PLANS DE SONDAGE A DISPERSION MINIMALE

Jean-Claude DEVILLE, Mohammed El HAJ TIRARI

Ensai/Crest, Laboratoire de Statistique d'Enquête

Entre ces deux échantillons d'unités primaires tirés indépendamment, il a été demandé aux Chefs Locaux de choisir celui qu'ils pensent le plus adapté, à la fois en termes de " représentativité ", mais aussi en tenant compte de leur commodité.

Cette procédure de choix, qui respecte le caractère aléatoire du tirage (vu le très grand nombre de combinaisons possibles) permet de parvenir à un tirage " adapté " au mieux. Il s'agit d'une optimisation (*on doit sans doute comprendre " amélioration "*) par rapport à ce que l'on faisait auparavant, où un seul échantillon était tiré.

(traduction en français du texte original)

Soit X une variable aléatoire (=dépendante de s, l'échantillon) qui en mesure la commodité – distances à parcourir, zones pourries, facilité d'accès, bref qqchose de plutôt urbain et pas trop mal famé.

Soit F sa fonction de répartition. On a donc deux échantillons s_1 et s_2 indépendants . On tire $s=argmax(X(s_1)=X_1 , X(s_2)=X_2)$ ce qui est un échantillon tout à fait probabiliste.

On va s'intéresser 5 minutes à $X^+=max (X_1 ,X_2)$ (et $X^-=min(X_1 ,X_2)$).

La fonction de répartition est F^2 de sorte que la médiane de X n'est autre que le premier quartile de X^+. La médiane de X^+ est le quantile à 0.707 de X.

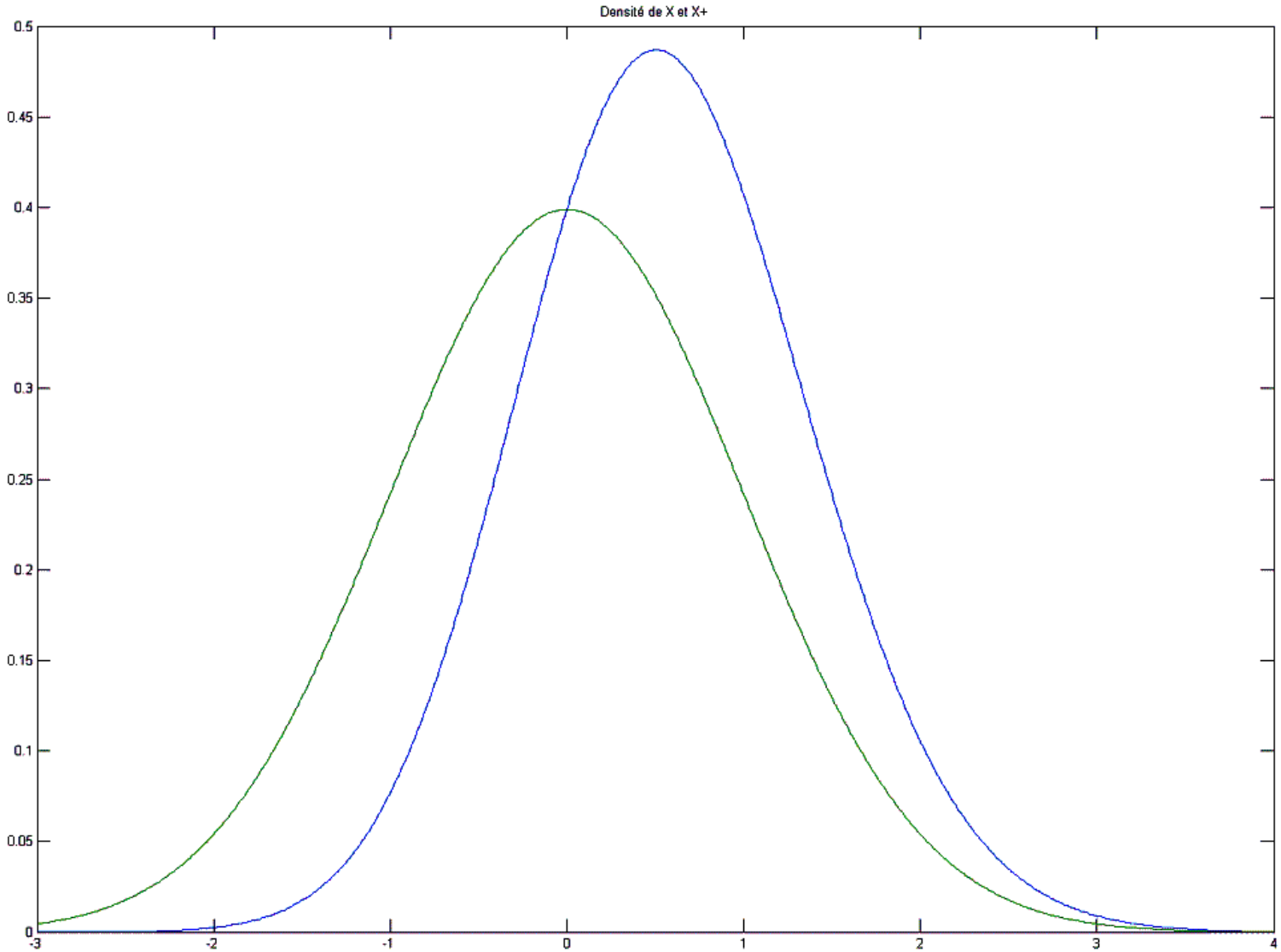
On peut s'attendre à un biais…

Bin oui. Il vaut

$$B = \int_{-\infty}^{+\infty} F(x)(1 - F(x))\, dx$$

Sous l'hypothèse où X suit une loi de gauss, on calcule qu'il vaut 0.5642 ($\pi^{-1/2}$) fois l'écart type. On calcule aussi le mode (la valeur la plus probable de l'échantillon!).

