F. NAKHLÉ

Sur l'analyse d'un tableau de notes dédoublées

Les cahiers de l'analyse des données, tome 1, n° 4 (1976), p. 367-379

<http://www.numdam.org/item?id=CAD_1976__1_4_367_0>

Les Cahiers de l'Analyse des Données
Vol. I - 1976 - n° 4 - p. 367-379

SUR L'ANALYSE D'UN TABLEAU DE NOTES DÉDOUBLÉES [DEDOU II]

Exposé des résultats de la thèse de F. Nakhlé (1)

Les méthodes considérées dans un précédent article ([Dédou I], Vol. I, n° 3 pp.243-257) sont ici appliquées au concours d'admission à l'Ecole Polytechnique. Les résultats obtenus par F. Nakhlé sont présentés de telle sorte que les lecteurs intéressés par l'analyse du concours puissent la suivre sans avoir fait des formules générales une étude approfondie.

3. Analyse statistique des notes obtenues par les candidats à l'Ecole Polytechnique :

Le modèle chinois d'une hiérarchie sociale fondée sur la réussite à des concours avait séduit des penseurs de notre XVIII° siècle. Si la France n'a pas instauré une suite d'épreuves conduisant par degrés solennels du rédactoriat à la magistrature suprême, du moins l'admission à l'Ecole Polytechnique - à l'X dit-on avec une familiarité qu'on ne prendra pas pour de la bonhomie! - est-elle depuis 180 ans comme le saut de haie au départ de plus d'une course illustre. On s'intéressera donc à l'analyse des notes obtenues par les candidats au concours de l'X.

Après avoir précisé les données traitées (§ 3.1), nous donnerons les résultats de l'analyse des correspondances du tableau des notes dédoublé (§ 3.2; cf § 1). Ces résultats sont accusés jusqu'à sembler caricaturaux : on tentera de les tempérer tout en les confirmant, d'une part en calculant selon l'usage des moyennes des écarts-type et des corrélations (§ 3.3), d'autre part en transformant les échelles de quelques notes (§ 3.4; cf 2).

3.1 Les données analysées :

3.1.1. Le concours de l'X :

Le concours comporte un écrit et deux oraux, appelés le petit "o" et le grand "O". Nous énumérons d'abord ces épreuves en introduisant au passage les sigles trilittères qui les désignent dans nos analyses (cf tableau 1). L'écrit comporte 9 épreuves.

Math.1 (EM1), Math.2 (EM2), Physique (EPH), Chimie (ECH), Français 1 (FR1), Français 2 (FR2), Calcul numérique (CUN), Dessin industriel (DEG), Langue vivante (ELV). La première épreuve de Français (FR1) est une dissertation; la deuxième est un résumé (FR2). Pour la langue vivante, il

(1) *Docteur de 3ème cycle en statistique - Université Pierre et Marie Curie.*

(*) *Suite de l'article publié dans les Cahiers de l'analyse des données vol. I, 1976, n° 3 pp. 243-257.*

y a choix entre l'allemand, l'anglais et le russe; les slavisants sont très rares; un candidat sur 5 est germaniste; la grande majorité (4/5) choisissent l'anglais.

Le petit "o", dit aussi barrière des mathématiques, comporte selon les candidats une ou deux épreuves de cette matière.

Le grand "O" cumule des épreuves intellectuelles, des épreuves sportives et la majoration :

Epreuves intellectuelles : Oral Math.1 (OM1), Oral Math.2 (OM2), Physique (OPH), Chimie (OCH), langue vivante (OLV). Epreuves sportives : Course de 100 mètre (100), 800 mètres (800), Saut en hauteur (HAU), Saut en longueur (LON), Lancer de poids (LAN), Natation (NAT). Enfin la majoration, qui additionne suivant une formule variant au cours des années, les marques des talents les plus divers dont les candidats ont pu faire la preuve : équitation, escrime, version grecque au baccalauréat... Le Dessin d'imitation, épreuve obligatoire jusqu'en 1970, fait depuis l'objet d'une majoration.

Le jeu de ces diverses séries de notes est assez complexe. Le total général, d'après lequel est décidée chaque année l'admission définitive des quelque 300 candidats qui "intègrent l'X", ne comporte que les épreuves de l'écrit et du grand "O", (notes sur 20, affectées des coefficients qu'on lira au tableau 3 et majoration ou bonification supplémentaire), à l'exclusion de celles du petit "o" que tous les candidats admis n'ont pas passées. En effet sont admis directement au grand "O" d'une part les candidats qui y ont déjà été admis une précédente année (anciens admissibles) et d'autre part les candidats dits "hyper-A" ayant totalisé aux cinq premières épreuves écrites (1° écrit) plus de 420/600, soit une moyenne supérieure à 14/20. Les autres candidats n'accèdent au grand "O" que pour autant qu'après avoir obtenu au premier écrit un total suffisant pour être admissibles au petit "o" (c'est ce qu'on appelle la sous-admissibilité), ils ont pu franchir ensuite cette barrière des mathématiques (être "grand-A").

3.1.2. *Les notes disponibles* :

Grâce à l'obligeante introduction du Pr. P. Robert, examinateur de Mathématiques au grand "O", F. Nakhlé a eu accès aux archives des concours 1970, 1971, 1972. Les données qu'il a relevées concernent exclusivement les candidats ayant passé la totalité des épreuves de l'écrit et du grand "O" (anciens admissibles, hyper-A, et grand-A). Pour 1970 et 1971 le détail des épreuves sportives avait été conservé, mais non pour 1972. La formule en ayant changé chaque année, F. Nakhlé n'a pas relevé le nombre de points de majoration acquis par les candidats : on regrettera cette omission lors de l'interprétation, où un élément supplémentaire, même incertain, oriente souvent une conjecture. Signalons pour mémoire que quelques candidats étrangers, dispensés des épreuves littéraires (Français et langue étrangère) n'ont pas été comptés. Relativement aux candidats, l'anonymat le plus strict a été imposé par la direction des études de l'E.P., bien qu'en leur temps toutes les notes aient été affichées en détail avant la proclamation des résultats. Sur les tableaux constitués par F. Nakhlé, les candidats sont donc désignés par leur numéro; et dans la suite, tous les résultats statistiques présentés - cartes et facteurs fournis par l'analyse de correspondance, moyennes, écarts-types, etc. - concernent exclusivement les épreuves.

