Krigeage sur graphes et groupes

Jean-Philippe Vert

Ecole des Mines de Paris, France
Jean-Philippe.Vert@mines.org

Journees de Geostatistiques, Ecole des mines de Paris, 19 septembre 2003

Plan

1. Motivations
2. Analyse harmonique et covariances sur les graphes
3. Covariances sur des groupes

Part 1

Motivations

Contexte

- Estimation d'une fonction $f : \mathcal{X} \to \mathbb{R}$ a partir d'observations ponctuelles $f(x_1), \ldots, f(x_n)$ ou $x_i \in \mathcal{X}$
- Ce dont on a besoin: une fonction de covariance $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ qui definit un processus de Gaussien de moyenne nulle par:

$$Ef(x)f(x') = k(x, x').$$

- On peut alors estimer les lois du processus conditionnellement aux observations, etc...

Cas classique

- $\mathcal{X} = \mathbb{R}^p$, avec $(p = 2, 3)$.
- Hypothese de stationarite:

$$k(x, x') = C(x - x').$$

- Parfois, hypothese d'isotropie:

$$k(x, x') = C(||x - x'||).$$

Conditions sur k et C

- k est une fonction de covariance ssi elle est symetrique:

$$\forall x, x' \in \mathcal{X}, \quad k(x, x') = k(x', x),$$

et definie positive:

$$\forall n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}, a_1, \ldots, a_n \in \mathbb{R}, \quad \sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0.$$

- Cas stationnaire: k est symetrique definie positive ssi C est la transformee de Fourier d'une mesure (reelle) positive.

Exemples de covariances stationnaire isotropiques

- Gaussienne:

$$C(x) = e^{-ax^2}, \quad \hat{C}(\omega) = \sqrt{\frac{\pi}{a}} e^{\pi^2 k^2 / a}.$$

- Exponentielle:

$$C(x) = e^{-2\pi\omega_0|x|}, \quad \hat{C}(\omega) = \frac{1}{\pi} \frac{\omega_0}{\omega^2 + \omega_0^2}$$

Notre probleme

- Comment generaliser ces methodes a des espaces $\mathcal{X}$ differents?
- example 1: $\mathcal{X}$ est un graphe
- example 2: $\mathcal{X}$ est un groupe

Part 2

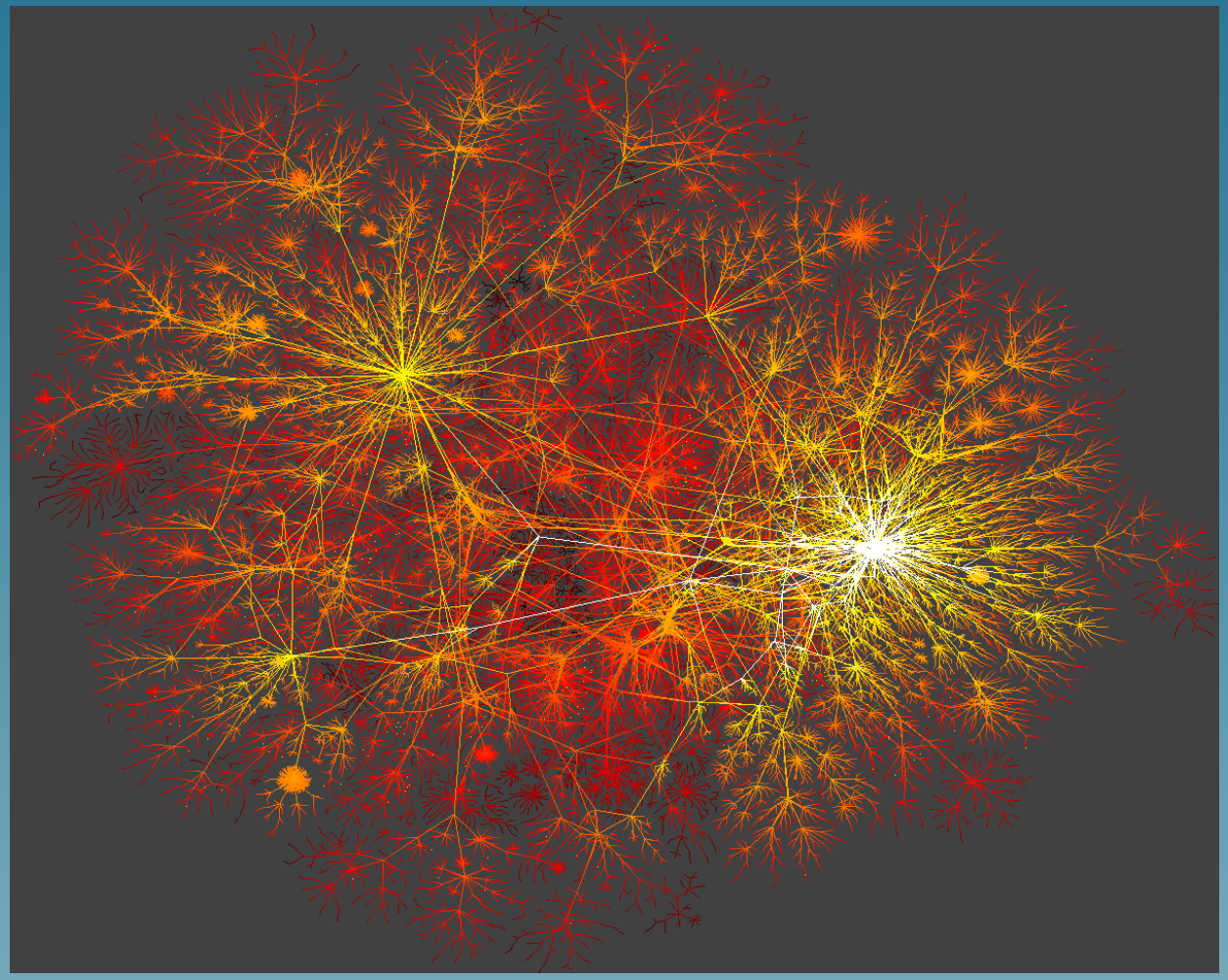
Analyse harmonique et covariances sur les graphes

