



Régression ridge a noyau dans un graphe : Application au réseau de distribution d'eau potable

Christophe Dumora, Jérémie Bigot, Vincent Couallier, David Auber, Cyril Leclerc

► To cite this version:

Christophe Dumora, Jérémie Bigot, Vincent Couallier, David Auber, Cyril Leclerc. Régression ridge a noyau dans un graphe : Application au réseau de distribution d'eau potable. 2017. hal-01515313

HAL Id: hal-01515313

<https://hal.science/hal-01515313>

Preprint submitted on 9 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

REGRESSION RIDGE À NOYAU DANS UN GRAPHE : APPLICATION AU RÉSEAU DE DISTRIBUTION D’EAU POTABLE

Christophe Dumora ^{1,2,3} & Jérémie Bigot ¹ & Vincent Couallier ¹ & David Auber ² & Cyril Leclerc ³

¹ *Institut de Mathématiques de Bordeaux, Université de Bordeaux, Talence, France*

² *Laboratoire Bordelais de Recherche en Informatique, Université de Bordeaux, Talence, France*

³ *LyRE, Centre R&D SUEZ, Talence, France*

Résumé. Ce travail aborde le problème de l’inférence sur les noeuds d’un très grand graphe, représentant un réseau de distribution d’eau potable, à partir d’une observation partielle de quelques données, possiblement chronologiques, sur un faible nombre de noeuds. Nous utilisons une approche de prédiction par noyau reposant sur un estimateur pénalisé de type ridge qui soulève des problèmes d’analyse spectrale d’une très grande matrice creuse.

Mots-clés. Analyse de graphe, regression ridge à noyau, réseau de distribution d’eau potable

Abstract. The aim of this paper is to present the methodology of statistical inference and prediction for processes defined on network graphs when those processes are only observed on a small number of nodes. The kernel regression approach presented here already raise the question of eigenvectors computation of a very large adjacency matrix.

Keywords. Graph network analysis, Kernel Ridge Regression, water distribution network

1 Introduction

Ce travail se place dans le cadre de l’application de méthodes statistiques dans un réseau de graphe. Nous présentons un estimateur pénalisé basé sur une régression ridge à noyau dans un graphe, appliqué au cas concret d’un réseau de distribution d’eau potable. Nous cherchons ici à estimer des paramètres clés de l’exploitation du réseau de distribution d’eau potable de la métropole de Bordeaux.

La fonction principale d’un réseau de distribution d’eau potable est d’assurer la fourniture d’une eau saine et bonne à boire, depuis son extraction aux différentes sources jusqu’aux robinets des consommateurs. Cette tâche est basée sur une bonne gestion des paramètres hydrauliques du réseau (tels que le débit ou la pression de service), sur la

maîtrise de la qualité de l'eau (par le biais de traitement et purification telle que la chloration), mais aussi sur le contrôle de la demande en eau. Pour cela le réseau de distribution est équipé de capteurs qui mesurent au niveau des usines de production et en certains points stratégiques du réseau de distribution ces paramètres clés sur un pas de temps fin ($\approx 6\text{min}$). Un des enjeux principaux pour la maîtrise et la gestion du réseau de distribution est d'être en mesure d'estimer ces paramètres en tout point du réseau composé de plus 60 000 canalisations, représentant plus de 3 000 km de linéaire.

Puisque le réseau de distribution est un système complexe représenté sous la forme d'un réseau de graphe, notre objectif est de reconstruire (prédire) un vecteur de caractéristiques numériques sur l'ensemble des noeuds à partir d'une observation partielle, possiblement chronologique, sur un faible nombre de noeuds.

Nous présentons dans un premier temps les notations ainsi que les méthodes, tirées en grande partie de Kolaczyk (2009), nous permettant de définir un noyau à partir du graphe représentant le réseau de distribution d'eau potable. Dans un second temps nous présenterons les résultats obtenus pour l'estimation des débits dans un sous-graphe de plus petite dimension.

