

Inférences statistiques sur un graphe : estimation et prédiction

Thibault Espinasse

10 octobre 2008

Table des matières

1	Remerciements	3
2	Introduction	4
3	Modélisation par automates cellulaires probabilistes	4
3.1	Un modèle de microéconomie : interactions locales hierarchisées	4
3.2	Deux modèles du trafic routier	5
4	Inférences statistiques sur un graphe	8
4.1	Structure de covariance sur un graphe	8
4.2	Krigage	10
4.3	Estimation	11

Cette introduction au domaine de recherche fait suite à un mémoire de Master 2 établi sous la direction de Fabrice Gamboa et Jean-Michel Loubes, Professeurs au Laboratoire de Statistique et de Probabilité à l'Université Paul Sabatier (Toulouse III).

1 Remerciements

Je tiens à remercier bien chaleureusement mes trois directeurs, Patrick Cattiaux, Fabrice Gamboa, et Jean-michel Loubes, ainsi que Gilles Stoltz, Wendelin Werner, Jean-François Le Gall, Pascal Massart et Elisabeth Gassiat qui se sont toujours montrés disponibles, pour répondre à mes questions techniques, comme aux hésitations concernant mon orientation. Merci aussi à tous les acteurs du laboratoire de Toulouse, qui m'ont accueilli à bras ouverts.

2 Introduction

En statistiques, il est courant, lorsque l'on travaille sur des données réelles, de ne faire qu'une observation partielle. Aussi, de nombreuses techniques ont été développées pour traiter avec des données manquantes. On peut entre autre citer le krigeage [4], qui considère les prédictions linéaires, les algorithmes type EM [7], SAEM [12] ou imputation [9]... On se propose d'appliquer ces résultats à des processus indexés par des graphes, et n'interagissant qu'entre sommets voisins. On s'intéresse à ce genre de processus dans le but de traiter les données venant d'une part des domaines où il existe structurellement un graphe explicite sur lequel on observe les variables (en particulier pour le trafic routier), et d'autre part de phénomènes modélisables par automates cellulaires. Il s'agira donc en premier lieu de définir un champ ou un processus aléatoire sur un graphe. La difficulté est de bien choisir une classe décrivant des champs ou processus dont les interactions n'ont lieu qu'entre voisins.

Un premier modèle "jouet" peut permettre d'appréhender les outils et les problèmes qui vont se poser.

3 Modélisation par automates cellulaires probabilistes

Dans cette partie, on donne deux exemples auxquels je me suis intéressé, pour mon M1, puis pour mon M2, de modèles utilisant des automates cellulaires probabilistes. On peut ici préciser que ce ne sera pas le seul cadre d'étude : on s'intéresse en général à un champ aléatoire sur un graphe.

Définition 3.1

Un automate cellulaire probabiliste est la donnée d'un ensemble de cellules S , sommets d'un graphe G , de leurs étiquettes $(e_s)_{s \in S} \in E^S$, et d'une famille de probabilité (ou de densités) de transitions $p_{\hat{s}}(e_{\hat{s}}(t) | (e_s(t-1))_{s \in V(\hat{s})})$ où $V(s)$ désigne l'ensemble des voisins de s . L'évolution se fait donc de manière synchrone et indépendante.

Remarque : Il s'agit en fait d'une chaîne de Markov dont la densité de transition peut s'écrire comme produit de densité de transition de chaque cellule, qui de plus ne dépend que des voisins de celle-ci. On pourrait tout à fait généraliser ce genre de considérations à des processus temporels.

3.1 Un modèle de microéconomie : interactions locales hiérarchisées

Lors du mémoire de M1, on s'était intéressé à des systèmes d'agents en interactions locales hiérarchisées. Il s'agit d'un automate cellulaire probabiliste dont le nombre d'états est infini, et les transitions sont à densité.