Motivation

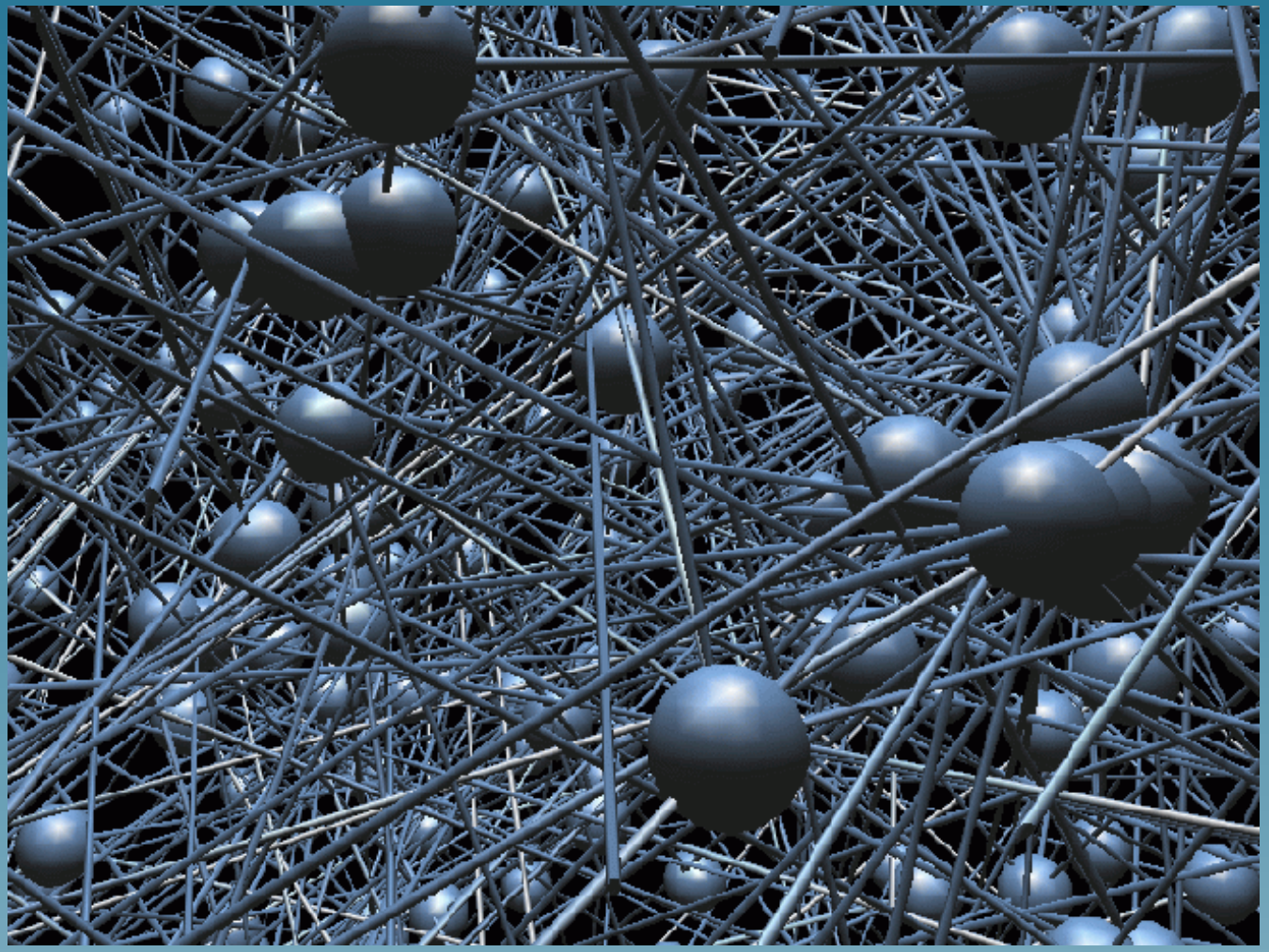
De nombreuses donnees peuvent se representer comme les noeux d'un graphe:

