C. R. Acad. Sci. Paris, Ser. I 349 (2011) 469–472

Contents lists available at ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistique

Conditions minimales de consistance pour la sélection de variables en grande dimension

Minimal conditions for consistent variable selection in high dimension

Laetitia Comminges

LIGM/IMAGINE, 6, avenue Blaise-Pascal, cité Descartes, Champs-sur-Marne, 77455 Marne-la-Vallée cedex, France

I N F O A R T I C L E

Historique de l'article :
Reçu le 10 janvier 2011
Accepté après révision le 14 février 2011
Disponible sur Internet le 24 février 2011

Présenté par Paul Deheuvels

R É S U M É

On s'intéresse à la sélection de variables ou, plutôt, à l'estimation de l'ensemble des variables pertinentes dans le modèle de bruit blanc gaussien. On suppose que la dimension du dispositif expérimental d est très grande mais que la fonction de régression f dépend d'un nombre d^* bien plus petit de variables. On présente des conditions suffisantes portant sur la relation entre d, d^* et l'intensité du bruit qui permettent d'estimer l'ensemble des variables pertinentes de façon consistante. Ces conditions sont prouvées d'être minimales (à une constante multiplicative près) dans le cadre où d^* n'augmente pas lorsque le niveau de bruit diminue, et presque minimales lorsqu'on autorise d^* à grandir quand le bruit diminue.

A B S T R A C T

We are interested in the variable selection task in the Gaussian white noise model. We suppose the dimension of the input variable is very large but the regression function depends on a much smaller number d^* of coordinates. We propose two methods based on thresholding that select the correct subset of variables with high probability, and get minimal conditions od consistency when d^* is a constant, and nearly minimal conditions when d^* is large.

1. Introduction

On considère le modèle d'observations idéalisé du bruit blanc gaussien : la donnée observée $(X(\boldsymbol{t}), \boldsymbol{t} \in [0,1]^d)$ est une réalisation du processus X défini par la différentielle stochastique

$$\mathrm{d}X(\boldsymbol{t}) = f(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} + n^{-1/2}\,\mathrm{d}W(\boldsymbol{t}), \quad \boldsymbol{t} \in [0,1]^d$$

où f est la fonction de régression inconnue, $W = (W(\boldsymbol{t}), \boldsymbol{t} \in [0,1]^d)$ est le processus de Wiener standard en dimension d et n un entier. On suppose que f dépend en fait d'un nombre d^* de ses coordonnées, avec d^* bien plus petit que d. En d'autres termes, il existe un sous ensemble $J \subset [\![1, d]\!]$ de cardinal d^* et une fonction $g : [0,1]^{d^*} \to \mathbb{R}$ telle que $f(\boldsymbol{t}) = g(\boldsymbol{t}_J)$, $\forall \boldsymbol{t} \in [0,1]^d$, où $\boldsymbol{t}_J$ représente le vecteur obtenu à partir de $\boldsymbol{t}$ en supprimant toutes les coordonnées dont l'indice n'est pas dans J. On s'intéresse à l'estimation de J dans le cadre non-paramétrique, en supposant uniquement que f est régulière

Adresse e-mail : commingl@imagine.enpc.fr.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2011.02.014

d'ordre deux. De plus, on ne suppose pas la connaissance exacte de d^* mais seulement celle d'un majorant s. La particularité de notre approche est que l'on permet à d, s et d^* de dépendre de n.

On suppose que f appartient à l'ensemble suivant, où $(\theta_{\boldsymbol{k}}[f])_{\boldsymbol{k}\in\mathbb{Z}^d}$ est la suite de ses coefficients de Fourier dans la base trigonométrique :

$$\Sigma(L) \triangleq \left\{ f\colon \max_{i=1,\dots,d} \sum_{\boldsymbol{k}\in\mathbb{Z}^d} k_i^4 \theta_{\boldsymbol{k}}[f]^2 \leqslant L \right\}.$$

De plus, afin que l'ensemble J soit identifiable, on suppose que f vérifie, avec $\kappa > 0$:

$$Q^j \triangleq \sum_{\boldsymbol{k}\colon k_j\neq 0} \theta_{\boldsymbol{k}}[f]^2 \geqslant \kappa, \quad \forall j \in J. \tag{1}$$

