.2 and 83.2 s in PC1, between 13.9 and 60.1 s in PC2, and between 16.1 and 51.2 s in PC3. The WineSips are not discriminated in PC4; WineSip differences between 15.0 and 17.0 s are fleeting and probably spurious. Overall, discriminability derived from the PB method identifies differences between WineSips at 5% more time slices than discriminability derived from the TTB method.

Fig. 6. Reciprocal index of discriminability of the six WineSip scores for each of the four PCs over time based on the PB method (grey line) and the TTB method (black line). Lower R_d values indicate higher discriminability. To show the R_d values more clearly, the y-axis is shown on a binary (base 2) logarithmic scale and $R_d > 8$ are suppressed.



4. Discussion

4.1. *Uncertainty of WineSip perception dynamics*

4.1.1. Interpretation of WineSip perception dynamics – Suppl. Video 1 shows that the perception of the six WineSips varies systematically with time and that ethanol level has a strong influence on sensory perceptions. The CEs enhance the interpretability of the PCA biplot by providing a visual indication of the uncertainty of the WineSip scores over time. Uncertainty is highest (CEs are largest) in the first half of the evaluation when perception of the samples changes the most and when the TCATA citation rates are highest. During this time interval, the six WineSips CEs are often visually well-separated in the PC1 vs. PC2 plane (mostly in PC2). The six WineSips are separated, albeit less well, in the PC3 vs. PC4 plane (mostly in PC3), particularly in the first third of the evaluation. Uncertainty of the WineSip scores in both planes is very low in the second half of the evaluation as the TCATA citation rates decline gradually toward zero. PC1 captures some WineSip differences (Section 3.2), but it also captures variance related

to a time-dependent category signature (Meyners & Castura, 2019) that is common to all the WineSips. This is why PC2 and PC3 (which were the focus of analysis by Castura et al., 2016b) seem to separate the six WineSips so well (Fig. 6) in spite of extracting only 13.0% of the variance, which is six times less than the %VAF in PC1.

4.1.2. Comparison of PB and TTB methods – Based on previously published research by Cadoret and Husson (2013), we thought we might find systematic differences in the both 95% CI widths and 95% CE areas from the PB and the TTB methods, but we did not. For most of the evaluation duration, the WineSip CIs and CEs for scores are sometimes larger using the TTB method and other times larger using the PB method (Section 3.1.3). Either method is potentially adequate for the Syrah data.

4.1.3. Effect of Procrustes rotations in the TTB method – In the TTB method, we found that virtual-panel scores in PC1 were most similar to the real-panel scores in PC1 even before Procrustes rotation. It was in PC2 and PC3 that Procrustes rotations tended to resolve mixing of virtual-panel scores that corresponded to the real-panel scores in PC2 and PC3; an explanation is that %VAF by these PCs were close in the real panel (Section 3.1.2). This also occurred, but to a lesser degree, in PC3 and PC4. Mixing of virtual-panel scores beyond PC4 that corresponded to the real-panel scores could not be resolved by Procrustes rotation because both the real- and virtual-panel scores were truncated before being rotated.

4.1.4. Comparison of PB-derived CEs here vs. Castura et al. (2016b) – Castura et al. (2016b) also constructed PB-derived 95% CEs. They obtained $\mathbf{Y}^{(b)}$ from $\mathbf{Y}$ only by resampling assessors with replacement. We took an additional step of resampling replicates with replacement for each assessor instance. Subsequent steps were the same. The total area of all 95% CEs from assessor-replicate resampling done here are larger than assessor-only resampling (Castura et al., 2016b) by 30.1% in the PC1 vs. PC2 plane and 39.7% in the PC3 vs. PC4 plane. Average differences in area are more pronounced (32.8% and 44.3%) up to 90 s, when most sensory changes occur. The approach we use here is preferred. It yields larger CEs because resampling assessors incorporates variability due to assessor disagreement and resampling replicates incorporates variability due to assessor non-repeatability.

4.1.5. Relationship between CI widths for scores vs. size of eigenvalues –Since PCs are often evaluated according to the relative sizes of their eigenvalues, it might be assumed that a direct relationships exists between CI widths and the eigenvalue sizes. We illustrate now why this assumption is wrong. Consider a non-discriminating TCATA panel whose panelists are non-repeatable during the time when citation rates are far from both zero and one (see Meyners & Castura, 2018). The panel is non-discriminating, so the range of the WineSip scores within each time slice will tend to be narrow. The panel is non-repeatable, so WineSip CIs will be wide. Yet after PCA, the first few PCs might have large eigenvalues if there is a strong time-dependent category signature (see Meyners & Castura, 2019); otherwise, these eigenvalues might be small. This is just one example that illustrates the lack of a simple, direct relationship between WineSip CI widths and the sizes of the eigenvalues. We observe this lack of relationship in the first four PCs of our results: both the PB- and TTB-derived 95% CIs were largest in PC1 and smallest in PC2 (Section

3.2.1). This occurs because PCA eigenvalues are based solely on the variability in $\mathbf{X}$: each eigenvalue is the sum of squares that the PC extracts from $\mathbf{X}$. On the other hand, the WineSip CIs are based on variability in the data set $\mathbf{Y}$: they provide information about the variability across the virtual panels' results ($\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(B)}$).

4.2. Discriminability of WineSips

4.2.1. TTB-derived discriminability of WineSips in the four PCs – Overall, we find the discriminability to be quite good, particularly in the first three PCs (Fig. 6). The panel shows slight discriminability ($R_d < 1$) in at least one of the first three PCs as early as 13.9 s and as late as 83.2 s, more than one minute after expectoration. PC4 shows negligible evidence of discriminability between 16 and 17 s and not thereafter. These conclusions are not readily gleaned from the overlapping TTB-derived 95% CEs in Suppl. Video 1.

4.2.2. Comparison of PB- and TTB-derived discriminability of WineSips – We reach similar conclusions regarding WineSip discriminability whether it is derived from the PB or the TTB method (Fig. 6). When preparing Suppl. Video 1 we showed only the TTB-derived discriminability because the PB-derived discriminability often coincides, except at times when the discriminability was quite bad for both approaches. The reason we reach similar conclusions from PB- and TTB-derived discriminability is that differences between the PB- and TTB-derived 95% CIs for scores are relatively small (Section 3.2.2). So our conclusion that both the PB and TTB methods are adequate for investigating the uncertainty in PCA scores for this type of data (in Section 4.1.2) extends also to how discriminability of PCA scores is derived. Differences between the PB- and TTB-derived discriminability are somewhat larger for sip-to-sip comparisons (Section 3.2.1) than for comparisons of all six WineSips (Section 3.2.2); the reason is that we based h on CIs for the difference between only two WineSips, not on multiplicity-adjusted CIs based on all WineSip pairs.

4.2.3. Method of getting the pooled standard error – An intermediate step for getting h is to obtain the unadjusted 95% CI for the pooled WineSip paired difference distribution (Section 2.4.2.4). We also tested an alternative approach for getting this unadjusted 95% CI. In the alternative approach, we first approximate the SE of each WineSip pair separately by the method described in Section 2.4.2.3. The pooled SE was obtained by summing the squares of these 15 SEs, dividing by one less than the number of SEs (14), and taking the square root. The unadjusted 95% CI for the pooled WineSip paired difference distribution was then obtained by multiplying the pooled SE by four. Although these two approaches differ, the alternative approach gives similar values of pooled SE and h as the approach that we used in the manuscript. The two approaches had similar cumulative durations of discriminability ($R_d < 1$), within two seconds. This was the case in every PC and for both the PB and TTB methods. From a practical perspective, the results are the same, but we prefer to get the pooled SE from pooled observations (as in Section 2.4.2.4) because it is analogous to the way that the mean squared error pools residuals from all observations in analysis of variance.

4.2.4. Discriminability in multiple PCs – In this paper, we investigate and visualize discriminability in each PC independently. A superior approach would provide an elliptical confidence region that follows the probability contours of the distribution of virtual-panel paired differences (Sections 2.4.2.3 & 2.4.2.4). In two PCs, if the 95% CE of the paired difference between two WineSips excludes zero, then the WineSips are significantly different in that plane of PCs (Castura et al., 2016b); however, considering two of four PCs simultaneously is incomplete, just as considering one of four PCs at a time is incomplete. It is not trivial to make multiplicity corrections, nor is it straightforward to quantify and visualize discriminability of multiple WineSips in two or more PCs. Differences in the shape and size of CEs need to be accounted for in two dimensions, which introduces complexity beyond what exists in one dimension. For this reason, we leave these as topics for future research. For now, we note a heuristic interpretation of an h-cross (Section 3.2.2): it is the half-size of the rectangular 90% confidence region (CR) in a plane of PCs. Each 95% CI is obtained from the percentile method, so each CI excludes 5% of the paired differences. When the two CIs are combined, they create a CR that jointly excludes at most 10% of the paired differences. This 90% CR may be conservative for two reasons. First, if some paired differences are excluded from both 95% CIs, then these paired differences will be double-counted in the bivariate rejection region, so the CR will overlap more than 90% of the paired differences. Second, a rectangular 90% CR often has areas that are sparse or empty because the distribution of virtual-panel WineSip paired differences are not equally distributed in the rectangular region; rather, they are roughly bivariate normally distributed in the plane of PCs. Visualizing differences among WineSips in each plane of PCs is a potential methodological refinement.

4.3. Future research

In this paper, we used the Syrah data, which have the special temporal structure of a TCATA data set, in which rows are combinations of WineSips and time slices. The methods that we describe could be adapted for the analysis of other types of temporal sensory data, static sensory data, or non-sensory data. Doing so offers the opportunity to use these methods in a different way. For example, if analyzing a conventional sensory QD sensory data set in which rows are products, then R_d can be used to quantify discriminability in each PC, and only the PCs that discriminate the products retained.

