New approaches for high-dimensional multivariate GARCH models

Benjamin Poignard

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'Université Paris-Dauphine

Approches nouvelles des modèles GARCH multivariés en grande dimension

École Doctorale de Dauphine — ED 543

Spécialité **Sciences**

COMPOSITION DU JURY :

M. Pierre ALQUIER
ENSAE ParisTech
Rapporteur

Mme Cristina BUTUCEA
Université Marne-la-Vallée
Membre du jury

M. Jean-David FERMANIAN
ENSAE ParisTech
Directeur de thèse

M. Marc HOFFMANN
Université Paris-Dauphine
Membre du jury

M. Ostap OKHRIN
Technische Universität Dresden
Rapporteur

M. Jean-Michel ZAKOÏAN
Université Lille III
Président du jury

Soutenue le 15.06.2017
par Benjamin POIGNARD

Dirigée par **Jean-David FERMANIAN**

DAUPHINE UNIVERSITÉ PARIS | PSL RESEARCH UNIVERSITY PARIS

“Non fui, fui, non sum, non curo.”

“Je n’étais pas - J’ai été
Je ne suis plus - Ça m’est égal.”

Épitaphe dans l’esprit de la philosophie du Jardin, qui, en grec ou latin, se lisait sur d’innombrables tombes romaines. Cette formule devint si répandue que le marbrier gravait la plupart du temps les initiales *NF.F.NS.NC.* (Lucien Jerphagnon, *Histoire de la pensée*).

Acknowledgements

Si les lecteurs veulent bien me le permettre, le présent document commencera par des remerciements. Et tout d'abord ceux-ci vont à Jean-David Fermanian, mon directeur de thèse. Ces années de thèse, en sus de ton suivi à l'ENSAE, ont été fondatrices pour moi. Tes judicieux conseils et soutiens aux moments importants ainsi que ta grande disponibilité et ton écoute ont toujours été présents, que ce soit durant ces années de thèse ou durant des travaux antérieurs (OFPR notamment). C'est surtout ta rigueur de travail qui a été indispensable pour la poursuite de cette thèse et m'a permis d'acquérir les fondamentaux nécessaires. Ces facteurs ont grandement contribué à développer mon goût pour la recherche et à poursuivre dans ce domaine. Ce sera un grand plaisir que de poursuivre nos travaux.

Je suis très sensible à l'honneur que m'ont fait les professeurs Pierre Alquier et Ostap Okhrin en acceptant d'être les rapporteurs de cette thèse. Je leur suis très reconnaissant du temps qu'ils ont consacré à l'étude de ce présent document. Par ailleurs, je remercie les professeurs Cristina Butucea, Jean-Michel Zakoïan et Marc Hoffmann pour leur participation au jury de thèse et leur intérêt pour les sujets abordés dans ce travail.

Mes années de thèse passées au CREST ont été une excellente période. J'ai pu y bénéficier d'un cadre de travail exceptionnel et de discussions enrichissantes avec Jean-Michel Zakoïan, Christian Francq et Christian Gouriéroux. Le laboratoire m'a toujours soutenu dans mes démarches, notamment ma période de visite à l'université d'Osaka. Le CREST fut aussi un lieu où j'ai pu rencontré d'autres "thèsards" devenus amis, je pense en particulier à Nicolas et Yiyi. Les fameux japonais illimités avec Nicolas. J'attends Yiyi à Osaka pour goûter des Takoyaki et Okonomiyaki. J'ai une pensée amicale pour Gulten, Sébastien et Anna.

En effet, une nouvelle aventure est sur le point de débuter au Japon, à Osaka. *I warmly thank the Center of Mathematical Modeling and Data Science of Osaka University and Masaaki Fukasawa, who supported my application for the JSPS scholarship. I look forward to working with my Japanese colleagues. That is another reason to improve my level in Japanese. This is going to be an exciting new life in the Kansai area.*

Cette thèse fut également possible grâce à l'orientation et l'enseignement dont j'ai pu bénéficier à Sainte Marie à la Verpillière. Je pense en particulier à Didier Tourrette et M.Paturel, qui ont passé du temps à me conseiller et m'orienter: je n'oublierai pas d'où je viens! Et qu'il est loin (mais j'ai l'impression que c'était hier) le temps de cette guerre futile avec M.Lavialle. J'adresse un remerciement chaleureux à ceux qui m'ont intéressé aux mathématiques et m'ont enseigné les bases solides dans cette matière: je pense en particulier à M.Rémy, qui fut essentiel dans la réussite de ma terminale, et Denis Pennequin, qui fut d'un soutien continu à l'université Paris 1. Enfin je pense à Guillaume et Emmanuel: c'est à leurs côtés que j'ai énormément progressé durant mes études à l'ENSAE.

Je tiens aussi à remercier Jean-Michel Beacco, Julien Pénasse et François Dezorme pour l'opportunité de poursuivre des travaux de recherche à l'Institut Louis Bachelier. Je les remercie pour leur confiance et j'espère que nous allons poursuivre pendant encore quelques années.

J'ai aussi une pensée pour mes amis de CS avec qui j'ai passé des heures dessus: RoiMoMo et Scrambled eggs, qui sauront se reconnaître! La fameuse Lego Map. Je repense aux discussions existentielles avec Jean-François, et ai une pensée pour Bastien et Yann.

Je ne saurai trop remercier ma famille, petite certes, mais qui sera toujours un soutien sur lequel je sais que je pourrai compter: Séverine, ma (belle) mère, Margaux, ma soeur, et Lionel, mon père. Ce sera écrit noir sur blanc: la question de rester avec toi ne s'est même pas posée. Cette volonté s'est naturellement imposée du fait du lien que nous avons: je revois ces matchs de football à Bourgoin-Jallieu auxquels tu assistais, ou aux fameuses parties de Sonic sur la Megadrive. La manière dont je m'étais exprimé pour justifier mon choix de rester n'était pas la bonne, et j'espère que ces lignes sauront dissiper cet étrange sentiment qui a dû perdurer sur la façon dont je m'étais expliqué. J'espère que vous êtes fiers de moi. J'en ai abattu du travail pour en arriver là aujourd'hui.

Enfin, j'ai une pensée particulière pour Chih-Ying, qui occupe une place très spéciale dans ma vie. Je te remercie pour ta patience et ton soutien tout au long de ces années. J'ai pu prendre la mesure de tout ce que tu as pu m'apporter.

Contents

List of Figures

List of Tables

Présentation générale

La modélisation statistique tend à un équilibre entre une paramétrisation parcimonieuse et suffisamment riche afin de décrire et prédire des processus stochastiques. Ce compromis entre complexité statistique et parcimonie doit répondre au besoin d'obtenir à la fois une bonne représentation statistique et des interprétations intuitives. Cet équilibre est souvent précaire dans une modélisation (semi-)paramétrique multivariée, comme l'illustre les modèles en temps discret pour les dynamiques matricielles, généralement gourmands en termes de paramètres.

Ce projet de thèse fut tout au long mûri par cette problématique d'équilibre. Il a pour but de proposer des dynamiques répondant à celle-ci, notamment via des méthodes de réduction de complexité. Être capable de proposer de nouveaux processus suffisamment riches et facilement estimables apporterait un gain important à la fois théorique et aussi pratique, avec la volonté de modéliser la dynamique de processus aléatoires multivariés. Le développement d'outils dits de "sparsité" ou de "régularisation", idest encourageant la réduction de dimension avec l'idée qu'un sous ensemble inconnu de variables est pertinent pour décrire un phénomène, est au coeur de ce présent travail et fait l'objet d'une analyse théorique approfondie. Des applications de ces outils sont proposées pour décrire la dépendance entre composantes de vecteurs de grande taille.

En temps discret, deux grandes familles de modèles de variance-covariance ont fait l'objet de développements conséquents dans la littérature: la famille des modèles à volatilité stochastique et la famille des modèles GARCH multivariés. Ces approches permettent de modéliser la dépendance temporelle du vecteur aléatoire d'intérêt par son moment conditionnel d'ordre deux de façon dynamique. Elles offrent des applications en gestion de portefeuille par exemple, en prenant en compte les risques de volatilité et de corrélation. Ce travail se place dans la seconde classe de modèles. La principale difficulté liée à cette approche est dûe au caractère non-linéaire des dynamiques générées, ce qui rend complexe toute étude probabilisite (difficulté d'extraire

des conditions de stationnarité, absence de formules de prédiction exactes pour les dynamiques de corrélations). De plus, la complexité statistique est inhérente notamment à cause du nombre de paramètres à estimer. En notant N la dimension du vecteur correspondant au nombre de composantes, la complexité est de l'ordre de $O(N^2)$. Ceci implique le plus souvent une incertitude sur la significativité statistique des paramètres estimés, pouvant impacter le pouvoir prédictif du modèle considéré. De cette remarque découle l'idée du fléau de la dimension ce qui contraint les études empiriques à considérer des tailles de vecteur relativement faibles, avec au plus une dizaine de variables. Or la grande dimension ne peut être occultée tant elle occupe une place prépondérante: par exemple en gestion d'actifs, les portefeuilles contiennent souvent plusieurs centaines de variables. En sus d'élaborer des modèles de prédiction fiables, la contrainte de temps nécessite le développement d'approches parcimonieuses en vue de résolutions rapides.

Le premier chapitre de ce document propose une nouvelle méthode pour générer des processus conditionnels de matrices de corrélation. Celles-ci vont être spécifiées à partir d'un sous-ensemble de corrélations partielles dont la structure est décrite par un graphe non dirigé appelé "vine" régulière défini dans la section 1.2. Cette approche fournit des processus multivariés très flexibles et potentiellement parcimonieux dans la mesure où les processus de corrélations partielles peuvent être spécifiés séparemment, le problème multivarié pouvant être considéré comme un système de dynamiques univariées liées par le graphe. Lewandowski, Kurowicka et Joe (2009) développent une approche dans laquelle toute matrice de corrélation peut être obtenue à partir d'une matrice de corrélation partielle, et vice versa, grâce à un algorithme itératif. Une fois le choix des indices entrant dans les corrélations partielles fixé, une "vraie" matrice de corrélation est générée pour des valeurs arbitraires de corrélations partielles. Contrairement aux dynamiques fort usitées issues du Dynamic Conditional Correlation (DCC) de Engle (2002), des séquences de matrices de corrélation sont obtenues sans étapes de normalisation en générant des processus univariés de manière indépendante. Ce chapitre introduit une nouvelle classe de processus dits vine-GARCH. Cette approche novatrice spécifie une dynamique de corrélations partielles données par la "vine" régulière où ses $N(N-1)/2$ branches sont associées à des nombres compris dans $]-1,1[$ et représentant les corrélations partielles correspondantes. En utilisant la propriété d'injection entre ces $N(N-1)/2$ corrélations partielles et les $N(N-1)/2$ corrélations "usuelles", une vraie matrice de corrélation est ainsi générée. Ces corrélations partielles sont empilées dans un vecteur noté Pc_t et ordonnées de manière lexicographique, du plus petit au

plus grand ensemble d'indices, tandis que les corrélations usuelles correspondent aux composantes de la matrice de corrélation conditionnelle notée R_t. Ainsi, la dynamique vine-GARCH proposée est

$$\begin{aligned} H_t &= D_t R_t D_t, \\ \Psi\left(Pc_t\right) &= \Omega + \sum_{k=1}^{p} \Xi_k \Psi\left(Pc_{t-k}\right) + \sum_{l=1}^{q} \Lambda_l \zeta_{t-l}, \\ R_t &= \text{vechof}\left(F_{\text{vine}}\left(Pc_t\right)\right), \quad \text{où} \end{aligned}$$

- H_t est la matrice de variance-covariance obtenue par le produit de la matrice diagonale D_t, dont les composantes correspondent aux variances conditionnelles univariées, et de la matrice de corrélation R_t.

- Pc_t est le vecteur des corrélations partielles définies par la structure de "vine" régulière.

- vechof$(\cdot)$ est l'opérateur de "devectorisation", transformant un vecteur en matrice symmétrique. Il s'agit de la transformation inverse de l'opérateur vech$(\cdot)$.

- Les quantités Ξ_k and Λ_l correspondent aux matrices $N(N-1)/2 \times N(N-1)/2$, de coefficients inconnus, et Ω un vecteur $N(N-1)/2$ de composantes inconnues. Ainsi est défini le vecteur des paramètres inconnus de corrélation $\theta_c = (\Omega, \Xi_1, \ldots, \Xi_p, \Lambda_1, \ldots, \Lambda_q)$. Ces matrices sont choisies de manière arbitraire, où en particulier la propriété de définie-positivité n'est pas imposée.

- Le vecteur ζ_{t-1} est $\mathcal{F}_{t-1}$-mesurable et correspond à l'innovation dans la dynamique des corrélations partielles. Il est défini de telle sorte que $\mathbb{E}[\zeta_{t-1}] \simeq \mathbb{E}[Pc_{t-1}]$, procédure qui est conforme avec les équations de mise à jour dans les modèles de type GARCH. La construction du vecteur ζ_{t-1} est décrite dans la sous-section 1.3.2.

- $\Psi(.)$ est une transformation déterministe de Pc_t afin de conserver des dynamiques de corrélations partielles dans $]-1,1[$. Par soucis de simplification, $\Psi(.)$ est connue et définie de $]-1,1[^{N(N-1)/2}$ dans $\mathbb{R}^{N(N-1)/2}$ telle que

$$\Psi(Pc_t) = \left(\psi(\rho_{1,2,t}), \ldots, \psi(\rho_{N,N-1|L_{N-1,N},t})\right)', \ \psi\left(x\right) = \tan\left(\pi x/2\right).$$

- La fonction $F_{\text{vine}}(.)$ correspond à l'injection du vecteur des corrélations partielles Pc_t vers les corrélations (dans R_t) en utilisant l'algorithme de Lewandowski,

Kurowicka et Joe (2009). Elle est définie de $]-1,1[^{N(N-1)/2}$ dans $]-1,1[^{N(N-1)/2}$ par $F_{\text{vine}}(\rho_{1,2,t},\ldots,\rho_{N-1,N|L,t}) = (\rho_{1,2,t},\ldots,\rho_{N-1,N,t})'$.

L'approche vine-GARCH encourage la parcimonie et donc la réduction du nombre de paramètres car des contraintes sur les corrélations partielles peuvent être imposées à tout niveau du graphe "vine" sans modifier les autres corrélations partielles. En effet, il est pertinent d'annuler (ou de laisser au moins constante) toutes les corrélations partielles associées à la "vine" à partir d'un niveau r donné. Lorsque les corrélations partielles sont supposées nulles à partir de ce niveau, il est nécessaire de savoir si les corrélations usuelles correspondantes dépendent de la structure de la vine à partir de celui-ci. Pour ce faire, le concept de "r vine-Free" est introduit. Une "vine" est "r vine-free" si, lorsque toutes les corrélations partielles sont nulles à partir du niveau r, les corrélations usuelles ne dépendent pas de la manière dont la "vine" est construite à partir de ce niveau. Cette propriété est vérifée par toute "vine" régulière. Ainsi la dimension du problème statistique peut potentiellement être réduite en utilisant cette propriété, seuls les r premiers niveaux des dynamiques de corrélations partielles devant être estimés. En outre, ce chapitre introduit une procédure d'estimation du modèle vine-GARCH par quasi-maximum de vraisemblance en plusieurs étapes. Celle-ci peut être menée équation par équation, passant en revue les noeuds successifs du graphe. Ceci fournit une solution au fléau de la dimension.

Une étude théorique approfondie est menée pour obtenir les conditions d'existence et d'unicitié de solutions stationnaires strictes de la dynamique proposée. En effet, prouver ces propriétés probabilistes est un préliminaire nécessaire avant de développer une théorie asymptotique (typiquement les propriétés de consistance et de normalité asymptotique de l'estimateur du quasi-maximum de vraisemblance) dans la mesure où les lois fortes des grands nombres ou les théorèmes centraux limites sont facilement obtenus dans ce cas. Par exemple, Boussama, Fuchs et Stelzer (2011) établissent ces résultats de stationnarité pour la famille des modèles BEKK. Dans le cas du processus vine-GARCH, le passage des corrélations partielles aux corrélations usuelles est non-linéaire par l'injection $F_{\text{vine}}(.)$. Cette transformation rend complexe toute étude visant à établir les conditions de stationnarité, à l'instar du modèle DCC. Pour établir les conditions d'existence et d'unicité de solutions stationnaires strictes du processus vine-GARCH, celui-ci est écrit comme une chaîne de Markov non linéaire. La difficulté majeure est l'impossibilité d'extraire une fonction déterministe explicite reliant la chaîne de Markov au processus d'innovation supposé stationnaire et ergodique.

Ainsi est utilisé le critère de Tweedie (1988) fournissant l'existence d'une mesure de probabilité invariante pour la dynamique vine-GARCH écrite comme une chaîne de Markov. Une fois établies les conditions de stationnarité, les propriétés asymptotiques de l'estimateur du quasi-maximum de vraisemblance en deux étapes sont étudiées et les conditions de consistance faible et de normalité asymptotique sont fournies. Enfin les performances empiriques du modèle vine-GARCH sont analysées au travers d'études simulées et sur données réelles.