Voila le dessin.



Cas gaussien : mode 0.5061 , médiane 0.5446 , moyenne (biais) 0.5642

Et la variance vous allez me dire!

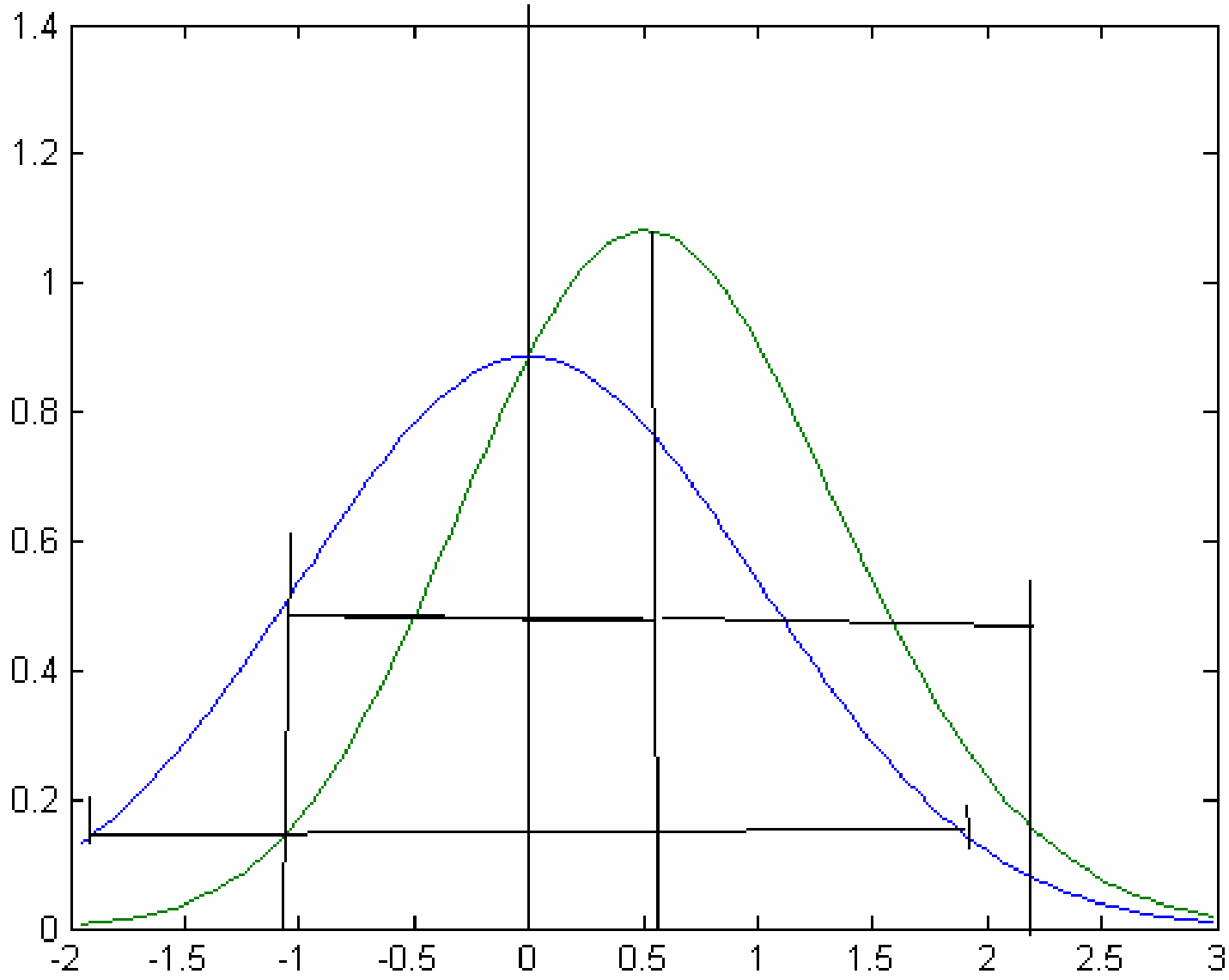
Et bien en utilisant le fait que $f(X_1, X_2) = f(X^+, X^-)$ pour toute fonction symétrique (addition, produit, différence au carré…) on arrive à y voir assez clair: si X à une loi symétrique alors X^+ et X^- ont la même loi, et donc leur écart quadratique moyen vaut la variance de X *!!!*

Bingo!!

L'intervalle de confiance dans la cas gaussien, passe de (-1.96,+1,96)

à (-1.022, 2.23). Il est nettement plus petit que pour le tirage naïf du bon vieux temps. Un peu bancal, mais qui y verra quelque chose?

Pour une variable d'intérêt estimée par Y^, le biais dépend de la corrélation avec X. Il vaut $B\ Cov(Y\hat{},X)$. La encore on arrive à tout calculer, mais je ne suis pas venu pour parler de ça.


1.4
1.2
1
0.8
0.6
0.4
0.2
0
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3

Plan de la présentation:

- **Plans à dispersion minimale: c'est quoi?**
- **Propriétés des plans à dispersion minimale; échantillonnage de type Metropolis**
- **Plan à variance minimale et méthode de Midzuno**
- **Plan à variance minimale : résolution**
- **Echantillonnage pour le plan à variance minimale**
- **Critère plus général : plan de sondage**
- **Critère plus général : échantillonnage**
- **Illustrations et commentaires**

1-Plans à dispersion minimale: c'est quoi?

- *U* population finie de taille *N*.
- Individus notés par un index *k* variant de 1 à *N*.
- Plan de sondage :
  - ➔ loi de probabilité sur l'ensemble ***S*** des parties de *U*
  - ➔ nombres $p(s)$ associés à chaque *s* de ***S***- ou échantillon- vérifiant $p(s) \geq 0$ et $\sum_{s \in \mathbf{S}} p(s) = 1$.
- Le support $supp(p)$ est l'ensemble des *s* tels que $p(s) > 0$.