- par nature,
- par discretisation/echantillonage d'un espace continu,
- par necessite

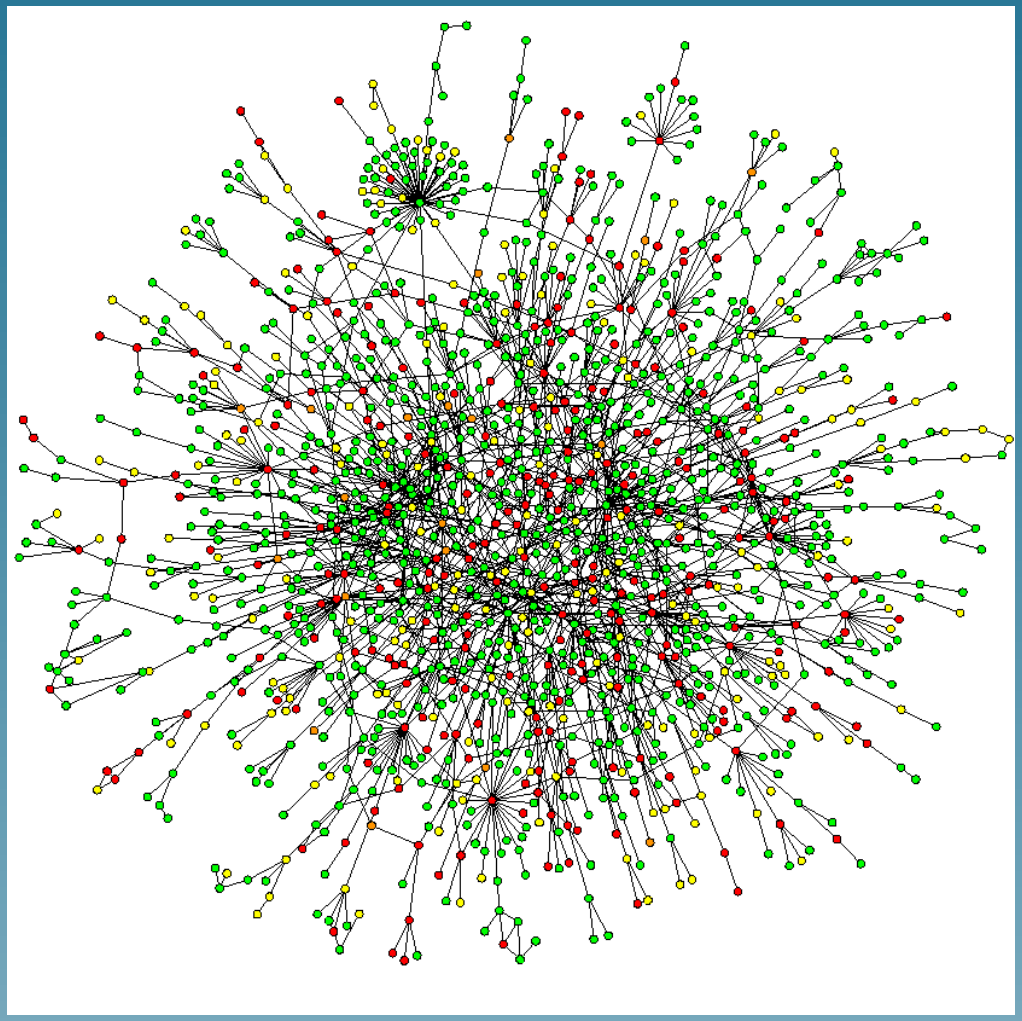
Internet (par nature)