Cette nouvelle approche pour générer des dynamiques de matrices de corrélation suppose d'imposer des contraintes a priori dans le graphe "vine" afin d'être parcimonieuse. Il s'agit de contraindre le nombre de paramètres en excluant certaines variables et groupes de variables - les corrélations partielles passées ou les corrélations partielles traitées selon le niveau dans le graphe "vine"- traitées comme non pertinentes pour décrire la corrélation conditionnelle instantannée. Cette réduction correspond à une approche en forme réduite a priori, la condition étant que le modèle conserve une flexibilité suffisante afin de capturer des dynamiques hétérogènes et de proposer de bonnes performances prédictives.

Le traitement de la parcimonie dans les modèles vine-GARCH est basé sur des choix a priori de niveaux "limites" au delà-desquels les corrélations partielles sont négligeables. De manière plus générale, développer des approches dites de pénalisation ou de régularisation, plus rigoureuses et moins artisanales, est souhaitable. C'est ce qui a motivé l'étude relative à la pénalisation "Sparse Group Lasso" ainsi que ses applications aux modèles dynamiques multivariés. En ajoutant à une fonction objectif une fonction de pénalité singulière en zéro, une procédure statistique réalise à la fois de la sélection de variable et de l'estimation. Le concept clé de régularisation intervient dans le cadre de statistiques en grande dimension, l'idée étant de contraindre les paramètres et donc les variables correspondantes, pour éviter les problèmes de surapprentissage. Le besoin de régularisation peut facilement être perçu en considérant le cas dans lequel il y a exactement le même nombre de variables que d'observations. La méthode des moindres carrés linéaires expliquera parfaitement les données, la statistique R^2 étant égale à un. En revanche, il est fort probable que l'utilisation du modèle estimé produise de faibles performances prédictives hors-échantillon dans la mesure où le modèle estimé est caractérisé par le surapprentissage. D'une part, les moindres carrés capturent le signal quant à la manière selon laquelle les variables prédictives doivent être utilisées pour prédire la variable de sortie; mais d'autre part les moindres carrés capturent le bruit inhérent à l'échantillon, ce qui implique que le modèle ne peut être utilisé pour produire

des prédictions hors-échantillon utiles. Ainsi dans ce cadre, un modèle de prédiction gagne en pertinence en recourant à de la régularisation ou réduction de dimension. Cela signifie que les estimateurs doivent être contraints de telle sorte que le surapprentissage soit évité. Pour ce faire, de nombreuses fonctions de régularisation ont été proposées dans la littérature, selon le problème que l'on cherche à décrire. L'intuition principale de la pénalisation est d'identifier le vrai support sous-jacent sparse inconnu, c'est-à-dire l'ensemble des indices pour lesquels les variables correspondantes sont conservées pour décrire une dynamique, la taille de cet ensemble étant plus petit que l'ensemble de toutes les variables potentielles (il est formé par les multiples manières de transformer et de faire interagir les variables).

Des procédures de régularisation sont détaillées par exemple par Hastie, Tibshirani et Wainwright (2015). Quant aux propriétés théoriques des estimateurs pénalisés, deux types d'analyse sont possibles. D'une part, les approches en échantillons finis traitent de la grande dimension en considérant la taille du vecteur des paramètres à estimer potentiellement plus grand que le nombre d'observations, supposé fixe. L'analyse théorique vise à établir des bornes en probabilité ou en espérance, pour une métrique donnée, telles l'erreur de prédiction ou l'erreur d'estimation du paramètre. Ces bornes seront valables avec une probabilité grande et sont fonctions du vrai support sparse. Ces types de résultats sont résumés par exemple par Bühlmann et van de Geer (2011). L'autre point de vue est asymptotique, cas dans lequel la taille de l'échantillon tend vers l'infini. Les premiers résultats asymptotiques pour l'estimateur Lasso ont été établis par Knight et Fu (2000). Fan et Li (2001) ont développé un cadre général de vraisemblance pénalisée et ont analysé les propriétés de consistance et de normalité asymptotique de l'estimateur SCAD. Le cas de la grande dimension est traité lorsque la taille de l'échantillon ainsi que la taille du vecteur des paramètres tendent simultanément vers l'infini. Par exemple, Fan et Peng (2004) traitent de cet asymptotique double pour des fonctions de vraisemblances pénalisées.

Le second chapitre contribue à cette littérature dite de "régularisation" ou "pénalisation statistique". J'y propose une étude théorique approfondie d'une généralisation de l'estimateur Sparse Group Lasso (SGL), initialement proposé par Simon, Friedman, Hastie et Tibshirani (2013). Dans le contexte de données dépendantes, un cadre de M-estimateur est développé dans lequel la fonction objectif - non pénalisée - est convexe et la pénalisation étudiée est du type "adaptive Sparse Group Lasso". Celle-ci fait intervenir deux pénalités, la composante l^1-Lasso et la composante l^1/l^2 pour le Group Lasso, pondérées par des coefficients stochastiques de première étape. De

plus, deux paramètres de régularisation sont introduits pour chaque composante de l'"adaptive Sparse Group Lasso", le Lasso et le Group Lasso. Le principal avantage de cette pénalisation est de favoriser la sparsité au niveau d'un groupe de paramètres, ce qui écartera le groupe de covariables concerné, ainsi que la sparsité à l'intérieur d'un groupe de paramètres lorsque celui-ci est considéré comme statistiquement significatif pour décrire la variable de sortie.

Dans ce cadre, le vecteur des paramètres θ de taille d est décomposé en m groupes $\mathcal{G}_k, k = 1, \cdots, m$ avec $\text{card}(\mathcal{G}_k) = \boldsymbol{c}_k$ et $\sum_{k=1}^{m} \boldsymbol{c}_k = d$. Ainsi $\theta = (\theta_i^{(k)}, k \in \{1, \cdots, m\}, i = 1, \cdots, \boldsymbol{c}_k)$. L'objet d'intérêt est l'ensemble $\mathcal{A} := \{j : \theta_{0,j} \neq 0\}$ qui correspond au support sous-jacent sparse, θ_0 étant le vrai paramètre inconnu. Cet ensemble inconnu est par hypothèse plus petit que l'ensemble de toutes les variables potentielles. L'objet principal de ce chapitre est de prouver d'un point de vue asymptotique la capacité de la pénalisation "adaptive SGL" à identifier le support $\mathcal{A}$ et d'établir les vitesses de convergence des paramètres de régularisation pour obtenir cette propriété. Plus précisément, le problème statistique consiste à minimiser dans l'espace convexe des paramètres Θ un critère pénalisé de la forme

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \{\mathbb{G}_T \varphi(\theta)\},$$

où

$$\mathbb{G}_T \varphi(\theta) = \mathbb{G}_T l(\theta) + \mathcal{R}(\lambda_T, \gamma_T, \tilde{\theta}, \theta),$$

avec (ϵ_t) le vecteur des observations; $\mathbb{G}_T l(\theta) = \frac{1}{T}\sum_{t=1}^{T} l(\epsilon_t; \theta)$ est la fonction objectif non pénalisée, supposée convexe par rapport aux paramètres pour toute réalisation de ϵ_t; le modèle d'intérêt entre dans le critère $l(\epsilon_t; \theta)$; $\mathcal{R}(\lambda_T, \gamma_T, \theta)$ est la fonction de régularisation (ou pénalité) "adaptive Sparse Group Lasso", définie par

$$\mathcal{R}(\lambda_T, \gamma_T, \tilde{\theta}, \theta) = \boldsymbol{p}_1(\lambda_T, \tilde{\theta}, \theta) + \boldsymbol{p}_2(\gamma_T, \tilde{\theta}, \theta),$$

avec

$$\begin{aligned} \boldsymbol{p}_1(\lambda_T, \tilde{\theta}, \theta) &= \lambda_T T^{-1} \sum_{k=1}^{m} \sum_{i=1}^{\boldsymbol{c}_k} \alpha(\tilde{\theta}_i^{(k)}) |\theta_i^{(k)}|, \\ \boldsymbol{p}_2(\gamma_T, \tilde{\theta}, \theta) &= \gamma_T T^{-1} \sum_{l=1}^{m} \xi(\tilde{\theta}^{(l)}) \|\theta^{(l)}\|_2. \end{aligned}$$

La quantité $\tilde{\theta}$ est un estimateur de première étape supposé $\sqrt{T}$-consistant. Les paramètres de régularisation λ_T et γ_T varient avec T. Enfin, $\alpha(\tilde{\theta}^{(k)}) \in \mathbb{R}_+^{\boldsymbol{c}_k}$, $\xi(\tilde{\theta}^{(l)}) \in \mathbb{R}_+$ sont définis

par

$$\alpha_T^{(k)} := \alpha(\tilde{\theta}^{(k)}) = (|\tilde{\theta}_i^{(k)}|^{-\eta}, i = 1, \cdots, \boldsymbol{c}_k),\ \xi_{T,l} := \xi(\tilde{\theta}^{(l)}) = \|\tilde{\theta}^{(l)}\|_2^{-\mu},$$

pour des constantes $\eta > 0$ et $\mu > 0$. Ces derniers jouent un rôle clé pour satisfaire la propriété oracle dans la mesure où ces poids impactent les convergences des paramètres de régularisation.

Dans un premier cadre asymptotique, pouvant être qualifié d'asymptotique simple, où seule la taille de l'échantillon diverge, sont établies en particulier la consistance et la distribution asymptotique de l'estimateur SGL dans sa version non "adaptive", cas dans lequel n'intervient pas d'estimateur de première étape et donc les poids α et ξ sont non stochastiques. Il est également prouvé dans le Théorème 2.4.16 que la version "adaptive" du SGL satisfait la propriété oracle au sens de Fan et Li (2001): l'estimateur sparse identifie le vrai support sparse sous-jacent et sa loi est asymptotiquement normale. Sur la base des travaux de Fan et Peng (2004) et de Zou et Zhang (2009), la grande dimension est également traitée avec un asymptotique double où la dimension du vecteur à estimer diverge avec la taille de l'échantillon. Ainsi la taille du vecteur des paramètres dépend de T avec $d := d_T = O(T^c)$ avec $0 < c < 1$. Le principal résultat de la section 2.5 est la propriété oracle en asymptotique double. Les vitesses de convergence des paramètres de régularisation sont explicitement établies dans le théorème 2.5.24, notamment via un compromis entre les pénalisations de la composante Lasso et de la composante Group Lasso. Cette analyse met en évidence le fait que ce cadre général de M-estimateur ne favorise pas la flexibilité dans le comportement de d_T, autrement dit c ne peut être compris dans tout l'ensemble $]0, 1[$. Ce problème a été rencontré par Fan et Peng (2004) dans un cadre i.i.d. et sans estimateur "adaptive". Ce manque de flexibilité provient de la nécessité de contrôler le terme d'ordre trois dans les développements de Taylor. Ce problème n'apparaît pas si la fonction objectif correspond aux moindres carrés, dans la mesure où ce terme d'ordre trois disparaît. Par exemple, Zou et Zhang (2009) ont prouvé la propriété oracle pour l'estimateur "elastic-net" d'un point de vue asymptotique double pour des modèles linéaires avec $0 < c < 1$. Enfin, les propriétés asymptotiques de l'"adaptive Sparse Group Lasso" sont illustrées par des expériences simulées et soulignent que cet estimateur offre de meilleures performances que d'autres méthodes oracles - adaptive Lasso, adaptive Group Lasso - tant en termes de précision statistique que de sélection de variables.

Le cadre général développé dans le second chapitre englobe d'importantes familles

de modèles paramétriques et semi-paramétriques: par exemple les modèles linéaires généralisés; les modèles de type Cox; le problème d'estimation des matrices de précision dans un cadre gaussien; etc. Le modèle linéaire pénalisé est le plus fréquemment usité dans cette littérature dans la mesure où la fonction de perte convexe qui lui est rattachée, les moindres carrés ordinaires, est directement manipulable pour des études théoriques du type échantillon fini avec les bornes oracles - bornes d'erreur de l'estimateur pénalisé valables avec une forte probabilité pour un certain choix de paramètre de régularisation et exprimées en fonction du support sparse sous-jacent inconnu - ou de type asymptotique. En effet, le développement d'ordre trois étant nul, l'analyse en est grandement facilitée. En outre, d'un point de vue empirique, beaucoup d'algorithmes de résolution ont été proposés dans ce cadre des moindres carrés pénalisés (algorithme du gradient typiquement).

L'idée du troisième chapitre est de développer des dynamiques linéaires pour les processus multivariés de variance-covariance afin d'utiliser la méthode des moindres carrés et d'illustrer l'utilité de la méthodologie "adaptive Sparse Group Lasso" développée dans le chapitre 2. Dans le cas univarié, le modèle GARCH ne peut-être estimé par moindres carrés ordinaires, contrairement au modèle ARCH. Cette caractéristique peut-être étendue à un système multivarié sous la contrainte de définir une paramétrisation générant des matrices définies-positives. En notant le vecteur des observations (ϵ_t), avec $H_t = \mathbb{E}[\epsilon_t \epsilon_t' | \mathcal{F}_{t-1}]$ et $\mathcal{F}_t := \sigma(\epsilon_s, s \leq t)$ la filtration naturelle, la dynamique ARCH multivariée est donnée par

$$\epsilon_t \epsilon_t' = A + \sum_{k=1}^{q} (I_N \otimes \epsilon_{t-k}') B_k (I_N \otimes \epsilon_{t-k}) + \zeta_t, \quad \mathbb{E}[\zeta_t | \mathcal{F}_{t-1}] = 0,$$

avec A et B_k, $k = 1, \ldots, q$, symmétriques et définies-positives. La contrainte majeure à intégrer est la convexité du problème statistique par rapport aux paramètres. D'abord d'un point de vue empirique, la convexité assure de bonnes propriétés de convergence des algorithmes de résolution. En outre, d'un point de vue théorique, la propriété oracle de l'estimateur "adaptive SGL" repose sur la convexité du problème. C'est la raison pour laquelle les processus ARCH multivariés et Cholesky-GARCH exposés respectivement dans les sections 3.3 et 3.4 satisfont la propriété de linéarité par rapport aux paramètres. En outre, en vue de la définie positivité des processus matriciels, les contraintes imposées sur ceux-ci sont au plus linéaires.

Ainsi ce chapitre propose d'utiliser le cadre M-estimateur pénalisé développé dans le chapitre 2 pour les moindres carrés pénalisés. En utilisant les paramétrisations

exposées dans les sections 3.3 et 3.4, les processus peuvent être estimés par la méthode des moindres carrés ordinaires. Ceci est un avantage crucial par rapport aux fonctions objectifs non-linéaires car les méthodes de résolution sont rapides et les procédures de régularisation peuvent être aisément mises en oeuvre dans ce cadre. Le problème statistique s'exprime ainsi

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \{\mathbb{G}_T \varphi(\theta)\},$$

où

$$\mathbb{G}_T \varphi(\theta) = \mathbb{G}_T l(\theta) + \mathcal{R}(\lambda_T, \gamma_T, \tilde{\theta}, \theta),$$

avec $\mathcal{R}(\lambda_T, \gamma_T, \tilde{\theta}, \theta)$ la pénalité "adaptive Sparse Group Lasso" étudiée dans le chapitre 2; la partie non pénalisée est

$$\begin{cases} \mathbb{G}_T l(\theta) &= \dfrac{1}{T}\sum_{t=1}^{T} l(\epsilon_t; \theta), \\ l(\epsilon_t; \theta) &= \|\text{Vech}(\epsilon_t \epsilon_t') - \Psi(\underline{\epsilon_{t-1}})\theta\|_2^2, \end{cases}$$

où $\Psi(\underline{\epsilon_{t-1}})$ est une matrice dont les composantes, correspondant à une transformation des éléments des vecteurs $\text{Vech}(\epsilon_{t-k}\epsilon_{t-k}')$, $k \geq 1$, sont $\mathcal{F}_{t-1}$-mesurables et dont la structure dépend des spécifications ARCH multivariées données dans les sections 3.3 et 3.4. Certaines spécifications de $\Psi(\underline{\epsilon_{t-1}})$ introduiront des matrices de variance-covariance définies-positives. Dans ce cadre d'estimation linéaire, $\tilde{\theta}$ est un estimateur de première étape des moindres carrés ordinaires.

Cette approche ne peut pas être développée pour les dynamiques MGARCH en présence de termes autorégressifs. Néanmoins, ces processus peuvent être approchées avec q élevé. Pour les MGARCH, la méthode d'estimation par quasi-maximum de vraisemblance gaussien est la plus fréquemment usitée. Cette méthode peut difficilement être utilisée pour des vecteurs stochastiques de grande taille N dans la mesure où la complexité est de l'ordre $O(N^2)$. C'est la raison pour laquelle la majorité des applications se limitent à des vecteurs de taille faible (typiquement $N \leq 10$) ou se placent dans des processus scalaires, tels que les DCC ou BEKK scalaires, non adaptés pour des problèmes de grande dimension en présence de composantes hétérogènes.