Le problème de sélection de variables en grande dimension que l'on considère ici a été étudié, dans le cadre du modèle de régression, dans les travaux [1,5] où des algorithmes de sélection basés sur des méthodes locales ont été proposés. Cependant nos résultats indiquent que les conditions de consistance de leurs méthodes sont sous-optimales. En effet, nos deux estimateurs sont consistants sous des conditions plus faibles concernant l'ordre de d ainsi que l'ordre de d^*. Tandis que dans ces deux articles, d^* ne peut pas être d'ordre supérieur à une constante, et d d'ordre supérieur à $\log n$, nous présentons deux méthodes qui permettent à d d'être d'un ordre beaucoup plus élevé—la seule restriction étant que $\log d$ est négligeable face à n—quand d^* est une constante. Nous fournissons également dans la dernière section des résultats du type « borne inférieure » (voir [7, Section 2.6]) qui prouvent que les restrictions imposées ne peuvent pas être significativement allégées.

2. Sélection de variables par seuillage

Soit $\{\varphi_{\boldsymbol{k}}; k \in \mathbb{Z}^d\}$ la base trigonométrique de $L^2([0,1]^d;\mathbb{R})$. On translate le modèle fonctionnel original en le modèle séquentiel suivant :

$$y_{\boldsymbol{k}} = \theta_{\boldsymbol{k}} + n^{-1/2}\,\xi_{\boldsymbol{k}}, \quad \boldsymbol{k}\in\mathbb{Z}^d, \tag{2}$$

où $y_{\boldsymbol{k}} = \int_{[0,1]^d} \varphi_{\boldsymbol{k}}(\boldsymbol{t})\,\mathrm{d}X(\boldsymbol{t})$, $\theta_{\boldsymbol{k}} = \int_{[0,1]^d} f(\boldsymbol{t})\varphi_{\boldsymbol{k}}(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}$ et les $\{\xi_{\boldsymbol{k}}; \boldsymbol{k}\in\mathbb{Z}^d\}$ forment une famille de variables aléatoires indépendantes gaussiennes centrées réduites. Pour tout vecteur $\boldsymbol{x}$ de coordonnées $x_1,\dots,x_k$ et pour tout $q>0$, on utilisera les notations $\|\boldsymbol{x}\|_q = (\sum_j |x_j|^q)^{1/q}$, $\|\boldsymbol{x}\|_0 = \mathrm{Card}\{j\colon x_j\neq 0\}$ et $\|\boldsymbol{x}\|_\infty = \max_j |x_j|$. Si l'on fixe deux entiers $\ell, i \in [\![1,d]\!]$ et un réel $m>0$, on note $S_{m,\ell} = \{\boldsymbol{k}\in\mathbb{Z}^d\colon \|\boldsymbol{k}\|_4 \leqslant m,\ \|\boldsymbol{k}\|_0 \leqslant \ell\}$ et $S^i_{m,\ell} = \{\boldsymbol{k}\in S_{m,\ell}\colon k_i\neq 0\}$. On remarque que l'on a $J = \{j\in[\![1,d]\!]\colon \max_{\boldsymbol{k}\in S^j_{\infty,s}} |\theta_{\boldsymbol{k}}| > 0\}$. C'est pourquoi on propose d'estimer J par l'ensemble suivant, pour des paramètres $\lambda>0$ et $m>0$ dont la valeur sera précisée plus tard :

$$\hat{J}_1(m,\lambda) = \left\{ j\in[\![1,d]\!]\colon \max_{\boldsymbol{k}\in S^j_{m,s}} |y_{\boldsymbol{k}}| > \lambda \right\}. \tag{3}$$

Proposition 2.1. *Pour un réel $B \geqslant 2L/\kappa$, on pose $m = (Bs)^{1/4}$, $\lambda^2 = \frac{16s\log(6md)}{n}$ et $N(d^*,m) = \mathrm{Card}(\{\boldsymbol{k}\in\mathbb{Z}^*\times\mathbb{Z}^{d^*-1}\colon \|\boldsymbol{k}\|_4\leqslant m\})$. Si*

$$\frac{sN(d^*,m)\log(6md)}{n} \leqslant \frac{\kappa}{32} \tag{4}$$

alors la probabilité de l'événement $\hat{J}_1(m,\lambda)\neq J$ est majorée par $\mathbf{P}(\max_{\boldsymbol{k}\in S_{m,s}} |\xi_{\boldsymbol{k}}| > n^{1/2}\lambda) \leqslant (6md)^{-s}$. Par conséquent, si (4) est satisfaite et $d = d_n$ tend vers $+\infty$ lorsque $n\to\infty$ alors $\hat{J}_1(m,\lambda)$ est un estimateur consistant de J.