3.1.3. *Les tableaux considérés* :

Ils contiennent des notes affectées de leurs coefficients : e.g. pour l'épreuve écrite de mathématiques EM1 dont le coefficient est 7, des notes pouvant varier de 0 à 140, (dans les notations du § 1.1 on a (EM1) = 140); et à ces notes sont adjointes leurs complémentaires désignées non par un signe - (q^- = complémentaire de q^+, cf § 1.1), mais par un sigle trilittère commençant par la lettre K : e.g. KM1 pour la complémentaire de EM1; pour tout individu i, k(i,KM1) = 140 - k(i,EM1). Ces sigles sont dans le tableau 1.

En réunissant les concours 1970 et 1971 on dispose de 911 séries (*) de 20 notes obtenues à l'ensemble des épreuves, écrites ou orales, intellectuelles ou sportives; d'où après dédoublement, un tableau 911 x 40. En réunissant les concours 1970-1971-1972,on a 1386 séries de 14 notes (sans les épreuves sportives); d'où un tableau 1386 x 28. Il est possible de faire des analyses partielles par année. Enfin, détail surprenant, bien que les notes de langue soient toutes conservées, il n'a été possible de savoir que pour la moitié des candidats environ quelle était la langue choisie (anglais, allemand ou éventuellement russe) : d'où un tableau 606 x 28, construit à partir des 606 séries de 14 notes pour lesquelles était connue la langue choisie.

Les corrélations mutuelles des épreuves étant très stables dans le temps, et les contributions des épreuves sportives infimes, toutes ces analyses fournissent sur l'ensemble des épreuves intellectuelles des résultats qui concordent. Nous rapporterons donc principalement les résultats de l'analyse la plus complète : celle du tableau 911 x 40 relatif à 1970 et 1971 et cumulant épreuves intellectuelles et sportives.

3.2. L'analyse de correspondance des concours de 1970 et 1971 :

Dans toutes les analyses effectuées, seuls les deux premiers facteurs se sont prêtés à une interprétation certaine : nous ne parlerons donc que de ceux-ci. Avec le tableau 911 x 40 des notes des concours de 1970 et 1971, on a :

$\lambda_1 = 0{,}013$; $\tau_1 = 18{,}03$ %; $\lambda_2 = 0{,}0097$; $\tau_2 = 13{,}36$ %.

(pourcentage et v.p. stables au cours des analyses, comme les facteurs eux-mêmes). Les facteurs sont visibles sur la figure 4, à laquelle le tableau 1 apporte d'utiles précisions numériques.

3.2.1. Le premier facteur :

Il oppose les points q^+ (notes brutes) aux points q^- (notes complémentaires) : c'est donc un facteur de niveau général. Un sujet i a des notes d'autant meilleures que le facteur $F_1(i)$ est plus élevé : car alors dans le profil dédoublé de i, les notes q^+ ont relativement plus d'importance qu'elles n'en ont dans le profil moyen. Ce résultat était tout à fait prévisible; mais certains détails arrêtent l'attention. Pourquoi les épreuves sportives ont-elles toutes un facteur G_1 positif ? (trois d'entre elles devancent même sur le premier axe la première épreuve écrite de mathématique EM1). Y-a-t-il une corrélation insoupçonnée entre l'aptitude au lancer du poids et l'endurance au concours ? Le parfait taupin, athlète et géomètre se conforme-t-il au modèle antique du καλοσ καγαθοσ ? L'explication est sans doute autre : les épreuves sportives se passent après toutes les autres; et les candidats ne se dépensent à celles-là que pour autant que le niveau de celles-ci leur laisse espérer une réussite : d'où la corrélation positive observée entre sport et intellect. Mais que font loin en tête sur le premier axe les deux épreuves de langues (écrite et orale : ELV, OLV) encadrant le résumé de texte français (FR2) et suivies du dessin industriel DEG : la raison en est-elle que tous les candidats admis au grand "O" étant assez forts en maths, ils ne sont classés que suivant leurs aptitudes littéraires, ou graphiques. Assurément les profils des notes de mathématiques sont très plats (cf 3.4) : à l'oral les pointes à 18 ou 19 sont rares; les défaillances sanc-

(*) *Il vaut mieux dire* série de notes *qu'individu : car un même individu peut fournir plusieurs séries de notes à des concours successifs.*

tionnées par une note inférieure à 10, inexistantes; au contraire le niveau en langue varie largement du pire au meilleur (et l'épreuve OLV est la seule épreuve orale à ne pas avoir un profil plat) : l'analyse de correspondance accuse ces différences de profils qui sont son objet propre. Cependant l'importance des mathématiques au grand "O" subsiste à côté de celle de la langue vivante; seulement le rapport d'importance n'est pas mesuré par le rapport des coefficients : une épreuve de langue à coefficient 10, notée carrément de 0 à 20, arrive à peser dans le total d'admission presque autant qu'une épreuve de maths dont le coefficient est 25 mais qui est notée de 10 à 17. L'analyse de correspondance se prête aussi à des changements d'échelle, laissant invariants les facteurs $F_\alpha(i)$ sur l'ensemble I (candidats) mais modifiant les $G_\alpha(j)$, coordonnées des épreuves; nous y reviendrons au § 3.4.