Dans ce cadre pénalisé, la régularisation Sparse Group Lasso est particulièrement adaptée dans la mesure où les groupes peuvent être définis par le vecteur des variables correspondant aux retards. En effet, pour un nombre q élevé de retards initialement spécifiés, auquel correspond l'ensemble des variables retardées, la régularisation vise à identifier un sous-ensemble des variables retardées d'ordre $\tilde{q} < q$. L'idée est que

les coefficients tendent vers zéro à partir d'un certain retard, les variables observées récemment ayant un effet plus significatif sur la covariance instantanée que des observations plus lointaines. De plus, pour éviter les problèmes de surapprentissage, il est nécessaire de contraindre les paramètres par une telle procédure de pénalisation. Les performances de cette procédure de régularisation lors de l'estimation des dynamiques ARCH multivariées proposées sont étudiées à travers des simulations. Il s'agit de mesurer l'écart entre la vraie matrice de variance-covariance connue et simulée et les matrices de variance-covariance estimées selon plusieurs spécifications. Parmi celles-ci se trouvent les modèles ARCH et Cholesky-GARCH pénalisés ainsi que le DCC scalaire. Ces simulations mettent en évidence le gain en termes de précision obtenue sur la mesure de la matrice de variance-covariance lorsque la procédure "adaptive Sparse Group Lasso" est utilisée.

Chapter 1

Dynamic Correlation Model based on Vines

1.1 Introduction

A multivariate setting is necessary for modeling the cross-sectional and temporal dependencies between N financial asset returns. It allows for developing relevant management tools, especially when the interactions between financial markets become stronger. This concerns areas such as asset pricing, portfolio allocation, risk management, and the like.

The usual modeling approach relies on the specification of the first two moments of vectors of returns conditional on their past (and current market information possibly). Once this is done, some assumed vectors of innovations close the model specification. The multivariate GARCH (MGARCH) and the multivariate stochastic volatility (MSV) models are the two main frameworks: see the surveys of Bauwens, Laurent and Rombouts (2006) and Asai, McAleer and Yu (2006) respectively. Such approaches allow for generating sequences of asset return covariance matrices (H_t), and then provide their correlations as a by-product. In financial econometrics, MGARCH models are most commonly used. Indeed, they induce some typical patterns as volatility clustering and complex dependencies (through copula-GARCH models, e.g.), without the necessity of complex inference procedures, contrary to most MSV models.

Nonetheless, the number of MGARCH parameters often increases dramatically with the number of underlying assets. Therefore, some simplified MGARCH specifications

have searched for parsimony fostering simple estimation and interpretation, but sometimes at the price of an over-simplification. Besides, MGARCH models have to guarantee the positive definiteness of the generated covariance matrices. This induces complexities, and more or less arbitrary model constraints. Our goal will be to stay inside the MGARCH family, without suffering from these drawbacks and with a focus on correlation dynamics.

But how are correlation managed in such MGARCH models ? The BEKK model (Engle and Kroner, 1995) specifies the dynamics of the underlying covariance matrices H_t directly as a deterministic quadratic function of past returns, but the number of parameters has a $O(N^2)$ complexity. Hence Engle, Ng and Rotschild (1990) proposed the Factor-GARCH model, following the intuition that comovements of asset returns are driven by a small number of common underlying variables. As a by-product and in both cases, conditional correlations may be obtained, but their expressions are not intuitive or easily explicable.

Other specifications focus on conditional correlations more directly. Intuitively, univariate GARCH dynamics (or others) may be chosen to get conditional variance processes. Then, based on these dynamics, a correlation process (R_t) could be built. This was the way proposed by Engle (2002) with the Dynamic Conditional Correlation (DCC) approach. But to cope with the positive definiteness of R_t, DCC-type models have to rely on a not intuitive normalization stage. This has been a source of difficulties and criticism (see Caporin and McAleer, 2013), in particular to obtain a sound theory for inference. Fermanian and Malongo (2016) pointed out these drawbacks when exhibiting some conditions for the stationarity of DCC model trajectories. Moreover, although these families may allow for generating high-dimensional correlation matrices, their estimation and forecasting are clearly challenging without additional restrictions. Several attempts tried to reduce significantly the number of parameters, such as the scalar DCC processes of Engle and Sheppard (2001), the Flexible DCC model of Billio and Caporin (2006), among others. But the ability of the latter models to capture complex and rich dynamics of heterogeneous series is limited.

Therefore, the discussions around correlations often remain fragile and partly "black-box", since neither standard MGARCH or DCC-type models work directly on explicit correlation dynamics. Indeed, the former ones set covariances when the latter ones depend on a normalization stage. In this paper, we propose to circumvent the problem with another method using partial correlations. This approach tends to be both

parsimonious and flexible, and will specify some correlation and partial correlation dynamics directly. Any $N \times N$ correlation matrix may be described by $N(N-1)/2$ partial correlations. Lewandowski, Kurowicka and Joe (2009) explained how to deduce a correlation matrix from partial correlations (or the opposite), through an iterative algorithm. With such techniques, once the indices of a family of partial correlations is chosen conveniently, a "true" correlation matrix is generated, whatever the values of these partial correlations are. This property will be crucial here: by producing univariate dynamics of partial correlations independently, we obtain sequences of correlation matrices without any normalization stage, contrary to DCC-type models.

An important practical question will be to choose the indices of the relevant partial correlations. Kurowicka and Cooke (2006) showed that the partial correlations of a random vector can be mapped to a so-called vine tree. Such objects are sets of connected undirected trees. They have been discovered recently due to their ability to built high-dimensional distributions through a set of bivariate copulas (one copula per node of the vine) and marginal cdfs'. See Aas, Czado, Frigessi and Bakken (2006) for an introduction. Here, we develop a class of MGARCH models based on regular vines, the so-called "vine-GARCH" models. The latter models are flexible enough by allowing independent specifications/estimations of partial correlation processes. It is also parsimonious as one can set constraints at any level of the vine tree without altering other correlations.

The rest of this paper is organized as follows: Section 1.2 develops some basic definitions/properties of trees, vines, partial correlations and the way they will be relevant for constructing nonnegative definite matrices. After having set the definitions and notations of usual MGARCH and DCC models, the new vine-GARCH framework is detailed in Section 1.3. In Section 1.4, we define the statistical inference of our new models by a quasi-maximum likelihood (QML) procedure. The conditions of existence and uniqueness of strictly stationary solutions and the asymptotic properties of the vine-GARCH model are provided respectively in Section 1.5 and Section 1.6. Section 1.7 contains an empirical study with simulated data and a database of stock returns, and then we conclude the study.

1.2 Vines and partial correlations

This section emphasizes how to specify a relevant set of partial correlations by considering a graphical approach based on vines.

1.2.1 Vines

Let $\mathcal{N}$ be a set of n elements. By definition, $T = (\mathcal{N}, \mathcal{E})$ is a tree with nodes $\mathcal{N}$ and edges $\mathcal{E}$ if $\mathcal{E}$ is a subset of unordered pairs of $\mathcal{N}$ with no cycle and if there is a path between each pair of nodes. Moreover, vines on n elements are undirected graphs that nest sets of some connected trees $T_1, \ldots, T_{n-1}$, where the edges of tree T_j are the nodes of tree T_{j+1}, $j = 1, \ldots, n-2$. A *regular vine* (R-vine) on n elements is a vine in which two edges in tree T_j are joined by an edge in tree T_{j+1} only if these edges share a common node, for any $j = 1, \ldots, n-2$. A formal definition is given below. See Kurowicka and Joe (2010) for a survey and additional results.

Definition 1.2.1. $V(n)$ is a labeled *regular vine* on n elements if:

1. $V(n) = (T_1, T_2, \ldots, T_{n-1})$.
2. T_1 is a connected tree with nodes $\mathcal{N}_1 = 1, 2, \ldots, n$ and edges $\mathcal{E}_1$. For $i = 2, \ldots, n-1$, T_i is a connected tree with nodes $\mathcal{N}_i = \mathcal{E}_{i-1}$, and the cardinality of $\mathcal{N}_i$ is $n-i+1$.
3. If a and b are nodes of T_i connected by an edge in T_i, where $a = \{a_1, a_2\}$ and $b = \{b_1, b_2\}$, then exactly one of the a_i equals one of the b_i. This is the *proximity* condition.

We consider only regular vines in this paper, and the properties we state hereafter are true for such vines implicitly. There are $n(n-1)/2$ edges in a regular vine on n variables. An edge in tree T_j is an unordered pair of nodes of T_j, or equivalently, an unordered pair of edges of T_{j-1}. The degree of a node is the number of edges incident with it.

Two particular cases of R-vines are important, traditionally. A regular vine is called a *canonical vine* (C-vine) if each tree T_i has a unique node of degree $n-i$, i.e. a node with maximum degree. A regular vine is called a *D-vine* if all nodes in T_1 have degree not higher than 2.

The variables reachable from a given edge via the membership relation are called the *constraint set* of that edge. When two edges are joined by an edge of the next tree, the intersection of the respective constraint sets are the *conditioning variables*, and the symmetric differences of the constraint sets are the *conditioned variables*. With the notations of point 3 of the previous definition, at tree T_i, say $a_1 = b_1$, and a_1 is a common element of a and b. This means that, at tree T_{i+1}, a_1 enters the conditioning set of (a_2, b_2). Thus, we define the conditioning and conditioned sets formally as follows.

Definition 1.2.2. For $e \in \mathcal{E}_i$, $i \leq n-1$, the *constraint set* associated with e is the complete union of the elements in $\{1, \ldots, n\}$ that are reachable from e by the membership relation. It is denoted by $U_e^\star$.

Definition 1.2.3. For $i = 1, \ldots, n-1$, if $e \in \mathcal{E}_i$, it connects two elements j and k in $\mathcal{N}_i$ and it can be written $e = \{j, k\}$. The *conditioning set* associated with e is $L_e := U_j^\star \cap U_k^\star$, and the *conditioned set* associated with e is a pair $\{C_{e,j}, C_{e,k}\} := \left\{U_j^\star \setminus L_e, U_k^\star \setminus L_e\right\}$.

Obviously, since the edges of a given tree T_i are the nodes of T_{i+1}, the same concepts of constraint/conditioning/conditioned sets apply to all the nodes in a vine.

Lemma 1.2.4. *(Bedford, Cooke, 2002)*
Let a regular vine on n variables. Then,

1. *the total number of edges is* $n(n-1)/2$;
2. *two different edges have different constraint sets;*
3. *each conditioned set is a doubleton and each pair of variables occurs exactly once as a conditioned set;*
4. *if* $e \in \mathcal{E}_i$, *then* $\#U_e^\star = i+1$, $\#L_e = i-1$;
5. *if two edges have the same conditioning set, then they are the same edge.*

In a regular vine, the edges of T_{m+1} (equivalently the nodes of T_{m+2}) will be denoted by $e = (a_j, a_k | b_1, \ldots, b_m)$, where a_j, a_k and the b_l, $l = 1, \ldots, m$ are different elements in $\{1, \ldots, n\}$. This notation means that the conditioning set of e is $L_e = \{b_1, \ldots, b_m\}$, and the conditioned set of e is $\{a_j, a_k\}$. Both C-, D- and R-vine and the concepts above can be visualized on Figures 1.1, 1.2 and 1.3.

To have the intuition, keep in mind that a node represents a random variable, and an edge between two nodes means we will specify the dependence between these two particular nodes, in general through a copula (that will be reduced to a partial correlation hereafter). Such copulae have to be defined afterwards, but, for the moment, assume this can be done easily. Typically, the goal is to describe the joint law of the n asset returns. For instance, in Figure 1.1, the five nodes in T_1 may be the asset returns r_i, $i = 1, \ldots, 5$, associated to stock indices. The first tree tells us we will specify the dependencies between r_1 and the other returns r_i, $i > 1$. Here, we select 1 as the core index (the "main factor") in this portfolio. Once we have controlled the T_1-related dependencies, the new nodes in T_2 are conditional asset returns given r_1. We select asset 2 given 1 as the "most relevant" one. The new edges tell us we focus now on conditional copulae between the latter node and the returns r_j given r_1, $j = 2, \ldots, 5$. And we go on with T_3, dealing with the asset returns r_j given r_1 and r_2, $j = 3, 4, 5$, etc. With such a C-vine and a set of convenient bivariate copulae, we obtain the joint law of $(r_1, \ldots, r_5)$ by gathering and multiplying conveniently all the (conditional) copulae we haver considered above. This is the simplest way of building vines. Obviously, more complex structures may be relevant too, as in the R-vine of Figure 1.3. With heterogeneous portfolios, for instance, it would be fruitful to particularize several nodes in T_1. See Aas et al. (2006) for other insights. In terms of model specification, the first chosen trees are crucial because they correspond to our intuitions (our "priors") about the most important linkages among the assets in the portfolio. Moreover, from some level on and in practice, it is often possible and useful to assume no dependencies: see the "r-vine free" property in Definition 1.2.10 below.

The next section focuses on how such vines can be related to some subsets of the partial correlations that are associated to a random vector.

1.2.2 Partial correlations

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a n-dimensional random vector, $n \geq 2$, with zero mean. For any indices i, j in $\{1, \ldots, n\}$, $i \neq j$ and any subset $L \subset \{1, \ldots, n\}$, for which i and j do not belong to L, $\rho_{i,j|L}$ is called the partial correlation of X_i and X_j, given X_k, $k \in L$. It is the correlation between the orthogonal projections of X_i and X_j on $< X_k, k \in L >^{\perp}$, the orthogonal of the subspace generated by $\{X_k,\, k \in L\}$. When L is empty, then $\rho_{i,j|\emptyset} = \rho\left(X_i, X_j\right) := \rho_{i,j}$ is the usual correlation. Note that, if the

random vector $\mathbf{X}$ is normal, then its partial correlations correspond to some conditional correlations.

Interestingly, partial correlations can be computed from usual correlations with a recursive formula. Let (i, j, k) be any set of distinct indices, and L be another (possibly empty) set of indices that is disjoint from (i, j, k). Following Lewandowski et al. (2009), we have

$$\rho_{i,j|k,L} = \frac{\rho_{i,j|L} - \rho_{i,k|L}\rho_{j,k|L}}{\sqrt{\left(1 - \rho^2_{i,k|L}\right)\left(1 - \rho^2_{j,k|L}\right)}}. \tag{1.2.1}$$

Assume we know the usual correlations $\rho_{i,j}$, for any couple (i, j), $i \neq j$. We check easily that any partial correlation can be calculated by invoking (1.2.1) several times with increasing subsets L. Actually, the opposite property is true if we start from a convenient subset of partial correlations. Indeed, the edges of a regular vine on n elements may be associated with the partial correlations of a n-dimensional random vector in the following way: for $i = 1, \ldots, n-1$, consider any $e \in \mathcal{E}_i$, the set of edges at tree T_i. Let $\{j, k\}$ be the two conditioned variables of e, and L_e its conditioning set. We associate the partial correlation $\rho_{j,k|L_e}$ to this node. Kurowicka and Cooke (2006) call this structure a *partial correlation vine specification*, that is simply a R-vine for which any edge is associated to a number in $]-1, 1[$. Actually, all positive correlation matrices may be generated by setting a (fixed) R-vine on n variables, and by assigning different partial correlations to all the nodes of this vine. This means setting ρ_e to any $e \in \bigcup_{i=1}^{n-1} \mathcal{E}_i$, and these partial correlations may be chosen in $]-1, 1[$ *arbitrarily*. This is the content of Corollary 7.5 in Bedford and Cooke (2002).

Theorem 1.2.5. *(Bedford, Cooke, 2002)*
For any regular vine on n elements, there is a one-to-one mapping between the set of $n \times n$ positive definite correlation matrices and the set of partial correlation specifications for the vine.

In other words, *any set* of $n(n-1)/2$ partial correlations that are deduced from a regular vine induce a true correlation matrix. Actually, the formulas (1.2.1) above enable to build such $n \times n$ correlation matrices based on $n(n-1)/2$ *arbitrarily chosen partial correlations* (see Kurowicka and Cooke, 2003, or Joe (2006)). For a given partial correlation vine, some explicit algorithms can be written to map the (usual) correlations and the underlying partial correlations: see Lewandowski et al. (2009). Such algorithms are available in the R-package called "vine-copula" (see Brechmann and Schepsmeier (2013), for instance).

Definition 1.2.6. Let a vine $V(n) = (T_1, T_2, \ldots, T_{n-1})$. The set of partial correlations associated to this vine is denoted by $\tilde{C}_{V(n)} := (C(T_1), C(T_2), \ldots, C(T_{n-1}))$. Denote by $R(\tilde{C}_{V(n)})$ the set of usual correlations that are deduced from $\tilde{C}_{V(n)}$.

Theorem 1.2.5 means that, whatever the values of the partial correlations $\tilde{C}_{V(n)}$ associated to a regular vine $V(n)$, we get a true correlation matrix with the coefficients $R(\tilde{C}_{V(n)})$. Since a standardized Gaussian random vector is fully specified by its correlation matrix, we obtain its joint law once we have chosen a partial correlation vine specification. At the opposite, for any Gaussian vector, there are many corresponding partial correlation vine specifications. In a Gaussian world, we recover the interpretation of vines as descriptors of random vector distributions. But more generally, partial correlation vine specifications can be associated to any random vector, just to describe its correlation matrix (when it exists).