1- Contraintes

1- **Taille fixe *n* =** $\mathrm{supp}(s) \subset \mathbf{S_n}$

2- $\pi_k = \sum_{s:k \in \mathbf{s}} p(s)$ **probabilités d'inclusion données pour chaque unité *k* de *U***

1-Quelle est l’idée?

En l’absence d’autres informations se traduisant par des contraintes, on n’accorde aucun privilège à quelque échantillon que ce soit.

En particulier, si cela est possible, c'est-à-dire quand toutes les probabilités d’inclusion sont égales (et donc égales à n/N), on rendra tous les échantillons équiprobables, et donc on choisira le sondage aléatoire simple.

Sinon on minimise la dispersion des nombres $p(s)$ compte tenu des contraintes.

1-Critère de dispersion?

On se base sur une fonction strictement convexe φ définie sur [0,1] qu'on suppose continument dérivable, la dérivée en 0 (resp 1) pouvant valoir $-\infty$ (resp $+\infty$) . Le critère :

$$\Phi(p) = \sum_{s \in \mathbf{S}_n} \varphi(p(s))$$

est une fonctionnelle sur l'ensemble des plans de taille n. Sa valeur minimum est obtenue pour l'échantillonnage aléatoire simple (SAS), c'est-à-dire pour la répartition uniforme p_u sur $\boldsymbol{S_n}$.

1-Critère de dispersion?(suite)

$$\Phi(p) = \sum_{s \in \mathbf{S_n}} \varphi(p(s))$$

- C'est une mesure de dispersion au sens où, si p n'est pas uniforme, les lois $p_t = p_u + t(p - p_u)$ sont, intuitivement, de plus en plus dispersées quand t croît et où $\Phi(p_t)$ est une fonction strictement convexe qui prend son minimum pour $t=0$.
- Ce critère reste le même (à une constante additive près) si on ajoute à une fonction affine arbitraire .
- 'Distance' au SAS.

1-Formalisme et conséquences..

$\boldsymbol{U}$ la Nx$\boldsymbol{S_n}$ matrice des $\boldsymbol{1_{k(s)}}$ (1 si k dans s , 0 sinon), et π le N-vecteur des probabilités d'inclusion. On écrira donc les contraintes sous la forme compacte :

$$\boldsymbol{U}p = \pi \ .$$

$\boldsymbol{1_N}$ et $\boldsymbol{1_S}$ = les vecteurs composés uniquement de 1 de tailles respectives N et $card(\boldsymbol{S_n})$.

On a : $\boldsymbol{1_N}'\boldsymbol{U} = n\,\boldsymbol{1_S}'$ et $\boldsymbol{U}\,\boldsymbol{1_S} = \binom{N}{n}\boldsymbol{1_N}$

En particulier:

$$\sum_U \pi_k = \mathbf{1}'_N \pi = \mathbf{1}'_N \mathbf{U}\ p = n\,\mathbf{1}'_S\ p = n$$

C'est le problème (P)……..

2- Propriétés du problème (P)

Propriété 1 : Le problème (P) possède toujours une solution unique.

La propriété suivante caractérise les solutions de (P) selon que tous les échantillons de $\boldsymbol{S}_n$ reçoivent une probabilité strictement positive (toutes les solutions sont ‘intérieures’) ou qu’il existe des vecteurs π menant à des solutions ‘au bord’, c’est à dire comportant des échantillons à probabilité nulle.

Propriété 2 : Une condition nécessaire et suffisante pour qu’il n’y ait jamais de solution ‘au bord’ est que : $\varphi'(0) = -\infty$

Corollaire : Si $n>1$ toutes les probabilités d’ordre deux $\pi_{kl} = \sum_{s:k\,et\,l\in s} p(s)$

sont strictement positives et il existe un estimateur sans biais de la variance d’échantillonnage.

2- Propriétés du problème (P) (suite)

Propriété 3 : Si $\varphi'(0) = -\infty$ il existe des systèmes de probabilités d'inclusion strictement positives tels que $p(s)=0$ pour certains échantillons de $\boldsymbol{S}_n$.

Définition: Comme φ est strictement convexe et continûment dérivable, φ' est croissante strictement et continue. Elle admet une fonction réciproque ψ qui croît continûment de 0 à 1 quand son argument croît de $\varphi'(0)$ à $\varphi'(1)$ (conjuguée de Young).

Propriété 4 : Si $\varphi'(0)=-\infty$, ψ est définie de $-\infty$ à 1 , toutes les contraintes sont inactives et

$$p(s) = \psi(\textstyle\sum_s \lambda_k)$$

Propriété 5 : Si $\varphi'(0)> -\infty$, la solution de (P) est caractérisée par un vecteur de N réels définis de façon unique en fonction des probabilités d'inclusion tels que :

$$p(s) = \begin{cases} \psi(\sum_{k \in s} \lambda_k) > 0 & si \sum_{k \in s} \lambda_k > 0 \\ 0 & si \sum_{k \in s} \lambda_k < 0 \end{cases}$$

Metropolis…

Propriété 6 : Les λ_k sont ordonnés comme les π_k .

2-Exemples de fonctions de distance :

- Entropie négative

$$\varphi_{ent}(p) = p\,log(p) \quad \text{(avec la convention } 0\,log(0)=0\text{)}$$

- Variance

$var(p)=(p-p_u)^2$. Par normalisation à une fonction affine près $var(p)=p^2$

- Type ‘puissance’

$\varphi_a(p)=p^a$ si $a>1$ ou $\varphi_a(p) = -p^a$ si $0<a<1$.