2 Notations

Soit $G = (V, E)$ le graphe non dirigé représentant le réseau de distribution d'eau potable de la métropole bordelaise. Les arrêtes E (resp. noeuds V) de G symbolisent les canalisations (resp. les objets sur lesquels les canalisations sont connectées, tels que des raccords, des vannes, des appareils de mesures etc.). G est composé de $N_E = 66\,482$ arrêtes et $N_v = 60\,097$ noeuds. Ainsi deux noeuds i et j sont connectés si une canalisation les relie. On rappelle que la matrice d'adjacence A de G est la matrice (symétrique) de taille $N_v \times N_v$ définie par

$$A_{ij} = \begin{cases} 1, & \text{si } \{i, j\} \in E \\ 0, & \text{sinon} \end{cases}$$

On définit également la matrice des degrés du graphe G $d_i = \sum_{j \neq i} A_{ij}$ comme $\mathbf{D} = \text{diag}[(d_i)_{i \in V}]$ et enfin le Laplacien de G , $\mathbf{L} = \mathbf{D} - \mathbf{A}$. Les propriétés de $\mathbf{L}$ sont bien connues, la plus petite valeur propre λ_1 est nulle, associée au vecteur propre $\mathbf{1}$, ce qui signifie que le graphe est connecté (voir Kimeldorf et Wahba (1971), Wu et Preciado (2013) pour plus de détails).

Soit $X = (X_1, \dots, X_{N_v})$ les attributs des noeuds du graphe G . On observe alors $X_i = x_i$ pour $i \in V^{obs} \subseteq V$. On note alors $n = |V^{obs}|$ le nombre de noeuds sur lesquels le processus X est observé (par exemple les noeuds représentant les appareils de mesures). Pour simplifier l'exposé, on suppose dans un premier temps que $X \in \mathbb{R}$. Notre objectif est de construire un prédicteur $\hat{h}$ de V dans $\mathbb{R}$ permettant de prédire les valeurs de X sur les noeuds non observés tout en s'ajustant au mieux aux données observées.

2.1 Un predicteur basé sur une régression pénalisée

La fonction $\mathbf{h}(\cdot)$ sur V peut être représentée comme un vecteur $\mathbf{h} \in \mathbb{R}$ avec le i -ème élément de $\mathbf{h}$ associé au noeud $i \in V$. Considérant les observations $\mathbf{x}^{obs}$, le graphe G et un noyau $\mathbf{K}$ ¹, la régression à noyau cherche un choix optimal de $\mathbf{h}$ dans la classe

$$\mathcal{H}_K = \{\mathbf{h} \in \mathbb{R}^{N_v} : \mathbf{h} = \Phi\beta \text{ et } \beta^T \Delta^{-1} \beta \leq \infty\} \quad (1)$$

Où Φ et $\Delta = \text{diag}[(\delta_j)]$ sont respectivement les matrices $N_v \times N_v$ orthogonales et diagonales, obtenues en effectuant la décomposition en éléments propres $\mathbf{K} = \Phi\Delta\Phi^T$. Ici K est une matrice de taille $N_v \times N_v$ obtenue en évaluant le noyau $\mathbf{K}$ sur tout le graphe G . Dans le but d'obtenir un choix optimal pour $\mathbf{h}$ dans $\mathcal{H}_K$ et dans l'optique d'une stratégie de régression pénalisée permettant de contraindre un choix de $\hat{\mathbf{h}}$ le plus proche des données observées et le plus lisse possible sur le graphe, on choisit une estimation de la forme $\hat{\mathbf{h}} = \Phi\hat{\beta}$ en choisissant $\hat{\beta}$ qui minimise

$$\sum_{i \in V^{obs}} (x_i - (\phi\beta)_i)^2 + \lambda \beta^T \Delta^{-1} \beta \quad (2)$$

avec λ un paramètre à choisir dictant l'importance relative entre l'attache aux données et la pénalisation. Si on s'intéresse au terme de pénalisation dans (2) on remarque que

$$\begin{aligned} \beta^T \Delta^{-1} \beta &= \beta^T \Phi^T \Phi \Delta^{-1} \Phi^T \Phi \beta \\ &= \beta^T \Phi^T K^{-1} \Phi \beta \\ &= h^T K^{-1} h \end{aligned}$$

