REPARLONS DE NOTES EN PATINAGE

Léo Gerville-Réache

REPARLONS DE NOTES EN PATINAGE

Léo Gerville-Réache[1]

[1]*Université de Bordeaux, CNRS, UMR 5218 IMS, France, leo.gerville-reache@u-bordeaux.fr*

Résumé. Les Jeux olympiques d'hiver de 2022 étant terminés, cette communication vous propose de replonger dans l'une des épreuves qui a permis à la France de capitaliser une médaille d'or de plus. Il s'agit de la danse sur glace. Comme d'habitude, c'est un jury composé de 9 juges qui, via un ensemble de notes, a classé les prestations des 23 couples en compétition. En retirant toujours la meilleure et la moins bonne note de chaque évaluation, il nous a semblé opportun de rappeler et réutiliser les possibilités statistiques qu'offre le modèle de Gauss-Markov.

Mots-clés. Jury, Patinage, Gauss-Makov, JO 2022

Abstract. With the end of the 2022 Winter Olympic Games, this talk invites you to dive back into one of the events that allowed France to capitalize on one more gold medal. This is ice dancing. As usual, it is a jury composed of 9 judges who, through a set of notes, ranked the performances of the 23 couples in competition. By always removing the best and worse score from each evaluation, it seemed appropriate to recall and reuse the statistical possibilities offered by the Gauss-Markov model.

Keywords. Jury, Ice skating, Gauss-Makov, Olympic Games 2022

1. Introduction

Nombre de sports fondent (pour tout ou partie) les résultats d'une compétition sur des notes attribuées par des juges. Dans les sports d'hiver, c'est le patinage artistique qui vient rapidement à l'esprit, mais le plongeon ou encore la gymnastique sont évidemment du même ressort…

Si vous regardez attentivement comment cela fonctionne, vous verrez que chaque juge (un expert reconnu de la discipline) attribue pour chaque prestation de la compétition, une ou plusieurs notes selon un barème de cotation. Pour autant, il est rare que les juges mettent exactement la même note à une même prestation. Il persiste une part de subjectivité (de variabilité) qui est bien normale.

La plupart du temps, sur l'ensemble des notes distribuées par le jury à une prestation, pour calculer la « note finale », on retire la note la plus basse et la note la plus haute. L'idée est de limiter l'impact d'une note, intentionnellement ou non, trop éloignée de la moyenne des notes attribuées. Mais est-ce une bonne d'idée ? Utilise-t-on bien toutes les possibilités de l'analyse statistique ?

Sans faire appel à l'Intelligence Artificielle pour une notation plus juste, via de la reconnaissance automatique liée à des caméras 3D [2], cette communication propose de revenir sur les résultats qui ont conduit le couple Gabriella PAPADAKIS / Guillaume CIZERON à remporter une médaille d'or aux JO de 2022. Via le modèle de Gauss-Markov déjà présenté aux JDS de 1999 [3], cette communication propose de rappeler cette approche statistique sur des données très actuelles.

2. Les données

En cherchant sur le site officiel https://olympics.com, on peut trouver le rapport officiel [1] de la compétition de patinage artistique où le couple français Gabriella PAPADAKIS / Guillaume CIZERON a été sacré champion olympique de danse rythmique sur glace. En moyennant les notes « Program Componen » de chaque jury pour chaque prestation, on obtient le tableau suivant (ici dans l'ordre du classement final qui prend en compte également les notes « Executed Element »). « CL1 » est le couple classé globalement premier et ayant donc obtenu la médaille d'or.