- Distance de Hellinger :

$$d_H(p,q) = \sum_s (p(s)^{1/2} - q(s)^{1/2})^2 \qquad \Rightarrow \varphi_{-1/2}$$

- Entropies (informations) de Renyi:

$$H_\alpha(p) = \frac{1}{1-\alpha} \log(\sum_s p(s)^\alpha)$$

2-Exemples numériques:

Exemple 1 :

Si entropie ou Hellinger tous les échantillons de taille n ont une probabilité strictement positive.

Exemple 2 :

Pour la variance on obtient facilement une solution ‘au bord’ :

$$N=4 \text{ et } n=2 \text{ , } \pi_1 = \pi_2 = 0.2 \text{ , } \pi_3 = \pi_4 = 0.8 \text{ .}$$

On obtient

$$p(1 , 2) = 0, p(1 , 3) = p(2 , 3) = p(1 , 4) = p(2 , 4) = 0.1 \text{ et } p(3 , 4) = 0.6 \text{ .}$$

Si on minimise l’entropie de ces trois valeurs sont respectivement 0.0141, 0.0930 et 0.61141.

3-Plan à variance minimale et méthode de Midzuno

Cas où toutes les probabilités sont strictement positives :

$$0 < p(s) = \lambda' s = \sum_{k \in s} \lambda_k$$

Et donc $\lambda = Cte(\ \pi - (n-1)/(N-1)\mathbf{1}N\ )$

Propriété 7 : Le plan à variance minimale charge tous les échantillons de taille *n* si et seulement si $\sum_{s\,\min} \pi_k > n(n-1)/(N-1)$ où *smin* est l'échantillon contenant les *n* plus petites unités. La probabilité de *s* est alors proportionnelle à $\sum_s (\pi_k - (n-1)/(N-1))$.

Schéma de Midzuno-Lahiri: x_k positive ou nulle et $p_k = x_k/X$ où *X* est le total des *x* .

On tire une première unité dans loi des p_k et on complète l'échantillon par sondage simple de *n-1* unités parmi les *N-1* restantes. On voit que $p(s) = (\sum_s p_k) / \binom{N-1}{n-1}$

et que $$\pi_k = \frac{n-1}{N-1} + p_k \left(1 - \frac{n-1}{N-1}\right) \geq \frac{n-1}{N-1}$$

La propriété la plus amusante de ce plan est de rendre sans biais l'estimateur par ratio du total d'une variable *y* quelconque.

Même lorsqu'il charge tout S_n le plan à variance minimum est un peu plus général que le schéma de Midzuno car ce dernier demande que chaque π_k soit supérieur à $(n-1)/(N-1)$.

4-Plan à variance minimale: résolution

Le problème (P) s'écrit :

$$Min\ p'p \qquad \text{sous les contraintes } Up=\pi \text{ et } p \geq 0 \quad .$$

Problème banal de programmation quadratique? algorithmes peu efficace!
Méthode ad hoc basée sur la:

Propriété 8 : Soit p_0 la solution de $Min\ p'p$ sous les contraintes $U_0\, p=\pi$ où U_0 est la restriction de U à un ensemble de colonnes $\boldsymbol{S}_0 \subset \boldsymbol{S}_n$. Soit $\boldsymbol{S}_{0+}$ l'ensemble des s vérifiant $p(s)>0$. Le plan p_* optimal obtenu en ajoutant les contraintes $p(s) \geq 0$ vérifie $p_*(s)>0$ sur un ensemble $\boldsymbol{S}_{*+}$ contenu $\boldsymbol{S}_{0+}$.

D'où un algorithme qui semble (?) avoir déjà été utilisé sans justification par:

[9] Joe, H., A Winning Strategy for Lotto Games , *The Canadian Journal of Statistics*, vol 18 , pp 233-244 , 1990

5-Plan à variance minimale: échantillonnage

Propriété 9 : Le plan de variance minimale est caractérisé par un N-vecteur λ_k tel que :

$p(s)= 0$ si $\Sigma_s \lambda_k < ou = 0$

$p(s)= \Sigma_s \lambda_k$ si $\Sigma_s \lambda_k > 0$.

Soit $w_k = sup(\lambda_k , 0)$.

Le schéma de tirage est le suivant (extension du schéma de Midzuno):

Etape 1 : on tire un échantillon s selon le schéma de Midzuno en utilisant la variable w_k .

Etape 2 : soit $r = sup\ (0 , \ \sum_s \lambda_k / \sum_s w_k \ )$. On accepte s avec la probabilité r ou on retourne à l'étape 1 avec la probabilité $1-r$.

Remarque: Si $\Sigma_s \lambda_k < ou = 0$ ou si tous les λ_k sont positifs, on ne génère pas de r .

6- Critère plus général : plan de sondage

On se débrouille numériquement pour résoudre (P) à partir de la méthode de Newton parfois modifiée. Les difficultés dépendent des fonctions $\psi = \varphi'^{-1}$ (domaine de définition, convexité,….) et de la dispersion des probabilités d'inclusion. On a développé une technique assez générale, qui marche très bien pour l'entropie et la variance, de façon plus chaotique pour d'autre critères (voir les exemples *in fine*).

Remarque sur le calcul des probabilités d'inclusion d'ordre 2 : Puisque le vecteur *p* peut être calculé, la matrice des probabilités s'inclusion d'ordre 2 est donnée par $U\,diag(p)\,U'$ et ce calcul ne pose aucune difficulté nouvelle puisque l'ensemble des échantillons a été énuméré. Ceci étant, le critère d'entropie a l'avantage de permettre un calcul récursif de ces probabilités d'inclusion dont les limitations en volume de stockage sont beaucoup moindres, ce qui permet d'utiliser de plus grandes valeurs de *n* et *N* .