3.2.2. *Le deuxième facteur* :

Dans le plan 1 x 2, de par le principe du bras de levier (cf § 1.2.2°, fig. 1) une note et sa complémentaire (point q^+ et point q^-) sont alignées avec l'origine de part et d'autre ce celle-ci; de plus la moyenne sur 20 de la note à une épreuve étant comprise entre 7 et 14, les points q^+ et q^- ne diffèrent pas en poids de plus d'un facteur 2, et leurs distances à l'origine sont de même dans un rapport compris entre 0,5 et 2. Globalement, l'ensemble J des notes et de leurs complémentaires apparait dans le plan 1 x 2 comme une figure à peu près symétrique à quatre pointes deux du côté $G_1 > 0$ (notes q^+) et deux du côté $G_1 < 0$ (notes q^-). Il suffit à l'interprétation de considérer le demi plan $\{F_1 > 0, G_1 > 0\}$, où se trouvent avec les notes q^+ les candidats reçus.

Du côté du deuxième axe positif, une pointe conduite par DEG, le dessin industriel, suivi de EPH, l'écrit de physique, CUN, le calcul numérique, EM2 la deuxième épreuve écrite de mathématiques; et avec ces notes, celles de toutes les épreuves scientifiques et aussi de l'athlétisme (c'est-à-dire toutes les épreuves sportives exceptée la natation).

Du côté du deuxième axe négatif, une autre pointe, ELV, l'écrit de langue vivante, suivi de OLV, l'oral de langue, FR2, le résumé de texte français, puis FR1, la dissertation, avec NAT la natation sans rien d'autre. Il est sous tous les regards, le type de ce polytechnicien aux évolutions aisées, habile à réduire les théories en arguments et en formules qui emportent l'assentiment ! Moins en vue, même s'il est pilote d'essai ou spécialiste des fusées, le polytechnicien ingénieur ($F_2 > 0$) existe pourtant et son rôle dans les chefs d'oeuvre techniques français mérite le respect.

N'est-il pas étonnant qu'après le facteur de niveau général (1° facteur), le seul facteur sûrement interprétable représente non l'opposition entre oral et écrit, ou entre mathématiques pures et physique-chimie, mais celle omni-présente entre littéraire et scientifique ? Plusieurs années d'une sévère préparation mathématique fournissent à l'Ecole des promotions dont la diversité reproduit celle de l'élite des classes du second degré.

Mais s'agit-il exactement de l'opposition entre lettres et sciences ? Que la natation aille avec le français et la langue étrangère offre matière à des allégories sur le verbe nager qui pour être impertinentes ne sont peut-être pas totalement hors de sens. Nous n'en dirons pas plus : citons plutôt *in extenso* l'interprétation de F. Nakhlé :

"Le deuxième axe oppose les disciplines littéraires et la natation aux disciplines scientifiques et à l'athlétisme. La natation acquise se conserve aisément sans entraînement intensif tandis que l'athlétisme exige un entraînement plus régulier et plus intensif. Toutefois, on prendra en considération que les candidats ne sont pas des *sportifs pro-*

fessionnels. Ce n'est pas aux classes préparatoires que les étudiants améliorent leurs performances sportives, du fait qu'ils y pratiquent peu le sport. Or il est beaucoup plus grave de rester deux ans sans faire de sport de stade que de la natation. En ce qui concerne les performances intellectuelles, disons que la littérature est aux sciences et au dessin ce que la natation est à l'athlétisme (toujours dans le cas de sportifs non-professionnels). L'une acquise se conserve aisément à l'esprit sans entretien intensif, tandis que les autres demandent un renouvellement et une adaptation constante. Et il semble qu'aux classes préparatoires les étudiants s'intéressent moins aux disciplines littéraires qu'aux disciplines scientifiques et au dessin industriel, discipline qui demande un *soin* particulier.

En bref selon F. Nakhlé, s'opposent sur le 2° axe d'un côté l'*acquis* antérieur - lettres et natation, $G_2 < 0$ - de l'autre les matières requérant une préparation *immédiate* intensive, sciences avec athlétisme, $G_2 > 0$. Selon nous il est vrai que du côté $G_2 < 0$ l'on trouve un *acquis* antérieur que les années de préparation à l'X ont pu entamer plus souvent qu'enrichir. Cependant le point DEG - dessin industriel - à l'extrémité positive du 2° axe représente peut-être aussi un acquis antérieur. Pas plus qu'ils ne sont des champions olympiques les candidats ne sont des dessinateurs professionnels; la culture scientifique d'un étudiant de vingt ans s'ajoutant à des habitudes graphiques acquises cinq ans plus tôt dans un collège d'orientation technique peuvent procurer une excellente note de dessin industriel. On lira donc sur l'axe 2 une diversité d'orientation d'esprit et de compétence, à laquelle contribuent outre les qualités innées du candidat, l'éducation secondaire qu'il a reçue et aussi son origine sociale (une certaine aisance offrant à la fois des vacances à la mer et des séjours à l'étranger : $F_2 < 0$).

Pour préciser cette interprétation et avoir dans le plan 1 x 2 une image vivante d'une promotion, il faudrait ajouter au tableau des données de nombreuses informations supplémentaires qui permettraient aussi de poursuivre l'interprétation au delà du 2° facteur. Nos conjectures quant à l'origine sociale des candidats seraient éclairées si figuraient les centres de gravité des classes d'individus dont le père exerce une profession déterminée (médecin, militaire, artisan, etc.). Leur culture personnelle et leurs talents apparaîtraient si les diverses majorations (cf § 3.1.1) étaient de même inscrites sur la carte 1 x 2. Il serait plus instructif et plus fructueux encore de compléter l'analyse 2 ans, 5 ans, 10 ans après le concours, en marquant les centres de classes définies par les destinées ultérieures des candidats : situer parmi les admissibles ayant échoué au concours considéré ceux qui ont réussi un an après; et parmi les reçus ceux qui ont avec des succès divers embrassé diverses carrières. Ainsi l'ensemble des notes d'une ligne du tableau ne serait pas perçu que par un seul total, déterminant le rang du candidat; mais on distinguerait diverses dimensions aidant à un pronostic; il serait possible de modifier les coefficients des épreuves, voire l'orientation de celles-ci, pour que les candidats reçus fussent plus conformes à ce qu'on attend d'eux : car on ne se bornerait pas à élever ou à abaisser la barre fatidique qui sépare les reçus des échoués : on pourrait l'incliner. L'analyse aiderait au perfectionnement du concours.