Si on veut choisir $\mathbf{K}$ tel que $\mathbf{h} = \Phi\beta$ soit localement lisse en fonction de la topologie de G , Belkin, Matveeva et Niyogi (2004) montrent que la pseudo-inverse de la matrice Laplacienne de G , $\mathbf{K} = \mathbf{L}^{-}$ est un bon candidat. En effet $h^T K^{-1} h = h^T L h = \sum_{i,j \in E} (h_i - h_j)^2$. Ainsi, choisir la base de fonctions $(\Phi_k)_{k=1 \dots N_v}$ et la pénalisation $\beta^T \Delta^{-1} \beta$ permet une régression adaptée à la structure du graphe. On cherche à lisser les valeurs de $\mathbf{h}$ assignées aux noeuds i et j adjacents dans G . Néanmoins choisir l'inverse de la matrice Laplacienne du graphe comme noyau soulève le problème du calcul des vecteurs propres d'une grande matrice (dans notre cas $60\,000 \times 60\,000$).

En outre, il faut remarquer que, puisque $\mathbf{1}$ est vecteur propre de K associé à la valeur propre nulle, toutes les combinaisons $h = \phi\beta$, vérifient $h^T \mathbf{1} = 0$. Les prédictions forment alors des vecteurs nécessairement centrés. On choisit donc de construire la régression des données x^{obs} en centrant au préalable les observations.

¹ $\mathbf{K}$ est une matrice symétrique semi-définie positive telle que toute sous-matrice K_n sélectionnant n lignes et n colonnes de $\mathbf{K}$ est également semi-définie positive, voir théorème de Mercer (1909)

2.2 Un argument RKHS pour expliciter le prédicteur

Du fait que $\mathcal{H}_K$ soit un espace de Hilbert à noyau reproduisant Kimeldorf et Wahba (1971) ont montré que la solution $\hat{\mathbf{h}}$ pouvait s'écrire sous la forme $\mathbf{h} = \mathbf{K}^{(\mathbf{N}_v, \mathbf{n})} \alpha$, avec $\mathbf{K}^{(\mathbf{N}_v, \mathbf{n})}$ la matrice $N_v \times n$ représentant le noyau $\mathbf{K}$ évalué sur $V \times V^{obs}$. Ainsi l'optimisation (2) s'obtient en choisissant α qui minimise

$$\sum_{i \in V^{obs}} (x_i - (K^{(n)} \alpha)_i)^2 + \lambda \alpha^T K^{(n)} \alpha \quad (3)$$

Où $\mathbf{K}^{(n)}$ est la matrice $n \times n$ résultant de l'évaluation de $\mathbf{K}$ sur $V^{obs} \times V^{obs}$. L'optimisation (3) permettant de définir $\hat{\alpha}$ prend la forme

$$\min_{\alpha} [(x^{obs} - K^{(n)} \alpha)^T (x^{obs} - K^{(n)} \alpha) + \lambda \alpha^T K^{(n)} \alpha] \quad (4)$$

Notons $\Phi^{(n)}$ la matrice des vecteurs propres et $\Delta^{(n)} = \text{diag}[(\delta_i^{(n)})]$ la matrice des valeurs propres obtenues en effectuant la décomposition en éléments propres $\mathbf{K}^{(n)} = \Phi^{(n)} \Delta^{(n)} \Phi^{(n)T}$ et $\beta^{(n)} = \Delta^{(n)} \Phi^{(n)T} \alpha$ ainsi que $\mathbf{M} = \Phi^{(n)} \Delta^{(n)1/2}$.

D'après Kolackzyk (2009) et Schölkopf et Smola (2002), la solution de (4) donne $\hat{\mathbf{h}} = \mathbf{K}^{(N_v, n)} \hat{\alpha}$ avec $\hat{\alpha} = \Phi^{(n)} \Delta^{(n)-1/2} \hat{\theta}$ et $\hat{\theta} = (\mathbf{M}^T \mathbf{M} + \lambda \mathbb{I})^{-1} \mathbf{M}^T x^{obs}$.