Monte Carlo testing could be used to further explore the different methods for investigating uncertainty described in this manuscript. The proportion of CEs enclosing the true location parameters of scores could allow for an objective comparison. However, it is unclear what data-generating function appropriately models the dependencies between WineSips, attributes, and times that occur in real TCATA data. Unless there is a theoretical statistical distribution that is representative of data observed in real sensory studies, such simulation results might lack practical value.

5. Conclusions

When investigating the uncertainty in the six WineSip scores over time, neither the PB- nor the TTB-derived confidence ellipses were consistently larger. We reach the similar conclusions regardless of which of the two methods we used to construct the confidence intervals. Both methods seem to be adequate to investigate the uncertainty of PCA scores (WineSips) and their discriminability over time. We also investigated and discussed the lack of simple relationships between the size of eigenvalues, the sizes of confidence intervals and confidence ellipses, and discriminability.

Using an objective, numerical index was effective in drawing our attention to moments of discriminability in the WineSips that might otherwise have gone unnoticed. Our discriminability index also helped to identify sip-to-sip differences that have not been identified previously. We show how to make multiplicity adjustments for evaluating discriminability in each PC independently. In the future, these methods can be extended to evaluate discriminability in multiple PCs simultaneously.

Interpretation of PCA biplots requires experience and training. Incorporating the discriminability index and other graphical features into our PCA biplots helps to increase interpretability and complements the conventional ways of visualizing PCA results in a way that goes beyond visualizing the uncertainty of scores. Collectively, these approaches can help to avoid over-interpretation of potentially spurious results. It can also encourage discovery and stimulate discoveries that might otherwise be missed.

Acknowledgements

We have benefited tremendously from excellent feedback from the editor and reviewers throughout the peer-review process. Detailed critical feedback led us to make substantial improvements to this manuscript. The data used in this study were part of Allison Baker's PhD research; we appreciate her feedback on early drafts of this manuscript. We also acknowledge financial support from the Research Council of Norway.

References

Babamoradi, H., van den Berg, F., & Rinnan, Å. (2013). Bootstrap based confidence limits in principal component analysis–a case study. *Chemometrics and Intelligent Laboratory Systems*, *120*, 97–105.

Baker, A.K., Castura, J.C., & Ross, C.F. (2016). Temporal check-all-that-apply characterization of Syrah wine finish. *Journal of Food Science*, *81*, S1521–S1529.

Baker, A.K., Ross, C.F., & Castura, J.C. (2019). Temporal Check-All-That-Apply characterization of Syrah wine finish. *Mendeley Data*, V2, doi: 10.17632/jv22xn66ky.2

Bivand, R., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R*, 2nd ed. New York: Springer. https://asdar-book.org.

Cadoret, M., & Husson, F. (2013). Construction and evaluation of confidence ellipses applied at sensory data. *Food Quality and Preference*, *28*, 106-115.

Castura, J.C. (2020). `tempR`: Temporal Sensory Data Analysis. R package version 0.9.9.16. http://www.cran.r-project.org/package=tempR

Castura, J.C., Antúnez, L., Giménez, A., & Ares, G. (2016a). Temporal check-all-that-apply (TCATA): A novel temporal sensory method for characterizing products. *Food Quality and Preference*, *47A*, 79–90.

Castura, J.C., Baker, A.K., & Ross, C.F. (2016b). Using contrails and animated sequences to visualize uncertainty in dynamic sensory profiles obtained from temporal check-all-that-apply (TCATA) data. *Food Quality and Preference*, *54*, 90-100.

Colonna, A.E., Adams, D.O., & Noble, A.C. (2004). Comparison of procedures for reducing astringency carry-over effects in evaluation of red wines. *Australian Journal of Grape and Wine Research*, *10*, 26-31.

Courcoux, P., Qannari, E. M., Taylor, Y., Buck, D., & Greenhoff, K. (2012). Taxonomic free sorting. *Food Quality and Preference*, *23*, 30-35.

Dunn, O.J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52-64.

Efron, B., & Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. New York: CRC press.

Fox., J., & Weisberg, S. (2019). An R Companion to Applied Regression, Third Edition. Thousand Oaks, CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Greenacre, M. (2007). *Correspondence Analysis in Practice*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press.

Husson, F., Lê, S., & Pagès, J. (2005) Confidence ellipse for the sensory profiles obtained by principal component analysis. *Food Quality and Preference*, 16, 245–250.

Johnson, R.A., & Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*, 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall.

Josse, J., Wager, S., & Husson, F. (2016). Confidence areas for fixed-effects PCA. *Journal of Computational and Graphical Statistics*, 25, 28-48.

Lawless, H.T., & Heymann, H. (2010). Data relationships and multivariate applications. In: *Sensory Evaluation of Food*, Food Science Text Series (pp. 433-449). New York: Springer.

Lebart, L. (2007). Which bootstrap for principal axes methods? In P. Brito, P. Bertrand, G. Cucumel, and F. de Carvalho (eds.): *Selected Contributions in Data Analysis and Classification*, pp. 581-588. New York: Springer.

Legendre, P., & Legendre, L. (2012). *Numerical Ecology*. Oxford, UK: Elsevier.

Lenfant, F., Loret, C., Pineau, N., Hartmann, C., & Martin, N. (2009). Perception of oral food breakdown. The concept of sensory trajectory. *Appetite*, *52*, 659-667.

Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979). *Multivariate Analysis*. Toronto: Academic Press.

Meyners, M., & Castura, J.C. (2016). Randomization of CATA attributes: Should attribute lists be allocated to assessors or to samples? *Food Quality and Preference*, *48*, 210–215.

Meyners, M., & Castura, J. C. (2018). The analysis of temporal check-all-that-apply (TCATA) data. *Food Quality and Preference*, *67*, 67-76.

Meyners, M., & Castura, J.C. (2019). Did assessors select attributes by chance alone in your TDS study, and how relevant is it to know? *Food Research International*, *119*, 571-583.

Næs, T., Brockhoff, P.B., & Tomic, O. (2010). *Statistics for Sensory and Consumer Science*. West Sussex, UK: John Wiley & Sons Ltd.

Ooms, J. (2021). `av`: Working with Audio and Video. R package version 0.6.0. https://CRAN.R-project.org/package=av

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Revelle, W. (2020) `psych`: Procedures for Personality and Psychological Research, Northwestern University, Evanston, IL, USA. R package version = 2.0.9. https://CRAN.R-project.org/package=psych

Ross, C.F., Hinken, C., & Weller, K. (2007). Efficacy of palate cleansers for reduction of astringency carryover during repeated ingestions of red wine. *Journal of Sensory Studies*, *22*, 293-312.

Schönemann, P.H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika 31*, 1-10.

Timmerman, M.E., Kiers, H.A.L., & Smilde, A.K. (2007). Estimating confidence intervals for principal component loadings: a comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology*, *60*, 295–314.