We now turn to the significant results that ensure the positive definiteness of the correlation matrices when using vine representations. By recalling equation (1.2.1), the following result ensures that any correlation computed from arbitrary partial correlations (belonging to $]-1,1[$, obviously) is still an element in $]-1,1[$.

Lemma 1.2.7. *(Kurowicka, Cooke ,2006)*
If $z, x, y \in]-1,1[$, then also $w \in]-1,1[$ with

$$w = z\sqrt{(1-x^2)\,(1-y^2)} + xy.$$

The next theorem enables the easy generation of sequences of correlation matrices. It will constitute an attractive feature of the vine-GARCH models introduced in Section 1.3.

Theorem 1.2.8. *(Kurowicka, Cooke, 2006)*
Let $D_n > 0$ be the determinant of the n-dimensional correlation matrix $\Sigma_n := [\rho_{i,j}]_{i,j=1,\ldots,n}$. For any set of partial correlations generated by a regular vine,

$$D_n = \prod_{i=1}^{n-1} \prod_{e \in E_i} \left(1 - \rho^2_{j,k|L_e}\right),$$

where (j,k) and L_e are respectively the conditioned set and the conditioning set of an edge e.

Corollary 1.2.9. *Whatever the values of set of partial correlations generated by a regular vine on* $\{1,\dots,n\}$*, the associated matrix* $[\rho_{i,j}]$ *is nonnegative definite.*

Proof of Corollary 1.2.9. By Theorem 1.2.8, D_n is nonnegative whatever the values of the partial correlations in $\mathcal{P}_n := \{\rho_{j,k|L_e}\}$, that induce the correlations $\rho_{i,j}$, $i,j = 1,\dots,n$. But the same result applies for every matrix Σ_k, $k = 1,\dots,n-1$ too. Indeed, given $\mathcal{P}_n$, we are able to calculate all the $\rho_{i,j}$, $i,j = 1,\dots,n$ (that belong to $[-1,1]$ by Lemma 1.2.7), and then any set of partial correlations associated to any new vine on $\{1,\dots,k\}$, $k < n$ by invoking (1.2.1). And Theorem 1.2.8 can be applied to Σ_k. But a symmetrical matrix for which all the main block diagonal submatrices have nonnegative determinants is nonnegative. □

To illustrate these ideas, let us revisit Figure 1.1 under a partial correlation point of view: an associated partial correlation vine will specify the set of partial correlations $\left\{\rho_{12},\rho_{13},\rho_{14},\rho_{15},\rho_{23|1},\rho_{24|1},\rho_{25|1},\rho_{34|12},\rho_{35|12},\rho_{45|123}\right\}$, that is sufficient to recover the correlation matrix between the five assets. To interpret such numbers, we can consider linear regressions of some conditioned sets on their conditioning sets. For instance, the node $(1,2)$ and the node $(1,3)$ are connected, and the model will specify the partial correlation $\rho_{12|3}$. This is the correlation between the residuals of the linear regressions of r_2 and r_3 on r_1. Roughly, this measures to what extent r_2 an r_3 are "dependent" given r_1. In practical terms, an econometrician could classify the portfolio components by their (a priori) order of importance. This order may depend on the final phenomenon that is modelled. For instance, if the portfolio payoff depends strongly on emerging markets, it may be relevant to select "Russia" or "Brazil" first instead of "the USA". Intuitively, the latter strategy is intermediate between a factor model where we would regress any asset return on a few pre-specified ones, and a PCA where the factors are linear combinations of all returns.

This way of interpreting C-vines has to be revisited with D-vines or even general R-vines. Roughly, D-vines are based on an ordered vision of dependencies across asset returns: any asset is associated to one or two neighbors, with whom correlations are relatively strong. Once they are controlled, the main remaining risk is measured by the correlation with (one or) two other known assets, etc. Such a linear view of the strength of dependencies is probably unrealistic in finance. At the opposite, R-vines allow very general and flexible hierarchies and orders among the sequences of partial correlations

of interest. Virtually, they allow to integrate any a priori "prior" information, as long as it is consistent with the proximity condition.

For the sake of parsimony, it would be interesting to cancel (or to leave constant, at least) all partial correlations associated to a vine, after some given level r. When zero partial correlations are assumed after the latter level, we would like to know whether the corresponding (usual) correlations depend on the trees $T_r, T_{r+1}, \ldots, T_{n-1}$ that could be built above.

Definition 1.2.10. We say that a vine is r-VF (VF for vine-Free) if

$$R(C(T_1), C(T_2), \ldots, C(T_{n-1})) = R(C(T_1), C(T_2), \ldots, C(T_{r-1}), C(T'_r), \ldots, C(T'_{n-1})),$$

for any alternative vine $V'(n) := (T_1, T_2, \ldots, T_{r-1}, T'_r, \ldots, T'_{n-1})$, where the partial correlations associated to the edges of T'_k, $k \geq r$, are zero.

If a vine is r-VF, once the partial correlations are zero above the level r, the correlations are independent on the way this vine has been built from this level. This r-VF property actually holds for any R-vine. This is a consequence of Theorem 2.3 in Brechmann and Joe (2015). They observed that the density of an underlying Gaussian vector is not altered when choosing arbitrary trees $T_{r+1}, \ldots, T_{n-1}$ with associated zero partial correlations.

1.3 vine-GARCH correlation dynamics

1.3.1 The usual DCC-GARCH framework

When dealing with correlation dynamics, the Dynamic Conditional Correlation model (DCC) of Engle (2002) is probably the most commonly used approach, inside the MGARCH family. We denote by $(\epsilon_t)_{t=1,\ldots,T}$ a sequence of N-dimensional vectorial stochastic process, whose dynamics is specified by θ, a finite-dimensional parameter. Denote by $(\mathcal{F}_t)$ the natural filtration, i.e. $\mathcal{F}_t := \sigma(\epsilon_s, s \leq t)$ and $\mathbb{E}_{t-1}[X] := \mathbb{E}[X|\mathcal{F}_{t-1}]$ for any random quantity X. The key model assumption is

$$\epsilon_t \;=\; H_t^{1/2}(\theta)\,\eta_t, \tag{1.3.1}$$

where the series $(\eta_t)_{t\geq 1}$ is supposed to be a strong white noise s.t. $\mathbb{E}[\eta_t] = 0$ and $\mathrm{Var}(\eta_t) = I_N$. We suppose $H_t(\theta) := H_t := \mathrm{Var}_{t-1}(\epsilon_t)$ is a $N \times N$ positive definite matrix. At this stage, the model is semi-parametric. Its specification is complete when the law of η_t and the dynamics of $(H_t(\theta))$ are specified. In this paper, we focus on the latter point mainly.

The matrix H_t represents the unobserved time-dependent conditional covariance matrix of the process (ϵ_t). A brute-force inference of all model parameters seems unfeasible even when the dimension N is small. To avoid this problem, a common approach consists of splitting the problem into two simpler ones: modelling conditional volatilities on one side, the correlation dynamics on the other side. This is the key idea of DCC models that we detail now.

Denote by $h_{i,t}$ the conditional variances of $(\epsilon_{i,t})$ and $\rho_{ij,t}$ the conditional correlations between $\epsilon_{i,t}$ and $\epsilon_{j,t}$, for $i, j = 1, \ldots, N$, $i < j$. In matrix notation, $H_t = D_t R_t D_t$ where $D_t = \mathrm{diag}(h_{1,t}^{1/2}, \ldots, h_{N,t}^{1/2})$ is the diagonal matrix of the conditional volatilities, and $R_t = [\rho_{ij,t}]$ is the matrix of the conditional correlations. By construction, R_t is the conditional covariance matrix of the vector of the standardized returns $u_t = (u_{1,t}, \ldots, u_{N,t})$ with $u_{i,t} = \epsilon_{i,t}/\sqrt{h_{i,t}}$. Both volatility and correlation dynamics depend on a specific set of parameters given by $\theta = (\theta_v, \theta_c)' \in \Theta_v \times \Theta_c$, where θ_v (resp. θ_c) is the set of parameters determining the volatility processes (resp. correlation process).

Let us assume that, for every $i = 1, \ldots, N$ and t, there exists a function h_i s.t.

$$h_{i,t} = h_i\left(\theta_v^{(i)}; \epsilon_{i,t-1}, \ldots, \epsilon_{i,t-q_i}; h_{i,t-1}, \ldots, h_{i,t-p_i}\right) \tag{1.3.2}$$

for some positive integers p_i and q_i and some parameter $\theta_v^{(i)} \in \mathbb{R}^{p_i+q_i+1}$. Once stacked, the parameters $\theta_v^{(i)}$ provide θ_v. Typically, we could assume GARCH(p_i,q_i) processes in (1.3.2), or even other univariate GARCH-type models (EGARCH, GJR-GARCH, T-GARCH, etc). Since our vine-GARCH framework only needs consistent estimates of conditional volatilities, as deduced from this first stage, there is a large amount of liberty to specify the individual volatility dynamics.

Note that we have supposed no spill-over effects between different asset volatilities in Equation (1.3.2). This assumption simplifies the estimation of θ_v by allowing an equation-by-equation inference procedure, and it is almost unavoidable when N is large. This absence of spill-over effects is commonly used in the DCC literature, even

if it may be questionable. Indeed, some studies have exhibited significant spill-over effects empirically: see Hamao, Masulis and Ng (1990), Koutmos and Booth (1995), Liao and Williams (2004), among others. We stress that this point if not crucial for our vine-GARCH model, and this assumption could be removed: see Remark 1.4.12 below.

Several (R_t) dynamics have been proposed in the literature. All of them have to cope with the positive definiteness of the correlation matrix and should not depend on too many parameters. The time-varying correlation model of Tse and Tsui (2002) and the DCC model (Engle and Sheppard, 2001) were the first attempts to model dynamic correlations. In this study, we consider the latter as our benchmark.

The DCC model specifies dynamics of the covariance matrix of the de-garched returns u_t directly. In its full form, called "Full DCC", the model belongs to the MARCH family of Ding and Engle (2001) and is specified as

$$Q_t = (\iota\iota' - A - B) \odot S + A \odot u_{t-1}u_{t-1}' + B \odot Q_{t-1},\ R_t = Q_t^{\star -1/2} Q_t Q_t^{\star -1/2},$$

where $Q_t = [q_{ij,t}]$ and $Q_t^\star = \text{diag}(q_{11,t}, q_{22,t}, \ldots, q_{NN,t})$. Above, S, A and B denote $N \times N$ symmetric matrices of unknown parameters and $\odot$ is the usual Hadamard product of two identically sized matrices. Following Ding and Engle (2001), if $(\iota\iota' - A - B) \odot S$, A and B are positive semi-definite, then the matrix Q_t is positive semi-definite. The significant downside of the full DCC model is its intractability as the (Q_t) process encompasses $3N(N+1)/2$ coefficients. In most empirical studies, the scalar DCC-GARCH is considered instead, where A and B are replaced by non negative scalars α and β times the identity matrix.

Billio and Caporin (2006) devised the Quadratic Flexible DCC (QFDCC), which reduces the size of the problem while remaining flexible. In the general form of a QFDCC model, the correlation driving process (Q_t) is defined as

$$Q_t = C'SC + A'u_{t-1}u_{t-1}'A + B'Q_{t-1}B,\ R_t = Q_t^{\star -1/2} Q_t Q_t^{\star -1/2},$$

where S, A, B and C are unknown matrices, S being symmetric positive. This model allows for interdependence across groups of assets. The correlation matrices are positive definite if the eigenvalues of $A + B$ are less than one in modulus. This model is parsimonious when the matrices A, B and C are diagonal. This yields to a model with $3N$ unknown parameters, after correlation targeting.

The set of correlation parameters of the DCC is $\theta_c = (S, A, B)$, whereas for the QFDCC it is $\theta_c = (S, C, A, B)$. In the literature, DCC-GARCH models with correlation targeting are implemented generally by considering the matrix S as the unconditional covariance matrix of the standardized residuals. However, in the case of a scalar DCC, Aielli (2013) has shown that this procedure produces biased estimates in general and proposed a corrected version of the model called cDCC. Actually, the scalar DCC and cDCC specifications provide empirically very close results. Therefore, in our empirical study, we consider the scalar DCC and the diagonal QFDCC.

1.3.2 Our model specification

In a DCC-type model, one has to rely on intricate normalizations to build sequences of ϵ_t correlation matrices. This makes the interpretation of the (R_t) dynamics not intuitive, because it is deduced from another underlying process (Q_t). Another drawback of the DCC is the lack of parsimony because the number of parameters grows rapidly, as in general BEKK models. Most of the time, DCC ones are used in a scalar form, but this modeling often fails in capturing fine-tuned and heterogeneous correlation dynamics. In this paper, we develop a method that ensures both parsimony and positive definiteness without relying on any normalization.

The idea is based on the modeling of a set of partial correlations, which parameterizes any correlation matrix. We use a partial correlation vine specification, i.e. a given regular vine and $N(N-1)/2$ numbers in $]-1, 1[$ to specify the corresponding partial correlations. And we invoke the one-to-one mapping between these $N(N-1)/2$ partial correlations and the $N(N-1)/2$ "usual" correlations. The former are stacked in a vector Pc_t and the latter are the coefficients of R_t. We order partial correlations lexicographically, from the shortest to the longest sets of indices. Then we propose the following "partial correlation" dynamics

$$H_t = D_t R_t D_t, \tag{1.3.3}$$

$$\Psi(Pc_t) = \Omega + \sum_{k=1}^{p} \Xi_k \Psi(Pc_{t-k}) + \sum_{l=1}^{q} \Lambda_l \zeta_{t-l}, \tag{1.3.4}$$

$$R_t = \text{vechof}\left(F_{\text{vine}}(Pc_t)\right), \quad \text{where} \tag{1.3.5}$$

- The vector Pc_t is the "partial correlation vector" deduced from a given R-vine structure.

- vechof$(\cdot)$ denotes the operator "devectorization", that transforms a vector into a symmetric matrix. It is the opposite of the usual operator vech$(\cdot)$.

- The Ξ_k and Λ_l denote $N(N-1)/2 \times N(N-1)/2$ matrices of unknown parameters, and Ω is a $N(N-1)/2$ unknown vector. Set the vector of parameters $\theta_c = (\Omega, \Xi_1, \ldots, \Xi_p, \Lambda_1, \ldots, \Lambda_q)$. Note that these matrices are *arbitrarily chosen*, and we do not impose non negativeness, in particular.

- The vector ζ_{t-1} is $\mathcal{F}_{t-1}$-measurable and updates the selected partial correlations at time t. Such ζ_{t-1} must be built so that $\mathbb{E}[\zeta_{t-1}] \simeq \mathbb{E}[Pc_{t-1}]$. This procedure is in line with usual updating equations in GARCH-type models.

- We apply a deterministic transformation $\Psi(.)$ to Pc_t. It twists the univariate dynamics to manage the constraint that partial correlations stay in $(-1, 1)$. For the sake of simplicity, $\Psi(.)$ will be known [1]. To fix the ideas, Ψ is defined from $]-1,1[^{N(N-1)/2}$ to $\mathbb{R}^{N(N-1)/2}$ as

$$\Psi(Pc_t) = \left(\psi(\rho_{1,2,t}), \ldots, \psi(\rho_{N,N-1|L_{N-1,N},t})\right)', \ \psi(x) = \tan(\pi x/2).$$

  Alternatively, $\Psi(.)$ could be chosen among the sigmoïd functions for instance, for which $\psi(x) = (\exp(\alpha x) - 1)/(\exp(\alpha x) + 1)$ for some $\alpha \in \mathbb{R}$.

- The function $F_{\text{vine}}(.)$ corresponds to the one-to-one mapping from the vector of partial correlations Pc_t to correlations (in R_t) by using the algorithm of Lewandowski et al. (2009). It is defined from $]-1,1[^{N(N-1)/2}$ to itself by

$$F_{\text{vine}}\left(\rho_{1,2,t}, \ldots, \rho_{N-1,N|L,t}\right) = (\rho_{1,2,t}, \ldots, \rho_{N-1,N,t})'.$$

Partial correlations are expectations of products of the two different quantities $\upsilon_{k|L,t}$, for some $L \subset \{1, \ldots, N\}$ and $k \notin L$, which are defined as

$$\upsilon_{k|L,t} = \frac{\epsilon_{k,t} - \mathbb{E}_{t-1}\left[\epsilon_{k,t}|\epsilon_{L,t}\right]}{\sqrt{h_{k|L,t}}},$$

where $\epsilon_{L,t} = (\epsilon_{i,t})_{i \in L}$, and $\mathbb{E}_{t-1}\left[\epsilon_{k,t}|\epsilon_{L,t}\right]$ corresponds to the orthogonal projection of the variable $\epsilon_{k,t}$ on the space spanned by the vector $\epsilon_{L,t}$. The variance of the "residual" $\epsilon_{k,t} - \mathbb{E}_{t-1}\left[\epsilon_{k,t}|\epsilon_{L,t}\right]$ is denoted by $h_{k|L,t}$. The variables $\upsilon_{k|L,t}$ are not observable, but

[1] There is no doubt the methodology could be extended to deal with a parametric function Ψ, i.e. that would depend on an unknown finite-dimensional additional parameter. Nonetheless, this would complicate the proofs in Section 1.6, while it is not a key point here. Such an extension is left to the reader.

we can evaluate $\mathbb{E}_{t-1}[\epsilon_{k,t}|\epsilon_{L,t}]$ and $h_{k|L,t}$ to get $\hat{v}_{k|L,t}$, an approximated value of $v_{k|L,t}$. Then, *by construction*, the $N(N-1)/2$-sized vector ζ_t will stack the variables $\hat{v}_{i|L,t}\hat{v}_{j|L,t}$, when $(i,j|L)$ is an edge of the underlying vine. The order of these edges in ζ_t will be the same as for Pc_t.