Notes	J1	J2	J3	J4	J5	J6	J7	J8	J9
CL1	9,85	9,70	9,90	9,85	9,85	9,60	9,80	9,70	9,85
CL2	9,75	9,65	9,45	9,65	9,65	9,70	9,65	9,75	9,50
CL3	9,85	9,55	9,75	9,70	9,65	9,45	9,45	9,50	9,35
CL4	9,65	9,15	9,60	9,50	9,50	9,45	9,55	9,35	9,10
CL5	9,60	9,15	9,05	9,25	9,25	9,40	9,25	9,40	9,10
CL6	9,40	9,40	9,15	9,45	9,50	9,30	9,25	9,45	9,05
CL7	9,15	9,20	9,25	9,40	9,15	9,35	9,15	9,25	8,85
CL8	8,65	8,65	8,70	8,95	8,80	8,50	8,65	8,75	8,45
CL9	8,85	8,55	8,70	8,85	8,70	8,65	8,65	8,85	8,70
CL10	8,60	8,50	8,85	8,65	8,85	8,70	8,55	8,65	8,90
CL11	8,50	8,45	8,60	8,80	8,50	7,90	8,30	8,15	8,60
CL12	8,10	8,40	8,40	8,50	8,20	8,15	8,40	8,45	8,15
CL13	8,35	8,15	8,25	8,45	8,30	8,05	7,90	8,10	8,40
CL14	8,05	7,90	8,05	8,30	8,15	8,00	7,90	8,35	8,15
CL15	7,30	7,40	7,70	7,55	7,85	7,80	7,65	8,00	8,00
CL16	8,10	7,55	7,75	7,75	7,70	7,70	7,90	7,95	7,60
CL17	7,40	7,75	7,70	7,60	7,35	7,45	7,35	7,40	7,85
CL18	7,35	7,40	7,45	7,55	7,45	7,25	7,55	7,55	7,95
CL19	7,15	7,50	7,95	8,10	7,55	7,45	7,35	7,40	7,85
CL20	7,65	7,40	7,65	7,40	7,50	7,25	7,65	7,75	7,40
CL21	7,60	7,35	7,50	7,35	7,90	7,25	7,45	7,65	7,45
CL22	7,75	7,40	7,70	7,50	7,55	7,25	7,15	7,25	7,65
CL23	5,70	6,20	7,30	6,10	5,75	6,85	6,50	6,55	7,25

3. Le modèle de Gauss-Markov

On suppose que I compétiteurs sont notés par J juges. Chaque juge note chaque compétiteur indépendamment les uns des autres avec la même échelle. Plus la performance est de qualité, plus la note est haute. Soit X_{ij} la note du juge j sur la prestation i. On suppose que la note X_{ij} est la réalisation de la variable aléatoire suivante :

$$X_{ij} = a_i + b_j + Y_{ij}, i = 1, \dots, I \text{ et } j = 1, \dots, J.$$

Où a_i est la « vraie » valeur de la prestation i, b_j est « l'erreur systématique » du juge j et Y_{ij} est une erreur aléatoire. Dans le modèle de Gauss-Makov, les erreurs Y_{ij} sont des variables aléatoires indépendantes et identiquement distribuées $N(0; \sigma^2)$.

Les valeurs a_i, b_j et σ^2 sont inconnues mais ce n'est pas difficile de les estimer. En posant que $b_1 + b_2 + \cdots + b_J = 0$, on obtient alors des estimations des MCO particulièrement simples des a_i et b_j :

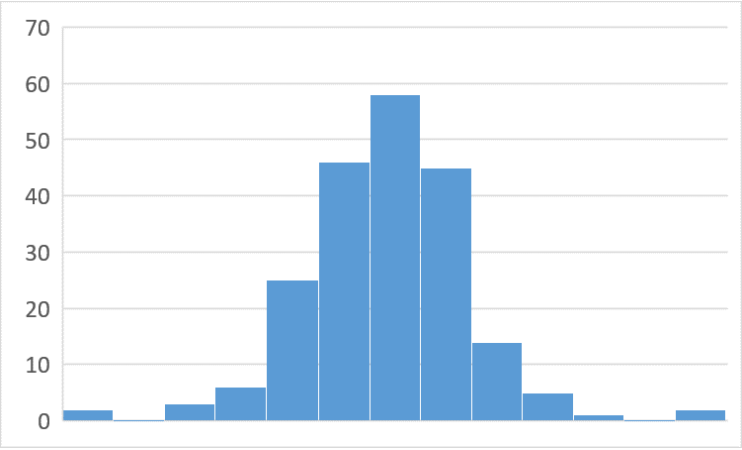
$$\hat{a}_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij} \text{ et } \hat{b}_j = \frac{1}{I}\sum_{i=1}^{I} x_{ij} - \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}$$