3.2.3. Considération de l'allemand :

Les données manquent pour enrichir le tableau des notes des nombreux éléments supplémentaires que nous voudrions adjoindre à N(I) et N(J). Nous n'en avons qu'un seul fourni par le choix de la langue : nous connaissons la langue choisie dans 606 des 1386 séries de notes fournies par les concours 1970-1971-1972; dans 124 cas sur 606 cette langue est l'allemand. F. Nakhlé a analysé le tableau dédoublé 606 x 28 des 606 séries de notes aux épreuves intellectuelles, en y adjoignant un individu supplémentaire i_{ALL}, dont la ligne est la moyenne (ou équivalemment la somme : puisque seul compte le profil de la ligne) des 124 lignes où est choisi l'allemand. Il vient :

$F_1(i_{ALL}) = 0{,}011$; $F_2(i_{ALL}) = -\ 0{,}018$.

Ce qui selon l'interprétation acceptée ci-dessus, signifie que les germanistes sont d'un niveau un peu supérieur à la moyenne ($F_1 > 0$), et plutôt littéraires ($F_2 < 0$). Ce résultat suggère une suite de réflexions.

a) Sur l'ensemble I des individus, le facteur F_α a moyenne nulle, variance λ_α, et écart type $\lambda_\alpha^{1/2}$; la moyenne du facteur F_α sur un sous-échantillon de 124 individus tirés au hasard est donc une variable aléatoire de moyenne nulle, variance $(\lambda_\alpha/124)$ et écart-type $(\lambda_\alpha/124)^{1/2}$. Comme cette dernière variable de par le théorème central limite a une distribution à peu près normale il est facile d'apprécier si l'écart entre $F_\alpha(i_{ALL})$ et zéro est significatif. On trouve :

$F_1(i_{ALL})/(\lambda_1/124)^{1/2}$ # 1,1;

$F_2(i_{ALL})/(\lambda_2/124)^{1/2}$ # 1,9;

l'écart sur l'axe 1 ne serait donc guère significatif; celui sur l'axe 2 l'est nettement (e.g. au seuil de 5 %).

b) Il n'était pas nécessaire de refaire une analyse factorielle à propos du tableau 606 x 28; il suffisait d'adjoindre la ligne supplémentaire i_{ALL} au tableau dédoublé 1386 x 28 de toutes les séries de notes intellectuelles que nous possédons; il eût encore été possible d'adjoindre une ligne i_{ALL} (moyenne des germanistes) au tableau dédoublé 911 x 40 des séries complètes (notes intellectuelles et notes de sports); la moyenne des notes de sports étant calculée sur les germanistes (les 2/3 environ du total) des années 1970-1971. L'intérêt de tous ces essais est de confirmer la stabilité des résultats.

c) Comme le remarque F. Nakhlé le choix de la langue peut fournir outre l'individu supplémentaire i_{ALL}, un caractère supplémentaire j_{ALL} défini comme suit :

$k(i,j_{ALL})$ = si la série i comporte une note d'allemand alors 1, sinon zéro.

Les lignes du tableau analysé ayant toutes même poids, F. Nakhlé démontre que :

$$F_\alpha(i_{ALL}) = \lambda_\alpha^{1/2}\, G_\alpha(j_{ALL})$$

d) D'autres éléments supplémentaires s'offrent encore. On posera :

$$k(i_{AL+}, j) = \Sigma\{k(i,j) \times k(i, j_{AL+}) \mid i \in I\};$$

$$k(i_{AL-}, j) = \Sigma\{k(i,j) \times k(i, j_{AL-}) \mid i \in I\};$$

où $k(i, j_{AL+})$ est la note de langue (e.g. la note d'écrit ELV) de l'individu i si celui-ci a choisi l'allemand et zéro sinon; et de même $k(i, j_{AL-})$ est la complémentaire de la note de langue de i, si celui-ci est germaniste et 0 sinon. Autrement dit j_{AL+} et j_{AL-} sont des colonnes incomplètes de notes de langue allemande, complétées par des zéros pour les non-germanistes; et ces colonnes fournissent les coefficients de pondération des lignes supplémentaires i_{AL+} et i_{AL-}. Comme en c) on a :

$F_\alpha(i_{AL\varepsilon}) = \lambda_\alpha^{1/2} \; G_\alpha(j_{AL\varepsilon})$; $\varepsilon = \pm$.

Les trois points AL+, ALL, AL- sont alignés; avec ALL au barycentre de AL+ et AL-, celui-là ayant pour masse la moyenne des notes d'allemand, celui-ci la complémentaire de cette moyenne. Il serait instructif de comparer j_{AL+} à ELV et j_{AL-} à KEL. Si par exemple j_{AL+} est plus écarté que ELV sur le 2° axe négatif, c'est que les meilleurs germanistes ont un caractère littéraire plus accusé que celui des meilleurs linguistes pris ensemble.

Nous nous sommes attardé sur la considération de l'allemand parce qu'elle donne un exemple de méthode pour d'autres études.

3.3. *Moyennes, Ecarts-type, Corrélations* :

Dans ce § nous considérons les résultats de calculs assez simples pour avoir été effectués depuis longtemps sans le secours de l'ordinateur. Ces calculs devenus aujourd'hui plus faciles sont toujours utiles; parfois, tout en permettant une vue d'ensemble des données, l'analyse factorielle suggère l'intérêt de certaines études de détail dont le choix eût été difficile *a priori*. Ici nous nous bornerons à étudier la contribution de chaque note prise séparément au résultat général; sans considérer les corrélations des notes deux à deux, même pour les paires les plus intéressantes qu'a distinguées l'analyse factorielle (e.g. OM1 et ELV).