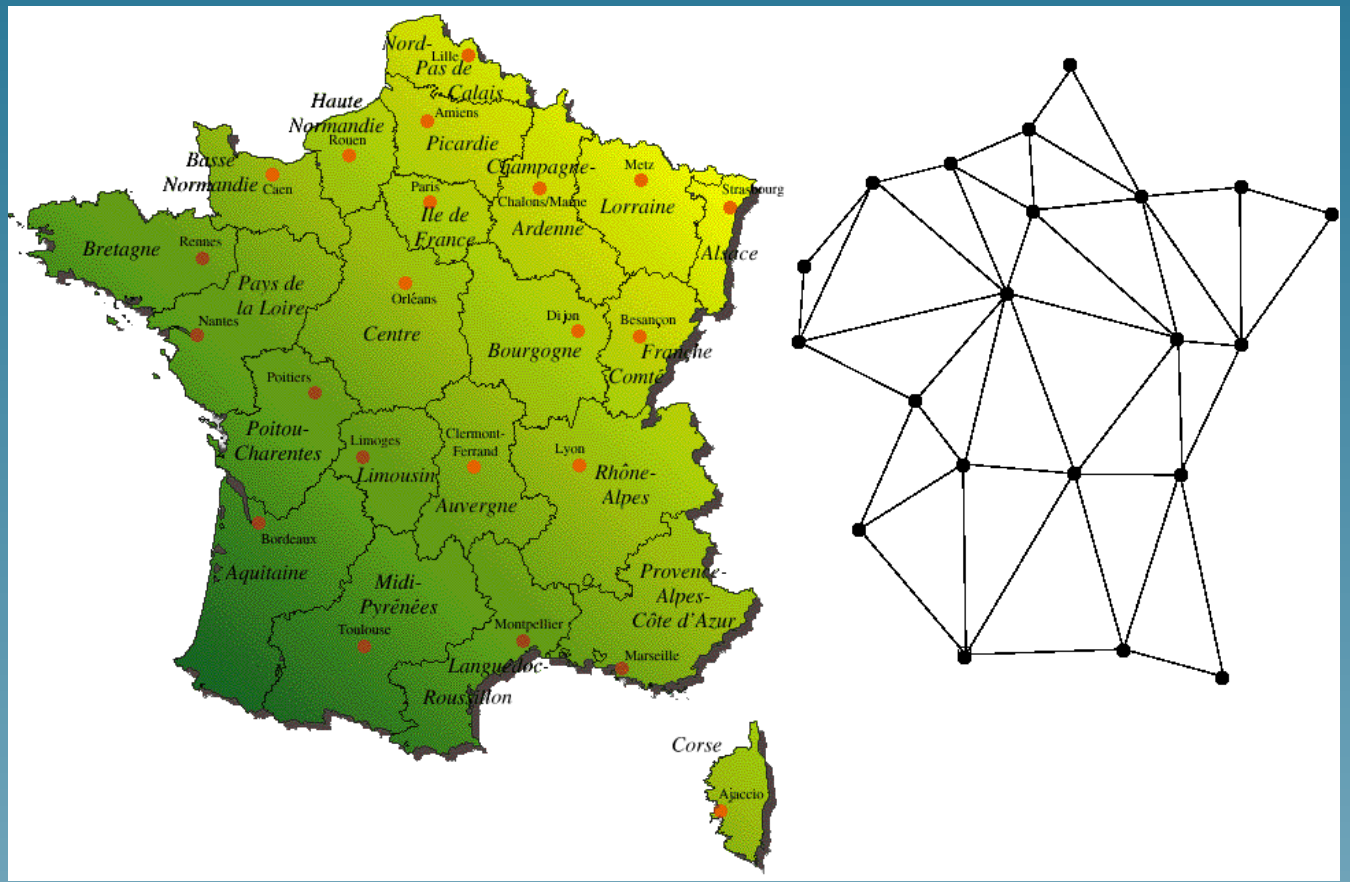
Reseau social (par nature)