7- Critère plus général : échantillonnage

Plutôt que Metropolis, on peut en général utiliser une extension de l'extension du schéma de Midzuno. La encore, ça marche bien pour de bonnes fonctions ψ, 'convexes mais pas trop'.

8-Illustrations et commentaires

On a expérimenté assez systématiquement cinq critères de dispersion, qui sont, par « ordre de convexité croissante » :

- la distance de Hellinger ($\psi(u)=u^{-2}$, $\varphi(p)=-sqrt(p)$),

- l'entropie ($\psi=exp$, $\varphi(p)=p\ log(p)$) ,

- le « carré » ($\psi(u)=u^2$, $\varphi(p)= p^{3/2}$) ,

- la variance ($\psi(u)=max\ (0,u)$, $\varphi(p)=p^2$)

- la « racine » ($\psi(u)=sqrt(u)$ si $u>0$, 0 sinon et $\varphi(p)=p^3$).

Tailles de populations allant jusqu'à 20,25 et tailles d'échantillon allant jusqu' à 10, 12. (problèmes liés à la technique d'énumération des échantillons).

8-Illustrations et commentaires(suite1)

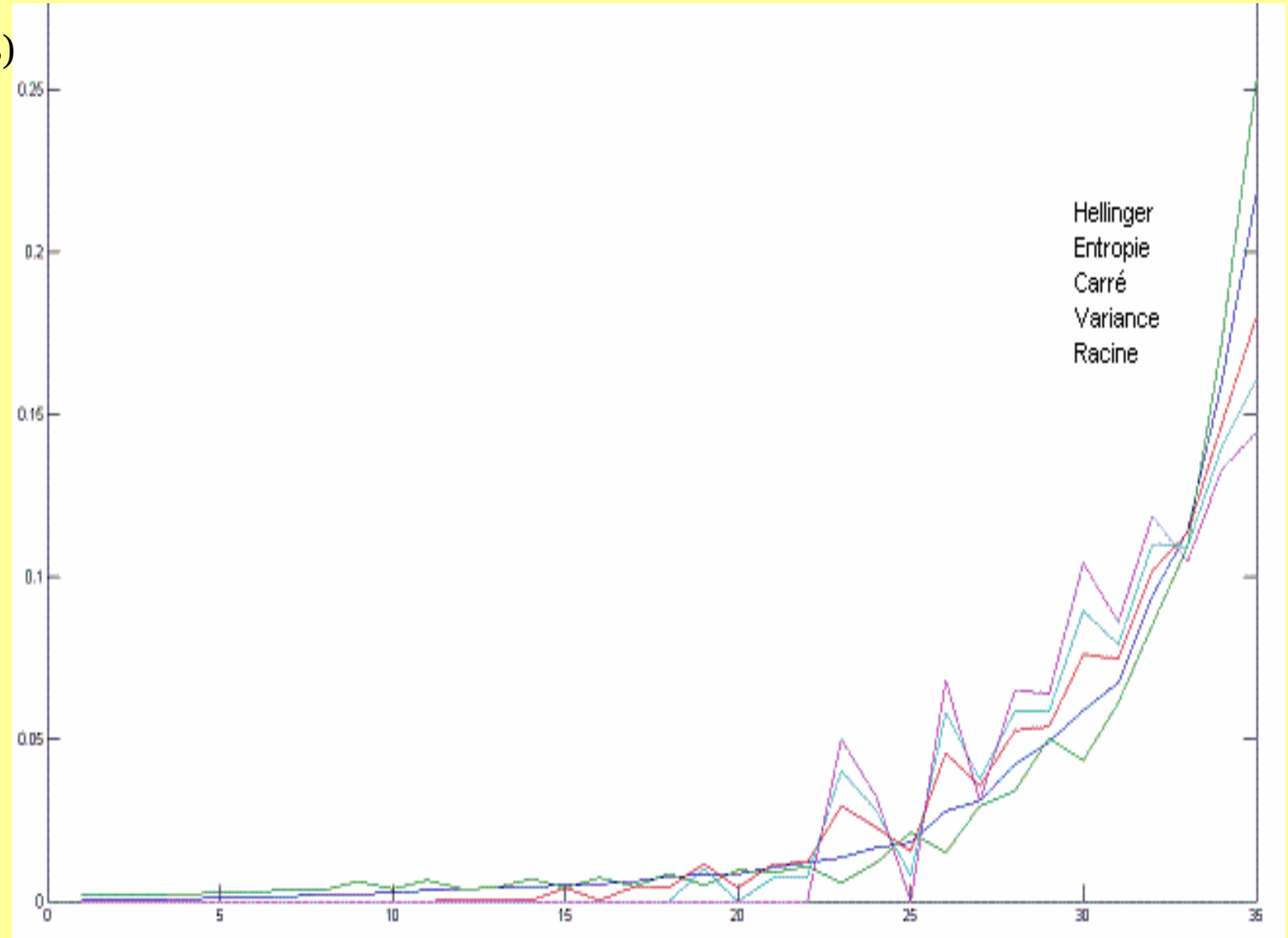
La distance de Hellinger et le critère d'entropie octroient des probabilités strictement positives à tous les échantillons de $\boldsymbol{S}_n$. La distance de Hellinger privilégie les échantillons extrêmes, petits ou gros.

Les trois autres critères font apparaître, sur cet exemple choisi ad hoc, des probabilités nulles pour les échantillons les plus petits (le critère de classement est la probabilité obtenue pour le critère d'entropie). Cet ensemble a tendance à augmenter avec la convexité du critère.

Illustration: Pour rester lisible mais néanmoins présenter les principales caractéristiques des plans à dispersion minimale, nous avons choisi le cas $N=7$et $n=3$. Le vecteur des probabilités d'inclusion est π=[.05 .1 .35 .45 .55 .7 .8] , soit quelque chose d'assez dispersé avec deux petits éléments. Au total le support de p comporte au plus 35 éléments, ce qui permet une lecture 'à la main' si on peut oser cette champignaquerie.