On rappelle la forme du modèle : on choisit pour ensemble de cellule l'ensemble \mathbb{Z} qui désigne l'ensemble des agents rationnels que nous considérons. Chaque agent $a \in \mathbb{Z}$ possède pour étiquette une action $x_a(t) \in \mathbb{R}$. Les probabilités de transition sont données par un type aléatoire $\theta_a \in \Theta$ de loi ν et une fonction de décision $g_a : \mathbb{R}^2 \times \Theta \rightarrow \mathbb{R}$. L'action se choisit en posant $x_a(t) = g(x_a(t-1), x_{a+1}(t-1), \theta_a(t))$ ce qui nous donne une famille de probabilité de transitions $p_a(x_a(t)|x_a(t-1), x_{a+1}(t-1))$. La question que l'on se posait alors était assez particulière : on se demandait, si étant donnée une fonction d'utilité $u : \mathbb{R}^3 \times \Theta$, prenant en argument $x_a(t), x_a(t-1), x_{a+1}(t)$ et θ_a les problèmes d'optimisation de u (avec anticipation, information partielle et prévention) avaient une solution stationnaire.

3.2 Deux modèles du trafic routier

Modèle de Nagel Schrekenberg

À l'instant t , chaque véhicule $i \in [1, n]$ a pour position $x_i(t)$ et pour vitesse entière $v_i(t) \in [1, v_{max}]$, $d_i(t) = x_{i-1}(t) - x_i(t)$ désigne la distance séparant le véhicule i de son prédécesseur. Chaque itération de l'algorithme comprend quatre étapes :

- *Accélération* $v_i(t+1) \leftarrow \min(v_i(t) + 1, v_{max})$
- *Décélération* $v_i(t+1) \leftarrow \min(d_i(t) - 1, v_i(t))$
- *Randomisation* $v_i(t+1) \leftarrow \max(v_i(t) - 1, 0)$ avec proba p
- *Mouvement* $x_i(t+1) \leftarrow x_i(t) + v_i(t)$

Modèle hybride

L'algorithme agit comme suit : le temps est découpé en tranches de dt , l'espace en tranches de $\frac{1}{n_x}$ et les vitesses en tranches de $\frac{1}{n_v}$

On choisit d'abord l'accélération pour chaque véhicule i , précédé par le véhicule j de la forme "follow the leader" :

$$p_{acc}(i, t+1) = n_v \left(c_1 * \frac{V(j, t-T) - V(i, t-T)}{(X(j, t-T) - X(i, t-T))^\alpha} + dv - c_2 \max(V(i, t) - v_{legal}, 0)^f \right)$$

La forme de cette accélération vient de modèles physiques appelés "Follow the leader". Le fait d'introduire un temps de réaction permet l'apparition des phénomènes d'"accordéon" que l'on peut observer dans les embouteillages. On a donc ici affaire à des chaînes de Markov d'ordre supérieur à 1.

On accélère alors :

$$V(i, t+1) = V(i, t) + \text{signe}(p_{acc}(i, t+1)) \frac{X}{n_v}$$

où X est une Bernouilli de paramètre $p_{acc}(i, t + 1)$

On avance ensuite : Si $X(j, t) - X(i, t) > dsécurité$ alors

$$X(i, t + 1) = X(i, t) + \frac{Y}{n_x}$$

où Y est une Bernouilli de paramètre $\frac{V(i, t+1)}{n_x dt}$ sinon

$$X(i, t + 1) = X(i, t)$$

On peut préciser que

- Ces modèles sont bien Markoviens : si on considère un état comme la donnée de la position et de la vitesse de tous les véhicules.
- Il s'agit même d'automates cellulaires : chaque voiture constitue une cellule qui n'a comme seuls voisins la voiture qui la précède et celle qui la suit. Les étiquettes sont de la forme (position, vitesse), et l'évolution se fait bien de manière synchrone et indépendante. (Ceci ce complique lorsqu'on autorise les changements de files).
- On peut s'intéresser à l'irréductibilité ou non de cette chaîne, et aux possibles convergences. On peut aussi se demander ce que donne ce genre de modèle lorsqu'on passe à la limite hydrodynamique, et la fréquence des événements "rares" qui surviennent plus souvent à cause du milieu aléatoire.