Pour identifier une note aberrante, il est possible de mettre en place un test statistique spécifique à ce modèle de Gauss-Markov. Une note aberrante est nécessairement une note qui s'écarte fortement de

la note « attendue » par un juge donné à une prestation donnée. C'est en regardant les y_{ij} (qui sont par définition des écarts (ou erreurs) entre le modèle et les observations que cela est possible. Voici le tableau des $y_{ij} = x_{ij} - \hat{a}_i - \hat{b}_i$.

Y_{ij}	J1	J2	J3	J4	J5	J6	J7	J8	J9	a_i
CL1	0,06	0,00	0,02	-0,02	0,05	-0,10	0,07	-0,12	0,03	**9,79**
CL2	0,11	0,10	-0,28	-0,07	0,00	0,15	0,07	0,08	-0,17	**9,64**
CL3	0,27	0,06	0,08	0,04	0,06	-0,05	-0,07	-0,12	-0,27	**9,58**
CL4	0,23	-0,19	0,09	-0,01	0,06	0,11	0,18	-0,11	-0,36	**9,43**
CL5	0,33	-0,03	-0,31	-0,10	-0,03	0,21	0,04	0,09	-0,20	**9,27**
CL6	0,08	0,16	-0,26	0,04	0,16	0,06	-0,02	0,09	-0,31	**9,33**
CL7	-0,04	0,10	-0,03	0,13	-0,05	0,24	0,02	0,02	-0,38	**9,19**
CL8	-0,02	0,06	-0,06	0,19	0,11	-0,09	0,03	0,04	-0,26	**8,68**
CL9	0,13	-0,08	-0,11	0,05	-0,03	0,01	-0,01	0,09	-0,05	**8,72**
CL10	-0,09	-0,10	0,07	-0,12	0,15	0,09	-0,08	-0,08	0,17	**8,69**
CL11	0,08	0,12	0,09	0,30	0,07	-0,44	-0,06	-0,31	0,15	**8,42**
CL12	-0,20	0,18	0,01	0,12	-0,12	-0,07	0,16	0,11	-0,19	**8,31**
CL13	0,14	0,02	-0,05	0,16	0,07	-0,08	-0,26	-0,15	0,15	**8,22**
CL14	-0,04	-0,10	-0,13	0,13	0,05	-0,01	-0,13	0,22	0,02	**8,09**
CL15	-0,39	-0,20	-0,08	-0,22	0,15	0,19	0,02	0,27	0,27	**7,69**
CL16	0,33	-0,14	-0,11	-0,11	-0,09	0,01	0,18	0,14	-0,21	**7,78**
CL17	-0,14	0,30	0,07	-0,02	-0,20	0,00	-0,13	-0,17	0,28	**7,54**
CL18	-0,15	-0,01	-0,14	-0,03	-0,06	-0,16	0,11	0,02	0,42	**7,50**
CL19	-0,44	0,00	0,27	0,43	-0,05	-0,05	-0,18	-0,22	0,23	**7,59**
CL20	0,14	-0,03	0,05	-0,19	-0,03	-0,18	0,19	0,20	-0,15	**7,52**
CL21	0,10	-0,06	-0,09	-0,23	0,39	-0,16	0,01	0,12	-0,08	**7,50**
CL22	0,29	0,02	0,15	-0,04	0,07	-0,13	-0,26	-0,25	0,15	**7,47**
CL23	-0,76	-0,18	0,75	-0,44	-0,73	0,47	0,09	0,05	0,75	**6,47**
b_j	**0,00**	**-0,09**	**0,09**	**0,08**	**0,01**	**-0,09**	**-0,06**	**0,03**	**0,03**	

Le plus gros écart en valeur absolue est pour le juge 1 sur la prestation du couple 23. L'écart est de précisément -0,76 et la note donnée de 5,7. Ce juge a mis une note très élevée à cette prestation.



