Apprentissage de modèles CHARME avec des réseaux de neurones profonds

José G. Gómez-García, Jalal Fadili, Christophe Chesneau

Apprentissage de modèles CHARME avec des réseaux de neurones profonds

José G. Gómez-García [1] · Jalal Fadili [2] · Christophe Chesneau [3]

[1] *Université Paris-Est Créteil, LAMA, UMR CNRS 8050.*
ESIPE, 71 rue Saint-Simon, 94000 Créteil.
jose-gregorio.gomez-garcia@u-pec.fr

[2] *Normandie Université, ENSICAEN, UNICAEN, GREYC, UMR CNRS 6072.*
ENSICAEN, 6 Bd du Maréchal Juin, 14050 Caen.
Jalal.Fadili@ensicaen.fr

[3] *Normandie Université, UNICAEN, LMNO, UMR CNRS 6139.*
LMNO, Sciences 3, Campus 2, Bd du Maréchal Juin, 14000 Caen.
christophe.chesneau@unicaen.fr

Résumé. Dans cette note, nous considérons un modèle appelé CHARME (Conditional Heteroscedastic Autoregressive Mixture of Experts). En quelques mots, c'est un modèle de mélange généralisé de séries chronologiques non linéaire et non paramétrique AR-ARCH. Nous garantissons la stabilité (ergodicité et stationnarité) du modèle sous certaines conditions de type Lipschitz pour les fonctions d'autorégression et de volatilité, lesquelles sont beaucoup plus faibles que celles présentées dans la littérature existante. Ce résultat et la propriété d'approximation universelle de réseaux de neurones (RN), possiblement avec des architectures profondes (RNP), nous fournit les bases pour développer une théorie d'apprentissage pour les fonctions d'autorégression-basées-sur-RN du modèle. En outre, la consistance forte et la normalité asymptotique de l'estimateur des poids et des biais des RN considéré sont garanties sous de faibles conditions.

Mots-clés. modèle AR-ARCH non-paramétrique ; réseaux de neurones profonds ; modèles de mélange ; séquence à changement de régime markoviens ; dépendance τ-faible ; ergodicité ; stationnarité ; identifiabilité ; consistance ; signaux d'EEG.

Abstract. In this note, we consider a model called CHARME (Conditional Heteroscedastic Autoregressive Mixture of Experts). Roughly speaking, this is a class of generalized mixture of nonlinear nonparametric AR-ARCH time series. We guarantee the stability (ergodicity and stationarity) of the model under certain Lipschitz-type conditions on the autoregression and volatility functions, which are much weaker than those presented in the current literature. This result and the universal approximation property of neural networks (NN), possibly with deep architectures (DNN), provides us with the bases for developing a learning theory for the NN-based autoregressive functions of the model.By the way, the strong consistency and asymptotic normality of the considered estimator of the NN weights and biases are guaranteed under weak conditions.

Keywords. Nonparametric AR-ARCH ; deep neural network ; mixture models ; Markov switching ; τ-weak dependence ; ergodicity, stationarity ; consistency ; EEG signals.

1 Introduction

Dans l'analyse de séries chronologiques, il est commun d'étudier les modèles tels que : AR, ARMA, ARCH, GARCH, etc. ; ou plus généralement, le modèle CHARN

$$X_t = f(X_{t-1}, \ldots, X_{t-p}, \theta^0) + g(X_{t-1}, \ldots, X_{t-p}, \lambda^0)\epsilon_t, \qquad t \in \mathbb{Z}, \tag{1}$$

où f, g sont des fonctions inconnues et $(\epsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc indépendant. Cependant, dans la pratique, il n'est pas toujours réaliste de supposer que le processus observé ait la même tendance f et la même volatilité g à chaque instant t. Entre autre, c'est le cas des signaux d'EEG, voir Lo *et al.* (2009), où l'on peut observer des changements de comportement, même brusques, lesquelles on ne peut pas les modéliser même en utilisant les modèles localement stationnaires. C'est pour cela que nous nous concentrons sur un modèle plus général, appelé CHARME, qui prend en compte ces changement brusques de comportement.