Pour le moment, très peu de résultats théoriques probabilistes existent sur ce genre de modèles, du fait de l'irréversibilité du processus. Les problèmes combinatoires d'identification des classes de récurrence ne sont pas non plus résolus. Durant mon M2, je n'ai pas eu le temps de me pencher sur ces nombreux problèmes, ni même d'envisager un cadre d'hypothèses raisonnables sur des modèles à automates cellulaire pour espérer pouvoir obtenir des résultats. je me suis contenté d'explorer les modèles existants dans la littérature, et de chercher ceux qui pourraient permettre une analyse, tout en cherchant le plus d'adéquation possible. C'est comme cela qu'est apparu le cadre des automates cellulaires d'où les questions d'inférence statistiques sur des graphes.

Dans tous les articles traitant du trafic, les physiciens s'intéressent aux graphes suivant : La figure 1 donne la densité spatiale k en fonction de la densité temporelle q et la figure 2 le graphe d'occupation de l'espace temps (t, x) par les véhicules.

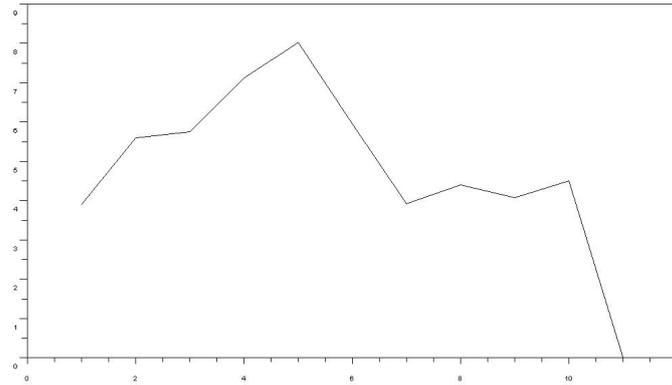


FIG. 1 – Figure 1

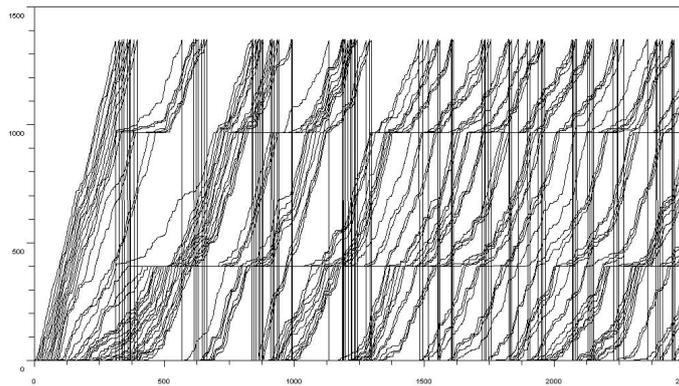


FIG. 2 – Figure 2

On discerne sur la figure 2 :

- la présence de bouchons qui se dissipent pour la densité utilisée
- la formation spontanée de bouchons après les intersections, mais non ailleurs.
- l'homogénéisation progressive de l'occupation de la route (tout les véhicules partent du même point en même temps)

4 Inférences statistiques sur un graphe

Le premier problème sera de définir une "bonne" classe de champs (ou de processus) aléatoires sur un graphe. Lors de mon M2, j'ai étudié un modèle jouet où l'on connaissait la forme explicite de la famille paramétrique de covariance.

4.1 Structure de covariance sur un graphe

Définition 4.1 (Graphe étiqueté) On appellera graphe étiqueté la donnée d'un ensemble G de sommets de cardinal n et d'une matrice de poids symétrique $W \in (G \times G)^{\mathbb{R}^+}$. On désignera par arête tout couple (i, j) avec $i \neq j$ tel que $W_{ij} > 0$, et on le notera $i \sim_{W_{ij}} j$.