By definition, Equations (1.3.1)-(1.3.5) define a so-called vine-GARCH(p,q) model.

In full generality, this simplified version of the vine-GARCH(p, q) model still encompasses $(p+q+1)N(N-1)/2$ parameters. However, this approach can become easily more parsimonious and would provide a nice alternative to full DCC-GARCH models. Indeed, since the r-VF property applies, one can set constraints to any level of the tree (say r), and choose zero partial correlations at and after the r-th tree in the underlying vine. We guess this should not modify significantly the (true) correlation dynamics, at least when r is large enough. This is due to the fact that partial correlations with non-empty conditioning subsets are correlations between residuals. In practice it is likely that these residuals tend to behave more and more as white noise when the number of conditioning variables increases and for a well-chosen R-vine. By canceling partial correlations after the step r, we get a particular model with less parameters than in the full vine-GARCH specification. And whatever the chosen structure of the vine is after level r, the reconstruction formulas (1.2.1) provide the same correlation matrices. This is a nice theoretical property. A slightly different simplification of our vine-GARCH models would be to assume constant (non zero) partial correlations after some level (say r) in the vine. But in this case, we cannot ensure a similar r-VF property.

Remark 1.3.11. Obviously, alternative dynamics could extend our vine-GARCH(p,q) specification (1.3.4). For instance, it could be possible to modify the model to include nonlinear features as asymmetries, switching regimes, time-varying parameters, exogenous variables, etc. A whole class of models is now open, based on partial correlations, exactly as the original GARCH framework of Bollerslev (1986) has been modified and revisited.

At time t, the vector ζ_t is a key information as it drives the shocks on the partial correlation processes. Here, we propose two ways of evaluating $\hat{v}_{k|L,t}$, and then ζ_t.

The first method is based on the linear regression of ϵ_t on $\epsilon_{L,t}$:

$$\epsilon_{k,t} = \alpha_{k|L} + \beta'_{k|L}\epsilon_{L,t} + \xi_{L,t},\ \mathbb{E}[\xi_{L,t}|\epsilon_{L,t}] = 0.$$

Then, we approximate $\epsilon_t - \mathbb{E}[\epsilon_t|\epsilon_{L,t}]$ by $\epsilon_t - \hat{\alpha}_{k|L} - \hat{\beta}'_{k|L}\epsilon_{L,t}$ and an empirical "rolling-window" estimator of $h_{k|L,t}$ can be defined by $\hat{h}_{k|L,t} := m^{-1}\sum_{i=1}^{m}(\epsilon_{k,t-i} - \hat{\alpha}_{k|L} - \hat{\beta}'_{k|L}\epsilon_{L,t-i})^2$, for some windows size m. Such size should increase with T in theory but trying to exhibit some "optimal" m is beyond the scope of the present work. We get $\hat{v}_{k|L,t} = (\hat{\alpha}_{k|L} + \hat{\beta}'_{k|L}\epsilon_{L,t})/\sqrt{\hat{h}_{k|L,t}}$, and then ζ_t. This approach may be termed "non parametric" in the sense that it does not rely on any hypothesis about the conditional distribution of ϵ_t.

The second method is based on the theoretical distribution of the residuals ϵ_t given $\mathcal{F}_{t-1}$, that is unknown at this stage. In accordance with our Gaussian QMLE, assume the latter distribution is elliptical. Then, its first two conditional moments can be calculated easily. Indeed, if a vector $(X, Y)'$ is elliptical with $\Sigma_{XX} = \mathrm{Var}(X)$, $\Sigma_{YY} = \mathrm{Var}(Y)$, $\Sigma_{XY} = \Sigma'_{YX} = \mathrm{Cov}(X, Y)$, then $\mathbb{E}[X|Y] = \mathbb{E}[X] + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mathbb{E}[Y])$ and $Var(X|Y) = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$: see Corollary 5 in Cambanis, Huang and Simons (1981). Hence we can calculate easily $\hat{v}_{k|L,t} = v_{k|L,t} = (\epsilon_{k,t} - \mathbb{E}_{t-1}[\epsilon_{k,t}|\epsilon_{L,t}])/\sqrt{h_{k|L,t}}$. To be specific, under these assumptions, we write

$$\mathbb{E}_{t-1}[\epsilon_{k,t}|\epsilon_{L,t}] = \mathrm{Cov}_{t-1}(\epsilon_{k,t}, \epsilon_{L,t})\,\mathrm{Var}_{t-1}(\epsilon_{L,t})^{-1}\,\epsilon_{L,t},$$

$$\begin{aligned} h_{k|L,t} &= \mathrm{Var}_{t-1}(\epsilon_{k,t} - \mathbb{E}_{t-1}[\epsilon_{k,t}|\epsilon_{L,t}]) \\ &= \mathrm{Var}_{t-1}(\epsilon_{k,t}) - \mathrm{Cov}_{t-1}(\epsilon_{k,t}, \epsilon_{L,t})\,\mathrm{Var}_{t-1}(\epsilon_{L,t})^{-1}\,\mathrm{Cov}_{t-1}(\epsilon_{L,t}, \epsilon_{k,t}), \end{aligned}$$

and the latter conditional covariances are $\mathcal{F}_{t-1}$ measurable, i.e. are known at t. In the theoretical part of the paper (Section 1.5, Section 1.6), this second method of calculation of ζ_t is used because the innovations η_t are assumed to be elliptical. In the empirical part (Section 1.7), this is the case too but only for convenience (higher speed of calculations).

1.3.3 Vine selection

The methodology above can be applied to any R-vine on N elements. Actually, the structure of the underlying R-vine may be seen as an additional parameter, independently of θ_c. Selecting a convenient R-vine may be useful to describe the dependence among the variables in a parsimonious and meaningful way. In particular, this would allow for the truncation of a given R-vine, once some important factors have been found in the first trees.

To do so with a C-vine, we can follow the sequential method developed by Dissmann, Brechmann, Czado and Kurowicka (2012). This method consists in starting by computing the Kendall's tau of all the couples of nodes, and selecting the variable, which induces the highest degree of dependence with the other ones. In the second tree, we compute a Kendall's tau per edge, but conditional on the variable chosen on the first tree. That is:

1. For tree T_1 and $N_1 = \{1, \cdots, N\}$, maximize the dependence criteria:

$$i_0 \leftarrow \arg\max_i \sum_{j \neq i} |\hat{\tau}_{ij}|,$$

   where $\hat{\tau}_{ij}$ is the empirical Kendall's tau and i_0 denotes the index of the variable, which maximizes this criterion. This variable is the root to build the edges on tree T_1, which are the nodes on tree T_2.

2. For $j = 2, \cdots, N-2$, $D_1 = p_1 = i_0$, maximize the dependence criteria:

$$p_j \leftarrow \arg\max_k \sum_{j \neq p_{j-1}, j \neq k} |\hat{\tau}_{jk|D_{j-1}}|,$$

   where $D_j = D_{j-1} \cup p_j$.

This sequential approach provides step-by-step the variable which should enter the conditioning set for the next tree.

We use nonparametric statistics proposed by Veraverbeke, Omelka and Gijbels (2011) to compute these quantities. We apply the same selection criteria to choose the convenient variable and proceed with the next trees similarly, until the last tree. The Kendall's tau is used as a dependence measure because it can be easily estimated, but other dependence measures are possible. This selection procedure is "bottom-up". Alternative methodologies exist, in particular the "top-down" procedure of Kurowicka (2011).

1.4 Statistical inference by QML

We can estimate vine-GARCH(p,q) models by maximizing a likelihood function that does not correspond to the true Data Generating Process necessarily, following the

Quasi-Maximum Likelihood (QML) methodology, as explained in Gourieroux, Monfort and Trognon (1984), Bollerslev and Wooldridge (1994) or White (1994), among others.

1.4.1 The QML estimator

We choose a standard Gaussian QML estimator: we do a MLE as if (η_t) were a Gaussian white noise, but for inference purpose only. Obviously, the "true" underlying distributions of these innovations may be different. Note that the η_t-law can be estimated empirically a posteriori from a sample of residuals $R_t(\hat{\theta})^{-1/2}\epsilon_t$. Using the assumed independence of the innovations η_t and developing H_t as $D_tR_tD_t$, the quasi-likelihood function of a path $(\epsilon_t)_{t=1,\dots,T}$ is written as

$$\begin{aligned}\mathcal{L}_T(\theta;\epsilon) &= \prod_{t=1}^{T}\exp\left\{-\frac{1}{2}\left(N\log(2\pi)+\log\left(|D_tR_tD_t|\right)+\epsilon_t'D_t^{-1}R_t^{-1}D_t^{-1}\epsilon_t\right)\right\}\\ &= \prod_{t=1}^{T}\exp\left\{-\frac{1}{2}\left(N\log(2\pi)+\log\left(|D_t^2|\right)+\epsilon_t'D_t^{-2}\epsilon_t-u_t'u_t+\log\left(|R_t|\right)+u_t'R_t^{-1}u_t\right)\right\},\end{aligned}$$

where $D_t = \mathrm{diag}(h_{1,t}^{1/2},\dots,h_{N,t}^{1/2})$, and $u_t = (\epsilon_{1,t}/h_{1,t}^{1/2},\dots,\epsilon_{N,t}/h_{N,t}^{1/2})' = D_t^{-1}\epsilon_t$ is the vector of GARCH standardized residuals. Thus, the quasi-log-likelihood function is the sum of two parts: the "variance part" of the likelihood, that depends on θ_v, and the "correlation part", that depends on both θ_v and θ_c. Therefore, our estimate $\hat{\theta}_{T,v}$ of θ_v is

$$\hat{\theta}_{T,v} = \arg\min_{\theta_v}\ \mathbb{G}_T l_1(\epsilon;\theta_v) := \frac{1}{T}\sum_{t=1}^{T} l_{1,t}(\theta_v;\epsilon_t) := \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\log(h_{i,t})+\frac{\epsilon_{i,t}^2}{h_{i,t}}\right]. \qquad (1.4.1)$$

The Newton-Raphson method is applied to solve such system. Note that $\hat{\theta}_{T,v}$ determines the (now estimated) variance processes $(h_{i,t})$ and then the (estimated) residuals u_t, denoted by $\hat{u}_t$. Given $\hat{\theta}_{T,v}$, a QML estimator of θ_c is obtained as

$$\hat{\theta}_{T,c} = \arg\min_{\theta_c}\ \mathbb{G}_T l_2(\epsilon_t;\hat{\theta}_{T,v},\theta_c) := \frac{1}{T}\sum_{t=1}^{T} l_{2,t}(\epsilon_t;\hat{\theta}_{T,v},\theta_c) := \sum_{t=1}^{T}\left[\log(|R_t|)+\hat{u}_t'R_t^{-1}\hat{u}_t\right]. \qquad (1.4.2)$$

Strictly speaking, all the likelihood equations above depend on the initial values ϵ_0, D_0 and R_0. To fix the ideas, we propose to initialize them by their sample counterparts: for all $i = 1,\cdots,N$, set $\epsilon_0 = 0$, $\tilde{h}_{i,0} = \frac{1}{T-1}\sum_{t=1}^{T}\epsilon_{i,t}^2$, $\tilde{D}_0 = \mathrm{diag}(\tilde{h}_{1,0}^{1/2},\dots,\tilde{h}_{N,0}^{1/2})$, and $\tilde{R}_0$ is

the empirical correlation matrix of the sample path $(\epsilon_1, \dots, \epsilon_T)$. To obtain convergence of $\hat{\theta}_T$, we will need the asymptotic irrelevance of these initial values (see Section 1.6).

Remark 1.4.12. The absence of volatility spill-over effects allows for the estimation of θ_v through N simple optimizations independently. Obviously, if we remove this assumption, such an estimator can still be obtained by (1.4.1). But, in general, this would require an optimization in a high-dimensional space, a task that becomes harder and harder with N.

Remark 1.4.13. It is possible to choose another QML parametric family that would be more adapted to fat tailed distributions typically (for instance the multivariate Student law, or any elliptical distribution). But then, we would loose the nice property of a two-stage estimation procedure, that is so important in practice.

1.4.2 Estimation strategy

Unfortunately, the underlying process (R_t) induces tricky computations of scores and Hessians for $\mathbb{G}_T l_2$. This is the case for both DCC and vine-Garch dynamics. Here, we propose two strategies depending on the dimensionality of the problem.

In this study, our DCC specifications are not highly parameterized: the scalar DCC (resp. diagonal QFDCC) requires the estimation of 3 (resp. $3N$) parameters, after correlation targeting. Consequently, the Sequential Quadratic Programming method is implemented for these dynamics, since it is well-suited for constrained optimization with a "reasonable" number of parameters.

As the general DCC model, the vine GARCH specification may suffer from the curse of dimensionality. However, when the matrices of parameters Ξ_j and Λ_k are diagonal (a usual situation), it is possible to weaken drastically this problem by proceeding sequentially. Indeed, in partial correlation R-vines, any partial correlation on tree T_k can be updated (through the ζ_t quantities) easily knowing the partial correlations on the previous trees $T_{k'}$, $k' \leq k-1$ [2].

Let us detail the sequential procedure for a C-vine, w.l.o.g. Instead of relying on a brute-force optimization in high dimension, a vine-GARCH model based on a C-vine

[2] To be specific, at a given node $(ij|L)$ of a R-vine, we need to calculate the $\zeta_{(ij|L),t-l}$, $l = 1, \dots, q$. As explained in 1.3.2, this necessitates the calculation of conditional covariances of the subvectors associated to the indices (iL) and (jL). Since L is the conditioning subset at this node, this is always possible once we have evaluated all node dynamics associated to the previous trees.

may be estimated by solving $N \times (N-1)/2$ simple optimization programs, related to the bivariate dynamics that are associated to any node. This means we estimate successively the dynamics of $(\epsilon_{i,t}, \epsilon_{j,t})$ where the $N \times (N-1)/2$ couples (i,j) describe the conditioned subsets of all the nodes in the underlying C-vine, starting from the bottom tree.

To be even more explicit, denote the nodes of the vine by $\{(ij|L)\}$, and the unknown matric parameters as $\Omega := [\omega_{(ij|L)}]$, $\Xi_k := \text{diag}(\xi_{(ij|L);k})$, $k = 1, \ldots, p$ and $\Lambda_l := \text{diag}(\lambda_{(ij|L);l})$, $l = 1, \ldots, q$. Assume the underlying C-vine is given in Figure 1.1. In particular, 1 is the root in the first tree. The $N-1$ first partial correlation dynamics are "usual" correlation processes and depend on the estimated volatility and the observations. The parameters of these $N-1$ first processes can be minimized *independently* based on the objective functions

$$\mathbb{G}_T l_2^{1j}(\epsilon, \hat{\theta}_{T,v}; \theta_{c,1j}) = \sum_{t=1}^{T} \left[\log |R_{(1j),t}| + \hat{u}'_{(1j),t} \left(R_{(1j),t}\right)^{-1} \hat{u}_{(1j),t} \right], \; j = 2, \ldots, N,$$

where $\hat{u}_{(1j),t} = [\hat{u}_{1,t}, \hat{u}_{j,t}]'$ and $R_{(1j),t}$ is the 2×2 correlation matrix of $(\epsilon_{1,t}, \epsilon_{j,t})$ given $\mathcal{F}_{t-1}$. With obvious notations, $\theta_{c,1j} = \left(\omega_{1j}, \boldsymbol{\xi}_{1j}, \boldsymbol{\lambda}_{1j}\right)$ are the remaining unknown parameters that are associated to the bivariate process $(\epsilon_{1,t}, \epsilon_{j,t})$.