3.3.1. *Moyennes* :

Toutes les épreuves intellectuelles ont pour leur note sur 20 (note non affectée de son coefficient) une moyenne supérieure à 10 : de 10, 17, pour ELV (écrit de langue) à 14,32 , OM1 (oral de Math.); toutes les épreuves d'athlétisme ont une moyenne inférieure à 10; seule entre les épreuves sportives la natation a une moyenne supérieure à 10 (12 ,76).

3.3.2 *Ecarts- type* :

L'écart-type de la note sur 20 varie considérablement d'épreuve à épreuve : comme l'a fait voir l'analyse factorielle (cf 3.2.1) certains profils sont très plats : ainsi les notes sur 20 des épreuves orales de mathématiques ont les plus faibles écarts-type ; 1,31 pour OM1; 1,54 pour OM2. Les examinateurs de mathématiques du grand "O" ne se permettent pas d'user pleinement du formidable coefficient qui leur est attribué (cf *infra*). Au contraire parmi les épreuves intellectuelles, le dessin industriel (DEG : σ = 4,42) et l'écrit de langue (ELV : σ = 4,11) se signalent par les plus forts écarts-type : et justement ces deux points se signalent aux deux extrémités de l'axe 2. Cependant dans le total général du concours, chaque note entre affectée d'un coefficient k : c'est le produit kσ qui mesure le mieux l'amplitude du rôle d'une épreuve dans les résultats du concours. Avec leur coefficient de 25, les deux épreuves orales OM2 et OM1 prennent alors la tête (kσ = 38,4 et 32,8); mais l'oral de physique OPH, l'oral de langue OLV et le dessin industriel malgré des coefficients plus faibles (respectivement 14, 10,6) suivent à peu de distance (kσ #27 pour les trois épreuves).

3.3.3. *Corrélations* :

Nous avons vu qu'une seule note au profil très plat (OM1, OM2) peut par un très fort coefficient, acquérir une amplitude de variation considérable dans le total général qui décide du destin des candidats. Une autre question est de savoir si la note d'une épreuve particulière X

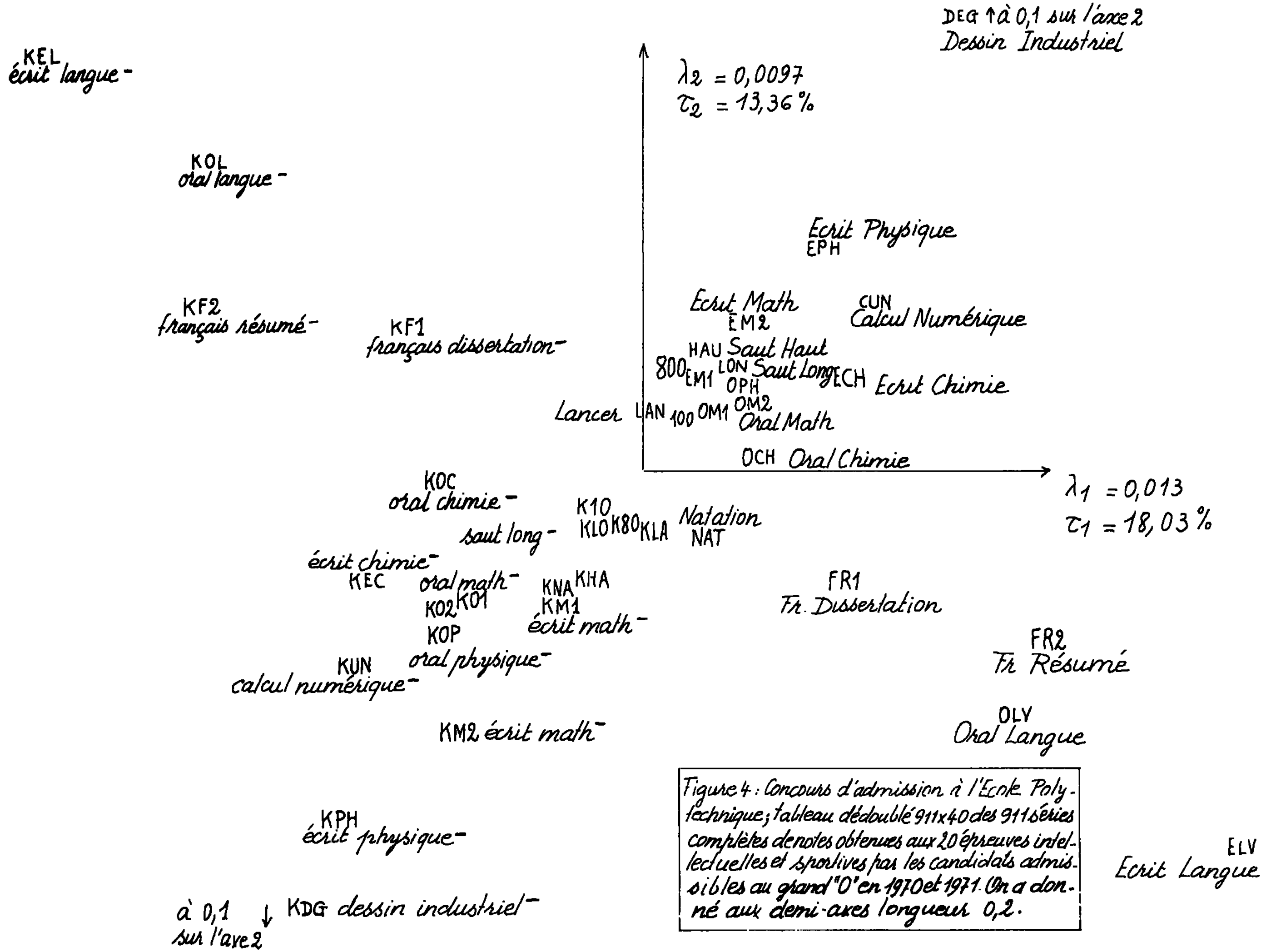

Figure 4 : Concours d'admission à l'Ecole Polytechnique; tableau dédoublé 911x40 des 911 séries complètes de notes obtenues aux 20 épreuves intellectuelles et sportives par les candidats admissibles au grand "O" en 1970 et 1971. On a donné aux demi-axes longueur 0,2.