NB : les échantillons sont ordonnés par probabilité croissante pour le critère d'entropie. Les trois dernières colonnes sont les numéros d'ordre des éléments de l'échantillon.

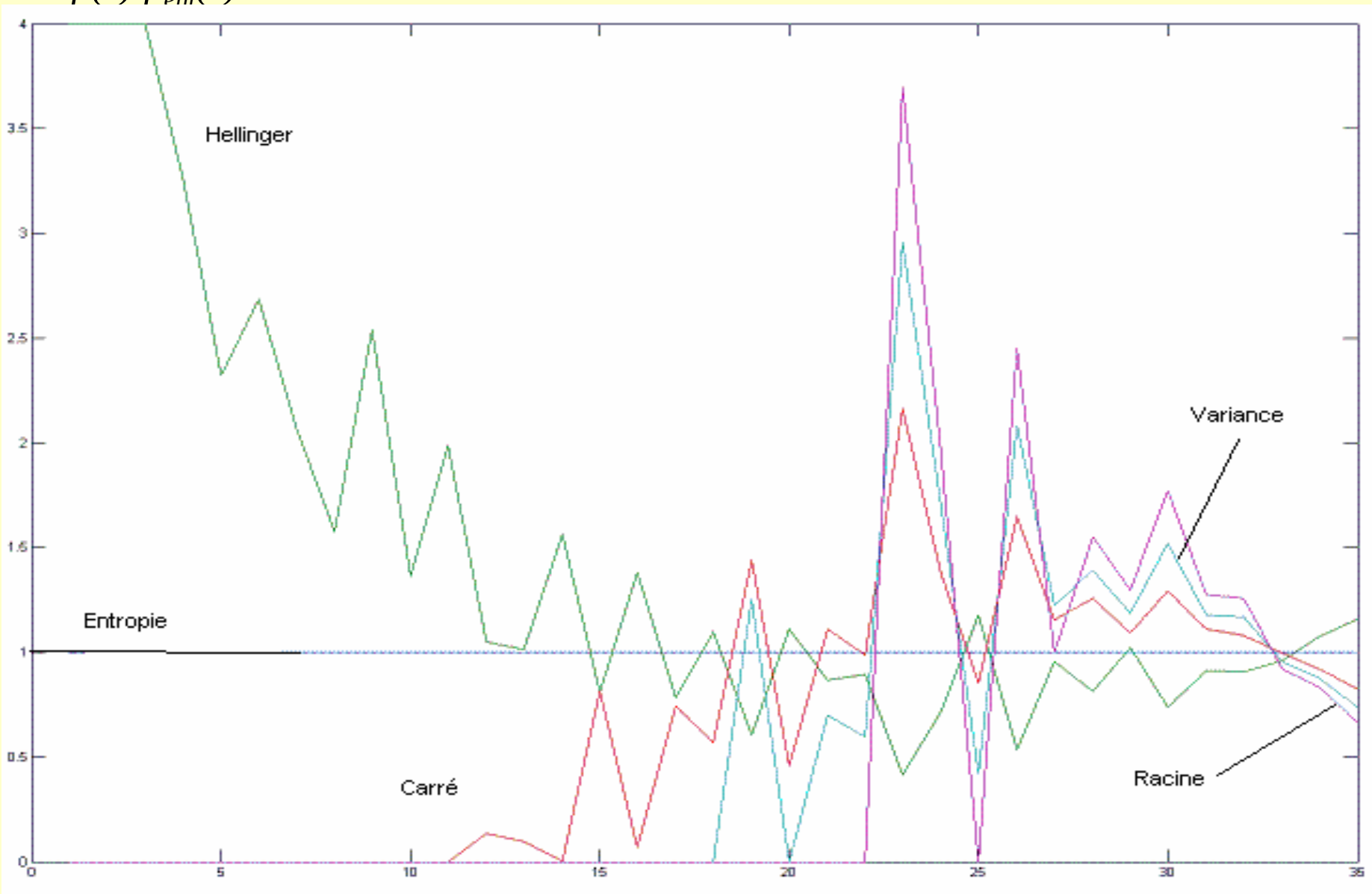
		p(s)				s	
Hellinger	Entropie	Carré	Linéaire	Racine			
0,00173	0,00019	0	0	0	1	2	3
0,00182	0,00027	0	0	0	1	2	4
0,00189	0,00037	0	0	0	1	2	5
0,00207	0,00063	0	0	0	1	2	6
0,00235	0,00101	0	0	0	1	2	7
0,00301	0,00112	0	0	0	1	3	4
0,00315	0,00153	0	0	0	1	3	5
0,00338	0,00215	0	0	0	1	4	5
0,00584	0,00230	0	0	0	2	3	4
0,00355	0,00260	0	0	0	1	3	6
0,00624	0,00314	0	0	0	2	3	5
0,00382	0,00364	0,00050	0	0	1	4	6
0,00421	0,00416	0,00039	0	0	1	3	7
0,00689	0,00441	0	0	0	2	4	5
0,00403	0,00498	0,00405	0	0	1	5	6
0,00737	0,00533	0,00036	0	0	2	3	6
0,00456	0,00583	0,00433	0	0	1	4	7
0,00821	0,00748	0,00425	0	0	2	4	6
0,00484	0,00797	0,01148	0	0	1	5	7
0,00947	0,00854	0,00392	0	0	2	3	7
0,00888	0,01022	0,01135	0,00714	0	2	5	6
0,01072	0,01197	0,01182	0,00714	0	2	4	7
0,00559	0,01353	0,02926	0,04000	0,05000	1	6	7
0,01173	0,01637	0,02252	0,02786	0,03199	2	5	7
0,02129	0,01810	0,01545	0,00762	0,00000	3	4	5
0,01479	0,02776	0,04578	0,05786	0,06801	2	6	7
0,02933	0,03070	0,03542	0,03762	0,03071	3	4	6
0,03409	0,04197	0,05268	0,05833	0,06498	3	5	6
0,05021	0,04915	0,05370	0,05833	0,06385	3	4	7
0,04341	0,05885	0,07596	0,08929	0,10431	4	5	6
0,06131	0,06720	0,07456	0,07905	0,08577	3	5	7
0,08535	0,09422	0,10185	0,11000	0,11839	4	5	7
0,10921	0,11398	0,11355	0,10905	0,10469	3	6	7
0,17214	0,15982	0,14673	0,14000	0,13274	4	6	7
0,25351	0,21851	0,18011	0,16071	0,14456	5	6	7