Pour définir ce modèle, considérons l'espace de Banach $(E, \|\cdot\|)$, doté de sa tribu borélienne $\mathcal{E}$. L'espace produit E^p est alors naturellement doté de sa tribu produit $\mathcal{E}^{\otimes p}$. Le modèle **CHARME(p)**, à valeurs dans E, est la série chronologique définie par

$$X_t = \sum_{k=1}^{K} \xi_t^{(k)} \left(f_k(X_{t-1}, \ldots, X_{t-p}, \theta_k^0) + g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k^0)\epsilon_t \right) \quad t \in \mathbb{Z}, \tag{2}$$

où

- pour chaque $k \in [K] := \{1, 2, \ldots, K\}$, $f_k : E^p \times \Theta_k \longrightarrow E$ et $g_k : E^p \times \Lambda_k \longrightarrow \mathbb{R}$ sont respectivement les fonctions d'autorégression et de volatilité, avec des espaces de paramètres respectifs Θ_k et Λ_k, $\mathcal{E}^{\otimes p} \times \mathcal{B}(\Theta_k)$- et $\mathcal{E}^{\otimes p} \times \mathcal{B}(\Lambda_k)$- mesurables, où $\mathcal{B}(\Theta_k)$ est la tribu borélienne sur Θ_k et pareillement pour Λ_k ;
- $(\epsilon_t)_t$ est un bruit blanc indépendant à valeurs dans E ;
- $\xi_t^{(k)} = \mathbb{I}_{\{R_t=k\}}$, avec $\mathbb{I}_{\mathcal{C}}$ désignant la fonction caractéristique de $\mathcal{C}$ (*i.e.*, elle vaut 1 sur $\mathcal{C}$ et 0 sinon), où $(R_t)_{t\in\mathbb{Z}}$ est une séquence de variables aléatoires indépendantes à valeurs dans l'espace fini $[K]$, qui est en plus indépendante du bruit blanc $(\epsilon_t)_{t\in\mathbb{Z}}$. Par la suite, on pose $\pi_k = \mathbb{P}(R_0 = k)$.

Le modèle (2) peut être étendu au cas $p = \infty$. Nous l'appellerons alors modèle CHARME à mémoire infinie et que nous désignerons par CHARME(∞). Dans ce cadre, l'espace d'états du modèle est le sous-ensemble de $E^{\mathbb{N}}$:

$$E^\infty := \left\{ (x_k)_{k>0} \in E^{\mathbb{N}} : \ x_k = 0 \text{ for } k > N, \text{ for some } N \in \mathbb{N}^* \right\},$$

doté de sa tribu produit $\mathcal{E}^{\otimes\mathbb{N}}$.

Il est clair que le modèle (2) contient le modèle (1) (cela correspond au cas $K = 1$ en (2)). D'ailleurs, des applications de ce modèle ont été traitées d'une manière directe ou indirecte dans plusieurs domaines de recherche. Voir par exemple : Tadjuidje-Kamgaing, J. (2005), Weigend, A.S. and Shi, S. (2000), Kirch, C. and Kamgaing, T. (2012) et Liehr *et al.* (1999).

2 Ergodicité et stationnarité des modèles CHARME

Le résultat suivant nous fournit des conditions pour avoir la stabilité du modèle dont la preuve est donnée dans la Section 8.1 de Gómez-García *et al.* (2020).

Théorème 1. *Considérons le modèle CHARME(∞), i.e.,* (2) *avec* $p = \infty$. *Supposons qu'il existe des séquences non-négatives* $(a_i^{(k)})_{i\geq 1,k\in[K]}$ *et* $(b_i^{(k)})_{i\geq 1,k\in[K]}$ *telles que, pour tout* $x, y \in E^\infty$ *et tout* $k \in [K]$,

$$\|f_k(x,\theta_k^0) - f_k(y,\theta_k^0)\| \leq \sum_{i=1}^{\infty} a_i^{(k)}\|x_i - y_i\|, \quad |g_k(x,\theta_k^0) - g_k(y,\theta_k^0)| \leq \sum_{i=1}^{\infty} b_i^{(k)}\|x_i - y_i\| \quad (3)$$

Notons $A_k = \sum_{i=1}^{\infty} a_i^{(k)}$, $B_k = \sum_{i=1}^{\infty} b_i^{(k)}$ *et* $C(m) = 2^{m-1}\sum_{k=1}^{K}\pi_k\left(A_k^m + B_k^m\|\epsilon_0\|_m^m\right)$. *Alors, nous obtenons les affirmations suivantes :*

(i) si $c := C(1) < 1$, *alors il existe une solution strictement stationnaire* $(X_t)_{t\in\mathbb{Z}}$ *du modèle CHARME(∞) appartenant à* $\mathbb{L}^1$.

