Padua 2017 Abstract Submission

I want to submit an abstract for:
Conference Presentation

Corresponding Author
Jing Cao

E-Mail
jcao@smu.edu

Affiliation
Southern Methodist University

Co-Author/s

Name	E-Mail	Affiliation
Lynne Stokes	slstokes@smu.edu	Southern Methodist University

Keywords
Ranking method, Shapley ranking, Squared-error loss, Percentile loss

Research Question
Which of the ranking methods (i.e., score-average, rank-average, and Shapley rank)provides more accurate ranking result in wine tasting?

Methods
We compare the ranking results of the three methods based on two criteria, the squared-error loss and the percentile loss.

Results
The study shows that the ranking based on score average in general is more accurate than the one based on rank average and the one based on Shapley ranking.

Abstract
Wine quality is an abstract measure that is difficult to define in absolute terms. It leads to debate over how to best aggregate wine tasting scores from a group of wine judges. Two methods are most commonly used in wine tasting due to their simplicity. One is score average which is a simple averaging of numerical scores assigned by the judges. The other is rank average which is the average based on the ranks of wines. The ranks of wines can either come from the conversion of the judges' scores or use the Borda count directly provided by the judges. Each of the methods has its pros and cons. Specifically, score average doesn't take into account the fact that each judge assigns scores using his or her own internal scale, while rank average may avoid distortion introduced by averaging scores assigned by individual judges. On the other hand, the difference in the original scores by individual judges does reflect their perception of difference in wine quality. For example, a 3-points difference in two wines with consecutive order certainly indicates a bigger difference in wine quality than a 1-point difference. However, such difference is no longer reflected if scores are converted in ranks, because the underlying assumption of using ranks is that wine quality changes with equal amount following the rank order.
Recently, another simple-to-use method for aggregating wine tasting scores is proposed, which is called Shapley ranking (Ginsburgh and Zang, 2012). It is a game-theory-based ranking method, where judges are not required to rank order or score all the wines, but only to choose a subset that they find meritorious. Specifically, each judge has one vote and the members of the subset of wines by a judge equally share her one unit of voting. Then the Shapley ranking of each wine is the sum of the shares added over all the judges. In addition to its simplicity, the Shapley method puts less burden on the judges as they do not have to rate or rank all the wines they evaluate. The limitation is that wines are not distinguishable within the subset.
As we can see from the comparison of the pros and cons among these methods, it is difficult to conclude which one is better. Statistically, we can conduct simulation study, where the wine quality is set to be known, to compare

the performance of different methods. In order to make the simulation study informative, the data generating scheme should mimic how the real data is produced. That is, the data-generating model should have a good fit to the real data.
Cao and Stokes (2010) developed a Bayesian ranking model. The model describes judges' different scoring pattern by three quantifiable characteristics: bias, discrimination ability, and random variation. Judge bias measures the systematic difference between a judge's score and the average score from all the judges. Judge discrimination measures a judge's ability to distinguish wines based on their quality. Judge variation measures the size of the random component of variability in a judge's assessment of wine quality. This model provides a way to adjust wine judges' individual internal scale of their assigned scores, yet still maintains the informative difference delivered by the numerical scores.
In this paper, we first investigate which of the following three models provides a better fit to the real data. Model 1 underlies the score average method, Model 2 underlies the rank average method, and Model 3 is the one proposed by Cao and Stokes that incorporates judges' scoring characteristics.
We use the deviance information criterion (DIC) to conduct model comparison. It is a measure of predictive power based on the trade-off between model fit and complexity. Lower DIC values indicate stronger models, and a generally accepted notion is that differences of more than 10 rules out the model with the higher DIC. A difference of less than 5 indicates that the two models are very comparable in terms of model fit. Using the Paris 1976 red wine tasting data, the DIC values for the three models are 296.8, 297, and 259, respectively. Using the Princeton 2012 red wine tasting data, the DIC values for the three models are 243.1, 246, and 207.1, respectively. Both of the real datasets show the same conclusion on the model comparison: Model 1 and Model 2 provide similar model fit to the data, while Model 3 yields much better model fit than the other two models.
Based on the model comparison result, we use Model 3 to simulate data following the setup of the Paris 1976 red wine tasting (i.e., 10 wines rated by 11 judges). To compute Shapley rankings, we use a similar simulation design in (Ginsburgh and Zang, 2012), where 3 cases are considered. The first case assumes that each judge would have chosen three wines; the second case starts with the wine with the highest score for each judge and then goes down until it reaches a gap of two points; the third case assumes that the subset of wines that each judge finds meritorious are the wines with a score that is higher than a certain cutoff (e.g., 15 points).
We compare the ranking results of the three methods based on two criteria. One is called the squared-error loss which calculates the sum of squared differences between the estimated ranks and the true ranks. It is a suitable measure when accurate ranking of all wines is of interest. The other is called the percentile loss which only considers whether the wines are correctly put in a certain subset, e.g., whether the selected best wine is indeed the best wine.
We have generated 1000 datasets and calculated the values from both of the loss functions. The following table summarizes the results. Note that for both of the loss functions, a smaller value indicates the method provides more accurate ranking result. The simulation study shows that the ranking based on score average in general is more accurate than the one based on rank average. More specifically, the score-average based ranking is about 25% more accurate than the rank-average based ranking when accurate ranking of all wines is of interest and 30% more accurate than that when the goal is choosing the very best wine among all wines. Shapley ranking, in general, has inferior performance compared with the other two methods.
Score average Rank average Shapley (case 1) Shapley (case 2) Shapley (case 3)
Squared-error loss 1.58 2.08 3.74 6.28 4.99
Percentile loss 0.25 0.36 0.41 0.26 0.39