Définition 4.2 On appellera matrice de connection la matrice diagonale D vérifiant $d_i = D_{ii} = \sum_k W_{ik}$. On appellera laplacien discret de G : $L := D - W$ et laplacien renormalisé $\tilde{L} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$.

L est par définition symétrique, on remarque de plus que

$$X^T L X = \frac{1}{2} \sum_{i,j} W_{ij} (X_j - X_i)^2$$

d'où L est une forme bilinéaire positive dont le noyau contient la droite

engendrée par le vecteur $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$. En fait dans le cadre d'un graphe connexe,

il s'agit du noyau, et l'écriture précédente montre que de plus le cône isotrope est alors réduit au noyau. On peut faire plusieurs autres remarques :

D'une part, on note que le Laplacien discret mesure bien une forme de régularité sur le graphe (on peut voir \mathbb{R}^n comme l'ensemble des fonctions de G dans \mathbb{R}). On impose à la fonction de peu varier sur entre deux sommets reliés par une arête, d'autant plus que le poids W_{ij} est élevé. On peut comparer l'égalité précédente à $\|f\|_{\Delta} = \int_{\Omega} (\nabla f)^2 d\omega$. Pour plus d'information, on peut lire [1].

D'autre part si on considère l'opérateur L restreint à l'espace quotienté par les fonctions constantes, il définit un produit scalaire, dont la norme associée mesure la régularité de f

On a $\text{Tr}(\tilde{L}) = n$

De plus, L (et donc \tilde{L}) est diagonalisable dans une b.o.n. et il est naturel de vouloir mieux connaître la forme de ses vap et vep. Le théorème suivant (cf [3]) donne une première indication.

Théorème 4.1 *Toutes les valeurs propres λ_i de \tilde{L} vérifient $0 \leq \lambda_i \leq 2$.*

On peut citer quelques exemples de vap pour des graphes classiques (on rappelle que l'on a toujours $\lambda_1 = 0$) :

- Pour un graphe complet à n sommets, on a $\lambda_2 = \dots = \lambda_n = \frac{n}{n-1}$
- Pour un graphe étoilé à n sommets, on a $\lambda_2 = \dots = \lambda_{n-1} = 1$ et $\lambda_n = 2$
- Pour un graphe chemin à n sommets, les valeurs propres sont de la forme $1 - \cos(\frac{\pi k}{n-1})$, $k = 0, \dots, n-1$
- Pour un graphe cycle à n sommets, les valeurs propres sont de la forme $1 - \cos(\frac{2\pi k}{n})$, $k = 0, \dots, n-1$
- Pour un n -cube à 2^n sommets, les valeurs propres sont de la forme $\frac{2k}{n}$ avec C_n^k comme ordre de multiplicité.

Dans le modèle étudié en M2, la covariance est de la forme suivante :

$$K_\sigma(i, j) = e^{-\frac{\sigma^2}{2} \tilde{L}(x_i, x_j)}$$

On pourra lire [2] pour comprendre pourquoi on choisira une covariance de cette forme. Néanmoins, on peut donner une heuristique qui permet de comprendre pourquoi on choisit une telle forme de matrice de covariance dans un cas de diffusion.

Considérons n variables aléatoires indépendantes $X_1^0 \dots X_n^0$ iid de loi $\mathcal{N}(0, 1)$. A chaque instant t , les variables vont un peu s'ajuster à leurs voisins. On voudrait donc écrire $X_i^t = (1 - \alpha)X_i^{t-1} + \sum_{i \sim j} (\frac{W_{ij}}{d_j} \alpha X_j^{t-1})$. Il est nécessaire de renormaliser α par le temps pour pouvoir regarder ce qui se passe en temps long ; on obtient une règle du type pour tout $t \leq k$, on calcule

$$\forall i, X_i^t = (1 - \frac{\alpha}{k})X_i^{t-1} + \sum_{i \sim j} (\frac{W_{ij}}{d_i} \frac{\alpha}{k} X_j^{t-1})$$