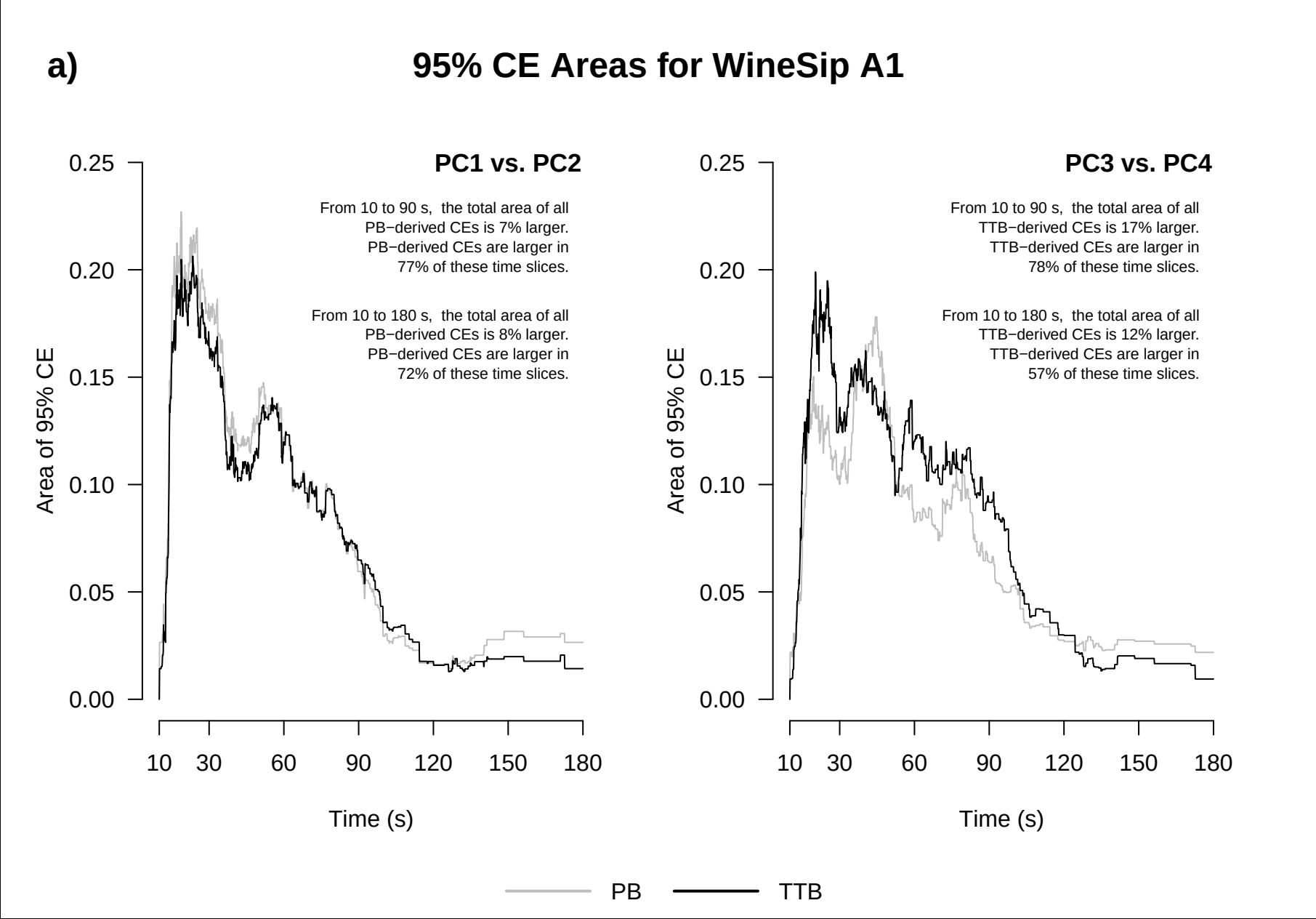
e-Component

Suppl. Video 1. The video shows the sensory evolution of six WineSips in the planes of PC1 vs. PC2 (left panel) and PC3 vs. PC4 (right panel). Uncertainty in perception dynamics within this subspace are investigated using both the PB (ellipse with dashed line) and the TTB (ellipse with solid line) methods. The $Range$ box is shown in blue. Each line of the "h-cross" is shown in blue when $R_d < 1$ and in grey otherwise. The discriminability of the six WineSips in each PC is indicated on the binary (base 2) logarithmic scale (on the right). Click on the video to review the animated sequence.

View video

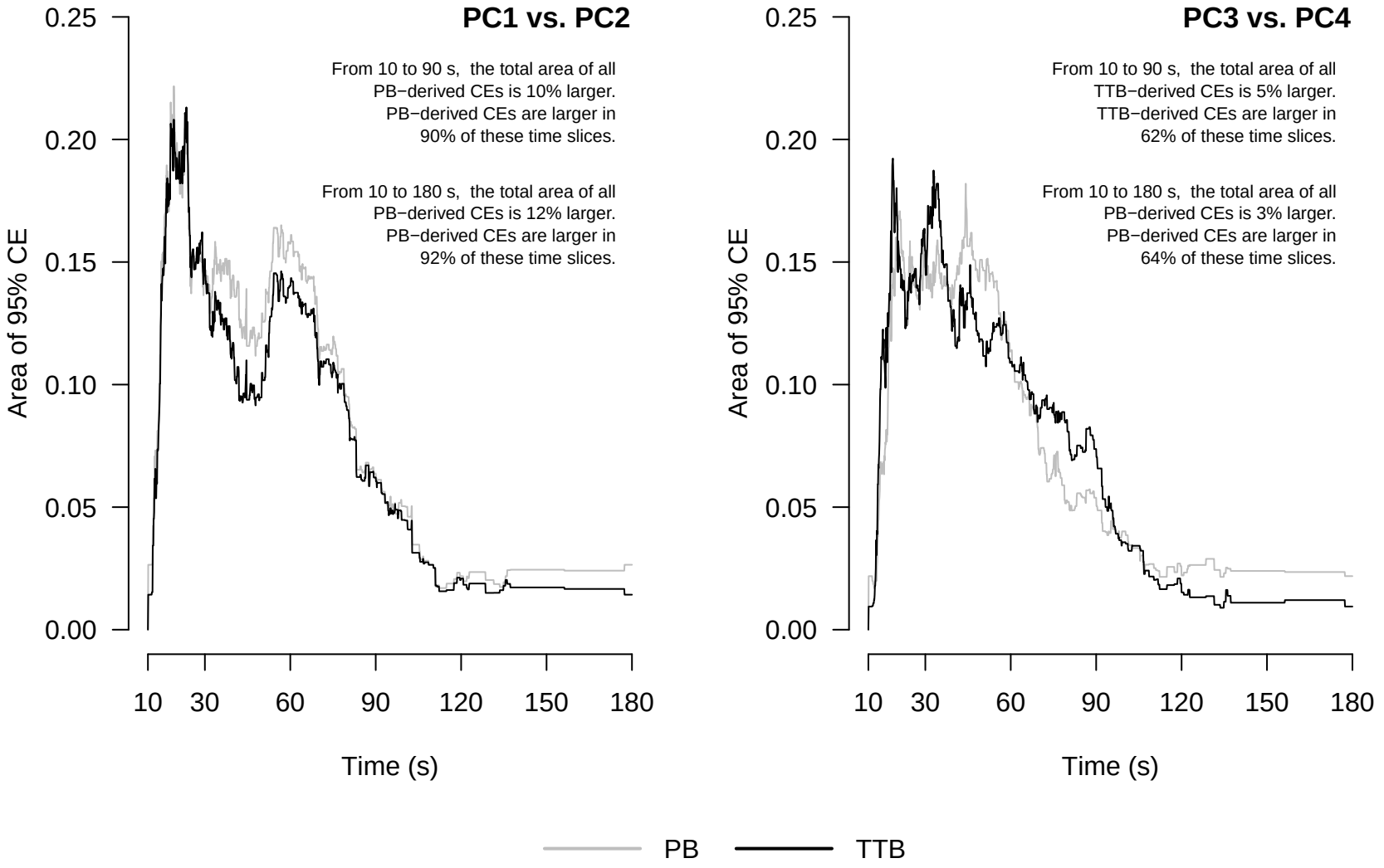
Suppl. Fig. 1.

WineSIp 95% CE areas based on the PB method (grey line) and the TTB method (solid line) in the planes of PC1 vs. PC2 (left panel) and PC3 vs. PC4 (right panel) for WineSips (a) A1, (b) H2, (c) L1, (d) A2, (e) H2, and (f) L2.


a)
95% CE Areas for WineSip A1
PC1 vs. PC2
From 10 to 90 s, the total area of all PB−derived CEs is 7% larger. PB−derived CEs are larger in 77% of these time slices.
From 10 to 180 s, the total area of all PB−derived CEs is 8% larger. PB−derived CEs are larger in 72% of these time slices.
PC3 vs. PC4
From 10 to 90 s, the total area of all TTB−derived CEs is 17% larger. TTB−derived CEs are larger in 78% of these time slices.
From 10 to 180 s, the total area of all TTB−derived CEs is 12% larger. TTB−derived CEs are larger in 57% of these time slices.
Area of 95% CE
0.25
0.20
0.15
0.10
0.05
0.00
10
30
60
90
120
150
180
Time (s)
PB
TTB

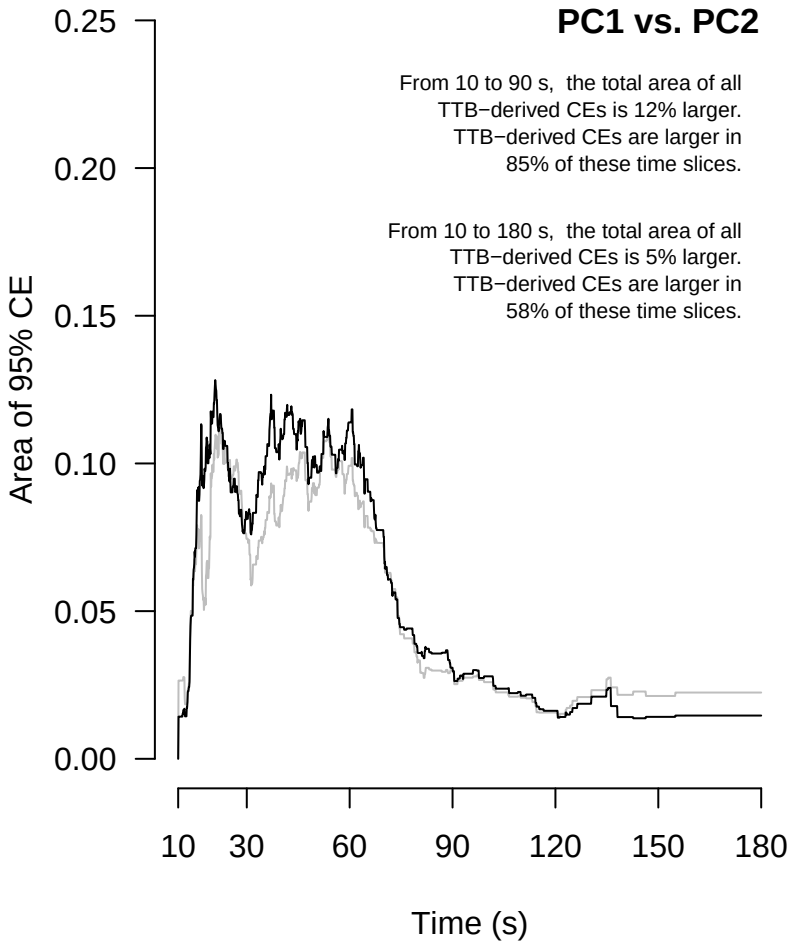
b)