We need to point out, however, the simulation setup is quite unfavorable to Shapley ranking, where judges have rated all the wines. In reality, judges would undertake quite different strategy when they are asked to use Shapley ranking. To produce this ranking, they only need to decide a meritorious subset: every wine in the subset is a candidate for the first place or a medal, while non-chosen wines are not. In other words, they don't need to distinguish the wines within the meritorious subset, nor the wines within the non-meritorious set. With this wine tasting rule, judges may place their attention more on dividing wines in two groups, instead of providing scores for all the wines, which may result in more accurate classification of wines in the meritorious subset. It is in our future research plan to conduct simulation study which follows the scheme of Shapley ranking more closely.

Reference
Cao, J. and Stokes, L. (2010). Evaluation of wine judge performance through three characteristics: bias,

discrimination, and variation. Journal of Wine Economics, 5(1), 132-142.

Ginsburgh, V., and Zang, I. (2003). Shapley Ranking of Wines. Journal of Wine Economics, 7(2), 169–180.

File Upload (PDF only)

- abstract-2017AAWE.pdf

Comparison of Different Ranking Methods in Wine Tasting

Jing Cao and Lynne Stokes

Southern Methodist University

Wine quality is an abstract measure that is difficult to define in absolute terms. It leads to debate over how to best aggregate wine tasting scores from a group of wine judges. Two methods are most commonly used in wine tasting due to their simplicity. One is score average which is a simple averaging of numerical scores assigned by the judges. The other is rank average which is the average based on the ranks of wines. The ranks of wines can either come from the conversion of the judges' scores or use the Borda count directly provided by the judges. Each of the methods has its pros and cons. Specifically, score average doesn't take into account the fact that each judge assigns scores using his or her own internal scale, while rank average may avoid distortion introduced by averaging scores assigned by individual judges. On the other hand, the difference in the original scores by individual judges does reflect their perception of difference in wine quality. For example, a 3-points difference in two wines with consecutive order certainly indicates a bigger difference in wine quality than a 1-point difference. However, such difference is no longer reflected if scores are converted in ranks, because the underlying assumption of using ranks is that wine quality changes with equal amount following the rank order.

Recently, another simple-to-use method for aggregating wine tasting scores is proposed, which is called Shapley ranking (Ginsburgh and Zang, 2012). It is a game-theory-based ranking method, where judges are not required to rank order or score all the wines, but only to choose a subset that they find meritorious. Specifically, each judge has one vote and the members of the subset of wines by a judge equally share her one unit of voting. Then the Shapley ranking of each wine is the sum of the shares added over all the judges. In addition to its simplicity, the Shapley method puts less burden on the judges as they do not have to rate or rank all the wines they evaluate. The limitation is that wines are not distinguishable within the subset.