Numéro d'ordre de l'échantillon

Graphiquement: N=7, n=3

Ratio $p(s)/p_{ent}(s)$:



Numéro d'ordre de l'échantillon

Graphiquement: N=7, n=3

Probabilités d'inclusion d'ordre 2 :

$$\begin{bmatrix} 0.0099 & & & & & \\ 0.0156 & 0.0306 & & & & \\ 0.0166 & 0.0335 & 0.1097 & & & \\ 0.0173 & 0.0356 & 0.1261 & 0.1603 & & \\ 0.0191 & 0.0413 & 0.1835 & 0.2569 & 0.3439 & \\ 0.0216 & 0.0491 & 0.2344 & 0.3230 & 0.4167 & 0.5552 \end{bmatrix}$$

← Hellinger

← Entropie

Carré

$$\begin{bmatrix} 0 & & & & & \\ 0.0004 & 0.0043 & & & & \\ 0.0048 & 0.0161 & 0.1045 & & & \\ 0.0155 & 0.0339 & 0.1427 & 0.1933 & & \\ 0.0338 & 0.0617 & 0.2020 & 0.2629 & 0.3242 & \\ 0.0455 & 0.0840 & 0.2461 & 0.3184 & 0.3905 & 0.5154 \end{bmatrix}$$

$$\begin{bmatrix} 0.0025 & & & & & \\ 0.0096 & 0.0195 & & & & \\ 0.0130 & 0.0264 & 0.1014 & & & \\ 0.0170 & 0.0345 & 0.1319 & 0.1777 & & \\ 0.0254 & 0.0514 & 0.1946 & 0.2605 & 0.3345 & \\ 0.0325 & 0.0656 & 0.2430 & 0.3210 & 0.4043 & 0.5336 \end{bmatrix}$$

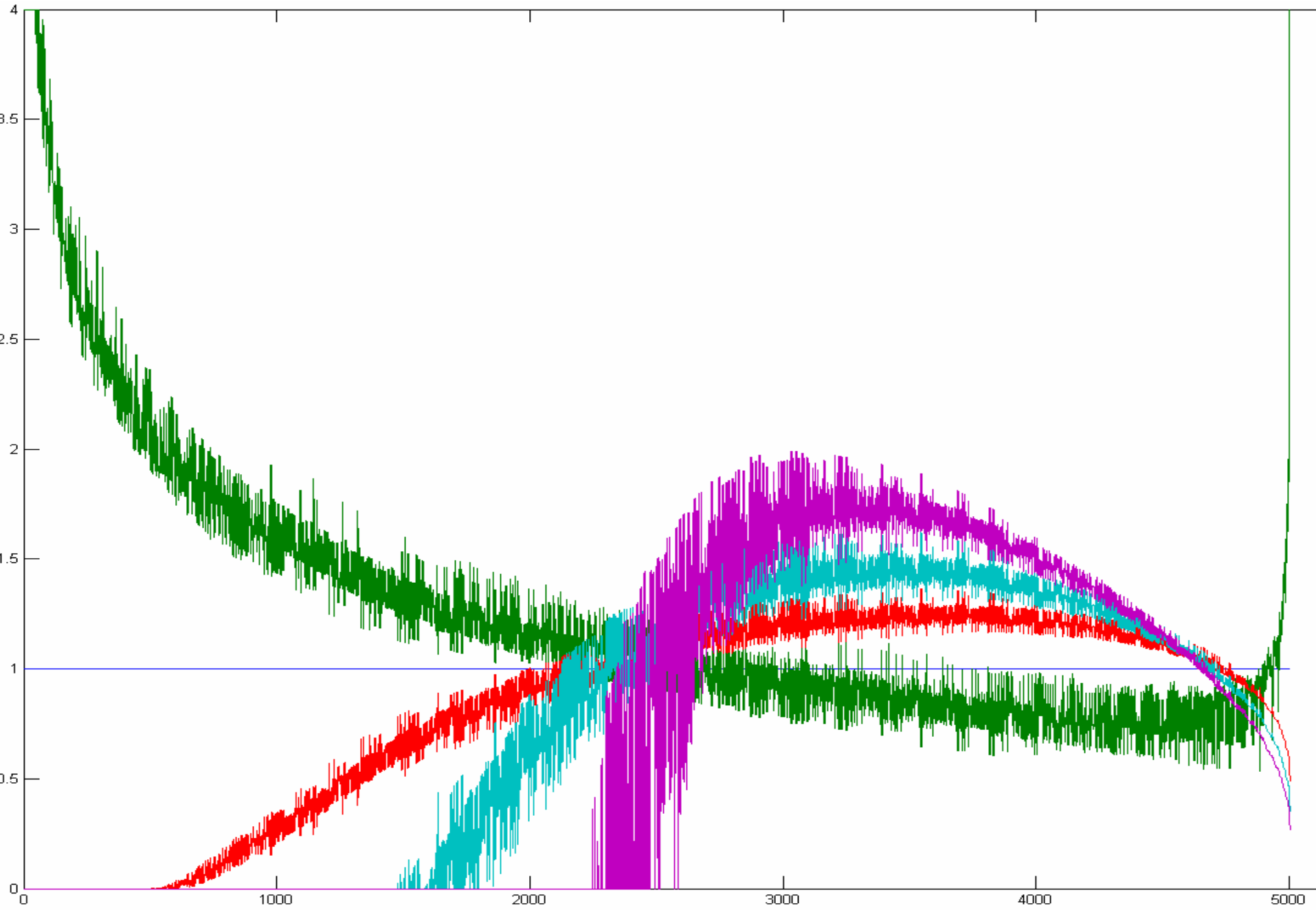
$$\begin{bmatrix} 0 & & & & & \\ 0 & 0 & & & & \\ 0 & 0.0071 & 0.1036 & & & \\ 0.0100 & 0.0350 & 0.1450 & 0.2069 & & \\ 0.0400 & 0.0650 & 0.2050 & 0.2669 & 0.3155 & \\ 0.0500 & 0.0929 & 0.2464 & 0.3155 & 0.3876 & 0.5076 \end{bmatrix}$$