Afin de présenter l'intérêt de ce résultat, considérons la situation où la dimension intrinsèque d^*, ainsi que son majorant s ne dépendent pas de n ou restent bornés lorsque $n\to\infty$. Une conséquence intéressante de la Proposition 2.1 est que si $d = d_n \to \infty$ lorsque $n\to\infty$ et $\log d_n = o(n)$ alors $\hat{J}_1$ est un estimateur consistant de J. Ce résultat n'est pas valide si l'on remplace $\hat{J}_1$ par les estimateurs de J proposés dans les travaux précédents sur ce sujet (voir [1,5]).

On propose maintenant une deuxième méthode de sélection qui utilise le seuillage par bloc et, lorsque d^* est une fonction croissante de n, parvient à estimer J de façon consistante sous des conditions légèrement plus faibles que celle de la Proposition 2.1. Pour tout $\ell\in[\![1,s_n]\!]$, toute partie $I\subset[\![1,d]\!]$ et tout $i\in[\![1,d]\!]$ on note $P^d_\ell = \{I\colon \mathrm{Card}(I)=\ell\}$ et $S^i_{m,I} = \{\boldsymbol{k}\in\mathbb{Z}^d\colon \|\boldsymbol{k}\|_4\leqslant m \text{ et } \{i\}\subset\mathrm{supp}(\boldsymbol{k})\subset I\}$. On vérifie aisément que, pour tout $j\in J$, $Q^j = \lim_{m\to\infty}\sum_{\boldsymbol{k}\in S^j_{m,J}}\theta^2_{\boldsymbol{k}} \geqslant \kappa$. Ainsi, $\forall j\in J$, $\forall\tau>1$ et pour m assez grand, il existe $I\in P^d_{d^*}$, tel que $Q^j_{m,I} \triangleq \sum_{\boldsymbol{k}\in S^j_{m,I}}\theta^2_{\boldsymbol{k}} \geqslant \kappa(1-\frac{1}{\tau})$. Cette propriété, combinée avec le fait que $\hat{Q}^j_{m,I} \triangleq \sum_{\boldsymbol{k}\in S^j_{m,I}}(y^2_{\boldsymbol{k}} - \frac{1}{n})$ est un estimateur sans biais de $Q^j_{m,I}$, nous mène à définir

$$\hat{J}_2\big(\{m_\ell,\lambda_\ell\}\big) = \left\{ j\in[\![1,d]\!]\colon \max_{I\in P^d_\ell} \hat{Q}^j_{m_\ell,I} > \frac{\lambda_\ell}{n}\ \forall \ell\leqslant s \right\}$$

où les paramètres (m_ℓ,λ_ℓ), $\ell=1,\dots,s$, seront définis dans la proposition suivante.

Proposition 2.2. *Soit $\tau > 1$ et soit B un réel satisfaisant $B \geqslant \tau L/\kappa$. On pose, pour tout $\ell \leqslant s$, $m_\ell = (B\ell)^{1/4}$ et $\lambda_\ell = 2(N(\ell, m_\ell) \times (s+1)\log d)^{1/2} + 2(s+1)\log d$. Si*

$$\frac{4(N(d^*, m_s)(s+1)\log d)^{1/2}}{n} + \frac{8(s+1)\log d}{n} < \kappa\left(1 - \frac{1}{\tau}\right) \tag{5}$$

alors $\mathbf{P}(\hat{J}_2(\{m_\ell, \lambda_\ell\}) \neq J) \leqslant 12/d$. Par conséquent, si (5) est satisfaite et $d = d_n$ tend vers $+\infty$ lorsque $n \to \infty$ alors $\hat{J}_2(\{m_\ell, \lambda_\ell\})$ est un estimateur consistant de J.

Remarque 1. Même si la complexité computationnelle des procédures de sélection n'est pas l'objet de cet article, il convient de noter qu'elle n'est pas toujours exponentielle en d ou d^*. Par exemple, pour déterminer $\hat{J}_1$, une stratégie raisonnable consiste à calculer étape par étape les coefficients $y_{\boldsymbol{k}}$. A chaque étape $\ell = 1, \ldots, s$, seuls les $y_{\boldsymbol{k}}$ avec $\|\boldsymbol{k}\|_0 = \ell$ sont calculés et comparés au seuil λ. En cas de dépassement de ce seuil par $y_{\boldsymbol{k}}$, toutes les variables correspondantes aux coordonnées non-nulles de $\boldsymbol{k}$ sont classées comme pertinentes. On arrête cette procédure itérative lorsque le nombre de variables déclarées pertinentes est $\geqslant s_n$. Cet algorithme a, dans les pires des cas, une complexité exponentielle, mais il y a un nombre important de situations où la complexité est polynomiale en d et en s. C'est le cas, par exemple, pour le modèle additif : $f(\boldsymbol{t}) = f_{i_1}(t_{i_1}) + \cdots + f_{i_{d^*}}(t_{i_{d^*}})$.