As we can see from the comparison of the pros and cons among these methods, it is difficult to conclude which one is better. Statistically, we can conduct simulation study, where the wine quality is set to be known, to compare the performance of different methods. In order to make the simulation study informative, the data generating scheme should mimic how the real data is produced. That is, the data-generating model should have a good fit to the real data.

Cao and Stokes (2010) developed a Bayesian ranking model. The model describes judges' different scoring pattern by three quantifiable characteristics: bias, discrimination ability, and random variation. Judge bias measures the systematic difference between a judge's score and the average score from all the judges. Judge discrimination measures a judge's ability to distinguish

wines based on their quality. Judge variation measures the size of the random component of variability in a judge's assessment of wine quality. This model provides a way to adjust wine judges' individual internal scale of their assigned scores, yet still maintains the informative difference delivered by the numerical scores.

In this paper, we first investigate which of the following three models provides a better fit to the real data. Model 1 underlies the score average method, Model 2 underlies the rank average method, and Model 3 is the one proposed by Cao and Stokes that incorporates judges' scoring characteristics.

We use the deviance information criterion (DIC) to conduct model comparison. It is a measure of predictive power based on the trade-off between model fit and complexity. Lower DIC values indicate stronger models, and a generally accepted notion is that differences of more than 10 rules out the model with the higher DIC. A difference of less than 5 indicates that the two models are very comparable in terms of model fit. Using the Paris 1976 red wine tasting data, the DIC values for the three models are 296.8, 297, and 259, respectively. Using the Princeton 2012 red wine tasting data, the DIC values for the three models are 243.1, 246, and 207.1, respectively. Both of the real datasets show the same conclusion on the model comparison: Model 1 and Model 2 provide similar model fit to the data, while Model 3 yields much better model fit than the other two models.

Based on the model comparison result, we use Model 3 to simulate data following the setup of the Paris 1976 red wine tasting (i.e., 10 wines rated by 11 judges). To compute Shapley rankings, we use a similar simulation design in (Ginsburgh and Zang, 2012), where 3 cases are considered. The first case assumes that each judge would have chosen three wines; the second case starts with the wine with the highest score for each judge and then goes down until it reaches a gap of two points; the third case assumes that the subset of wines that each judge finds meritorious are the wines with a score that is higher than a certain cutoff (e.g., 15 points).

We compare the ranking results of the three methods based on two criteria. One is called the squared-error loss which calculates the sum of squared differences between the estimated ranks and the true ranks. It is a suitable measure when accurate ranking of all wines is of interest. The other is called the percentile loss which only considers whether the wines are correctly put in a certain subset, e.g., whether the selected best wine is indeed the best wine.

We have generated 1000 datasets and calculated the values from both of the loss functions. The following table summarizes the results. Note that for both of the loss functions, a smaller value indicates the method provides more accurate ranking result. The simulation study shows that the ranking based on score average in general is more accurate than the one based on rank average. More specifically, the score-average based ranking is about 25% more accurate than the rank-average based ranking when accurate ranking of all wines is of interest and 30% more accurate

than that when the goal is choosing the very best wine among all wines. Shapley ranking, in general, has inferior performance compared with the other two methods.

	Score average	Rank average	Shapley (case 1)	Shapley (case 2)	Shapley (case 3)
Squared-error loss	1.58	2.08	3.74	6.28	4.99
Percentile loss	0.25	0.36	0.41	0.26	0.39

We need to point out, however, the simulation setup is quite unfavorable to Shapley ranking, where judges have rated all the wines. In reality, judges would undertake quite different strategy when they are asked to use Shapley ranking. To produce this ranking, they only need to decide a meritorious subset: every wine in the subset is a candidate for the first place or a medal, while non-chosen wines are not. In other words, they don't need to distinguish the wines within the meritorious subset, nor the wines within the non-meritorious set. With this wine tasting rule, judges may place their attention more on dividing wines in two groups, instead of providing scores for all the wines, which may result in more accurate classification of wines in the meritorious subset. It is in our future research plan to conduct simulation study which follows the scheme of Shapley ranking more closely.

Reference

Cao, J. and Stokes, L. (2010). Evaluation of wine judge performance through three characteristics: bias, discrimination, and variation. Journal of Wine Economics, 5(1), 132-142.

Ginsburgh, V., and Zang, I. (2003). Shapley Ranking of Wines. *Journal of Wine Economics*, 7(2), 169–180.