On peut en fait réécrire $\hat{\mathbf{h}}$ en fonction de x^{obs} : $\hat{\mathbf{h}} = \mathbf{H}(\lambda) x^{obs}$, la matrice $\mathbf{H}(\lambda)$ s'appliquant aux observations afin d'obtenir les estimations $\hat{x}$ sur G . Un calcul simple donne alors

$$\hat{\theta} = (\text{diag}[\frac{1}{\delta_i^{(n)} + \lambda}]) \mathbf{M}^T x^{obs} \quad (5)$$

et

$$\begin{aligned} \hat{\mathbf{h}} &= \mathbf{K}^{(N_v, n)} \Phi^{(n)} (\text{diag}[\frac{1}{\delta_i^{(n)} + \lambda}]) \Phi^{(n)T} x^{obs} \\ &= \mathbf{H}(\lambda) x^{obs} \end{aligned} \quad (6)$$

La ligne i de $\mathbf{H}(\lambda)$ contient les coefficients appliqués aux observations x^{obs} pour déterminer l'estimation $\hat{\mathbf{h}}$ du noeud i .

3 Cas d'étude - Estimation du débit sur un sous-graphe

Nous présentons dans cette partie l'application de la regression ridge à noyau à l'estimation du débit sur un sous graphe à partir de l'observation de cette variable sur un nombre de noeuds restreint. Le graphe Figure 1 représente le réseau de distribution de la commune d'Ambès près de Bordeaux. Il est composé de $N_v = 802$ noeuds et $N_E = 868$ arrêtes. Cette partie du réseau est composée de $n = 7$ capteurs mesurant le débit toutes les 6 minutes.

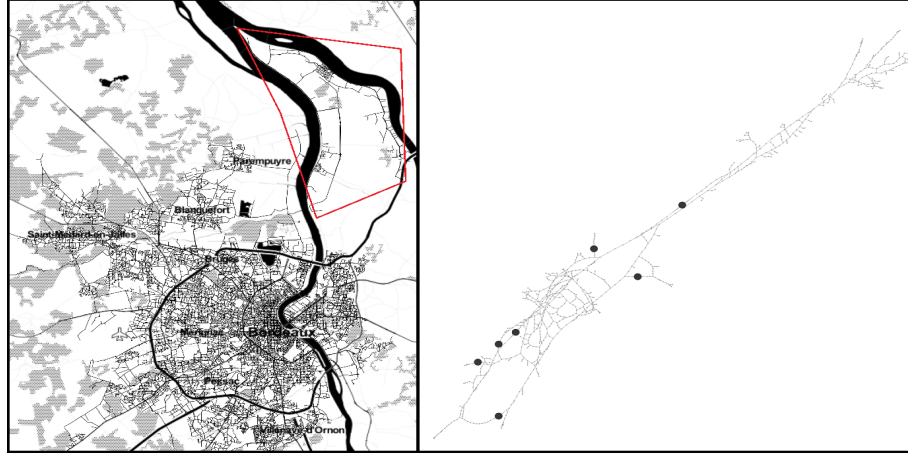


Figure 1: Réseau de distribution de la commune d'Ambès, à gauche encadré en rouge, à droite sous la forme d'un réseau de graphe (les points noirs représentant les points de mesures de débit)

Le graphe de droite sur la Figure 2 représente l'estimation $\hat{\mathbf{h}}$ obtenue à partir de l'observation du débit sur les 7 points de mesures, le paramètre λ est fixé à 0.11. Les noeuds sur lesquels les débits ont été observés sont représentés avec une forme carré. La taille des noeuds quant à elle, est proportionnelle à la valeur de débit estimée. On peut observer que de manière générale les estimations sont localement lisses. Le graphe de gauche sur la Figure 2 permet de représenter les coefficients $\mathbf{H}(\lambda)_i$ appliqués à chacune des observations pour estimer le débit en un noeud i ici encadré en rouge. On remarque que plus les noeuds sont éloignés du noeud considéré plus celui-ci est faible.