Puis en posant $Z_i^t = \sqrt{d_i} X_i^t$ i.e. $Z^t = D^{\frac{1}{2}} X^t$, on peut écrire

$$\forall t \leq k, \forall i, Z_i^t = (1 - \frac{\alpha}{k})Z_i^{t-1} + \sum_{i \sim j} (\frac{W_{ij}}{\sqrt{d_i} \sqrt{d_j}} \frac{\alpha}{k} Z_j^{t-1})$$

On a alors

$$Z^t = (I - \frac{\tilde{L}\alpha}{k})^t Z^0$$

En particulier, si on pose $Z^\infty = \lim_{k \rightarrow \infty} Z^k$, on obtient

$$Z^\infty = \exp(-\alpha \tilde{L}) Z^0$$

Puis

$$\begin{aligned}
 K^\infty(i, j) = \text{cov}(Z_i^\infty, Z_j^\infty) &= \mathbb{E} \left[\sum_k Z_k^0 \exp(-\alpha \tilde{L})(i, k) \sum_{k'} Z_{k'}^0 \exp(-\alpha \tilde{L})(j, k') \right] \\
 &= \mathbb{E} \left[\sum_k Z_k^0 \exp(-\alpha \tilde{L})(i, k) Z_k^0 \exp(-\alpha \tilde{L})(j, k) \right] \\
 &= \mathbb{E} \left[\sum_k Z_k^0 \exp(-2\alpha \tilde{L})(i, j) Z_k^0 \right] \\
 &= \text{dexp}(-2\alpha \tilde{L})(i, j)
 \end{aligned}$$

en utilisant l'indépendance des Z_i^0 . On peut dès lors comprendre le rôle que joue ce type de matrice dans la structure de covariance pour des processus liés localement.

Dans le cas général, la première étape consistera à spécifier un cadre d'étude. Il s'agit de définir des champs sur des graphes, et d'obtenir des propriétés pouvant indiquer si une famille paramétrique de covariance décrit bien de tels champs.

Plusieurs voies sont a priori à explorer (ce que je n'ai pas encore fait). On pourrait en particulier chercher une description utilisant la mesure spectrale, ou bien s'intéresser aux descriptions locales (par exemple type champs de Gibbs).

Une fois ce cadre spécifié, l'objectif est de prévoir, à partir d'une observation partielle des données, la valeur du champ aux sommets non observés. Pour cela, il faudra développer à la fois des méthodes de prévision, mais aussi d'estimation.

4.2 Krigeage

On observe un champ aléatoire Z de fonction moyenne m et de covariance K sur \mathbb{R}^d sur un ensemble $A \in \mathbb{R}^d$. Notre but est d'obtenir une prédiction linéaire de Z en un point x_0 (ou sur un ensemble $B \in \mathbb{R}$) de façon à minimiser l'erreur quadratique. En d'autres termes, on recherche la projection dans \mathbb{L}^2 de $Z(x_0)$ sur $\text{Vect}(1, Z(A))$;

On se placera dans le cadre restreint où A est fini : $A = \{x_1, \dots, x_n\}$, et les variables $Z(x_i)$ sont indépendantes au sens de l'algèbre linéaire, i.e. $K = (K(x_i, x_j))_{i,j=1,\dots,n}$ est inversible. Le but est de déterminer $\arg \max_{\lambda_0} \mathbb{E}[Z(x_0) - \lambda_0 - \sum \lambda_i Z(x_i)]^2 = \mathbb{E}[Z(x_0) - \lambda_0 - \lambda^T Z]$. On pose de plus $m = [m(x_1), \dots, m(x_n)]^T$. On pose $k = \text{cov}(Z, Z(x_0))$.

