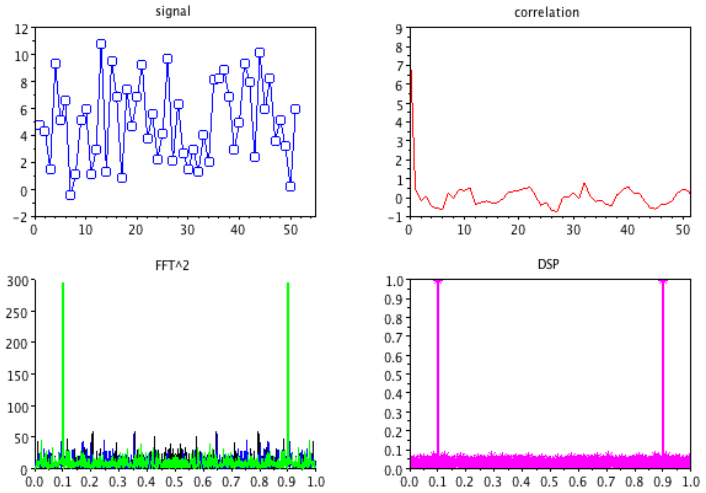
TD5 - Fonction de corrélation et DSP

Ce qu'on veut faire

Ce TD est dédié à l'introduction de deux outils statistiques importants pour l'étude des propriétés des séquences d'ADN : la fonction de corrélation et la Densité Spectrale de Puissance (DSP). Nous allons juste voir comment calculer ces grandeurs et appliquer cette méthode à des signaux aux propriétés connues, aléatoires ou déterministes.

Un exemple du résultat qu'on peut obtenir est tracé dans cette figure, où on a étudié un signal sinusoïdal avec bruit additif.



1 Bruit blanc discret – fonction corrélation

- Créer un signal aléatoire `z` (vecteur) de longueur 10^3.
- Calculer sa moyenne (`numpy.mean`) et sa variance (`numpy.var`).
- Construire le signal de moyenne nulle en soustrayant la moyenne du signal de départ.
- Calculer sa fonction de corrélation (`numpy.correlate`) et la tracer en fonction de la distance d le long de la séquence, pour des distances allant de zéro à 100
  Remarques :
  i. il faut utiliser l'option `mode='same'` pour indiquer qu'il s'agit d'une auto-correlation
  ii. `numpy.correlate` donne la fonction de correlation en fonction de la distance d pour d entre 0 et la longueur du signal. Pour des raisons de symétrie, la deuxième moitié du résultat est symétrique de la première, et il n'a pas donc d'intérêt de la tracer.
- Comparer. Comment peut-on retrouver la variance d'un signal à partir de sa fonction de corrélation ? Vérifier la cohérence de votre réponse avec la définition de fonction de corrélation.

2 Bruit blanc discret – DSP

- Calculer la FFT du signal `z` (`numpy.fft.fft`), puis construire l'estimateur de la DSP (que vaut Δt ?) :
$$|TF[z]|^2/T = |FFT(z)\Delta t|^2/(N\Delta t)$$
- Le tracer en fonction de la fréquence (vecteur entre 0 et $(N-1)\Delta f$: que vaut Δf) ?
- Calculer la DSP du signal `z` : pour ce faire, générer 1000 réalisations indépendantes du signal `z` et calculer la moyenne $\langle |FFT(z)\Delta t|^2/(N\Delta t)\rangle$ sur les réalisations.
- Tracer et interpréter le résultat.

3. D'autres signaux !

- **Signal aléatoire dichotomique :** A partir du signal `z`, créer un signal aléatoire *dichotomique* (composé de +1 et -1 seulement) de même longueur. Tracer le signal, sa fonction de corrélation et sa DSP. Interpréter le résultat.
- **Signal déterministe sinusoïdal :** Créer un signal sinusoïdal de longueur 10^3 dont la période est 10. Tracer le signal, sa fonction de corrélation et sa DSP. Interpréter le résultat.
- **Signal déterministe exponentiel :** Créer un signal de la forme `exp(-t/tau)` (avec par exemple `tau`=10) de longueur 10^3. Tracer le signal, sa fonction de corrélation et sa DSP *renormalisée par sa valeur maximale*. Interpréter le résultat. Quelle influence a le paramètre `tau` ?

3 Sommes et combinaisons linéaires

- Que deviennent DSP et fonction de corrélation pour la somme d'un signal sinusoïdal et d'un signal `50*exp(-t/tau)` ? Quelle propriété de la fonction corrélation et de la DSP on met en évidence par cet exemple ?
- Qu'observe-t-on pour la somme d'un signal sinusoïdal et d'un bruit blanc ? Etudier le rôle de l'amplitude du bruit.

4 Marche aléatoire

On s'intéresse à une marche aléatoire obtenue en sommant des contributions générés suivant une distribution uniforme (bruit blanc). Numériquement, on peut donc l'obtenir (dans sa version à temps discret) en sommant les premiers n pas d'un bruit blanc discret ξ_i :

$$z_n = \sum_{i=0}^{n} \xi_i \, ;$$

- Sous python, on peut obtenir la somme précédente par l'instruction "somme cumulative" `numpy.cumsum`.
- Construire un vecteur "marche aléatoire" de longueur 10^3 et étudier sa fonction de corrélation et sa DSP (comme moyenne sur plusieurs réalisations). Retrouve-t-on le comportement attendu DSP $\propto 1/f^2$?

5 Pour aller plus loin : séquence d'ADN

Lorsqu'on veut appliquer ce type de traitement à des séquences d'ADN, la première étape est de les convertir en séquences numériques. Typiquement, on applique un codage binaire du genre :

A, G $\rightarrow$ 1

C, T $\rightarrow$ -1

Trois séquences réelles sont disponibles (sous forme d'une ligne ou d'une colonne) sur le site de l'UE, et à l'adresse :

`http://www.lptmc.jussieu.fr/user/barbi/ENSEIGNEMENT/M2/sequences/`

La sequence `cytomegalovirus` est le génome complet d'un virus qui infecte l'homme, la séquence `human` est une séquence d'introns, et la séquence `lambda` vient d'un virus bactériophage qui infecte la bactérie Escherichia coli.

Vous pouvez essayer de convertir ces séquences en séquences binaires par vous même, ou bien de vous servir d'un script que Bertrand Caré a préparé pour nous, disponible également sur le site.

Une fois la conversion effectuée, vous pourrez appliquer à ces séquences les mêmes analyses statistiques que nous venons de mettre en place.

Qu'on observe-t-on ?