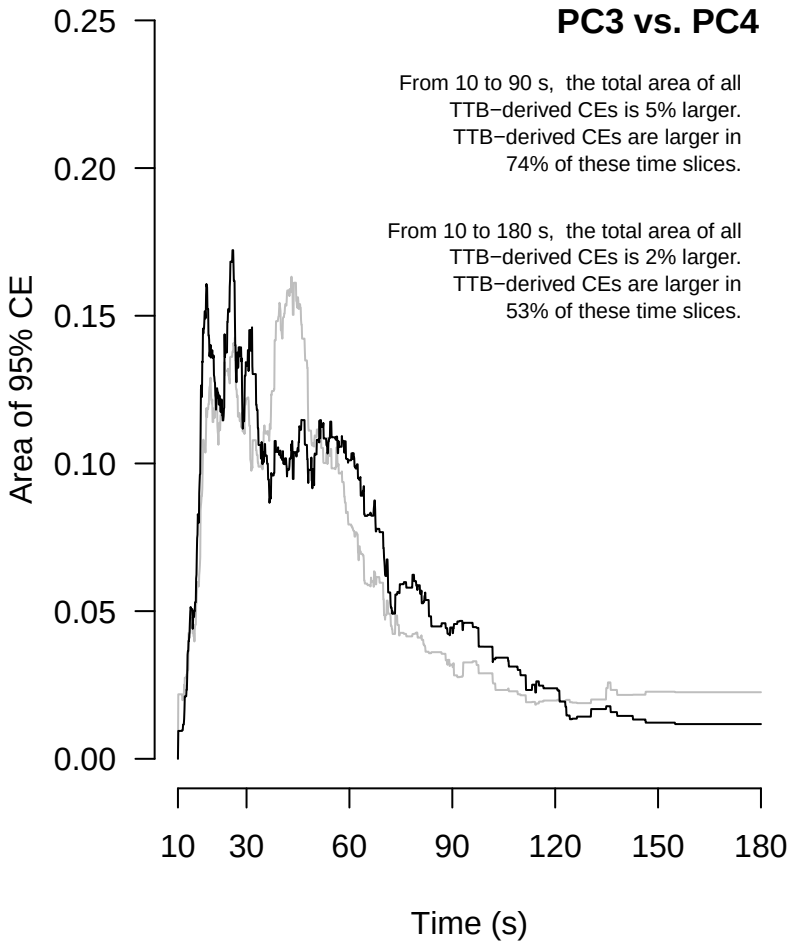
95% CE Areas for WineSip H1



c)

95% CE Areas for WineSip L1

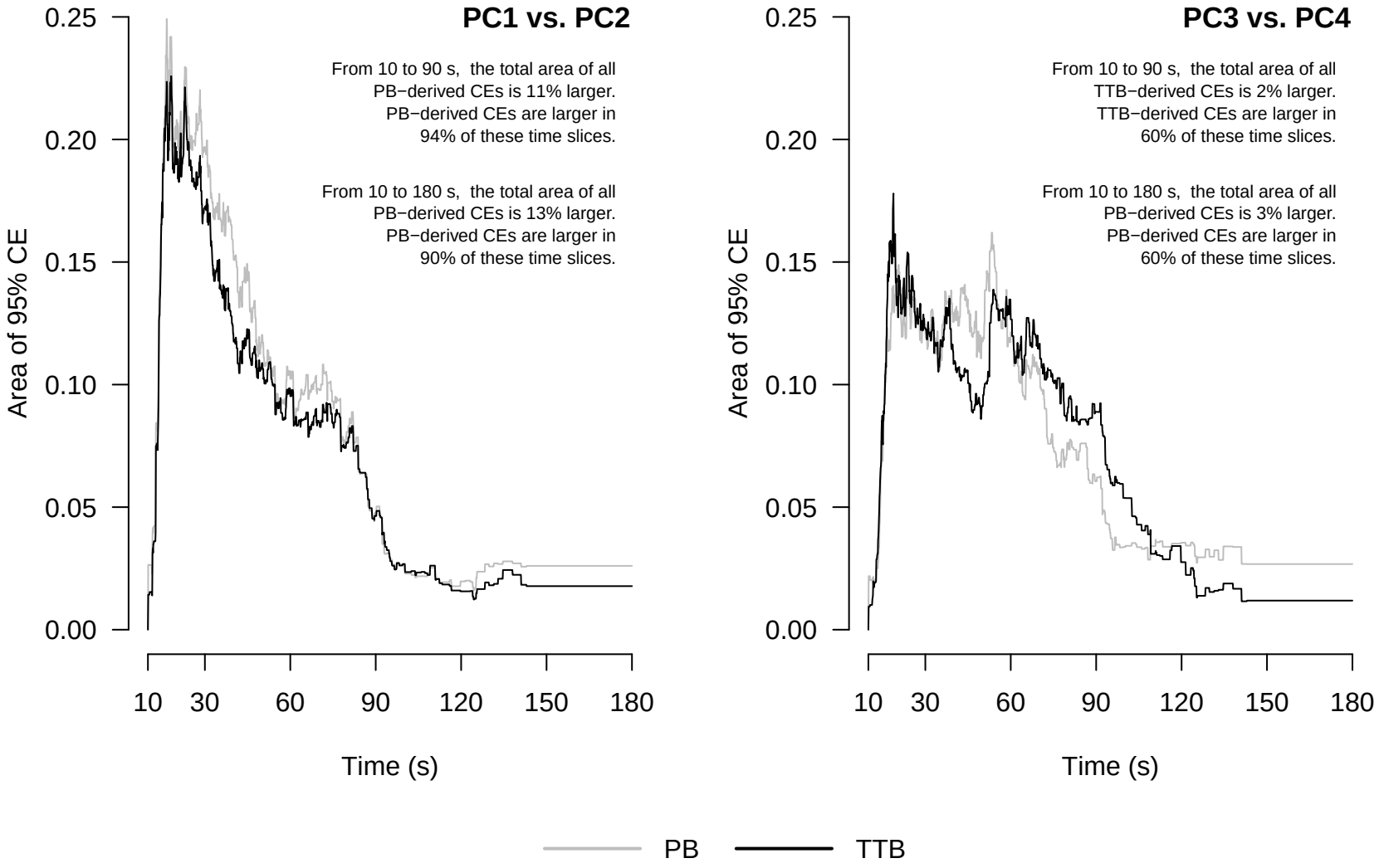




PB TTB

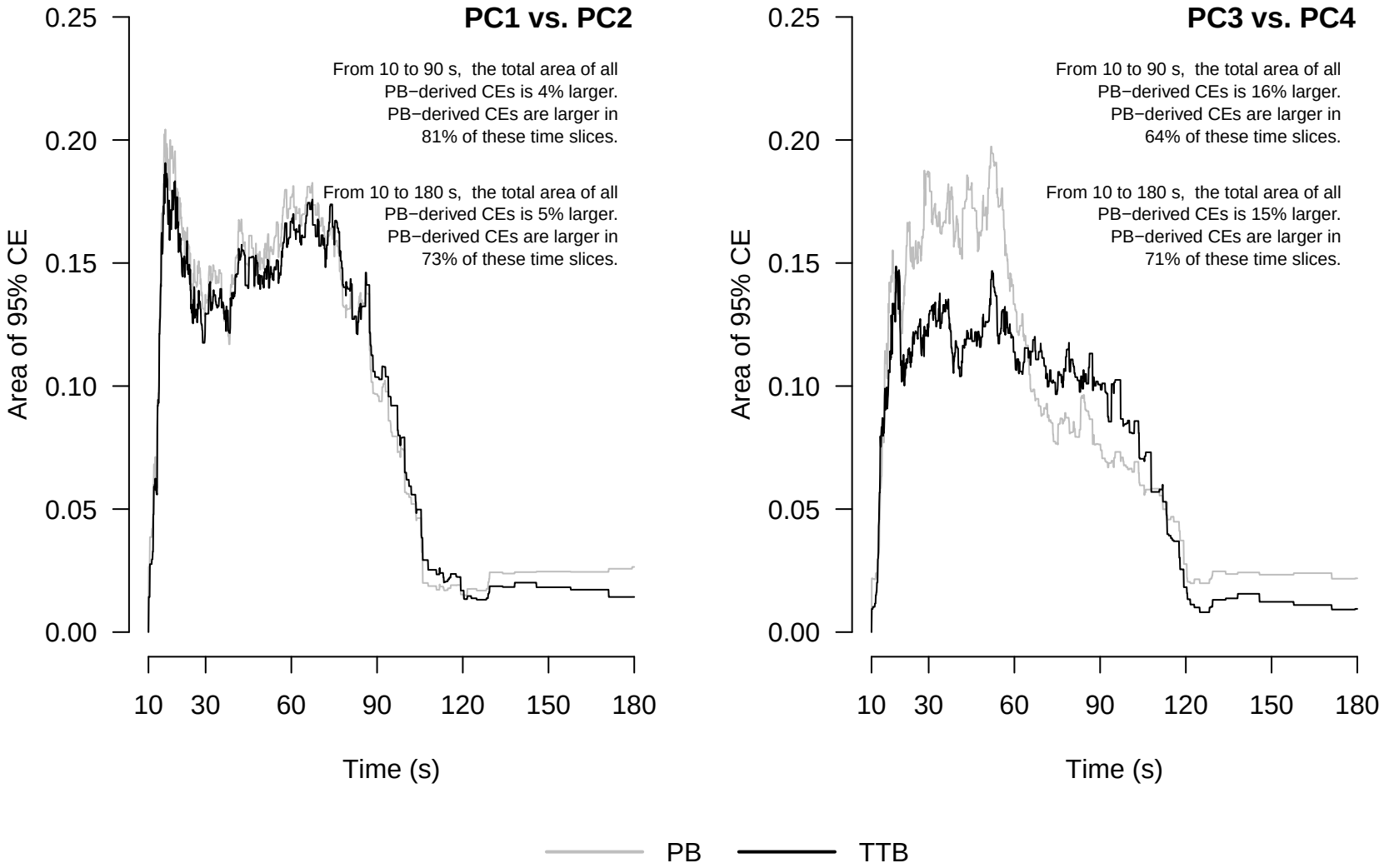
d)