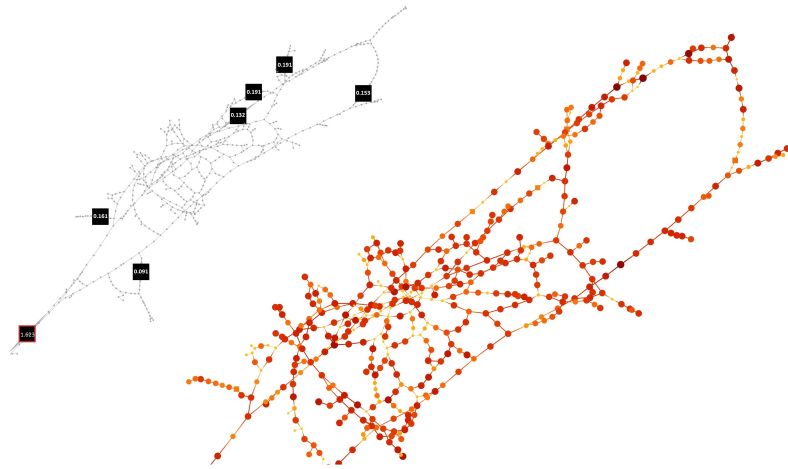


Figure 2: Représentations zoomées du graphe, à gauche avec $\mathbf{H}(\lambda)_i$ sur les noeuds observés et à droite avec les estimations $\hat{\mathbf{h}}$ sur l'ensemble des noeuds

4 Conclusion - Perspectives

La méthode présentée ici fonctionne bien sur des jeux de données “écoles” et sur des sous-graphes du jeu de données réelles. Le passage à l’échelle sur le jeu de données réelles est en cours, en effet la méthode proposée ici requiert le calcul d’une pseudo-inverse sur le graphe tout entier ($60\,000 \times 60\,000$) afin d’obtenir K_n , de taille $(n \times n)$, dont la diagonalisation n’est pas lourde. On pourra s’appuyer sur des méthodes computationnelles utilisées dans les dessins de graphes (Koren, 2004).

Une perspective afin d’améliorer ce travail, est d’effectuer cette régression lorsque que le processus X observé sur chaque noeud du graphe est multivarié. On pourra par la suite étendre cette méthode au cas de séries temporelles partiellement observées sur un graphe.

Enfin on pourra s’intéresser au choix du paramètre λ dont une des méthodes standard est la validation croisée généralisée (Amini et Roozbeh, 2015) qui peut être adaptée à notre cas de régression à noyau.

Bibliographie

- [1] Amini, M., & Roozbeh, M., (2015) *Optimal partial ridge estimation in restricted semi-parametric regression models*, J. Multivariate Anal., 136, pp. 26–40
- [2] Auber, D., (2004), *Tulip—A huge graph visualization framework*, Graph drawing software, Springer, Berlin, Heidelberg, pp. 105-126.
- [3] Belkin, M., & Matveeva, I., & Niyogi, P., *Regularization and semi-supervised learning on large graphs*, dans Shawe-Taylor, J., & Singer, Y., (dir.), *Learning Theory. (COLT)*, (2004). Lecture Notes in Computer Science, vol 3120. Springer, Berlin, Heidelberg, pp. 624-638.
- [4] Fei, H., & Huan, J., (2009), *L2 norm regularized feature kernel regression for graph data*, dans CIKM, pp. 593-600
- [5] Kimeldorf, G., & Wahba, G., (1971), *Some results on Tchebycheffian spline function*, Journal of Mathematical Analysis and Applications, vol33, pp.82-95
- [6] Kolaczyk, Eric D., (2009), *Statistical analysis of network data: methods and models*, Springer, New York; London.
- [7] Koren, Y., (2004), *Drawing graphs by eigenvectors: theory and practice**, Computers & Mathematics with Applications, Vol49
- [8] Mercer, J., (1909), *Functions of Positive and Negative Type, and their Connection with the theory of integral equations*, Philosophical Transactions of the Royal Society of London, Series A 209, pp. 415-446
- [9] Schölkopf, B., & Smola, A.J., (2002), *Learning with kernels*, the MIT Press
- [10] Wu, Z., & Preciado, Victor M., (2013), *Laplacian spectral properties of graph from random local samples*, CoRR