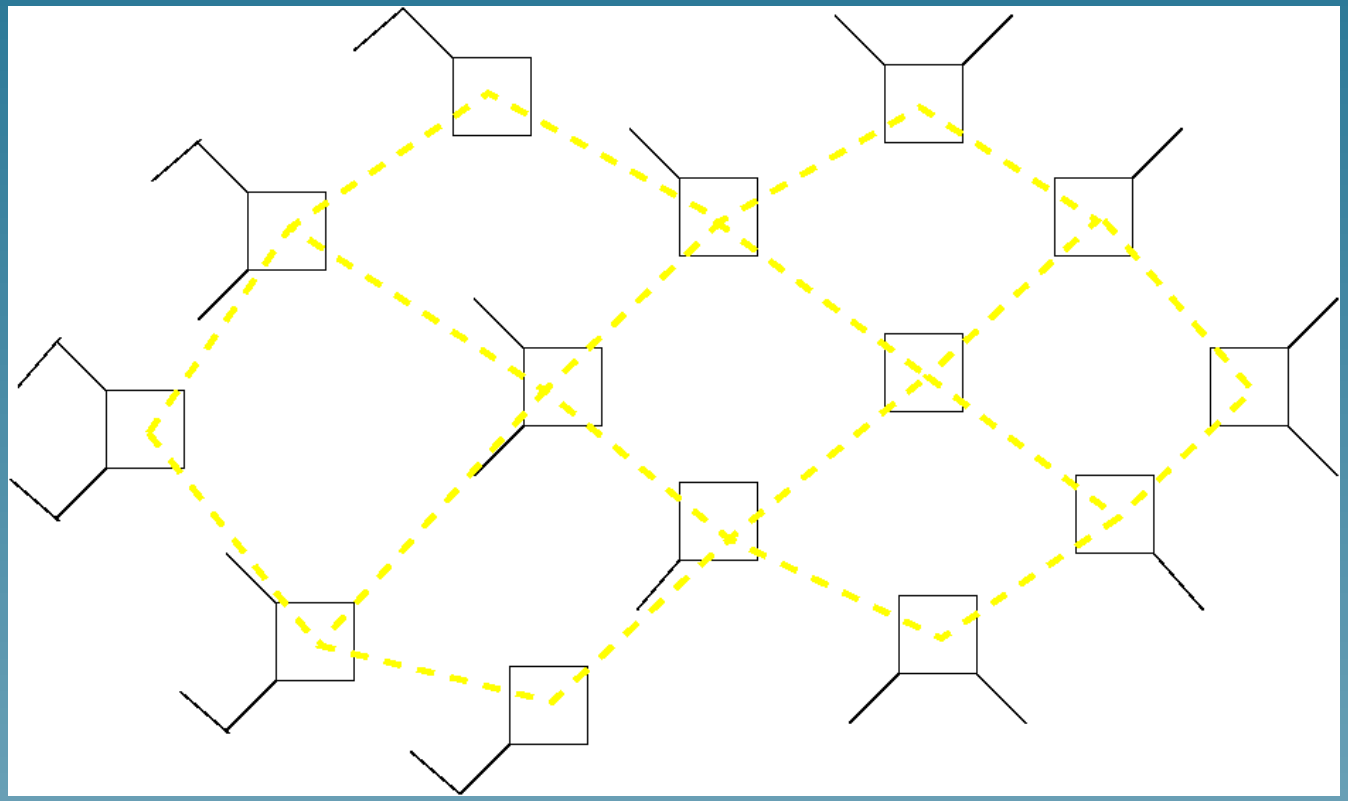
Interaction des proteines (par nature)



Regions (par discretisation)



Molecules (par necessite)



Covariance sur un graphe

- Graphe $G = (\mathcal{X}, E)$ avec $\mathcal{X}$ un ensemble fini de noeuds, $E \subset \mathcal{X} \times \mathcal{X}$ des liens entre les noeudx.
- $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ est une covariance ssi la matrice $K = [k(x, x')]_{x,x' \in \mathcal{X}}$ est symmetrique definie positive.
- Comment trouver une telle matrice qui contienne de l'information sur la structure du graphe

Premiere approche: distance

- Soit $d(x, x')$ une distance sur le graphe, par exemple la longueur du plus court chemin entre x et x'.
- Soit $k(x, x') = C(d(x, x'))$ ("stationnaire isotropique")
- Probleme: pas de condition generale sur C pour que k soit definie positive...

Deuxieme approche: analyse harmonique

- Dans le cas $\mathcal{X} = \mathbb{R}^p$, la condition sur C est:

$$C(h) = \int_{\omega \in \mathbb{R}^p} e^{2\pi i \omega h} g(\omega) d\omega$$

  avec $g \geq 0$.

- C se decompose sur la base de Fourier $\phi_\omega(h) = e^{i\omega h}$

- Faisons de l'analyse harmonique sur les graphe...