NB : La normalité des Y_{ij} est vérifiée via l'histogramme des y_{ij} ci-contre.

Mais cette note est-elle significativement trop élevée ?

Si l'on souhaite tester l'hypothèse H_0 selon laquelle il n'existe aucune note aberrante, on peut procéder ainsi [3,4]. On définit Λ :

$$\Lambda = \frac{IJ}{(I-1)(J-1)} \frac{\max_{ij} y_{ij}^2}{S^2} \quad \text{avec} \quad S^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}^2.$$

Ici, $\Lambda = 0{,}087$. Sous l'hypothèse H_0 la p-value est déterminée de la manière suivante :

$$p = IJ \left[1 - I_z \left(\Lambda; \frac{1}{2}; \frac{1}{2}(IJ - I - J) \right) \right]$$

Où I_z est la fonction Béta incomplète.

La p-value vaut environ 0,015. Cela signifie que la note attribuée par le juge 1 à la prestation du couple 23 est significativement trop basse. La raison n'est pas connue mais la p-value est nette ! Il est alors possible de proposer une correction de cette note grâce à la formule suivante :

$$\tilde{x}_{23;1} = x_{23;1} + \frac{IJ}{(I-1)(J-1)} y_{23;1}$$

La note corrigée $\tilde{x}_{23;1}$ serait alors de 6,60 (au lieu de 5,70).

Maintenant, testons l'hypothèse qu'un juge soit moins précis que les autres juges. Pour cela on commence par mesurer la précision de chaque juge avec la formule suivante [3,4] :

$$\Lambda_j = \frac{J}{(J-1)S^2} \sum_{i=1}^{I} y_{ij}^2$$

Sous l'hypothèse H_0 la p-value est déterminée de la manière suivante :

$$p = J\left[1 - I_z\left(\mathrm{Max}_j \Lambda_j; \frac{I-1}{2}; \frac{(I-1)(J-1)}{2}\right)\right]$$

On trouve ici que $\mathrm{Max}_j \Lambda_j = \Lambda_9 = 0{,}245$. La p-value pour ce juge 9 est de 0,003. Le juge 9 est significativement moins précis que les autres ! C'est à vous de prendre votre décision quant à la qualité de ce juge et la pertinence de ses notes…

Pour finir, qu'en est-il des résultats de la compétition ? En tous les cas, de la partie « Program Componen » de cette compétition. Qu'est-ce-que ce jury est en capacité de juger ?

Cette fois-ci, l'objectif est de tester si un ensemble de prestations donné est statistiquement indiscernable ou non. Pour reformuler : Le jury est-il suffisament précis pour affirmer une différence significative, par exemple, entre les couples CL1 et CL2 par exemple

On cherche à tester l'hypothèse selon laquelle la « vraie » valeur de k prestations d'indice $(i_1, \ldots, i_k) = \Phi$ est la même. C'est-à-dire que :

$$H_0 : a_{i_1} = a_{i_2} \ldots = a_{i_k}.$$

Pour cela, on calcule la statistique [3,4] :

$$Z(\Phi) = \frac{J(J-1)(I-1)}{(k-1)S^2} \sum_{i \in \Phi} \left(a_i - \frac{1}{k} \sum_{i \in \Phi} a_i\right)^2$$

Sous l'hypothèse H_0, $Z(\Phi)$ suit une loi de Fisher à $k-1$ et $(I-1)(J-1)$ degrés de liberté. La p-value du test est alors la suivante:

$$p = F^{-1}_{k-1,(J-1)(I-1)}\big(Z(\Phi)\big)$$

En testant, $a_1 = a_2$, c'est-à-dire que les prestations des couples CL1 et CL2 sont égales (ou « indiscernables ») on obtient $Z(\Phi) = 2{,}25$ et on trouve $p = 0{,}14$. Aussi la différence n'est pas statistiquement significative. Le jury n'est pas en capacité de produire une différence significative entre le couple CL1 et le couple CL2.