95% CE Areas for WineSip A2



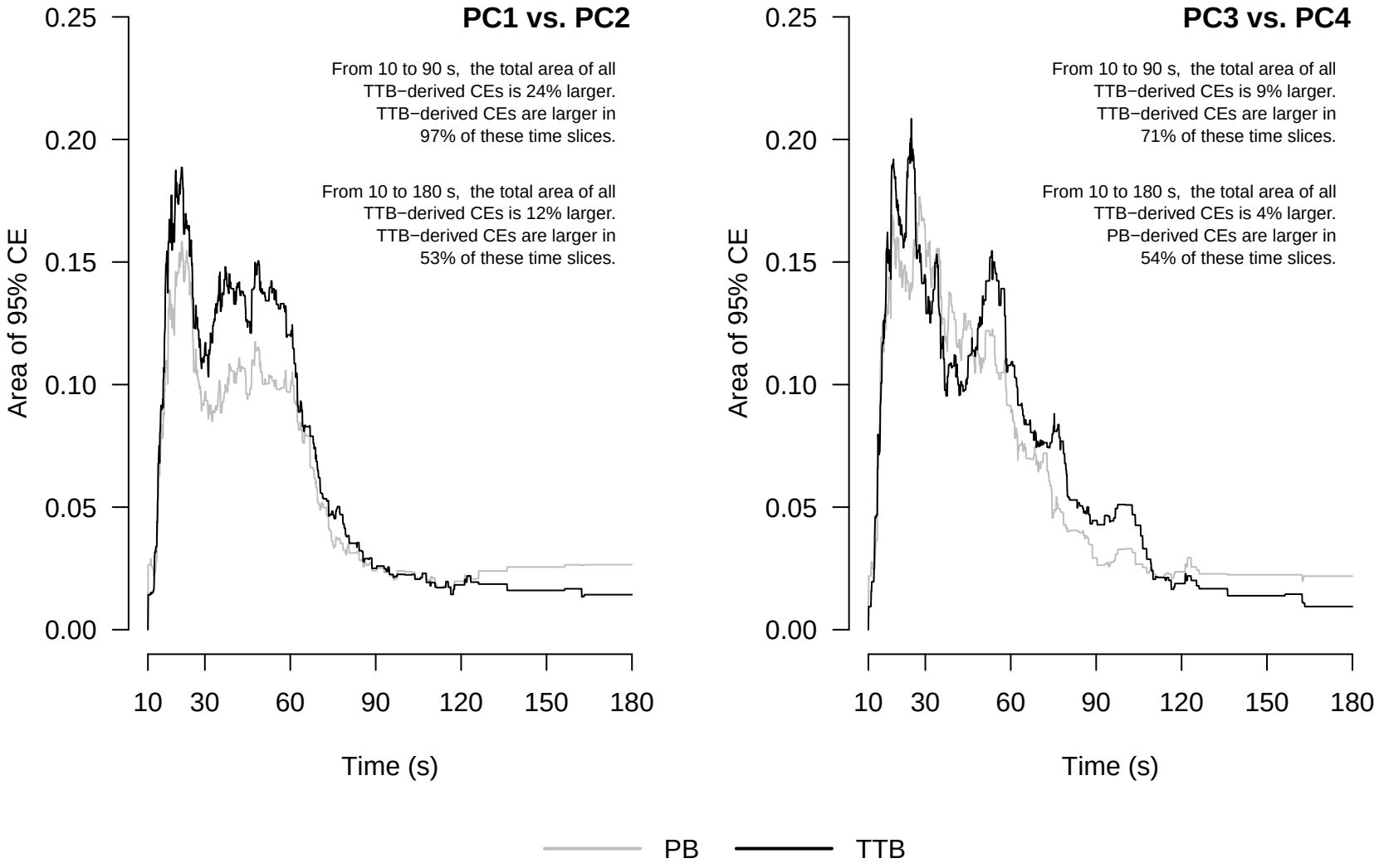
e)

95% CE Areas for WineSip H2



f)

95% CE Areas for WineSip L2

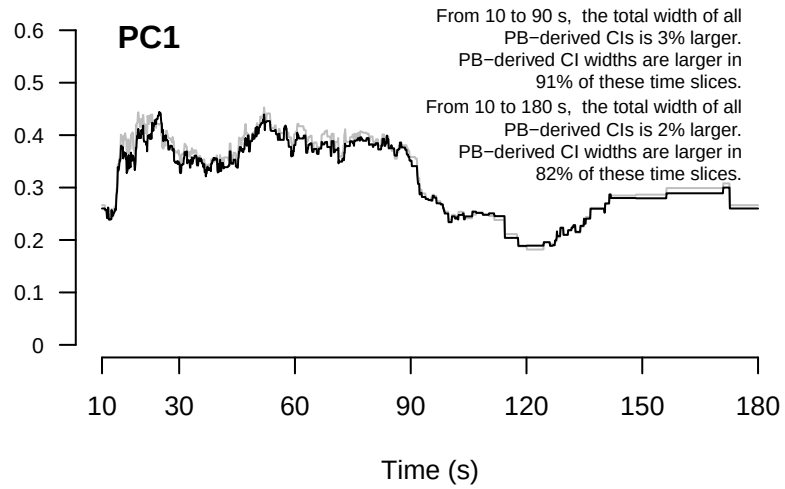


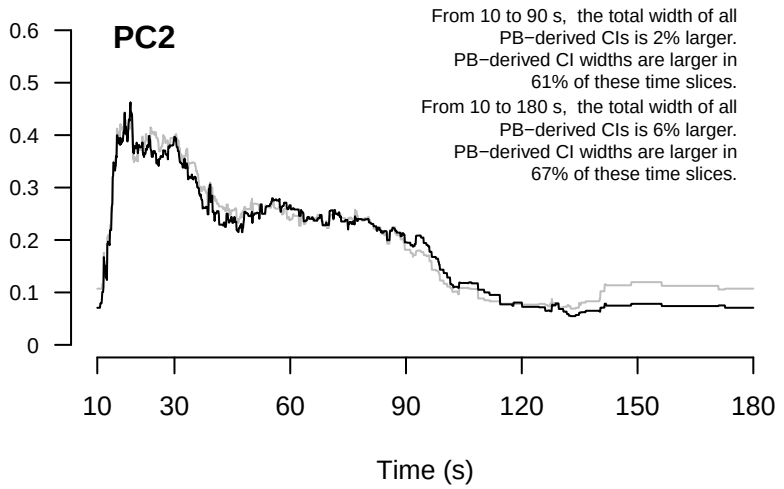
Suppl. Fig. 2.

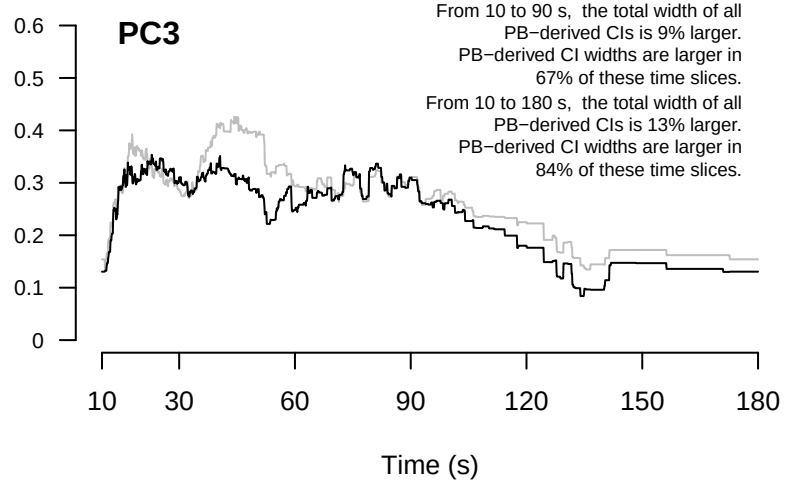
WineSIp 95% CI widths based on the PB method (grey line) and the TTB method (solid line) in PC1 (top left panel), PC2 (top right panel), PC3 (bottom left panel) and PC4 (bottom right panel) for WineSips (a) A1, (b) H1, (c) L1, (d) A2, (e) H2, and (f) L2.

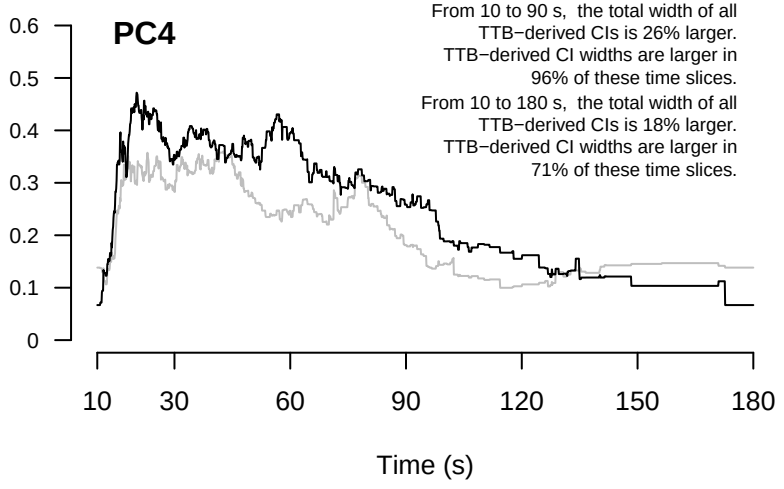
a)