TABLEAU 1

Concours d'admission à l'École Polytechnique : Années 1970 & 1971
Calculs de moyennes et de variances ; et résultats d'analyse factorielle –

	TITRE de l'épreuve	sigles X^+	sigles X^-	coeff x ($a=20x$)	moyenne de la note	écart-type σ(note)	écart-type $\sigma(x_* \text{note})$	corrélation corr(X,T)	corrélation corr$(X,T\text{-}X)$	premier facteur $G_1(X^+)$	premier facteur $G_1(X^-)$	premier facteur $\text{calb}_1(X^+)$	premier facteur $\text{calb}_1(X^-)$	deuxième facteur $G_2(X^+)$	deuxième facteur $G_2(X^-)$	deuxième facteur $\text{calb}_2(X^+)$	deuxième facteur $\text{calb}_2(X^-)$
premier écrit	Écrit Math 1	EM1	KM1	7	12,45	2,51	17,56	,055	,015	,025	–,042	,00002	,00003	,040	–,067	,00005	,00009
premier écrit	Écrit Math 2	EM2	KM2	7	13,06	2,64	18,49	,115	,051	,049	–,092	,00008	,00015	,067	–,126	,00015	,00028
premier écrit	Écrit Physique	EPH	KPH	7	12,43	3,27	22,91	,136	,049	,090	–,149	,00026	,00043	,101	–,167	,00033	,00054
premier écrit	Écrit chimie	ECH	KEC	3	11,+0	3,48	10,44	,104	,068	,102	–,135	,00013	,00017	,041	–,055	,00002	,00003
premier écrit	Français dissertat.	FR1	KF1	6	10,88	2,96	17,74	,067	,023	,098	–,118	,00023	,00028	–,054	,065	,00007	,00009
deuxième écrit	Français résumé	FR2	KF2	6	10,40	3,49	20,95	,180	,097	,200	–,216	,00091	,00099	–,082	,088	,00015	,00016
deuxième écrit	Calcul numér.	CUN	KUN	3	11,08	3,79	11,38	,100	,070	,112	–,139	,00015	,00019	,077	–,095	,00007	,00009
deuxième écrit	Examen industr.	DEG	KDG	6	10,30	4,42	26,50	,123	,038	,156	–,166	,00055	,00059	,295	–,313	,00198	,00210
deuxième écrit	Langue écrit	ELV	KEL	5	10,17	4,11	20,56	,207	,124	,292	–,302	,00160	,00165	–,183	,190	,00063	,00065
GRAND "O" – oral	Oral Math 1	OM1	KO1	25	14,32	1,31	32,84	,346	,165	,033	–,085	,00015	,00038	,025	–,063	,00008	,00021
GRAND "O" – oral	Oral Math 2	OM2	KO2	25	14,14	1,54	38,42	,286	,101	,038	–,093	,00019	,00046	,027	–,065	,00010	,00023
GRAND "O" – oral	Oral Physique	OPH	KOP	14	13,58	1,99	27,80	,172	,058	,045	–,096	,00014	,00030	,038	–,080	,00010	,00021
GRAND "O" – oral	Oral Chimie	OCH	KOC	7	12,66	2,22	15,55	,128	,066	,057	–,099	,00011	,00019	,004	–,007	,00000	,00000
GRAND "O" – oral	Oral Langue	OLV	KOL	10	10,81	2,76	27,65	,215	,091	,179	–,210	,00127	,00149	–,120	,141	,00057	,00067
GRAND "O" – sports	Course 100 m.	100	K10	0,75	9,49	4,19	3,14	,014	,011	,029	–,026	,00000	,00000	,024	–,021	,00000	,00000
GRAND "O" – sports	Course 800 m.	800	K80	0,75	7,76	4,62	3,46	,008	,005	,015	–,009	,00000	,00000	,045	–,028	,00000	,00000
GRAND "O" – sports	Saut hauteur	HAU	KHA	0,75	9,68	4,82	3,61	,010	,007	,028	–,026	,00000	,00000	,057	–,053	,00001	,00001
GRAND "O" – sports	Saut longueur	LON	KLO	0,75	7,98	4,87	3,65	,009	,006	,033	–,022	,00000	,00000	,048	–,031	,00001	,00000
GRAND "O" – sports	Lancer	LAN	KLA	0,75	9,97	3,19	2,39	,006	,004	,004	–,004	,00000	,00000	,029	–,029	,00000	,00000
GRAND "O" – sports	Natation	NAT	KNA	1,25	12,76	4,37	5,46	,010	,006	,024	–,041	,00000	,00001	–,033	,059	,00001	,00001