Théorème 4.2 *Le meilleur prédicteur linéaire s'obtient en posant :*

$$\begin{aligned}
 \lambda &= K^{-1}k \\
 \lambda_0 &= m(x_0) - k^T K^{-1}m
 \end{aligned}$$

Dans le cas gaussien, ce meilleur prédicteur est optimal au sens où il s'agit de l'espérance conditionnelle des prévisions en fonction des observations. En réalité on connaît même mieux : on connaît la loi de variables à prédire conditionnellement aux observations :

Théorème 4.3 Soit $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$ un vecteur gaussien.

Alors, $\mathcal{L}(X_1|X_2) \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$

On est donc capable, si l'on connaît la structure de covariance et le vecteur moyen, de donner la meilleure prédiction linéaire, qui dans le cas Gaussien est optimale car est l'espérance conditionnelle (qui de plus annule l'erreur quadratique). Il faut donc spécifier une structure de covariance, de préférence paramétrique, et trouver un moyen d'estimer les paramètres.

4.3 Estimation

Dans la plupart des cas usuels, on ne connaît ni la covariance ni la moyenne et nous devons les estimer avec l'échantillon même qui sera utilisé pour la prévision. On peut procéder par maximum de vraisemblance.

On rappelle qu'un estimateur désigne une fonction mesurable des observations. Il s'agit donc d'une variable aléatoire, si on considère que les observations sont juste des réalisations d'une variable aléatoire de loi inconnue.

Différents critères existent pour évaluer si un estimateur \hat{m} est un "bon" estimateur de m :

- Il est dit sans biais si on a $\mathbb{E}[\hat{m}] = m$
- Lorsque la taille de l'échantillon grossit, on considère les critères asymptotiques habituels (presque sûre, en probabilités)
- On peut considérer le minimiseur de l'erreur quadratique moyenne. En particulier, on peut chercher l'estimateur optimal (qui minimise la variance parmi les estimateurs non biaisés).

On peut définir selon les besoins d'autres critères. Un estimateur jouant un rôle très important en statistiques s'appelle l'estimateur du maximum de vraisemblance. Il ne possède pas de propriétés générales (non biaisé...) mais il est lié à l'estimateur optimal. Il s'agit en fait de l'estimateur le plus probable lorsque l'on ne connaît rien *a priori* sur la loi des variables à estimer.

Formellement, on observe n réalisations indépendantes X_1, \dots, X_n d'une variable aléatoire X . La loi de cette variable est inconnue mais on sait qu'elle appartient à une famille paramétrique de lois $(P_\theta)_{\theta \in \mathbb{R}^D}$ de densité $(f_\theta)_{\theta \in \mathbb{R}^D}$ (par exemple gaussiennes centrées de variance inconnue θ). Le but est d'estimer θ . La valeur la plus probable (estimateur de maximum de vraisemblance) est celle qui maximise

$$l(\theta, X_1, \dots, X_n) = f_\theta(X_1) \cdots f_\theta(X_n)$$

On cherche donc à maximiser la vraisemblance, ou la log-vraisemblance. Le problème est que les variables non observées interviennent dans cette

vraisemblance. C'est pourquoi plusieurs techniques ont été développées afin de faciliter cette maximisation.

L'algorithme EM a été formalisé pour la première fois dans [7]. L'idée principale est que pour estimer des paramètres à partir de données incomplètes, si on connaît la loi conditionnelle des données non-observées sachant les données observées et les paramètres, on peut créer un algorithme qui converge vers le maximum de vraisemblance.