Laplacien sur les graphe

- Les ϕ_ω sont les fonctions propres de l'operateur Laplacien:

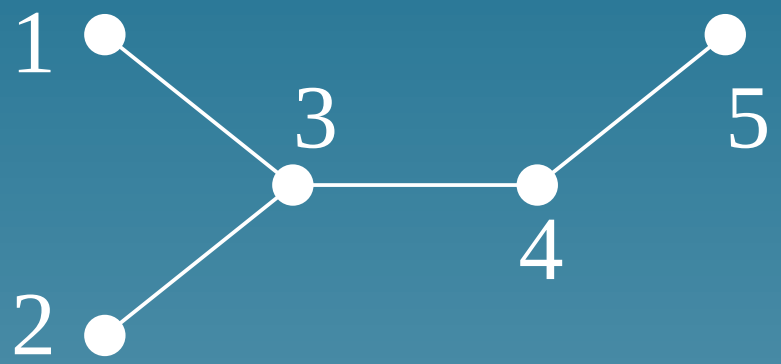
$$\Delta = \sum_{i=1}^{p} \frac{\partial}{\partial x_i}.$$

- Laplacien sur un graphe: si $f : \mathcal{X} \to \mathbb{R}$,

$$\Delta f(x) = \sum_{x' \sim x} [f(x') - f(x)]$$

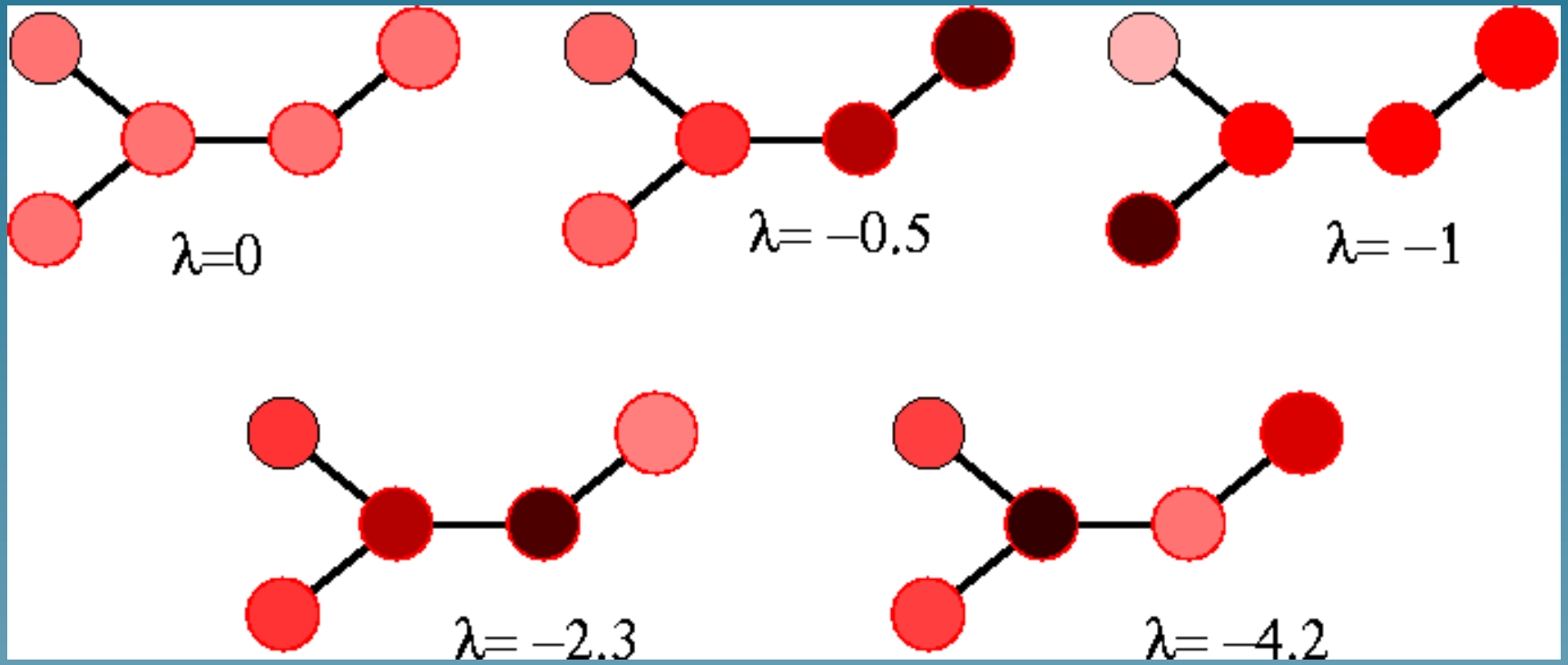
- Forme matricielle: $\Delta = A - D$, avec A la matrice d'ajacence et D la matrice diagonale des degres.