95% CI Widths for WineSip A1





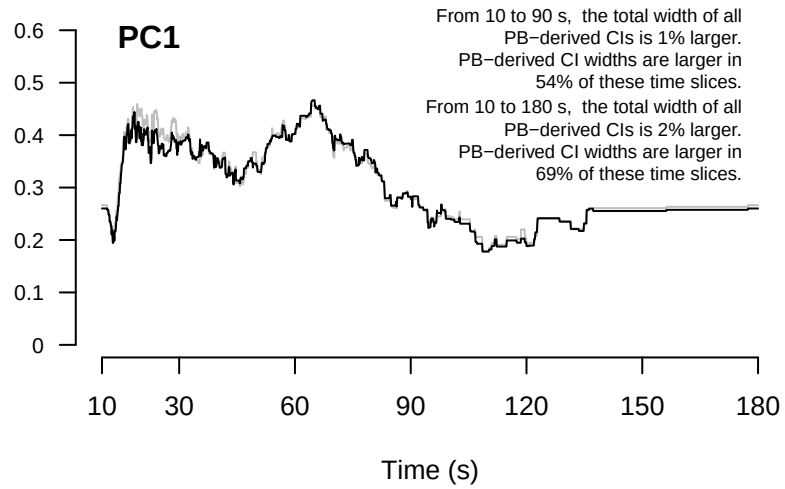


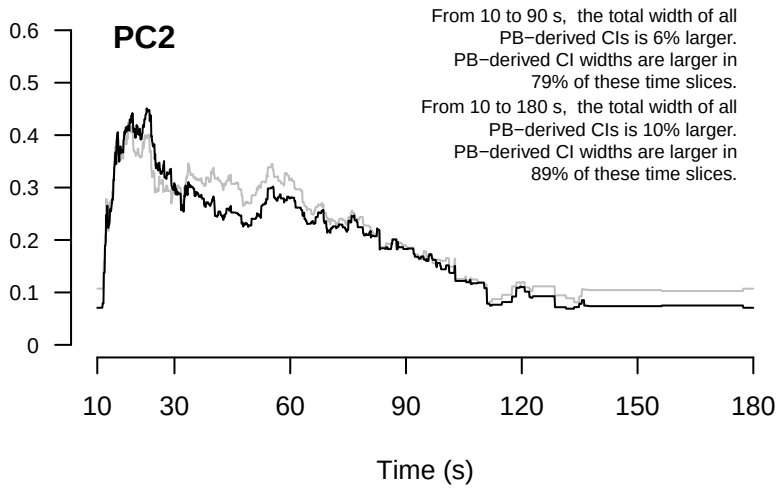


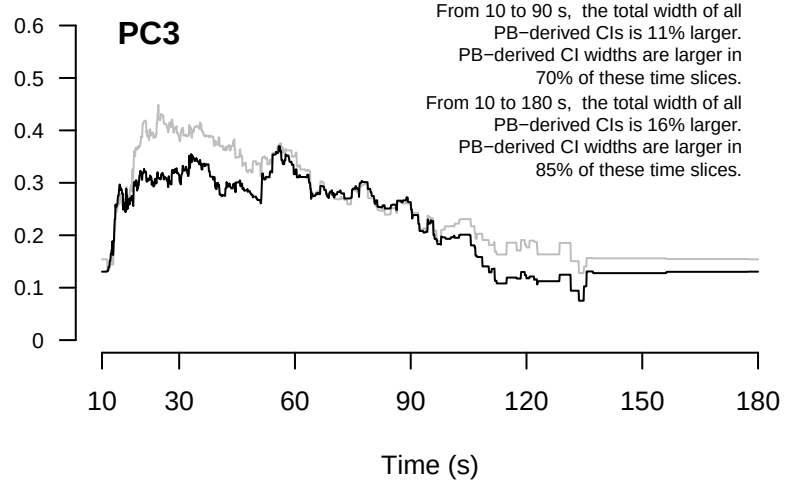
PB TTB

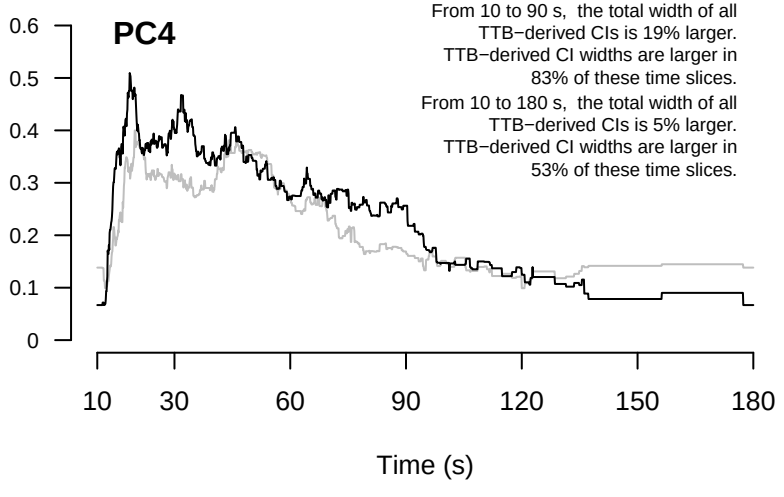
b)

95% CI Widths for WineSip H1





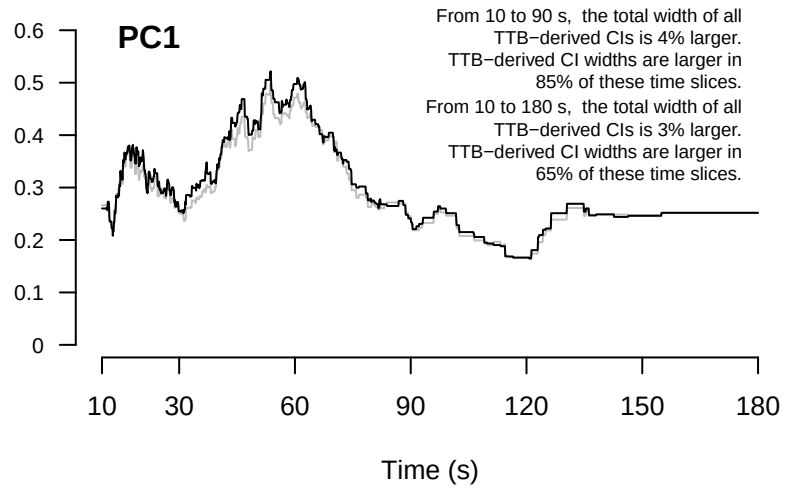


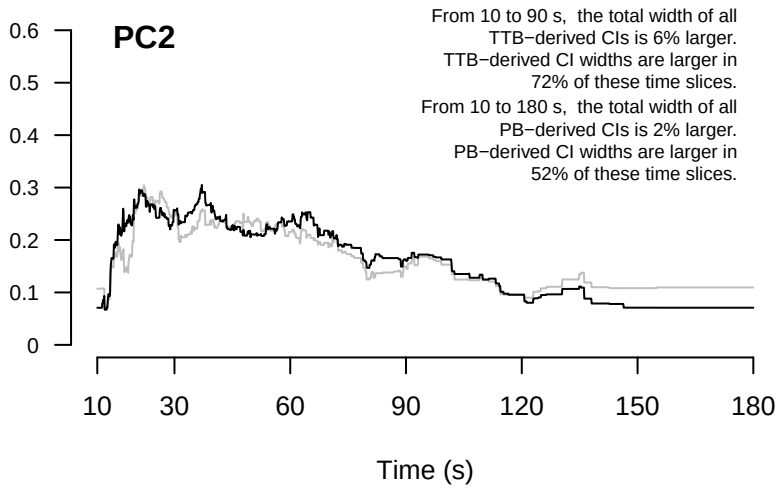


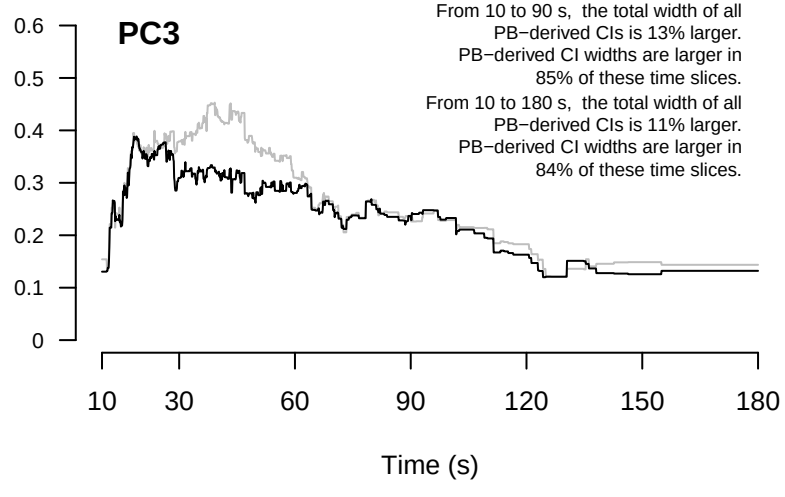
PB TTB

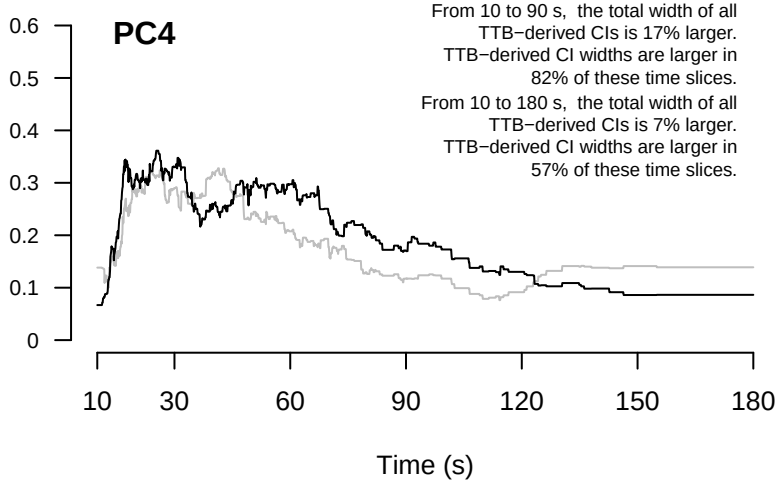
c)