(ii) si en plus $C(m) < 1$ *pour certain* $m > 1$, *alors cette solution appartient à* $\mathbb{L}^m$.

Remarque 1.

(1.1) Le résultat précédent est également valable dans le cas $p < \infty$. En effet, il suffit de prendre $a_i^{(k)} = b_i^{(k)} = 0$ pour tout $i > p$ et tout $k \in [K]$ dans les inégalités (3).

(1.2) Remarquons que le modèle CHARME(∞) (2) avec $p = \infty$ peut être réécrit comme une séquence de Markov $X_t = F(X_{t-1}, X_{t-2}, \ldots; \tilde{\xi}_t)$, $t \in \mathbb{Z}$, via la fonction

$$F(x; (\xi^{(0)}, \ldots, \xi^{(K)})) = \sum_{k=1}^{K} \xi^{(k)}\left(f_k(x,\theta_k^0) + g_k(x,\lambda_k^0)\xi^{(0)}\right), \quad (4)$$

avec des innovations $\tilde{\xi}_t := (\epsilon_t, \xi_t^{(1)}, \ldots, \xi_t^{(K)}) = (\epsilon_t, \xi_t) \in E\times B_e$, où $B_e := \{e_1, \ldots, e_K\}$ est la base canonique de $\mathbb{R}^K$. Sous les hypothèses du Théorème 1, la fonction F est continue car les fonctions $f_k(\cdot,\theta_k^0)$ et $g_k(\cdot,\lambda_k^0)$ sont continues par la condition (3). Il découle alors de (Doukhan, P. and Wintenberger, O, 2008, Lemma 5.5) et de la complétude de $\mathbb{L}^m$, qu'il existe une fonction mesurable H telle que le processus CHARME(∞) peut être écrit comme $X_t = H(\tilde{\xi}_t, \tilde{\xi}_{t-1}, \ldots)$. C'est-à-dire : le processus CHARME(∞) peut être représenté par un décalage de Bernoulli causal. En outre, sous ces hypothèses, $(X_t)_{t\in\mathbb{Z}}$ est le seul décalage de Bernouli causal, solution à (2) avec $p = \infty$. Donc, la solution $(X_t)_{t\in\mathbb{Z}}$ est automatiquement un processus ergodique. Enfin, le théorème ergodique implique la LFGN pour ce processus. Cette conséquence du Théorème 1 sera un résultat clé pour établir la consistance forte lorsqu'il s'agit d'estimer les fonctions d'autorégression et de volatilité du modèle CHARME(p).

(1.3) Stockis *et al.* (2010) montre l'ergodicité du modèle CHARME(p) avec $p < \infty$, sous réserve de multiple conditions. En particulier, les auteurs demandent la régularité du bruit blanc $(\epsilon_t)_{t\in\mathbb{Z}}$. En revanche, nous n'avons pas besoin de cette restriction ici.

3 Estimation des paramètres du modèle : consistance

Soit $(X_t)_{1-p\leq t\leq n}$ $n+p$ observations de la solution strictement stationnaire $(X_t)_{t\in\mathbb{Z}}$ du modèle (2) (cette solution existe grâce au Théorème 1). Supposons que le nombre d'états K est connu et que nous avons accès aux observations des variables cachées iid $(R_t)_{1-p\leq t\leq n}$, ou bien, des variables $(\xi_t^{(k)})_{1-p\leq t\leq n, k\in[K]}$. Une hypothèse similaire peut être trouvée dans la littérature pour des cas spécieux du modèle CHARME. Voir, *e.g.*, Tadjuidje-Kamgaing, J. (2005) et Stockis *et al.* (2010).

Notre objectif est d'étudier un estimateur non linéaire des paramètres

$$(\theta^0, \lambda^0) := (\theta_1^0, \ldots, \theta_K^0, \lambda_1^0, \ldots, \lambda_K^0)$$

du modèle CHARME(p) (2) à partir des observations $(X_t)_{1-p\leq t\leq n}$ et $(\xi_t^{(k)})_{1-p\leq t\leq n, k\in[K]}$. Cet objectif est atteint en résolvant le problème de minimisation

$$\begin{gathered}(\widehat{\theta}_n, \widehat{\lambda}_n) \in \operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda} Q_n(\theta,\lambda), \text{ où} \\ Q_n(\theta,\lambda) := \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{K} \xi_t^{(k)} \ell\big(X_t, f_k(X_{t-1},\ldots,X_{t-p},\theta_k), g_k(X_{t-1},\ldots,X_{t-p},\lambda_k)\big).\end{gathered} \tag{5}$$

Ici, $\ell : E\times E\times\mathbb{R} \longrightarrow \mathbb{R}\cup\{+\infty\}$ est une certaine fonction de coût. En général, ℓ devrait satisfaire $\ell(u,u,\tau)=0$, $\forall\tau$.