Le problème se présente alors ainsi :

On a une fonction $l(\Theta, z)$ de vraisemblance où Θ désigne l'ensemble des paramètres et z un vecteur de données totales. N'observant que y la projection de z sur un sous espace vectoriel, on cherche à maximiser le logarithme de la vraisemblance moyenne $L(\Theta) = \log \mathbb{E}[(l(\Theta, z)|y)]$. Pour cela on applique un algorithme dont la k -ième étape consiste en deux parties :

Espérance : On calcule $Q(\Theta, \Theta_k) = \mathbb{E}_{\Theta_k}[\log(l(\Theta, z))|y]$;

Maximisation : On choisit le Θ_{k+1} qui maximise $\Theta \mapsto Q(\Theta, \Theta_k)$;

Il s'agit en fait de maximiser en moyenne conditionnellement aux observations (en utilisant le paramètre de l'étape précédente) la log vraisemblance. Le succès de cet algorithme est due au fait que le logarithme de la vraisemblance moyenne L augmente à chaque étape.

Le but est de prouver que sous des hypothèses suffisantes, l'algorithme converge bien vers un $\arg \max L$. Les hypothèses contenues dans l'article séminal [7] sont plutôt restrictives. De plus la démonstration comporte une erreur, et on doit rajouter des hypothèses encore plus fortes (par exemple [8] suppose la convergence de Θ_p)

Plusieurs versions stochastiques de l'EM ont été développées. L'idée est d'introduire du bruit pour éviter la convergence vers un point stable (par exemple maximum local) qui ne soit pas le maximum global recherché. Ceci permet aussi d'alléger un peu les hypothèses.

On peut aussi pour des cas simples se demander ce que donne l'algorithme qui, au lieu de maximiser $\mathbb{E}_{\Theta_k}[\log(l(\Theta, z))|y]$ à chaque étape, calculerait $\arg \max_{\Theta} l(\Theta, z)$ ou z a été construit en complétant les variables observées z_{obs} par la moyenne conditionnelle $\mathbb{E}[z_{prev}|z_{obs}, \Theta_k]$.

Enfin, on pourrait vouloir chercher des intervalles de confiance sur les variables prédites : on cherche pour tout $\epsilon \geq 0$, un intervalle I_{ϵ} tel que

$$P(z_{prev} \in I_{\epsilon}) \geq 1 - \epsilon$$

Pour résumer, pour ce pencher sur ce problème de prédiction de champs sur des graphes, il faudra explorer plusieurs techniques :

- Pour définir la classe de champs sur laquelle on va travailler, il faudra se demander quels outils peuvent nous servir (structures locales de covariance, mesure spectrale...)
- Un travail sur la structure spectrale du Laplacien sera à faire, car il est très probable qu'il ait un rôle à jouer dans la définition des champs sur les graphes.
- Il faudra approfondir les résultats qui existent en prédiction. On s'intéressera en parallèle aux problèmes de régularisation dans les RKHS, qui sont de même nature.
- On pourra s'attarder sur les différents algorithmes qui existent, mais aussi s'intéresser à d'autres techniques d'optimisation
- Enfin on pourra chercher des moyens d'obtenir des intervalles de confiance sur des variables prédites à partir de paramètres eux-mêmes estimés.

Références

- [1] J. Simola and R. Kondor *Kernels and regularization on Graphs*
- [2] R. Kondor and J. Lafferty *Diffusion Kernels on Graphs and Other Discrete Structures*
- [3] F. Chung-Graham *Spectral Graph Theory*
- [4] M. L. Stein *Interpolation of Spatial Data*
- [5] C. K. I. Williams *Prédiction with Gaussian processes : from linear regression to linear prediction and beyond*
- [6] B. S. Kerner *The Physics of Traffic*
- [7] A. P. Dempster, N. M. Laird, D.B. Rubin *Maximum likelihood from incomplete data via the EM algorithm*
- [8] F. J. Wu *On the convergence properties of the EM algorithm*
- [9] J. Barnard, D. B. Rubin *Small-Sample Degrees of freedom with multiple imputation*
- [10] K. Nagel, M. Schreckenberg *A cellular automaton model for free-way traffic*
- [11] N. Aronszajn, *Theory of reproducing kernels*
- [12] A. Kuhn, M. Lavielle *Coupling a stochastic approximation version of EM algorithm with a MCMC procedure*