95% CI Widths for WineSip L1





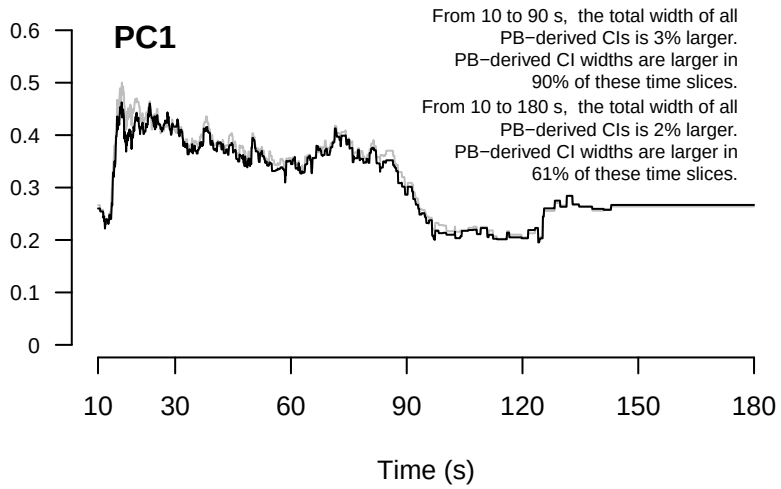


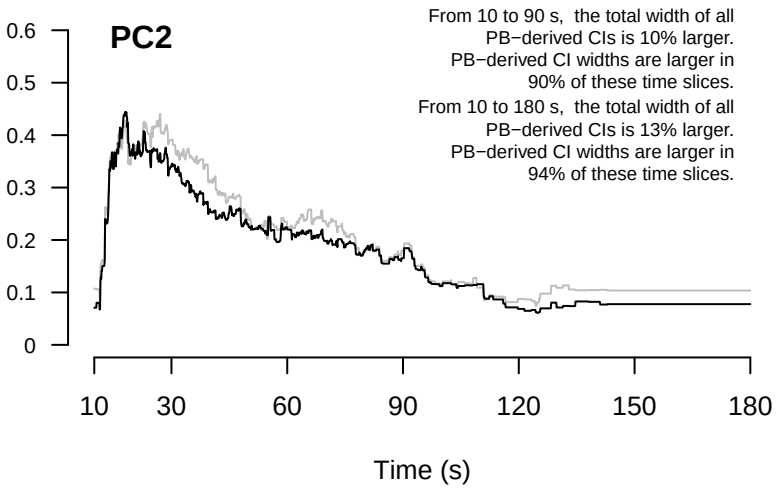


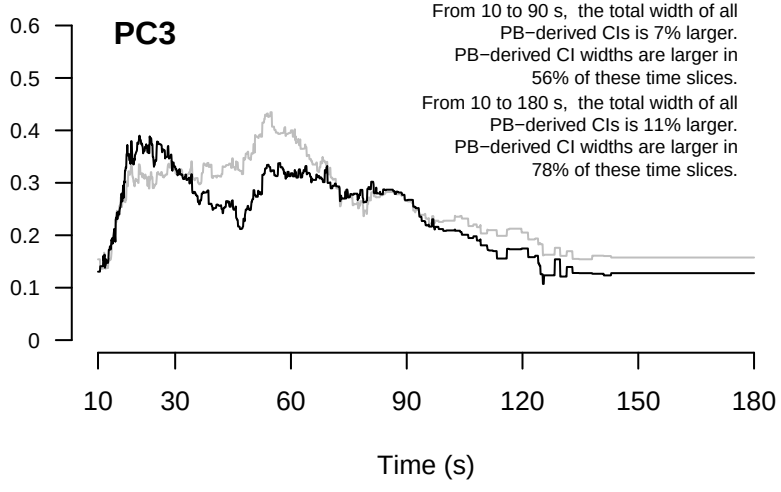
PB TTB

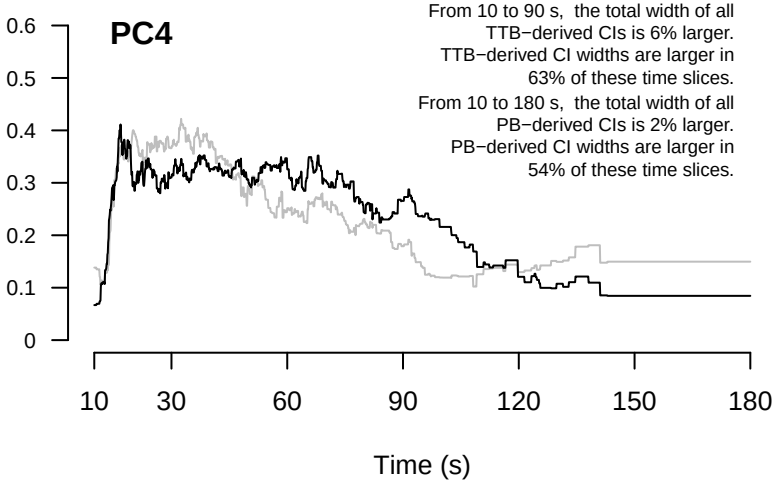
d)

95% CI Widths for WineSip A2





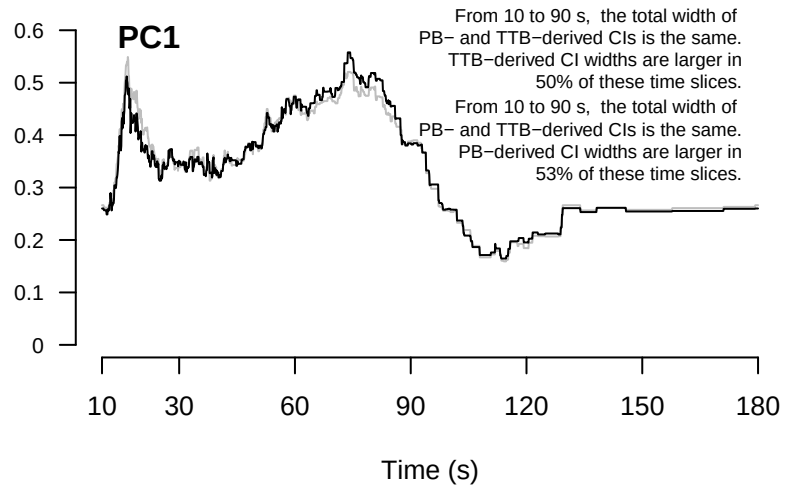


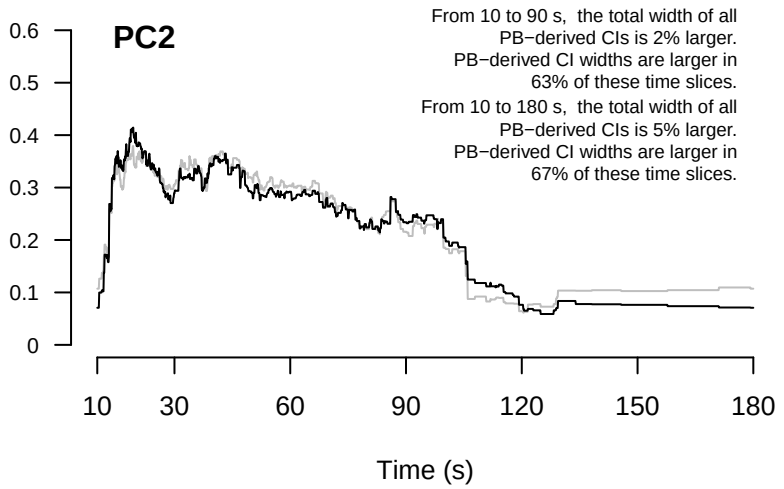


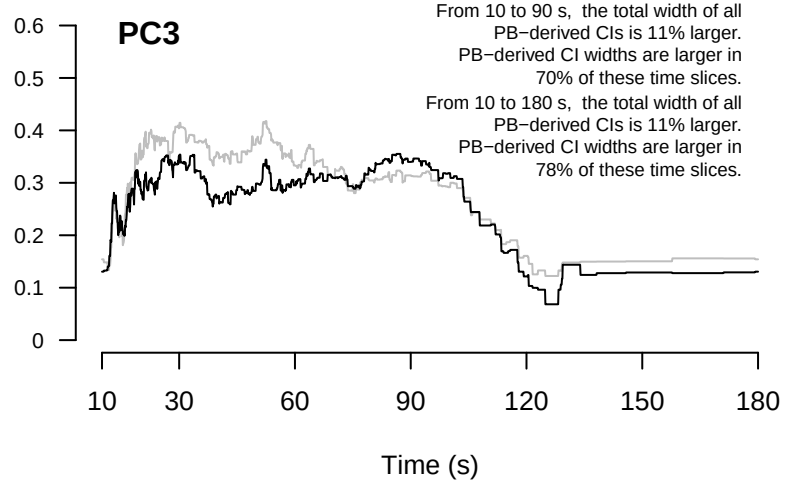
PB TTB

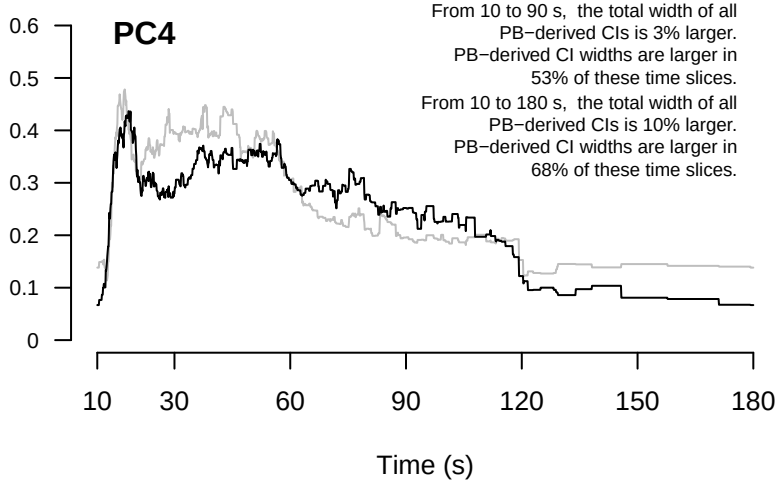
e)