Now, after conditioning by 1, there are $N-2$ dynamic partial correlations in T_2. Due to (1.7.2), they follow the ARMA-type dynamics

$$\psi\left(\rho_{2j|1,t}\right) = \omega_{2j|1} + \sum_{k=1}^{p} \xi_{2j|1;k} \psi\left(\rho_{2j|1,t-k}\right) + \sum_{l=1}^{q} \lambda_{2j|1;l} \hat{v}_{2|1,t-l} \hat{v}_{j|1,t-l}, \; j = 3, \ldots, N.$$

For QML inference purpose, we assumed $\epsilon_t | \mathcal{F}_{t-1} \sim \mathcal{N}(0, H_t)$. As explained in Subsection 1.3.2, $\hat{v}_{k|1,t-1}$ above depends on the volatility processes, the observations and the correlations calculated from tree T_1. Hence, we can estimate the partial correlations dynamics on tree T_2 by maximizing $N-2$ objective functions independently over each correlation parameter space of tree T_2, given the estimated correlations on T_1. The objective functions are, for all $j = 3, \ldots, N$

$$\mathbb{G}_T l_2^{2j|1}(\epsilon, \hat{\theta}_{T,v}, \hat{\rho}_{12}, \hat{\rho}_{1j}; \theta_{c,2j|1}) = \sum_{t=1}^{T} \left[\log |R_{(2j),t}| + \hat{u}'_{(2j),t} \left(R_{(2j),t}\right)^{-1} \hat{u}_{(2j),t} \right].$$

Here, $R_{(2j),t}$ is the correlation matrix of $(\epsilon_{2,t}, \epsilon_{j,t})$ given $\mathcal{F}_{t-1}$. Its coefficient $\rho_{2j,t}$ is computed from the estimated dynamic partial correlations $\hat{\rho}_{2j|1,t}$ and the (estimated) correlations $\hat{\rho}_{1l,t}$, $l = 2, \ldots, N$. Obviously, $\hat{u}_{(j,k),t} = [\hat{u}_{j,t}, \hat{u}_{k,t}]'$.

We apply the same reasoning for the next trees in the C-vine. There are $N-3$ objective functions to be maximized on tree T_3, $N-4$ on tree T_4, etc, until tree T_{N-1} where only one objective function needs to be maximized. The estimation of any partial correlation process of a tree T_k depends only on a subset of partial correlations associated to the nodes of T_{k-1} and before, invoking the recursive formula (1.2.1). Consider any node $(ij|L)$ in T_k and denote by $\theta_{c,ij|L} = \left(\omega_{ij|L}, \boldsymbol{\xi}_{ij|L}, \boldsymbol{\lambda}_{ij|L}\right)$ the associated subvector of θ_c. For instance, with our C-vine of Figure 1.1, L does not depend on the conditioned subsets and is $L := L_i = \{1, \ldots, i-1\}$, $k = 2, \ldots, N-1$. Our iterative algorithm can be summarized as

$$\hat{\theta}_{T,c,ij|L_i} = \underset{\theta_{c,ij|L_i}}{\arg\min}\, \mathbb{G}_T l_2^{ij|L_i}(\hat{\theta}_{T,v}, \epsilon, \hat{\rho}_{i-1,i|L_{i-1}}, \hat{\rho}_{i-1,j|L_{i-1}}; \theta_{c,ij|L_i}),$$

for every i and j in $\{1, \ldots, N\}$, $i < j$.

We denote this strategy *C-vine (D-vine, or even R-vine) iterative process*, which is particularly effective when N becomes "large" (say larger than 5 assets). At each node on a specific level, only $(p+q+1)$ parameters need to be estimated. Consequently, we also use the Sequential Quadratic Programming method when estimating the C-vine iterative process.

A drawback of the latter iterative process may be the propagation of estimation errors from one partial correlation level to the next one. It is still possible to estimate the vine-GARCH at once for reasonable portfolio sizes ($N \leq 5$) to avoid this iterative method. But the nonlinearity and the instability of the likelihood function in the vine-GARCH case require another approach to maximize $\mathbb{G}_T l_2(\epsilon; .)$. In such a case, we propose to use a stochastic algorithm, the simulated annealing, that prevents from falling in local maxima. Note that the simulated annealing algorithm can also be used when estimating model through the previous iterative methodology. However in this case, the Sequential Quadratic Programming is a lot quicker, which is the reason we used this method in the simulation study.

1.5 On the stationarity of the vine-GARCH process

We prove the existence of stationary solutions, which is the first step towards providing asymptotic results (consistency/asymptotic normality of QML estimates), because law of large numbers (potentially uniform) and some Central Limit Theorems are obtained easily in this case. In the GARCH literature, proving stationarity properties has been fulfilled notably by Bougerol and Picard (1992) for univariate GARCH models, by Ling and McAleer (2003) for multivariate ARMA-GARCH models, by Boussama et al. (2011) for BEKK models, notably.

After introducing some notations, we specify the vine-GARCH model. It is rewritten as an "almost linear" Markov chains in Subsection 1.5.2. The existence of strong and weak stationary solutions is stated in Subsection 1.5.3. Subsection 1.5.4 exhibits sufficient conditions to get their uniqueness. These probabilistic results are established for the $p = q = 1$ case.

1.5.1 Notations

Let $A \in \mathcal{M}_{n\times m}(\mathbb{R})$.

- If $n = m$, then $\text{diag}(A) = (a_{ij}\mathbf{1_{i=j}})_{\mathbf{1\le i\le m, 1\le j\le m}}$ and $\text{Vecd}(A) = (a_{ii})_{1\le i\le m} \in \mathbb{R}^m$.
- If $n = m$ and A symmetric, $\text{Vech}(A) \in \mathbb{R}^q$ with $q = m(m+1)/2$ such that the components are those of A column-wise without redundancy.
- If $n = m$, then $\rho(A)$ is the spectral radius of A, that is the largest of the modulus of the eigenvalues of A. We denote $\lambda_1(A)$ the smallest eigenvalue of A positive definite.
- The Kronecker product is denoted $\otimes$ and $A^{\otimes k} = A \otimes A \otimes \cdots \otimes A$ (k times). The Hadamard product is denoted $\odot$.
- In the following, we consider the submultiplicative norm

$$\|A\| := \sup\{\frac{\|Ax\|}{\|x\|},\ x \neq 0\},$$

where $x \in \mathbb{R}^m$ and $\|x\|$ is the Euclidean norm of vector x. For $B \in \mathcal{M}_{m\times n}(\mathbb{R})$, this norm satisfies

$$\|AB\| \leq \|A\|\|B\|, \quad \text{Trace}\,(AB) \leq (nm)^{1/2}\|A\|\|B\|.$$

We define the spectral radius norm for squared non-negative matrices, which is submultiplicative, as

$$\|A\|_s := \sup\{\sqrt{\lambda} : \lambda \in \text{Spect}\,(A'A)\}.$$

We also define the maximum absolute column sum of a matrix $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ as

$$\|A\|_\infty = \max_i \sum_j |A_{ij}|.$$

- For a N dimensional vectorial process $(\epsilon_t)_t$, we denote $\epsilon_t = (\epsilon_{1,t}, \cdots, \epsilon_{N,t})'$ and $\vec{\epsilon_t} := (\epsilon^2_{1,t}, \cdots, \epsilon^2_{N,t})'$.
- We denote by $C^0_b(\mathbb{E})$ the space of all continuous and bounded functions $f : \mathbb{E} \to \mathbb{R}$.

The quantity of interest is H_t, which is split between volatility terms contained in D_t and correlation terms in R_t as

$$H_t = D_t R_t D_t, \tag{1.5.1}$$

where $D_t = \text{diag}(\sqrt{h_{11,t}}, \cdots, \sqrt{h_{NN,t}})$ is the diagonal matrix of the conditional variances, which is $\mathcal{F}_{t-1}$ measurable. The $\mathcal{F}_{t-1}$ measurable (D_t) process contains components supposed to be univariate GARCH dynamics without cross-effects, such that

$$\text{Vecd}(D^2_t) = V + A.\text{Vecd}(D^2_{t-1}) + B.\vec{\epsilon}_{t-1}, \tag{1.5.2}$$

where the matrices A and B are diagonal and V is a positive vector of R^N.

The vine-GARCH specification parametrizes the correlation dynamics as

$$\begin{array}{lcl} R_t & = & \text{vechof}\,(F_{\text{vine}}\,(Pc_t))\,, \\ \Psi\,(Pc_t) & = & \Omega + \Xi\Psi\,(Pc_{t-1}) + \Lambda\zeta_{t-1}, \end{array} \tag{1.5.3}$$

In this section, we specify the Data Generating Process (DGP) differently from the specification given in (1.3.1). A significant quantity is the vector of standardized residuals, defined as $u_t = D_t^{-1}\epsilon_t$. We straightforwardly have $\mathbb{E}_{t-1}[u_t] = 0$ and $\mathbb{E}_{t-1}[u_t u_t'] = R_t$. This implies that u_t can be specified as $u_t = R_t^{1/2}\eta_t^*$, such that η_t^* is a centered random vector with $\mathbb{E}_{t-1}[\eta_t^*\eta_t^{*\prime}] = I_N$. Therefore, the "true" DGP will be the stationary process (η_t^*). The two "innovations" (η_t) and (η_t^*) are related to each other by the relation

$$H_t^{1/2}\eta_t = D_t R_t^{1/2}\eta_t^*.$$

Note that, if $\mathbb{E}_{t-1}[\eta_t^*] = 0$ and $\mathbb{E}_{t-1}[\eta_t^*\eta_t^{*\prime}] = I_N$, then $\mathbb{E}_{t-1}[\eta_t] = 0$ and $\mathbb{E}_{t-1}[\eta_t\eta_t'] = I_N$, and the opposite.

1.5.2 vine-GARCH as Markov Chains

The vine-GARCH specification can be written as a Markov chain, a representation that is relevant for studying stationary solutions. To do so, we define

$$X_t := \left(\vec{\epsilon}_t, \text{Vecd}(D_t^2), \Psi\left(Pc_t\right)\right)', \tag{1.5.4}$$

such that, for all $t > 0$, $(X_t)_t$ satisfies

$$X_t = T_t X_{t-1} + \nu_t. \tag{1.5.5}$$

This means $(X_t)_t$ follows an autoregressive form of order 1 with stochastic T_t. Let us focus on the first component of X_t. Setting $\vec{u}_t := (u_{1,t}^2, \ldots, u_{N,t}^2)$, we have

$$D_t^2\vec{u}_t = \vec{u}_t \odot \text{Vecd}(D_t^2) = \vec{\epsilon}_t = \vec{u}_t \odot V + \vec{u}_t \odot A.\text{Vecd}(D_{t-1}^2) + \vec{u}_t \odot B.\vec{\epsilon}_{t-1}. \tag{1.5.6}$$

Using the dynamics of $\text{Vecd}(D_t^2)$ and $\Psi\left(Pc_t\right)$, the matrix T_t satisfies

$$T_t = \begin{pmatrix} \vec{u}_t \odot B & \vec{u}_t \odot A & 0 \\ B & A & 0 \\ 0 & 0 & \Xi \end{pmatrix}, \tag{1.5.7}$$

and the vector of innovation ν_t is defined as

$$\nu_t = \begin{pmatrix} \vec{u}_t \odot V \\ V \\ \Omega + \Lambda\zeta_{t-1} \end{pmatrix}. \tag{1.5.8}$$

Note that $\zeta_t = \zeta(\chi_t, \eta_t)$ where $\chi_t = (Pc_t, D_t)$.

Assumption 1. The vectorial process $(\eta_t^*)_{t\in\mathbb{Z}}$ satisfies the Markov property with respect to $\mathcal{F}$, i.e

$$\forall t \in \mathbb{Z},\ \mathbb{E}[\eta_t^*|\mathcal{F}_{t-1}] = \mathbb{E}[\eta_t^*|X_{t-1}].$$

Besides, $\mathbb{E}_{t-1}[\eta_t^*] = 0$ and $\mathbb{E}_{t-1}[\eta_t^*\eta_t^{*\prime}] = I_N$.

As a consequence (and equivalently, in fact), the same property is fulfilled with the other "innovations" $(\eta_t)_{t\in\mathbb{Z}}$: the process $(\eta_t)_{t\in\mathbb{Z}}$ satisfies the Markov property with respect to $\mathcal{F}$, i.e

$$\forall t \in \mathbb{Z},\ \mathbb{E}[\eta_t|\mathcal{F}_{t-1}] = \mathbb{E}[\eta_t|X_{t-1}].$$

Moreover, $\mathbb{E}_{t-1}[\eta_t] = 0$ and $\mathbb{E}_{t-1}[\eta_t\eta_t'] = I_N$.

Proposition 1.5.14. *Under assumption 1, $(X_t)_t$ is a Markov Chain of order one.*

Proof of Proposition 1.5.14. Note that $u_t = D_t^{-1}H_t^{1/2}\eta_t$, where H_t is a deterministic function of X_{t-1}. Since η_t satisfies the Markov property with respect to $\mathcal{F}$, then $u_t|\mathcal{F}_{t-1} \overset{d}{=} u_t|X_{t-1}$. Furthermore, X_t can be rewritten as follows: there exists constant matrices Γ_1 and Γ_2 such that

$$X_t = (\Gamma_1.\xi_t) \odot T_0 X_{t-1} + (\Gamma_2.\vec{\chi}_t) \odot \nu_0,$$

where T_0 (resp. ν_0) is the T_t (resp. ν_t) matrix when $u_t = 1$, $\xi_t := (\vec{u}_t, 1)'$ and $\vec{\chi}_t := (\vec{u}_t, 1, \zeta(\chi_{t-1}, \eta_{t-1}))'$. Then X_t is a measurable function of $(\eta_t, X_{t-1}, \eta_{t-1})$, where η_t satisfies the Markov property by assumption 1. Consequently, $(X_t)_t$ is Markovian.

□

1.5.3 Existence of stationary vine-GARCH solutions

The recurrence equation (1.5.5) is stochastic through T_t and ν_t, i.e. through the innovations η_t (or η_t^*) and the $\mathcal{F}_{t-1}$-measurable matrix R_t. A consequence of this

parametrization is that T_t depends on subcomponents of X_t. Hence, we can not extract an expression such as $X_t = f(\eta_t, \eta_{t-1}, \cdots)$ nor $X_t = f\left(\eta_t^*, \eta_{t-1}^*, \cdots\right)$, for some explicit function $f(.)$. This comes from the nonlinear relationship between T_t and the past innovations (before and including t). Classical techniques such as Lyapunov exponent are not adapted in our framework.

The existence of stationary solutions -but not a unique solution- for the vine-GARCH model can be proved using the criterion of Tweedie (1988). Tweedie provides the existence of an invariant probability measure for the Markov chain defined in (1.5.5). Ling and McAleer (2003) used this criterion to establish the stationarity of vector ARMA-GARCH models.

The stationarity of the $(\vec{\epsilon}_t)_t$ process requires the control of T_t, which should avoid non-explosive patterns. The matrix T_t is a function of $(\vec{u}_t)_t$, which are dependent variables. Furthermore, the conditional law of $\vec{u}_t$ is a function of H_t and D_t, which in turn is a function of X_{t-1}. This is the reason we need the next hypothesis.

Assumption 2. For some $p \geq 1$, $\|T^\star\|_s < \infty$, where

$$T^\star := \sup_{\mathbf{x}\in\mathbb{R}^{\mathbf{d}}} \mathbb{E}\left[|T_t^{\otimes p}||X_{t-1} = \mathbf{x}\right].$$

Assumption 3. Denoting by λ the Lebesgue measure, the conditional kernel of η_t^* given $X_{t-1} = \mathbf{x}$ is defined as

$$d\mathbb{P}_{\eta_t^*}^{X_{t-1}=\mathbf{x}}(u) = f_{\eta_t^*}(u|\mathbf{x})\mathbf{d}\lambda(\mathbf{u}).$$

Furthermore, for all $u \in \mathbb{R}^m$, the mapping $\mathbf{x} \to \mathbf{f}_{\eta_t^*}(\mathbf{u}|\mathbf{x})$ is continuous and there exists an integrable function g such that, for all $u \in \mathbb{R}^m$,

$$\sup_t \sup_{\mathbf{x}\in\mathbb{R}^{\mathbf{d}}} f_{\eta_t^*}(u|\mathbf{x}) \leq \mathbf{g}(\mathbf{u}).$$

Moreover, $\forall t$, $\mathbb{E}\left[\|\eta_t^*\|^{2p}|X_{t-1} = \mathbf{x}\right] \leq \psi(\|\mathbf{x}\|)$ satisfying $\forall \alpha > 0$, $\lim_{v\to\infty} \dfrac{\psi(v)}{v^\alpha} = 0$.

Assumption 4. There exists a positive real number a such that, for almost every trajectory and every $\theta \in \Theta$, the partial correlations of our chosen vine (i.e. the components of the vectors $Pc_t(\theta)$) belong to the fixed interval $[-1 + a, 1 - a]$.

In particular, the latter assumption implies that, for every $\theta \in \Theta$, the determinant of almost every correlation matrices $R_t(\theta)$ are strictly larger than $a^{N(N-1)} > 0$ (apply

Kurowicka and Cooke, 2006, Theorem 3.2), and that the norm of $R_t^{-1}(\theta)$ is bounded from above a.e. [3]. Moreover, the function $F_{\text{vine}}(.)$ that maps partial correlations to usual correlations has a bounded derivative, when applied to the trajectories $(Pc_t(\theta))$ generated by the model.