Afin de présenter notre résultat de consistance, il est plus commode de définir les processus

$$Y_t = (X_{t-p}, X_{t-p+1}, \ldots, X_t) \quad \text{et} \quad \xi_t = (\xi_t^1, \ldots, \xi_t^K), \quad t\in\mathbb{Z}.$$

Soit $(E^{p+1}\times B_e, \mathcal{E}^{\otimes(p+1)}\otimes\Xi, P)$ l'espace de probabilité commun, dans lequel sont définis les vecteurs aléatoires Y_t et ξ_t. Adoptons la notation suivante :

$$h(Y_t,\xi_t,\theta,\lambda) := \sum_{k=1}^{K} \xi_t^{(k)} \ell\big(X_t, f_k(X_{t-1},\ldots,X_{t-p},\theta_k), g_k(X_{t-1},\ldots,X_{t-p},\lambda_k)\big). \tag{6}$$

En utilisant des arguments complexes du calcul des variations (en particulier sur les intégrands normaux et l'épi-convergence (voir Rockafellar, R.T. (1976) et Rockafellar, R.T. and Wets, R.J.B. (1998)), nous pouvons établir la consistance de l'estimateur (5) sous de faibles conditions. En particulier, sans la nécessité d'avoir un échantillon iid ni la différentiabilité de la fonction Q_n. Ceci est résumé dans le théorème suivant :

Théorème 2. *Soit $(X_t)_{t\in\mathbb{Z}}$ une solution strictement stationnaire et ergodique du modèle (2) (elle existe grâce au Théorème 1 avec $C(m)<1$ pour certain $m\geq 1$). Considérons les conditions raisonnables (**A.1**)-(**A.7**) de Gómez-García et al. (2020). Alors,*

(i) chaque point d'accumulation de $(\widehat{\theta}_n, \widehat{\lambda}_n)_{n\in\mathbb{N}}$ appartient à $\operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda} \mathbb{E}h(Y,\xi,\theta,\lambda)$ p.s.

(ii) si de plus la suite $(Q_n)_{n\in\mathbb{N}}$ est équi-coercitive, et $\operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda} \mathbb{E}h(Y,\xi,\theta,\lambda) = \{\theta^0,\lambda^0\}$, alors $(\widehat{\theta}_n, \widehat{\lambda}_n)\to(\theta^0,\lambda^0)$ et $Q_n(\widehat{\theta}_n, \widehat{\lambda}_n)\to \mathbb{E}h(Y,\xi,\theta^0,\lambda^0)$ p.s.

4 Apprentissage du modèle avec des RNP

Il est connu que, étant donné toute fonction cible continue f et une précision cible $\epsilon > 0$, les réseaux de neurones (RN) avec suffisamment paramètres (poids et biaises) judicieusement choisis donnent une approximation de la fonction pour une erreur de taille ϵ. Cette propriété d'approximation universelle des RN, nous permet de considérer le modèle CHARME(p) (2) avec les fonctions f_k et g_k exactement modélisées par des RN, avec $E = \mathbb{R}^d$. Pour une introduction de réseaux de neurones (profonds), voir Section 2.2 de Gómez-García *et al.* (2020). Avec les mêmes notations de cette dernière section, pour chaque $k \in [K]$, soit $\theta_k = \Big((W_k^{(1)}, b_k^{(1)}), \ldots, (W_k^{(L_k)}, b_k^{(L_k)})\Big)$, où $W_k^{(l)}$ et $b_k^{(l)}$ sont respectivement la matrice des poids et le vecteur de biaises de la l-ème couche du RN f_k. Similairement $\lambda_k = \Big((\bar{W}_k^{(1)}, \bar{b}_k^{(1)}), \ldots, (\bar{W}_k^{(\bar{L}_k)}, \bar{b}_k^{(\bar{L}_k)})\Big)$ pour le RN g_k. De plus, nous considérons la même fonction d'activation φ pour toutes les couches des RN f_k, g_k, avec $k \in [K]$.