3. Borne inférieure

On s'intéresse d'abord au cas où d^* ne dépend pas de n. On observe alors que les deux estimateurs $\hat{J}_1$ et $\hat{J}_2$ sont consistants dès que $s_n \log d_n = o(n)$ quand $n \to \infty$ et, avec la connaissance préalable de d^*, on aurait obtenu le même résultat avec la condition $\log d_n = o(n)$ quand $n \to \infty$. La proposition suivante dit que si cette condition n'est pas vérifiée, alors, pour tout estimateur, il existe au moins une fonction f telle que la probabilité d'erreur ne tend pas vers 0.

Pour tout $(L, \kappa) \in \mathbb{R}_+^2$ et $d^* \in \mathbb{N}$, on définit $\Sigma = \Sigma(L, \kappa, d^*)$ comme l'ensemble des fonctions de $\Sigma(L)$ telles que le cardinal de J soit au plus d^* et satisfaisant la condition (1). De façon à ce que Σ ne soit pas vide, on suppose que $\kappa \leqslant L$.

Proposition 3.1. *Si pour un certain $\alpha \in (0, 1/8)$ l'inégalité $d^*(\log d_n - \log d^*) \geqslant \alpha^{-1} n\kappa$ est vérifiée pour tout $n \geqslant 1$, alors il existe une constante $c > 0$ telle que $\inf_{\bar{J}_n} \sup_{f \in \Sigma} \mathbf{P}_f(\bar{J}_n \neq J_f) \geqslant c$, $\forall n \geqslant 1$, où l'infimum est pris sur la collection de tous les estimateurs $\bar{J}_n$.*

Intéressons-nous maintenant au cas où $d^* = d_n^*$ tend vers ∞ avec $\log d_n^* = o(\log d_n)$. Pour éviter les complications techniques peu informatives, on se focalise sur le cas où $s = d^*$ est connu. Sous cette hypothèse, la condition assurant la validité de la Proposition 2.2 est plus faible que celle de la Proposition 2.1. La condition (5) de la Proposition 2.2 fait apparaître deux termes, $(d_n^* \log d_n)/n$ et $(N(d_n^*, m_{d_n^*}) d_n^* \log d_n)^{1/2}/n$, qui ne doivent pas exploser lorsque $n \to \infty$. On s'intéresse maintenant à la deuxième, qui n'est pas loin d'être optimale à une constante près. Pour tout γ réel positif, on pose $N_1(\gamma) = \mathrm{Card}\{\boldsymbol{k} \in \mathbb{Z}^{d^*} : \|\boldsymbol{k}\|_4^4 \leqslant \gamma d^*\}$ et $N_2(\gamma) = \mathrm{Card}\{\boldsymbol{k} \in \mathbb{Z}^{d^*} : k_2^4 + \cdots + k_{d^*}^4 \leqslant \gamma d^*\}$ en remarquant au passage que $N(d^*, m_{d^*}) = N_1(\gamma) - N_2(\gamma)$ avec $\gamma = L\tau/\kappa$. On définit les fonctions h et l_γ sur $(0,1)$ par $h(z) = \sum_{k \in \mathbb{Z}} z^{k^4}$ et $l_\gamma(z) = \log(h(z) z^{-\gamma})$. Ces fonctions sont très étroitement liées aux quantités $N_1(\gamma)$ et $N_2(\gamma)$ (voir [6]). Soit $z(\gamma)$ l'unique solution dans $(0,1)$ de l'équation $l'_\gamma(z) = 0$.