Example



$$\Delta = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 1 & 1 & -3 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

Fonctions propres du Laplacien



Spectre du Laplacien

- Les fonctions propres $\phi_1, \ldots, \phi_n$ du Laplacien forment une base de Fourier.
- Valeurs propres $\lambda_1 = 0 \geq \ldots \geq \lambda_n$ decroissent quand la frequence augmente.
- Transformee de Fourier d'une fonction $C : \mathcal{X} \to \mathbb{R}$:

$$\forall x \in \mathcal{X}, \quad C(x) = \sum_{i=1}^{n} \hat{c}_i \phi_i(x)$$

Du Laplacien a la covariance

- Si $C(x)$ est une fonction "reguliere", la fonction $k(x, x') = C(x)C(x')$ est une covariance "reguliere".
- Bonne fonction de covariance:

$$K(x, x') = \sum_{i=1}^{n} \gamma(\lambda_i)\phi_i(x)\phi_i(x')$$

avec $\gamma : \mathbb{R}_- \to \mathbb{R}_+$, croissante.

- Matrices: si $\Delta = U^{-1}DU$ alors $K = U^{-1}\gamma(D)U$

Exemple: noyau de la chaleur

- On choisit $\gamma(\lambda) = e^{\beta\lambda}$, et on obtient

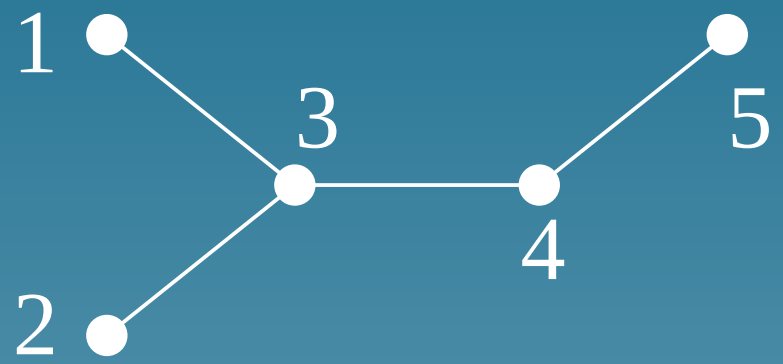
$$K_\beta = U^{-1} e^{\beta D} U = e^{\beta \Delta}.$$

- Ce noyau est la solution de l'equation de la chaleur:

$$\frac{\partial}{\partial \beta} K_\beta = \Delta K_\beta$$

- D'autres covariances sont bien sur possibles!

Example de ovariance



$$K = \exp(\Delta) = \begin{pmatrix} 0.49 & 0.12 & 0.23 & 0.10 & 0.03 \\ 0.12 & 0.49 & 0.23 & 0.10 & 0.03 \\ 0.23 & 0.23 & 0.24 & 0.17 & 0.10 \\ 0.10 & 0.10 & 0.17 & 0.31 & 0.30 \\ 0.03 & 0.03 & 0.10 & 0.30 & 0.52 \end{pmatrix}$$

Part 3

Covariance sur les groupes

(verifier l'heure...)

Motivations

- Un groupe est un ensemble G muni d'une operation $\circ$ associative, avec un element neutre e tel que $e \circ x = x \circ e = x$, et tel que chaque element $x \in G$ a un inverse $x^{-1} \in G$ tel que $x \circ x^{-1} = e$.
- $(\mathbb{R}^p, +)$ est un groupe (commutatif)
- L'ensemble des permutations de $(1, dots, n)$ muni de la composition est un groupe (non commutatif)
- Exemple: classez les 6 chaines de television par ordre de preference, et je predit vos revenus (?)

Covariance sur un groupe

- Sur $(\mathbb{R}^p, +)$, on prend $k(x_1, x_2) = C(x_1 - x_2)$ quand C a une transformee de Fourier reelle positive
- Sur $(g, \circ)$, soit $C : G \to \mathbb{R}$ et

$$k(x_1, x_2) = C(x_1 \circ x_2^{-1})$$

- Quelle condition sur C pour the k soit une covariance?
- Reponse: par l'analyse harmonique sur les groupes...

Cas des groupes finis (resume)

- Theorie classique de la representation lineaire des groupes finies (voir J.-P. Serres, 1966)
- La transformee de Fourier d'une fonction $f : G \to \mathbb{R}$ en un ensemble de matrices carres (on decompose f sur les caracteres des representations irreductibles)
- Theorem (Bochner): k est une covariance ssi la transformee de Fourier de C est composee de matrices semidefinie positive.
- De plus, C est constante sur les classes de conjugaison ssi c'est

une combinaison lineaire de caracteres irreuctibles : peu de degre de liberte!