Pour cette question de la différence significative entre CL1 et CL2, on aurait pu penser réaliser un test de Student sur deux échantillons appariés. C'est après tout, une approche tout à fait classique. La p-value vaut dans ce cas $p = 0{,}034$. C'est-à-dire qu'avec une approche via le test de Student, la différence entre CL1 et CL2 est significative. Il n'est pas rare que deux approches statistiques produisent des décisions différentes. Pour autant, dans notre cas, il semble que prendre en compte toute l'information contenue dans le tableau de données soit plus pertinente…

4. Discussion

Avec une approche via le modèle de Gauss-Markov, il a été possible de tester la qualité des juges et finalement mesurer la qualité du jury. Ce jury est capable de distinguer certaines prestations mais pas toutes. Par exemple, bien que les moyennes des couples CL1 et CL2 soient différentes (9,79 contre 9,64) celles-ci ne sont pas significativement différentes.

Mais cette approche demande d'attendre la fin de la compétition pour que le classement soit réalisé. Un peu comme une délibération (statistique) du jury... On perd cependant cette « magie » de l'attente du couple en tête de la compétition au fur et à mesure des prestations.

Maintenant, qu'en est-il de la procédure actuelle qui consiste à systématiquement éliminer (du calcul de la moyenne de chaque prestation) la note la plus basse et la note la plus haute ; procédure ayant clairement pour but d'éviter les notes anormales (dans un sens ou dans l'autre).

Le plus important est de comprendre qu'au lieu d'avoir dans notre cas, 9 notes par prestation, il n'en reste plus que 7. Pourtant, si les juges sont « honnêtes », c'est une perte d'information et de précision importante. En retirant systématiquement 2 notes de manière préventive, on se prive potentiellement d'une plus grande précision du jury.

Pour autant, chaque juge sachant que s'il note trop bas ou trop haut par rapport aux autres juges, sa note ne sera pas prise en compte a un avantage. Ce principe « oblige » les juges à maitriser leur notation qui, dans le cas contraire, pourraient voir leur note non prise en compte…

L'un des écueils de la procédure qui élimine systématiquement la meilleure et la moins bonne note, c'est qu'une note anormale peut parfaitement se glisser dans les notes conservées [2]. Clairement, un juge qui note bas comme le juge 2 et qui une fois note haut (mais pas trop) va voir sa note anormale conservée ! De plus, un juge qui aurait une bonne évaluation des prestations mais qui surnoterait systématiquement verrait ses notes retirées alors qu'il classe correctement les prestations.

La recherche d'une procédure « optimale » est délicate. Cette optimalité n'existe sans doute pas. Comme toute procédure d'évaluation, elle doit être honnête et claire. Il est sans doute difficile d'admettre l'idée que les notes soient retenues et « délibérées » en fin de compétition via le modèle de Gauss-Markov, même si statistiquement, cette approche est pertinente. Il est également difficile d'admettre qu'un jury, dont la précision n'est pas infaillible, classe des prestations alors qu'elles sont clairement indiscernables pour ce jury.

On dit souvent que le jury est souverain. C'est sans doute le principe qui permet d'accepter, le plus souvent, le résultat prononcé…

Bibliographie

[1] https://olympics.com/beijing-2022/olympic-games/static/owg2022/pdf/OWG2022/FSK/OWG2022_FSK_C77B_FSKXICEDANCE----------QUAL000100--.pdf

[2] *JO 2022*. En patinage artistique, l'Intelligence Artificielle pour une notation plus juste ?, *Journal Ouest France*, 13 Février 2022.

[3] Gerville-Réache L., Nikulin M.S. (1999). Analyse statistique de l'évaluation des sportifs par un jury, *Journées de statistique*, Grenoble, France, 51-54.

[4] Greenwood P.E., Nikulin M.S. (1996) *A Guide to Chi-Squared Testing* John Wiley & Sons. Chap. 24 : Statistical Evaluation of Ratings in Sports Compétitions Such as Skating or Gymnatics.