Theorem 1.5.15. *Under assumptions 1-4 the process (ϵ_t, D_t, R_t) as defined in equations (1.5.1), (1.5.2), and (1.5.3) possesses a strictly stationary solution such that $(\epsilon_t, D_t, R_t) \in \mathcal{F}_t$, the sigma field induced by the observations. Furthermore, the solution (ϵ_t) is second-order stationary and, when the innovations η_t^* are Gaussian given $\mathcal{F}_{t-1}$, then $\mathbb{E}\left[\|\epsilon_t\|^{2p}\right] < \infty$.*

The key result for the existence of an invariant probability measure for Markov chains is the criterion of Tweedie (1988). When using this approach, the irreducibility of (X_t) is not required to obtain stationarity.

Let $(X_t)_{t\in\mathbb{Z}}$ be a homogeneous Markov chain with a measurable state space $(E, \mathcal{E})$, such that its transition probability is $P(\mathbf{x}, \mathbf{B}) = \mathbb{P}(\mathbf{X_t} \in \mathbf{B} | \mathbf{X_{t-1}} = \mathbf{x})$, where $\mathbf{x} \in \mathbf{E}$ and $B \in \mathcal{E}$. Theorem 2 of Tweedie (1988) states the following:

Lemma 1.5.16. *Suppose $(E, \mathcal{E})$ is a locally compact separable state space and $(X_t)_{t\in\mathbb{Z}}$ is a Feller chain, that is for $h \in C_b^0(E)$, then $E\left[h(X_t)|X_{t-1} = \mathbf{x}\right]$ is also $C_b^0(E)$.*

1. *If for some compact set $B \in \mathcal{E}$, there exists a non negative mapping $g(.)$ and $\epsilon > 0$ such that*

$$\int_{B^c} P(\mathbf{x}, \mathbf{y})\mathbf{g}(\mathbf{y})\mathbf{d}\lambda(\mathbf{y}) \leq \mathbf{g}(\mathbf{x}) - \epsilon, \ \mathbf{x} \in \mathbf{B^c}, \tag{1.5.9}$$

   then there exists a σ-finite invariant measure μ for P such that $0 < \mu(B) < \infty$.

2. *Furthermore, if*

$$\int_B \left(\int_{B^c} P(\mathbf{x}, \mathbf{y})\mathbf{g}(\mathbf{y})\mathbf{d}\lambda(\mathbf{y}) \right) d\mu(\mathbf{x}) < \infty, \tag{1.5.10}$$

   then μ is finite and hence $\pi = \mu/\mu(E)$ is an invariant probability measure.

3. *Furthermore, if*

$$\int_{B^c} P(\mathbf{x}, \mathbf{y})\mathbf{g}(\mathbf{y})\mathbf{d}\lambda(\mathbf{y}) \leq \mathbf{g}(\mathbf{x}) - \mathbf{f}(\mathbf{x}), \ \mathbf{x} \in \mathbf{B^c}, \tag{1.5.11}$$

[3] Indeed, $\|R_t^{-1}\|_s \leq \lambda_{min}(R_t)^{-N} \leq a^{N^2(N-1)}$.

then μ admits a finite f-moment, i.e. $\mathbb{E}_\mu[f(X_t)] < \infty$.

The next Lemma is a specific version of Lemma A.2 in Ling and McAleer (2003). Its proof is omitted.

Lemma 1.5.17. *For a given squared matrix T, if $\rho(|T|) < 1$, then there exists a positive vector M such that $(Id - |T|)'M > 0$.*

Proof of Theorem 1.5.15. We first show that $(X_t)_{t\in\mathbb{Z}}$ is a Feller process. Let $h \in C_b^0(\mathbb{R}^d)$. We have

$$\begin{aligned}\mathbb{E}[h(X_t)|X_{t-1}=\mathbf{x}] &= \mathbb{E}[h(T_t\mathbf{x}+\nu_\mathbf{t})|\mathbf{X_{t-1}}=\mathbf{x}]\\ &= \mathbb{E}\left[h(\phi_1(u_t)\mathbf{x}+\phi_\mathbf{2}(\mathbf{u_t},\eta^*_\mathbf{t-1}))|\mathbf{X_{t-1}}=\mathbf{x}\right],\end{aligned}$$

for continuous transforms ϕ_1 and ϕ_2. By construction, $u_t = D_t^{-1}H_t^{1/2}\eta_t = R_t^{1/2}\eta_t^*$, where $R_t^{1/2}$ is a continuous mapping of X_{t-1}. Consequently, we obtain

$$\begin{aligned}\mathbb{E}[h(X_t)|X_{t-1}=\mathbf{x}] &= \mathbb{E}\left[h\circ\tilde{\phi}(\mathbf{x},\eta^*_\mathbf{t})|\mathbf{X_{t-1}}=\mathbf{x}\right]\\ &= \int h\circ\tilde{\phi}(\mathbf{x},\mathbf{u})\mathbf{d}\mathbb{P}_{\eta^*_\mathbf{t}}^{\mathbf{X_{t-1}=x}}(\mathbf{u})\\ &= \int h\circ\tilde{\phi}(\mathbf{x},\mathbf{u})\mathbf{f}_{\eta^*_\mathbf{t}}(\mathbf{u}|\mathbf{x})\mathbf{d}\lambda(\mathbf{u}),\end{aligned}$$

for some continuous transform $\tilde{\phi}$. Now, let $(\mathbf{x_n})_\mathbf{n}$ be a sequence such that $\mathbf{x_n} \underset{\mathbf{n}\to\infty}{\longrightarrow} \mathbf{x}$. As $h(.)$ is bounded and $\forall u, (h\circ\tilde{\phi}(\mathbf{x_n},\mathbf{u}))_\mathbf{n}$ is convergent, then $\lim_n \mathbb{E}[h(X_t)|X_{t-1}=\mathbf{x_n}] = \mathbb{E}[h(X_t)|X_{t-1}=\mathbf{x}]$ by the Lebesgue dominated convergence theorem under assumption 3. In other words, $\mathbf{x}\to\mathbb{E}[\mathbf{h(X_t)}|\mathbf{X_{t-1}}=\mathbf{x}]$ is continuous.

Second, we exhibit an explicit functional $g(.)$ to apply the Tweedie's criteria. To do so, take $g(\mathbf{x}) = \mathbf{1} + |\mathbf{x}^{\otimes\mathbf{p}}|'\mathbf{M}$, for any vector M, which will be explicit later. We have, for $p \geq 1$,

$$\mathbb{E}[g(X_t)|X_{t-1}=\mathbf{x}] = 1 + \mathbb{E}\left[|(T_t\mathbf{x}+\nu_\mathbf{t})^{\otimes\mathbf{p}}|'|\mathbf{X_{t-1}}=\mathbf{x}\right]M.$$

By some property of the Kronecker product and algebraic manipulations, let us rewrite $(T_t\mathbf{x}+\nu_\mathbf{t})^{\otimes\mathbf{p}} = (\mathbf{T_t x})^{\otimes\mathbf{p}} + \mathcal{B}(\mathbf{x}) = \mathbf{T_t^{\otimes p}}\mathbf{x}^{\otimes\mathbf{p}} + \mathcal{B}(\mathbf{x})$. We deduce that

$$\mathbb{E}[g(X_t)|X_{t-1}=\mathbf{x}] \leq 1 + \left(\mathbb{E}\left[|T_t^{\otimes p}\mathbf{x}^{\otimes\mathbf{p}}|'|\mathbf{X_{t-1}}=\mathbf{x}\right] + \mathbb{E}[\|\mathcal{B}(\mathbf{x})\||\mathbf{X_{t-1}}=\mathbf{x}]\right)M. \tag{1.5.12}$$

We focus on the first expectation in (1.5.12). As T_t is a function of u_t, its conditional distribution depends on R_t. Hence T_t is a function of X_{t-1}. Then, we obtain

$$\begin{aligned}\mathbb{E}\left[|(T_t\mathbf{x})^{\otimes\mathbf{p}}|'|\mathbf{X_{t-1}}=\mathbf{x}\right]M &\leq |\mathbf{x}^{\otimes\mathbf{p}}|'\mathbb{E}\left[|\mathbf{T_t}^{\otimes\mathbf{p}}|'|\mathbf{X_{t-1}}=\mathbf{x}\right]\mathbf{M}\\ &\leq |\mathbf{x}^{\otimes\mathbf{p}}|'\left(\sup_{\mathbf{x}\in\mathbb{R}^{\mathbf{d}}}\mathbb{E}\left[|\mathbf{T_t}^{\otimes\mathbf{p}}|'|\mathbf{X_{t-1}}=\mathbf{x}\right]\right)\mathbf{M}\\ &\leq |\mathbf{x}^{\otimes\mathbf{p}}|'(\mathbf{T}^{\star})'\mathbf{M}.\end{aligned}$$

As for the second expectation in (1.5.12), by taking any multiplicative norm $\|.\|$, we have

$$\mathbb{E}\left[\|\mathcal{B}(\mathbf{x})\||\mathbf{X_{t-1}}=\mathbf{x}\right]\leq K\mathbb{E}\left[\|\nu_t\|\|(T_t\mathbf{x})^{\otimes(\mathbf{p-1})}\|+\|\nu_\mathbf{t}\|^{\mathbf{2}}\|(\mathbf{T_t}\mathbf{x})^{\otimes(\mathbf{p-2})}\|+\cdots+\|\nu_\mathbf{t}\|^{\mathbf{p}}|\mathbf{X_{t-1}}=\mathbf{x}\right], \tag{1.5.13}$$

where K is a non-negative constant. In (1.5.13), we need to upper bound quantities of the type $\mathbb{E}\left[\|\nu_t\|^m\|T_t\|^n|X_{t-1}=\mathbf{x}\right]$, i.e. terms as $\mathbb{E}\left[(\|\zeta_{t-1}\|+\|\vec{u}_t\|)^m\|\vec{u}_t\|^n|X_{t-1}=\mathbf{x}\right]$ when $m+n\leq p$. First, we consider $\mathbb{E}\left[\|\vec{u}_t\|^{m+n}|X_{t-1}=\mathbf{x}\right]$. Recall that $u_t=R_t^{1/2}\eta_t^*$. Taking the spectral norm of $R_t^{1/2}$, we obtain a.s.

$$\|R_t^{1/2}\|=\rho\left(R_t^{1/2}R_t^{1/2\,\prime}\right)^{1/2}=\sqrt{\text{Trace}\left(D_t^{-1}H_tD_t^{-1}\right)}\leq\sqrt{N}.$$

Using the previous inequality and assumption 3, we have

$$\mathbb{E}\left[\|\vec{u}_t\|^{m+n}|X_{t-1}=\mathbf{x}\right]\leq\mathbb{E}\left[\|R_t^{1/2}\|^{2(m+n)}\|\vec{\eta}_t^*\|^{m+n}|X_{t-1}=\mathbf{x}\right]\leq N^{n+m}\mathbb{E}\left[\|\vec{\eta}_t^*\|^{m+n}|X_{t-1}=\mathbf{x}\right]. \tag{1.5.14}$$

By assumption, $\mathbb{E}\left[\|\eta_t^*\|^{2p}|X_{t-1}=\mathbf{x}\right]\leq\psi(\|\mathbf{x}\|)$. Then, we obtain

$$\mathbb{E}\left[\|\vec{u}_t\|^{m+n}|X_{t-1}=\mathbf{x}\right]\leq\alpha_{m,n}\psi\left(\|\mathbf{x}\|\right)^{(m+n)/p},$$

for some constant $\alpha_{m,n}$.

Another product element we shall bound is $\mathbb{E}\left[\|\zeta\left(\chi_{t-1},\eta_{t-1}\right)\|^m\|\vec{u}_t\|^n|X_{t-1}=\mathbf{x}\right]$. To do so, we take $n+m=p$, where $m\geq 1$. Using the conditional Hölder's inequality, we obtain

$$\mathbb{E}\left[\|\zeta\left(\chi_{t-1},\eta_{t-1}\right)\|^m\|\vec{u}_t\|^n|X_{t-1}=\mathbf{x}\right]\leq\mathbb{E}\left[\|\zeta\left(\chi_{t-1},\eta_{t-1}\right)\|^p|X_{t-1}=\mathbf{x}\right]^{m/p}\mathbb{E}\left[\|\vec{u}_t\|^p|X_{t-1}=\mathbf{x}\right]^{n/p}. \tag{1.5.15}$$

In (1.5.15), $\mathbb{E}\left[\|\vec{u}_t\|^p|X_{t-1}=\mathbf{x}\right]^{n/p}$ can be straightforwardly upper bounded using (1.5.14). We now focus on the conditional expectation of $\|\zeta\left(\chi_{t-1},\eta_{t-1}\right)\|^p$. Denoting $\tilde{v}_{k|L,t}=$

$\epsilon_{k,t} - \mathbb{E}_{t-1}\left[\epsilon_{k,t}|\epsilon_{L,t}\right]$, we have

$$\mathbb{E}\left[\|\zeta\left(\chi_{t-1},\eta_{t-1}\right)\|^p|X_{t-1}=\mathbf{x}\right] \leq \sup_{(i,j|L)\in E}\mathbb{E}\left[|\frac{\tilde{\upsilon}_{i|L,t-1}\tilde{\upsilon}_{j|L,t-1}}{\sqrt{h_{i|L,t-1}}\sqrt{h_{j|L,t-1}}}|^p|X_{t-1}=\mathbf{x}\right]. \tag{1.5.16}$$

For $p=1$, we apply the Cauchy-Schwartz inequality to (1.5.16) as

$$\mathbb{E}\left[|\frac{\tilde{\upsilon}_{i|L,t-1}\tilde{\upsilon}_{j|L,t-1}}{\sqrt{h_{i|L,t-1}}\sqrt{h_{j|L,t-1}}}||X_{t-1}=\mathbf{x}\right] \leq \mathbb{E}\left[\frac{\tilde{\upsilon}^2_{i|L,t-1}}{h_{i|L,t-1}}|X_{t-1}=\mathbf{x}\right]^{1/2}\mathbb{E}\left[\frac{\tilde{\upsilon}^2_{j|L,t-1}}{h_{j|L,t-1}}|X_{t-1}=\mathbf{x}\right]^{1/2}=1.$$

In this case, we obtain

$$\mathbb{E}\left[\|\mathcal{B}(\mathbf{x})\||\mathbf{X_{t-1}}=\mathbf{x}\right]=\alpha_1\mathbb{E}\left[\|\zeta_{t-1}\|+\|u_t\||X_{t-1}=\mathbf{x}\right]\leq\alpha_2\psi\left(\|\mathbf{x}\|\right)+\alpha_3,$$

for some constants α_k, $k=1,2,3$. Consequently for $p=1$, we deduce that (1.5.12) can be upper bounded as

$$\begin{aligned}\mathbb{E}\left[g(X_t)|X_{t-1}=\mathbf{x}\right] &\leq 1+\left(\mathbb{E}\left[|T_t\mathbf{x}|'|\mathbf{X_{t-1}}=\mathbf{x}\right]+\mathbb{E}\left[\|\mathcal{B}(\mathbf{x})\||\mathbf{X_{t-1}}=\mathbf{x}\right]\right)M\\ &\leq 1+|\mathbf{x}|'(\mathbf{T}^\star)'\mathbf{M}+\mathbf{O}\left(\|\mathbf{x}\|^{\mathbf{a}}\right),\end{aligned}$$

for any $a>0$. Let us now try to extend this result for $p>1$. The quantity given in (1.5.16) is a product of $\tilde{\upsilon}_{k|L,t-1}$ components, which can be decomposed as

$$\begin{aligned}\tilde{\upsilon}_{i|L,t-1} &= e_i'H_{t-1}^{1/2}(\theta)\{\eta_{t-1}-\mathbb{E}_{t-2}\left[\eta_{t-1}|\epsilon_{L,t-1},X_{t-1}=\mathbf{x}\right]\}\\ &= e_i'D_{t-1}R_{t-1}^{1/2}\{\eta^*_{t-1}-\mathbb{E}_{t-2}\left[\eta^*_{t-1}|\epsilon_{L,t-1},X_{t-1}=\mathbf{x}\right]\}\end{aligned}$$

Assuming all denominators are bounded from below a.s., this implies that (1.5.16) can be upper bounded as

$$\begin{aligned}\sup_{(i,j|L)\in E}\mathbb{E}\left[|\frac{\tilde{\upsilon}_{i|L,t-1}\tilde{\upsilon}_{j|L,t-1}}{\sqrt{h_{i|L,t-1}}\sqrt{h_{j|L,t-1}}}|^p|X_{t-1}=\mathbf{x}\right] &\leq Cst.\mathbb{E}\left[\|D_{t-1}\|^{2p}\|R_{t-1}\|^p\|\eta^*_{t-1}\|^{2p}|X_{t-1}=\mathbf{x}\right]\\ &\leq Cst.\mathbb{E}\left[\|\mathbf{x}\|^{\mathbf{p}}\|\eta^*_{\mathbf{t-1}}\|^{\mathbf{2p}}|\mathbf{X_{t-1}}=\mathbf{x}\right]\\ &\leq Cst.\|\mathbf{x}\|^{\mathbf{p}}\psi(\|\mathbf{x}\|).\end{aligned}$$

This upper bound is not of order $O(\|\mathbf{x}\|^{\mathbf{k}})$, for $k\leq p-1$. We rely on the Gaussian distribution hypothesis to circumvent this obstacle.