indique le niveau du total général T (quelle que soit d'autre part la fraction afférente à X dans T ou dans les variations de T). Pour en juger F. Nakhlé a calculé pour chaque note X corr(X,T), coefficient de corrélation avec le total général, et aussi corr(X,T - X), coefficient de corrélation avec le total des notes autres que X, affectées chacune de leur coefficient k : corr(X,T - X) est toujours inférieur à corr(X,T), et enseigne mieux que ce dernier nombre si l'épreuve X est orientée dans l'axe du concours. Première remarque : dans la colonne des corr (X,T-X) le plus grand nombre corr(OM1,T - OM1) est faible : 0,165; le plus petit corr(EM1,T - EM1) est très faible : 0,015 (pour ne rien dire les épreuves sportives : 0,004 < corr(X,T-X) < 0,011 : le concours de l'X est une épreuve d'endurance dont un candidat réussit inégalement les étapes successives. Il est plus frappant que dans les cinq plus forts coefficients on trouve avec les deux épreuves orales de math, les deux épreuves de langue ELV et OLV et le résumé de français FR2 dont l'analyse factorielle nous a déjà signalé l'importance sur le 1° axe. Non seulement les disciplines littéraires de par leur variance jouent amplement dans le total général, mais encore elles jouent dans le sens du concours, elles indiquent mieux que toutes les épreuves, oral de math excepté, le niveau d'un candidat. Admettons que FR2 devance la chimie : la chimie est une science particulière dont une connaissance générale peut suffire à l'ingénieur qui n'en est pas spécialiste, tandis que tous doivent lire et rédiger; mais la physique vient après la chimie; et la première épreuve écrite de mathématique EM1 est au dernier rang des épreuves intellectuelles, devançant de peu la course de 100 m ! Il est fâcheux que l'aptitude à poursuivre sur le papier durant quatre heures une question de mathématiques soit peu représentative des mérites qui distinguent un polytechnicien; que la solidité semble moins primée que le brillant. Car le brillant est tout autre que la véritable originalité qui requiert la persévérance dans l'effort fastidieux; tant il est vrai que le génie est une longue patience... Il faudrait ici confronter aux notes des candidats du concours, particulièrement à EM1 et OM1, les réalisations ultérieures de l'ingénieur.

3.4. *Essais de changement d'échelle* :

Dans ce § on cherche dans quelle mesure, des transformations linéaires des notes telles que celles étudiées au § 2, peuvent, tout en conservant les facteurs $F_\alpha(i)$ sur l'ensemble I (individus ou séries de notes), modifier les facteurs $G_\alpha(q^+)$ et $G_\alpha(q^-)$ sur l'ensemble J (épreuves q^+ et leurs complémentaires q^-).

3.4.1. *Rappel des résultats du § 2.*

Il sera commode de considérer simultanément les notes du tableau k analysé, notes affectées du coefficient κ_q de l'épreuve q; et les notes sur 20 notées v :

$$k(i,q^+) = \kappa_q\ v(i,q^+) ; \quad k(i,q^-) = \kappa_q\ v(i,q^-) ;$$

$$v(i,q^+) + v(i,q^-) = 20.$$

Sur l'histogramme des notes sur 20 obtenues par l'ensemble des candidats à l'épreuve q on définit les quatres paramètres x_q, y_q, X_q, Y_q, qui ont l'avantage d'être des fractions de l'unité.

$$\text{Moy } v(x,q^+) = 20\ x_q \ ; \quad y_q = 1 - x_q ;$$

$$\min v(x,q^+) = m_q = 20\ X_q ; \ \text{Max } v(x,\ q^+) = M_q = 20\ (1 - Y_q) .$$

Une transformation linéaire du tableau dédoublé des notes (nous dirons en bref un changement d'échelle; mais de façon précise, il y a pour l'échelle changement d'origine, d'extrémité et d'intervalle) est définie en fixant pour chaque épreuve trois paramètres ρ_q, ξ_q, η_q : en désignant par le sigle qn les nouvelles notes à l'épreuve q on pose :

$$k(i,qn^+) = \rho_q(k(i,q^+) + 20\ \kappa_q\ \xi_q);$$

$$k(i,qn^-) = \rho_q(k(i,q^-) + 20\ \kappa_q\ \eta_q).$$

Il revient au même de dire qu'on a de nouvelles notes sur 20 et un nouveau coefficient :

$$v(i,qn^+) = (v(i,q^+) + 20\ \xi_q)/(1 + \xi_q + \eta_q);$$

$$v(i,qn^-) = (v(i,q^-) + 20\ \eta_q)/(1 + \xi_q + \eta_q);$$

$$\kappa_{qn} = \kappa_q\ {}_q\ (1 + \xi_q + \eta_q).$$

Pour que le nouveau tableau ne comporte pas de notes négatives, il faut qu'on ait pour toutes les épreuves : $-X_q \leqslant \xi_q$; $-Y_q \leqslant \eta_q$. Il n'y a changement de profil pour une épreuve q que si l'un au moins des deux nombres ξ_q et η_q est non nul.

Aux changements d'échelles considérés, nous imposerons la condition de normalisation N que la somme des coefficients reste constante (et donc égale à 136, cf tableau 1) :

$$\Sigma\ \kappa_q\ \rho_q\ (1 + \xi_q + \eta_q) = \Sigma\ \kappa_q = 136; \qquad (N).$$

Cette condition étant posée, l'analyse factorielle du nouveau tableau fournit sur l'ensemble I (des individus) les mêmes facteurs que l'analyse du tableau initial si est vérifiée par les coefficients de changement d'échelle de chaque épreuve la condition C :

$$(1 + \xi_q + \eta_q)\ \rho_q\ x_q\ y_q = (x_q + \xi_q)(y_q + \eta_q). \qquad (C).$$

Sur l'ensemble des notes, les facteurs sont alors modifiés suivant les formules :

$$G_\alpha(qn^+) = G_\alpha(q^+)(x^q/(x^q + \xi^q));$$

$$G_\alpha(qn^-) = G_\alpha(q^-)(y^q/(y^q + \eta^q)).$$

Sous les conditions N et C, il est impossible de modifier les coefficients κ_q si on ne modifie le profil que d'une seule épreuve q; dans ce cas on a pour cette épreuve :

$$(1 + \xi_q + \eta_q)\ \rho_q = 1;\ x_q\ y_q = (x_q + \xi_q)(y_q + \eta_q);$$

ξ_q et η_q sont de signe contraire; et par exemple à partir de ξ_q on peut calculer η_q et ρ_q. Si on modifie les profils de plusieurs épreuves (deux au moins) il est possible de modifier le système des coefficients.