Variance

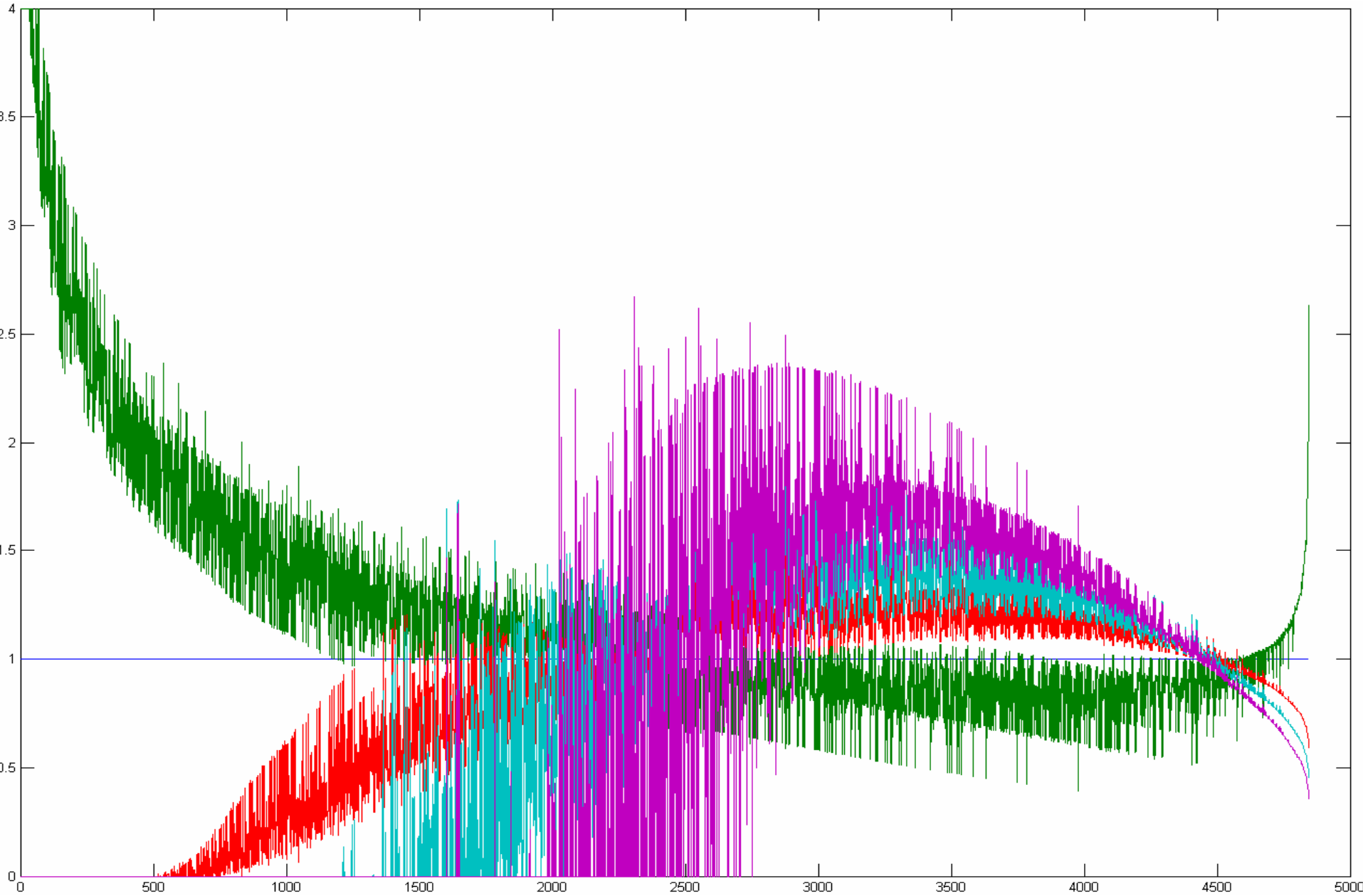
$$\begin{bmatrix} 0 & & & & & \\ 0 & 0 & & & & \\ 0 & 0 & 0.0946 & & & \\ 0 & 0.0320 & 0.1508 & 0.2227 & & \\ 0.0500 & 0.0680 & 0.2004 & 0.2678 & 0.3139 & \\ 0.0500 & 0.1000 & 0.2543 & 0.3150 & 0.3807 & 0.5000 \end{bmatrix}$$

Racine

Encore un pour la route: *N=12, n=6*



Et un petit dernier : *N=20, n=4*



Cette fois-ci c’est fini pour de bon.
Merci à tous,

Au revoir…

Ou adieu.