95% CI Widths for WineSip H2





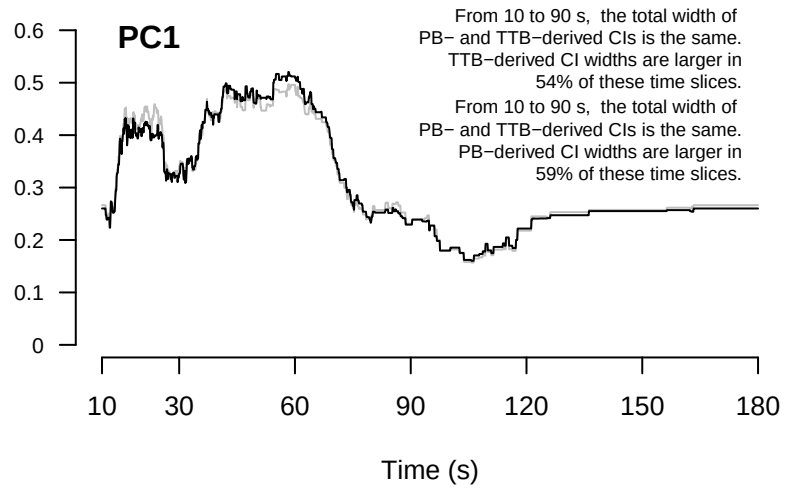


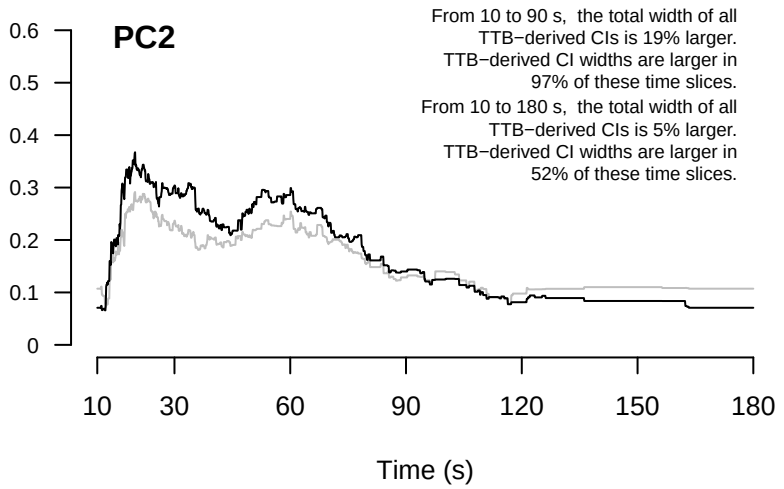


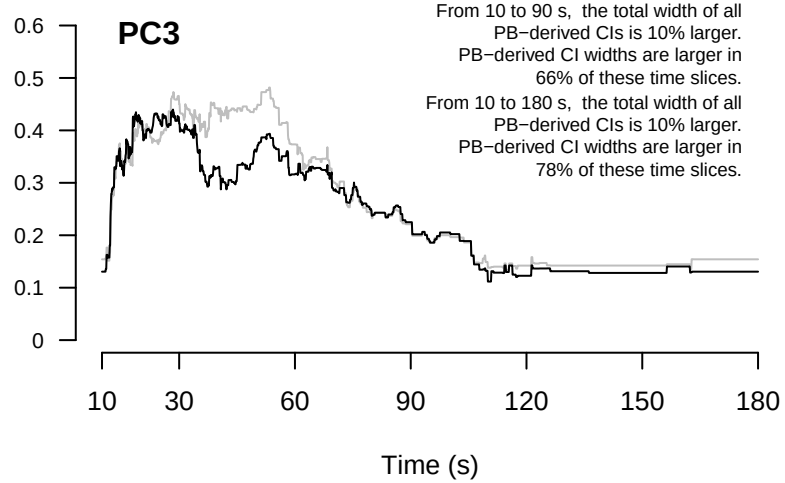
PB TTB

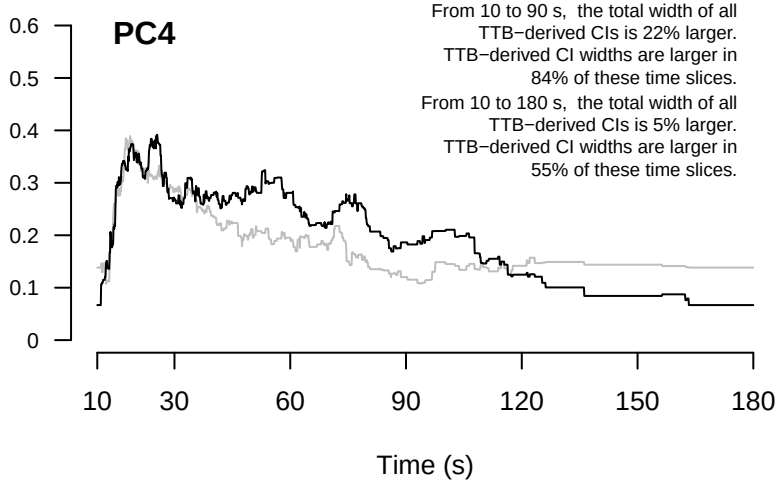
f)

95% CI Widths for WineSip L2





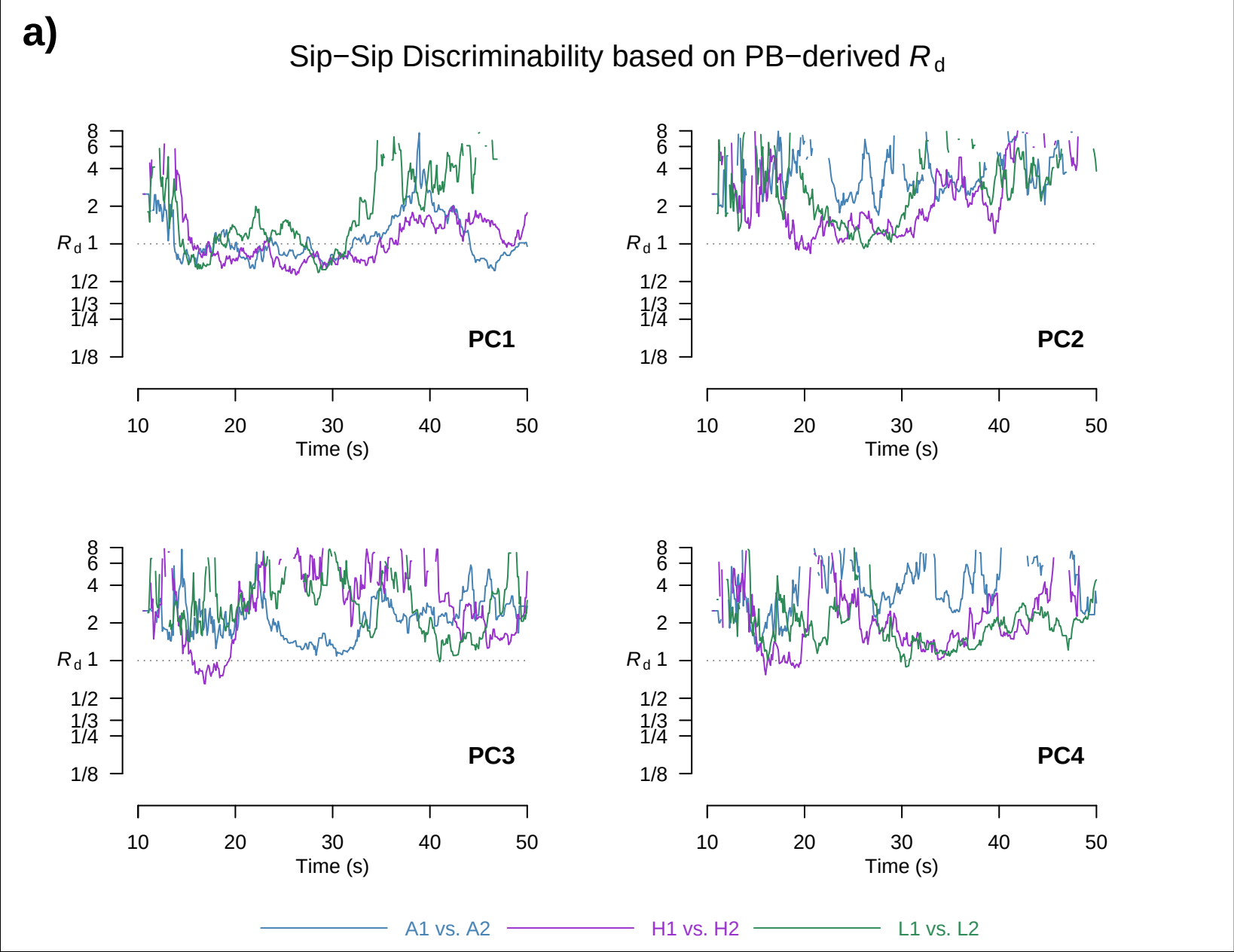




PB TTB

Suppl. Fig. 3.

Reciprocal of discriminability of sips within wine treatments is shown in each of the four PCs over time based on (a) the PB method and (b) the TTB method. Sips were not discriminated outside the times shown. To display the R_d values more clearly, the y-axis is shown on a binary (base 2) logarithmic scale and $R_d > 8$ are suppressed.


a)
Sip−Sip Discriminability based on PB−derived R_d
8
6
4
2
R_d 1
1/2
1/3
1/4
1/8
PC1
PC2
PC3
PC4
10
20
30
40
50
Time (s)
A1 vs. A2
H1 vs. H2
L1 vs. L2

b)