Dans ces changements d'échelle, la principale contrainte est celle de positivité des nouvelles notes; contrainte qu'on peut relacher un peu : 1 % de notes négatives est tolérable, car pourvu que les totaux des colonnes restent positifs le programme d'analyse factorielle est indifférent à la présence de nombres négatifs. Sous les conditions N et C, le coefficient d'une épreuve augmente si ξ_q et η_q sont positifs; il diminue d'autant plus que ξ_q et η_q sont plus négatifs. Si l'on change le système des coefficients, il faut que certains coefficients diminuent (cf N); si l'on change seulement un seul profil, il faut que ξ_q ou η_q

soit négatif. D'où l'importance des ξ_q et η_q négatifs : or la positivité requiert $- X_q \leq \xi_q$, $- Y_q \leq \eta_q$; en définitive plus les X_q et Y_q sont grands (i.e. moins les histogrammes des notes v s'étendent sur l'intervalle *a priori* possible (0,20)) plus on a de liberté pour modifier le tableau des notes sans altérer les facteurs sur I. Dans ce § nous nous bornerons à considérer deux épreuves que l'analyse factorielle nous a signalées : une épreuve orale de mathématiques, OM1, pour laquelle les notes vont de 10 à 19; et l'épreuve écrite de langue pour laquelle au contraire presque tout l'intervalle de notation est utilisé, les notes variant de 0 à 19,5.

3.4.2. *L'épreuve de mathématiques OM1* : On a :

$x = 14{,}37/20 = 0{,}718;$ $\qquad y = 1 - x = 0{,}282$

$X = m/20 = 10/20 = 0{,}5;$ $\qquad Y = (20 - M)/20 = 1/20 = 0{,}05$

Si l'on s'interdit d'introduire des notes négatives, la réduction maximale du coefficient de l'épreuve est obtenus pour $\xi = - X$, $\eta = - Y$: on a alors (cf C) :

$$(1 + \xi + \eta)\rho = (x + \xi)(y + \eta)/(xy)$$

$$= (x - X)(y - Y)/(xy) = 0{,}250;$$

$$\kappa_n = \kappa\ (1 + \xi + \eta)\rho = 25 \times 0{,}25 = 6{,}25$$

et pour les facteurs on a, en se référant à l'analyse globale 1970-71-72, (1386 séries de notes) d'après laquelle F. Nakhlé a étudié les histogrammes :

$$G_\alpha(\mathrm{OM1n}) = G_\alpha(\mathrm{OM1}) \times (x/(x + \xi)) = G_\alpha(\mathrm{OM1}) \times 3{,}3$$

$$G_1(\mathrm{OM1n}) = 0{,}029 \times 3{,}3 = 0{,}096$$

$$G_2(\mathrm{OM2n}) = 0{,}0325 \times 3{,}3 = 0{,}107$$

$$G_\alpha(\mathrm{KO1n}) = G_\alpha(\mathrm{KM1}) \times (y/(y + \eta)) = G_\alpha(\mathrm{KM1}) \times 1{,}22$$

$$G_1(\mathrm{KO1n}) = - 0{,}074 \times 1{,}22 = - 0{,}090$$

$$G_2(\mathrm{KO1n}) = - 0{,}083 \times 1{,}22 = - 0{,}100$$

On voit que du point de vue de l'analyse des correspondances on peut réduire au quart de sa valeur le coefficient de OM1, à condition de transformer les notes afin qu'elles s'étalent de 0 à 20 : les facteurs sont alors multipliés, mais n'atteignent encore que le tiers en valeur absolue de ceux des épreuves de langues, parce que non seulement les notes brutes de OM1 ne s'étendent que sur 9 points (de 10 à 19) mais encore relativement à cette étendue leur écart-type 1,33 est faible. En tolérant que soit enfreinte la contrainte de positivité pour les 14 notes de OM1 valant {10; 10,5; 18,5; 19} on peut poser $\xi = 11/20 = 0{,}55$; $\eta = 2/20 = 0{,}10$; et le coefficient de OM1 peut être réduit à 3,75 ! Du point de vue total général (qui n'est pas celui de l'invariance des facteurs) le coefficient peut être réduit dans le rapport de l'étendue à 20; i.e. à $(9/20) \times 25 = 11{,}25$ (ou $(7/20) \times 25 = 8{,}75$ si on néglige les 14 notes extrêmes).

3.4.3. *L'épreuve écrite de langue ELV* : On a :

$x = 10{,}4/20 = 0{,}52$; $\qquad y = 1 - x = 0{,}48;$

$X = 0$; $\qquad Y = 0{,}5/20 = 0{,}025.$

Si on pose $\xi = -X = 0$; $\eta = -Y = -0{,}025$ le nouveau coefficient diffère fort peu de l'ancien : il passe de 7 à 0,95 x 7 = 6,65. En revanche, on peut aplatir le profil de ELV en prenant ξ et η positifs; ce qui assurément ne nuira pas à la positivité, mais augmentera le coefficient de ELV; augmentation permise (par la condition N) toutefois puisque le coefficient de OM1 peut,lui,être réduit de 25 à 6,25. Par exemple, en posant $\xi = \eta = 0{,}40$, il vient pour nouveau coefficient 22,7 (au lieu de 7) et on a alors G_1(ELVn) = 0,17, valeur qui demeure considérable.

Que conclure de ces expériences de changement d'échelle ? L'interprétation des facteurs, donnée en § 3.2 n'est pas en jeu; mais on voit dans quelle mesure, comme nous le pressentions, la distance à l'origine des points q^+ et q^- dépend non seulement des mérites des candidats et des jugements des examinateurs, mais aussi de la liberté avec laquelle ceux-ci usent des points qu'il leur est permis d'attribuer : une très mauvaise copie de langue est notée 0 à 2,5 σ en dessous de la moyenne des notes de l'épreuve; une planche d'oral de math détestable reçoit un 10/20, mais elle est à 3,3 σ en dessous de la moyenne des notes de l'épreuve. Les grandes lignes de différenciation des candidats sont claires. Mais on sait que les notes individuelles pourraient être autres qu'elles ne sont sans cesser d'exprimer fidèlement la réalité; et chaque année le destin de dizaines de jeunes gens en serait changé ...