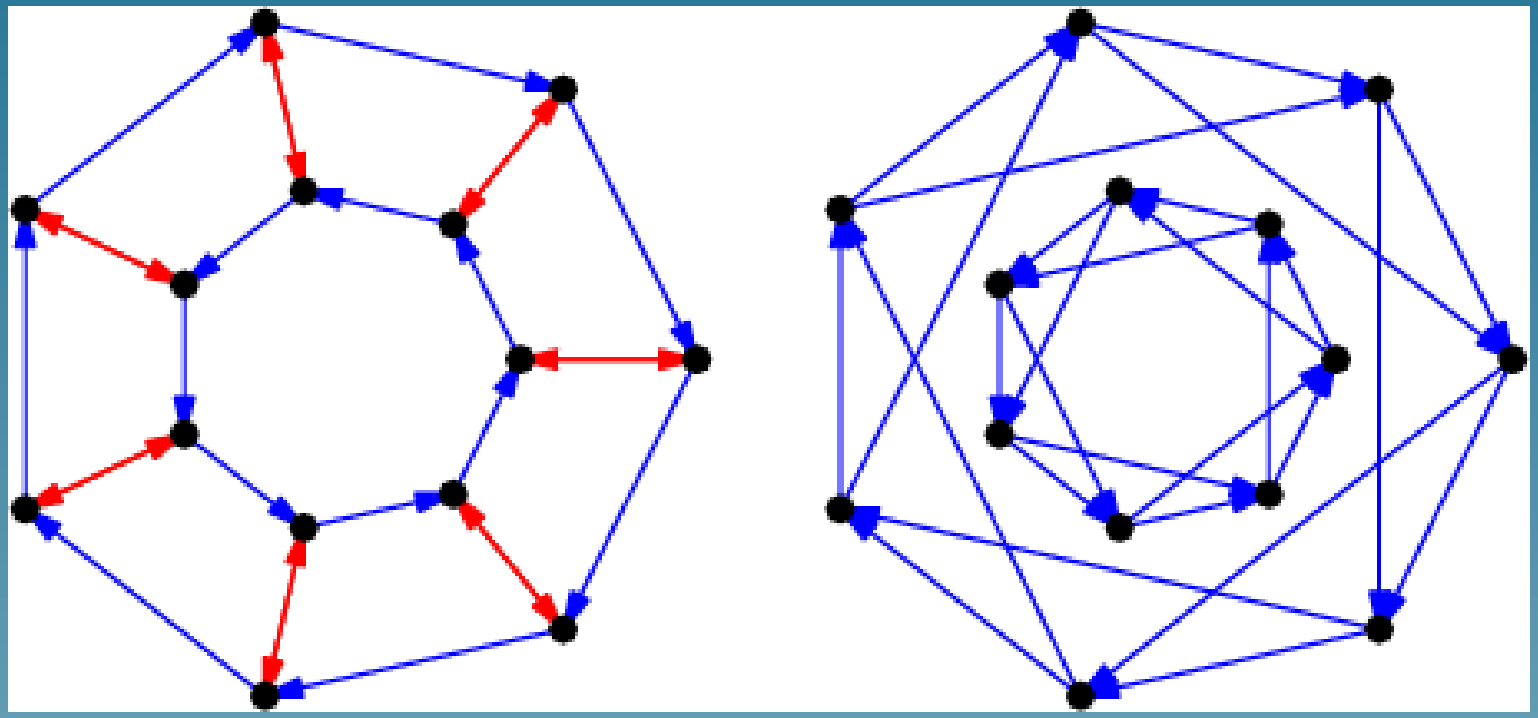
Calcul pratique d'une covariance sur un groupe

- Soit G un groupe
- Soit $S \subset G$ tel que $e \in S$ et si $x \in S$ alors $x^{-1} \in S$
- Le graphe de Cayley a pour neuds G et pour liens:

$$x_1 \sim x_2 \text{ ssi } x_1 \circ x_2^{-1} \in S.$$

- On peut faire un noyau sur le graphe de Cayley!

Graphes de Cayley



Groupe dihedral D_7, un groupe de permutation de 14 elements, avec differents ensembles S.

Conclusion

Conclusion

- On peut etendre la methodologie de krigeage (et processus Gaussien, machines a vecteurs de support, methodes a noyau...) a des espaces structures
- L'approche par analyse harmonique est tres generale pour les espaces avec une structure algebrique
- Idem pour la generalisation sur des varietes Riemannienne (analyse harmonique par calcul differentiel.
- Encore tres peu applique