Ergodicité et Stationnarité. En considérant les notations du Théorème 1, les précédentes notations et en notant $\|W_k^l\|$ la norme spectrale de la matrice correspondante, on peut démontrer que

$$A_k = \big(\mathrm{Lip}(\varphi)^{L_k - 1} \prod_{l=2}^{L_k} \|W_k^{(l)}\|\big) \sum_{i=1}^{p} \|W_{k,i}^{(1)}\| \quad \text{et} \quad B_k = \big(\mathrm{Lip}(\varphi)^{\bar{L}_k - 1} \prod_{l=2}^{\bar{L}_k} \|\bar{W}_k^{(l)}\|\big) \sum_{i=1}^{p} \|\bar{W}_{k,i}^{(1)}\|.$$

Par conséquence, si $C(m) = 2^{m-1} \sum_{k=1}^{K} \pi_k \left(A_k^m + B_k^m \|\epsilon_0\|_m^m\right) < 1$ pour un certain $m \geq 1$, il existe une solution strictement stationnaire du modèle CHARME(p)-basé-sur-RN.

Consistance. Les conditions (**A.1**)-(**A.7**) de Gómez-García *et al.* (2020) sont satisfaites pour les RN f_k et g_k, pour tout $k \in [K]$ (cela a été montré dans le cité article). Donc, l'existence de la solution stationnaire et ergodique du modèle CHARME(p)-basé-sur-RN, implique le Théorème 2(i).

Pour pouvoir appliquer le Théorème 2(ii), nous avons besoin d'une certaine équi-coercitivité et unicité des vrais paramètres (θ^0, λ^0). Ceux-ci sont discutées et assurées dans la Section 6.2.1 de Gómez-García *et al.* (2020).

5 Commentaires

Établir la normalité asymptotique de l'estimateur (5) est très complexe dans un cadre variationnel. C'est pour cela que nous nous restreignons aux arguments habituels de la théorie d'inférence statistique qui demandent, en particulier, la dérivabilité d'ordre trois de la fonction Q_n. En plus, pour simplifier les résultats, nous prenons $\ell(u, v, \tau) = \|u - v\|^2/\tau^2$ et $g_k \equiv 1$, pour tout $k \in [K]$. Sous ces conditions et restrictions, nous établissons

la normalité asymptotique de l'estimateur (5). Les détails peuvent être trouvés dans la Section 5 de Gómez-García *et al.* (2020) et seront aussi discutés lors de cette présentation.

Références

Doukhan, P. and Wintenberger, O. (2008) *Weakly dependent chains with infinite memory.* Stochastic Processes and their Applications, 118 :1997–2013.

Gómez-García, JG., Fadili, J. and Chesneau, C. (2020) *Learning CHARME models with (deep) neural networks.* arxiv preprint arxiv :2002.03237.

Kirch, C. and Kamgaing, T. (2012) *Testing for parameter stability in nonlinear autoregressive models.* Journal of Time Series Analysis, 33(3) :365–385.

Liehr, S., Pawelzik, K., Kohlmorgen, J. and Moler, K.R. (1999) *Hidden markov mixtures of experts with an application to eeg recordings from sleep.* Th. of Biosc, 118 :246–260.

Lo, M.T., Tsai, P.H., Lin, P.F., Lin, C. and Hsin, Y.L. (2009) *The nonlinear and nonstationary properties in eeg signals : probing the complex fluctuations by hilbert-huang transform.* Advances in Adaptive Data Analysis, 1(3) :461–482.

Rockafellar, R.T. (1976) *Integral functionals, normal integrands and measurable selections.* In J. Gossez and L. Waelbroeck, editors, *Nonlinear Operators nd the Calculus of Variations*, number 543 in Lecture Notes in Mathematics, pages 157–207. Springer.

Rockafellar, R.T. and Wets, R.J.B. (1998) *Variational Analysis.* Springer.

Stockis, J-P., Franke, J. and Tadjuidje Kamgaing, J. (2010) *On geometric ergodicity of charme models.* Journal of Time Series Analysis, 31 :141–152.

Tadjuidje-Kamgaing, J. (2005) *Competing neural networks as model for nonstationary financial time series.* PhD thesis, University of Kaiserslautern.

Weigend, A.S. and Shi, S. (2000) *Predicting daily probability distributions of s&p500 returns.* Journal of Forecasting, 19(4) :375–392.

Yarotsky, D. (2017) *Error bounds for approximations with deep relu networks.* Neural Networks, 94 :103–114, 2017.