Proposition 3.2. *Supposons que $\frac{L}{\kappa} \geqslant 1 + \frac{1}{2z(1)}$, et soit γ^* le plus grand entier γ satisfaisant $\frac{L}{\kappa} \geqslant \gamma(1 + \frac{1}{2z(\gamma)})$. Si $9(N_1(\gamma^*) - N_2(\gamma^*))^2 \log\binom{d_n}{d^*} \geqslant \kappa^2 n^2 N_1(\gamma^*)$ alors il existe une constante positive $c \geqslant 0.0875$ telle que, si d_n^* est assez grand, $\inf_{\hat{J}} \sup_{f \in \Sigma} \mathbf{P}_f \times (\hat{J} \neq J) \geqslant c$.*

Dans la section précédente, on a prouvé la consistance de l'estimateur $\hat{J}_2$ sous la condition que $\frac{d_n^* \log d_n}{n^2}(N_1 - N_2)(\tau L/\kappa)$ reste borné par une constante. Le but de la Proposition 3.2 est de prouver que cette condition est presque minimale. En effet, on peut vérifier que le terme $\frac{(N_1(\gamma^*) - N_2(\gamma^*))^2 \log\binom{d_n}{d_n^*}}{N_1(\gamma^*)}$ est du même ordre que $(N_1 - N_2)(\gamma) d_n^* \log d_n$. Il s'ensuit, au vu de la Proposition 3.2, que l'estimation consistante de J est impossible si $\frac{d_n^* \log d_n}{n^2}(N_1 - N_2)(\gamma^*)$ est supérieur à une constante. Par conséquent, on retrouve presque la condition de la Proposition 2.2 à la différence près que l'argument de $N_1 - N_2$ est γ^* au lieu de $\tau L/\kappa$. Cette différence tend à disparaître lorsque $\frac{L}{\kappa}$ devient de plus en plus grand, car γ devient de plus en plus proche de $\frac{L}{\kappa}$. Quant à l'ordre de $N_1(\gamma) - N_2(\gamma)$, il est approximativement de la forme $f(\gamma)^{d^*}$.

Remarque 2. Les résultats énoncés dans les Propositions 2.1, 2.2 and 3.1 impliquent des bornes non asymptotiques du même type pour le risque, lorsque la qualité d'un estimateur $\hat{J}$ de cardinal $\leqslant s$ est mesurée par la distance de Hamming $\Delta(\hat{J}, J)$. En effet, il est clair que $\Delta(\hat{J}, J) \leqslant 2s\mathbf{1}_{\hat{J} \neq J}$, impliquant que $\mathbf{E}_f[\Delta(\hat{J}, J)] \leqslant 2s\mathbf{P}_f(\hat{J} \neq J) \leqslant 2s/(6md)^s$ sous les conditions de la Proposition 2.1. On a donc $\mathbf{E}_f[\Delta(\hat{J}, J)] \to 0$ lorsque $d \to \infty$, quelle que soit la valeur de s. De la même façon, l'inégalité $\mathbf{1}_{\hat{J} \neq J} \leqslant \Delta(\hat{J}, J)$ implique que sous les conditions de la Proposition 3.1, il vient $\mathbf{E}_f[\Delta(\hat{J}, J)] \geqslant c$.

4. Conclusion

Dans cette Note, nous avons proposé deux estimateurs de l'ensemble des variables pertinentes d'une fonction dans le modèle du bruit blanc gaussien. Nous avons exhibé des conditions sur la relation entre la dimension ambiente, la dimension intrinsèque et le niveau de bruit qui garantissent la consistance des estimateurs proposés. La (presque) minimalité de ces conditions a été également prouvée. Dans l'avenir, nous voudrions étendre ces résultats aux modèles de diffusions, en utilisant les techniques de [4], ce qui permettrait de réduire la dimension et d'utiliser des méthodes d'adaptations clasiques [2]. Il serait également intéressant de mener une étude analogue pour les modèles à plusieurs directions révélatrices [3].

Références

[1] K. Bertin, G. Lecué, Selection of variables and dimension reduction in high-dimensional non-parametric regression, Electron. J. Statist. 2 (2008) 1224–1241.
[2] A. Dalalyan, Sharp adaptive estimation of the drift function for ergodic diffusions, Ann. Statist. 33 (6) (2005) 2507–2528.
[3] A. Dalalyan, A. Juditsky, V. Spokoiny, A new algorithm for estimating the effective dimension-reduction subspace, J. Mach. Learn. Res. 9 (2008) 1648–1678.
[4] A. Dalalyan, M. Reiß, Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case, Probab. Theory Related Fields 137 (1–2) (2007) 25–47.
[5] J. Lafferty, L. Wasserman, Rodeo: sparse, greedy nonparametric regression, Ann. Statist. 36 (2008) 28–63.
[6] J.E. Mazo, A.M. Odlyzko, Lattice points in high-dimensional spheres, Monatsh. Math. 110 (1991) 47–61.
[7] A. Tsybakov, Introduction to Nonparametric Estimation, Springer, New York, 2009.