Now, the vectors η_t^* (or η_t, equivalently) is supposed to be Gaussian, conditional to the past. By the Cauchy-Schwartz inequality, we have

$$\mathbb{E}\left[|\frac{\tilde{\upsilon}_{i|L,t-1}\tilde{\upsilon}_{j|L,t-1}}{\sqrt{h_{i|L,t-1}}\sqrt{h_{j|L,t-1}}}|^p|X_{t-1}=\mathbf{x}\right] \leq \mathbb{E}\left[\frac{\tilde{\upsilon}_{i|L,t-1}^{2p}}{h_{i|L,t-1}^p}|X_{t-1}=\mathbf{x}\right]^{1/2} \mathbb{E}\left[\frac{\tilde{\upsilon}_{j|L,t-1}^{2p}}{h_{j|L,t-1}^p}|X_{t-1}=\mathbf{x}\right]^{1/2}.$$

Since any $\tilde{\upsilon}_{i|L,t-1}/\sqrt{h_{i|L,t-1}}$ is a Gaussian random variable $\mathcal{N}(0,1)$, given X_{t-1}, the r.h.s. of the latter inequality is uniformly bounded wrt i, j, L and $\mathbf{x}$. We deduce that (1.5.16) can be upper bounded as

$$\mathbb{E}\left[\|\zeta\left(\chi_{t-1},\eta_{t-1}\right)\|^p|X_{t-1}=\mathbf{x}\right]=O(1),$$

for all $\mathbf{x}$.

This result is proved using $\forall t \geq 1, \sigma_{k|L,t}^2(\mathbf{x}) > \mathbf{0}$ a.s.. We need to prove that this holds almost surely for any $\mathbf{x} \in \mathbf{B^c}$. That means we need to control for the variance and correlation dynamics when $\mathbf{x}$ can take very large values. By contradiction, suppose $\forall k \notin L$

$$\sigma_{k|L,t}^2(\mathbf{x}) = \mathbb{E}\left[\left(\epsilon_{\mathbf{k,t}} - \mathbb{E}\left[\epsilon_{\mathbf{k,t}}|\epsilon_{\mathbf{L,t}}\right]\right)^{\mathbf{2}}|\mathbf{X_{t-1}}=\mathbf{x}\right] = \mathbf{0} \Rightarrow \epsilon_{\mathbf{k,t}} = \mathbb{E}\left[\epsilon_{\mathbf{k,t}}|\epsilon_{\mathbf{L,t}},\mathbf{X_{t-1}}=\mathbf{x}\right] \text{ a.s.} \tag{1.5.17}$$

Using the decomposition $\epsilon_t = H_t^{1/2}\eta_t$, relationship (1.5.17) becomes

$$\epsilon_{k,t} = Q'(\mathbf{x})\epsilon_{\mathbf{L,t}} \text{ a.s.}, \tag{1.5.18}$$

where $Q'(\mathbf{x})$ corresponds to a vector containing the coefficients of H_t used for computing the conditional expectation under the Gaussian distribution. As H_t is $\mathcal{F}_{t-1}$ measurable, then Q is a function of $\mathbf{x}$. (1.5.18) means that $\epsilon_{k,t}$ can be written as a linear combination of $\epsilon_{n,t}$, for $n \in L$, given $\mathbf{x}$. If there exists a linear relationship between the components of ϵ_t given $\mathbf{x}$, then the matrix $H_t(\mathbf{x})$ is not a full rank matrix. As $D_t(\mathbf{x})$ is a diagonal matrix, it is always nonsingular, $H_t(\mathbf{x})$ singular implies that $R_t(\mathbf{x})$ is not positive definite. This contradicts $\lambda_1(R_t(\mathbf{x})) > \mathbf{0}$ a.s.. We deduce that

$$\exists \mu > 0, \text{ such that } \forall k, \forall L, k \notin L, \sigma_{k|L,t}^2(\mathbf{x}) \geq \mu \text{ for almost all } \mathbf{x}.$$

Consequently, using assumption 3, we have obtained

$$\begin{aligned}\mathbb{E}\left[g(X_t)|X_{t-1}=\mathbf{x}\right] &\leq 1+|\mathbf{x}^{\otimes\mathbf{p}}|'(\mathbf{T}^{\star})'\mathbf{M}+\mathbf{O}\left(\|\mathbf{x}\|^{\mathbf{a}}\right)\\ &\leq g(\mathbf{x})-|\mathbf{x}^{\otimes\mathbf{p}}|'\left(\mathbf{I_N}-(\mathbf{T}^{\star})'\right)\mathbf{M}+\mathbf{O}\left(\|\mathbf{x}\|^{\mathbf{a}}\right),\end{aligned}$$

for all $a>0$. We denote $N(\mathbf{x}):=\sum_{\mathbf{i=1}}^{\mathbf{s}}|\mathbf{x_i}|^{\mathbf{p}}$. Since $(Id-(T^{\star})')\,M>0$ by Lemma (1.5.17), then there exists $m_0>0$ such that

$$(I_s-(T^{\star})')\,M\geq m_0N(\mathbf{x}),\ \forall\mathbf{x}\in\mathbb{R}^{\mathbf{s}}.$$

Similarly, $\exists m_1>0$ such that $\forall\mathbf{x}\in\mathbb{R}^{\mathbf{s}},\mathbf{g}(\mathbf{x})\geq\mathbf{m_1N}(\mathbf{x})$. Using the Hölder's inequality, we have $\forall k\leq p$

$$\sum_{j_1,j_2,\cdots,j_k}^{s}|x_{j_1}x_{j_2}\cdots x_{j_k}|=\left(\sum_{j=1}^{s}|x_j|\right)^k\leq\left(\sum_{j=1}^{s}|x_j|^p\right)^{k/p}s^k. \tag{1.5.19}$$

Hence using inequality (1.5.19), $\forall k\leq p,\exists m_2>0$ such that

$$g(\mathbf{x})\leq\mathbf{1}+\|\mathbf{M}\|\sum_{\mathbf{j_1},\mathbf{j_2},\cdots,\mathbf{j_k}}^{\mathbf{s}}|\mathbf{x_{j_1}x_{j_2}}\cdots\mathbf{x_{j_p}}|\leq\mathbf{1}+\mathbf{c_2N}(\mathbf{x}),$$

We deduce that

$$\begin{aligned}\mathbb{E}\left[g(X_t)|X_{t-1}=\mathbf{x}\right] &\leq g(\mathbf{x})\left(\mathbf{1}-\mathbf{m_0}\frac{\mathbf{N}(\mathbf{x})}{\mathbf{g}(\mathbf{x})}+\mathbf{O}\left(\frac{\mathbf{N}(\mathbf{x})^{\mathbf{a/p}}}{\mathbf{g}(\mathbf{x})}\right)\right)\\ &\leq g(\mathbf{x})\left(\mathbf{1}-\mathbf{m_0}\frac{\mathbf{N}(\mathbf{x})}{\mathbf{1}+\mathbf{m_2N}(\mathbf{x})}+\mathbf{O}\left(\frac{\mathbf{N}(\mathbf{x})^{\mathbf{a/p}}}{\mathbf{m_1N}(\mathbf{x})}\right)\right)\end{aligned}$$

We denote $B:=\{\mathbf{x}\in\mathbb{R}^{\mathbf{s}}|\mathbf{N}(\mathbf{x})\leq\mathbf{\Gamma}\}$, with $\Gamma>1$. For Γ large enough, $\forall\mathbf{x}\notin\mathbf{B}$, and $0<a<1$, we have

$$\mathbb{E}\left[g(X_t)|X_{t-1}=\mathbf{x}\right]\leq g(\mathbf{x})\left(\mathbf{1}-\frac{\mathbf{m_0}}{\mathbf{2m_2}}+\mathbf{O}\left(\mathbf{1}\right)\right)<\mathbf{g}(\mathbf{x})\left(\mathbf{1}-\frac{\mathbf{m_0}}{\mathbf{3m_2}}\right). \tag{1.5.20}$$

As $1\leq g(\mathbf{x})$, then $\mathbb{E}\left[g(X_t)|X_{t-1}=\mathbf{x}\right]\leq g(\mathbf{x})-\varepsilon$, for $\varepsilon>0$. This proves (1.5.9), idest $\exists\mu$ a σ-finite invariant measure for $(X_t)_t$ such that $0<\mu(A)<\infty$.

Now for any $\mathbf{x} \in \mathbf{B}$, (1.5.20) provides

$$\mathbb{E}\left[g(X_t)|X_{t-1} = \mathbf{x}\right] \leq g(\mathbf{x}) + \mathbf{O}\left(\|\mathbf{x}\|^{\mathbf{a}}\right) \leq \mathbf{K}, \tag{1.5.21}$$

for some constant $K > 0$. This implies

$$\int_B \left(\int_{B^c} P(\mathbf{x}, \mathbf{y}) \mathbf{g}(\mathbf{y}) \mathbf{d}\lambda(\mathbf{y}) \right) d\mu(\mathbf{x}) \leq \int_{\mathbf{B}} \mathbb{E}\left[\mathbf{g}(\mathbf{X_t})|\mathbf{X_{t-1}} = \mathbf{x}\right] \mathbf{d}\mu(\mathbf{x}) \leq \mathbf{K}\mu(\mathbf{B}) \leq \infty. \tag{1.5.22}$$

Consequently, (1.5.10) is proved and μ is finite and $\pi = \mu/\mu(E)$ is an invariant probability measure. Then there exists a strictly stationary solution of the stochastic recurrence equation (1.5.5).

Finally, using inequality (1.5.20), we obtain (1.5.11) for $f(\mathbf{x}) = \beta \mathbf{g}(\mathbf{x})$, where $\beta \in (0, 1)$. As $m_1 N(\mathbf{x}) \leq \mathbf{g}(\mathbf{x})$, then

$$\mathbb{E}_\pi\left[N(X_t)\right] < \infty.$$

□

1.5.4 Uniqueness of stationary vine-GARCH Solutions

Tweedie's criterion provides the existence of an invariant probability measure for Markov chains. However, the uniqueness of such a measure is not ensured. Uniqueness is a significant result as it provides the ergodicity of the stationary solution. This is a significant feature for inference purpose since asymptotic properties for M-estimators are based on Uniform Law of Large Numbers, or the ergodic theorem (see Billingsley, 1995).

Assumption 5. The sequence of innovations (η_t^*) is strongly stationary.

Assumption 6. There exist some strictly positive constant C_h s.t., for any stationary solution, for all t,

$$h_{i|L,t}^{-1} \quad \leq \quad C_h \quad \mathbb{P} - \text{a.s.},$$

where $(i|L)$ is associated to an arbitrary node $(i, j|L)$, $L \neq \emptyset$ of the underlying vine $V(n)$.

Note that, when L is empty, the model provides a lower bound for all conditional variances: for every i and t, $h_{i,t}^{-1} \leq C_v$. Let us introduce some intermediate quantities. We denote $C_F > 0$ (resp. $C_{\Psi^{-1}} > 0$) the Lipschitz constant of $F_{\text{vine}}(.)$ (resp. $\Psi^{-1}(.)$).

Let us consider two (arbitrarily chosen) stationary solutions (D_t, R_t, ϵ_t) and $(\tilde{D}_t, \tilde{R}_t, \tilde{\varepsilon}_t)$. They share the innovations (η_t^*) and the model parameters. The proof of uniqueness relies on some top Lyapunov exponent of a stochastic matrix process denoted by

$$M_t = \begin{pmatrix} \|\Xi\|_\infty + \|\Lambda\|_\infty \Upsilon_{2,t} & \|\Lambda\|_\infty \Upsilon_{1,t} \\ \Gamma_{2,t} & \Gamma_{1,t} \end{pmatrix},$$

where

$$\begin{cases} \Upsilon_{1,t} &= C_h\sqrt{N}\left(\|D_t\|_s + \|\tilde{D}_t\|_s\right)\{\alpha + \sqrt{N}\|\tilde{D}_t\|_s^2\|\eta_t^*\|_2^2 C_h^2\gamma\}, \\ \Upsilon_{2,t} &= C_h\sqrt{N}\left(\|D_t\|_s + \|\tilde{D}_t\|_s\right)\{\beta + \sqrt{N}\|\tilde{D}_t\|_s^2\|\eta_t^*\|_2^2 C_h^2\delta\}, \\ \gamma &= C_v^{1/2}N\{\|D_t\|_s + \|\tilde{D}_t\|_s\}\left[1 + \frac{NC_v\|D_t\|_s^2}{\lambda_1(R_t)} + \frac{NC_v\|\tilde{D}_t\|_s^2}{\lambda_1(\tilde{R}_t)} + \frac{N^2C_v^2\|D_t\|_s^2\|\tilde{D}_t\|_s^2}{\lambda_1(R_t)\lambda_1(\tilde{R}_t)}\right] \\ \delta &= \sqrt{N}C_F C_{\Psi^{-1}}\|D_t\|_s\|\tilde{D}_t\|_s\left[1 + \frac{NC_v\|D_t\|_s^2}{\lambda_1(R_t)} + \frac{NC_v\|\tilde{D}_t\|_s^2}{\lambda_1(\tilde{R}_t)} + \frac{N^2C_v^2\|D_t\|_s^2\|\tilde{D}_t\|_s^2}{\lambda_1(R_t)\lambda_1(\tilde{R}_t)}\right] \\ \alpha &= \sqrt{N}C_v^{1/2}\|\eta_t^*\|_s\left\{1 + \frac{N\|D_t\|_s C_h}{\lambda_1(R_t)}\{\|D_t\|_s + \|\tilde{D}_t\|_s\}\left[1 + \frac{N\|\tilde{D}_t\|_s^2 C_h}{\lambda_1(\tilde{R}_t)}\right]\right\}, \\ \beta &= \sqrt{N}C_F C_{\Psi^{-1}}\|\tilde{D}_t\|_s\|\eta_t^*\|_s\left\{\frac{\sqrt{N}\|D_t\|_s^2 C_h}{\lambda_1(R_t)}\left[1 + \frac{N\|\tilde{D}_t\|_s^2 C_h}{\lambda_1(\tilde{R}_t)}\right] + \frac{1}{\lambda_1^{1/2}(R_t) + \lambda_1^{1/2}\left(\tilde{R}_t\right)}\right\}. \end{cases}$$

and

$$\begin{cases} \Gamma_{1,t} &= \|A\|_\infty + N\|B\|_\infty\|\eta_{t-1}^*\|_2^2, \\ \Gamma_{2,t} &= \|B\|_\infty\|\tilde{D}_{t-1}\|_s^2\frac{2\|\eta_t^*\|_2^2}{\lambda_1^{1/2}(R_t) + \lambda_1^{1/2}\left(\tilde{R}_t\right)}NC_F C_{\Psi^{-1}}. \end{cases}$$

Assumption 7. (M_t) is a stationary stochastic process and $\mathbb{E}\left[\log\left(M_t\right)\right] < \infty$ such that its top Lyapunov exponent defined as

$$\gamma_M := \lim_{t\to\infty}\frac{1}{t}\log\left(M_t M_{t-1}\cdots M_1\right)$$

is strictly negative.

Theorem 1.5.18. *Under assumptions 1 and 5-7, a strictly stationary solution of the vine-GARCH model is unique and ergodic, given a sequence* $(\eta_t^*)_{t\in\mathbb{Z}}$.

Proof of Theorem 1.5.18. We remind that $\epsilon_t = D_t u_t = H_t^{1/2}\eta_t$ and $u_t = R_t^{1/2}\eta_t^*$. The

model equations define a solution (ϵ_t, D_t, R_t) given (η_t^*). The dynamic system is specified as

$$\left\{\begin{array}{lcl} \text{Vecd}\left(D_t^2\right) & = & V + A\text{Vecd}\left(D_{t-1}^2\right) + B\vec{\epsilon}_{t-1}, \\ R_t & = & \text{vechof}\left(F_{\text{vine}}\left(Pc_t\right)\right), \\ \Psi\left(Pc_t\right) & = & \Omega + \Xi\Psi\left(Pc_{t-1}\right) + \Lambda\zeta_{t-1}. \end{array}\right.$$

A key quantity is the vector of innovations (ζ_t) defined as

$$\left\{\begin{array}{lcl} \zeta_t & = & \left[v_{i|L,t}v_{j|L,t}\right]_{(i,j|L)\in V(N)}, \\ v_{i|L,t} & = & \dfrac{\epsilon_{i,t} - \mathbb{E}_{t-1}\left[\epsilon_{i,t}|\epsilon_{L,t}\right]}{\sqrt{h_{i|L,t}}}, \end{array}\right.$$

such that

$$\begin{array}{rcl} h_{i|L,t} & = & \text{Var}_{t-1}\left(\epsilon_{i,t}\right) - \text{Cov}_{t-1}\left(\epsilon_{i,t},\epsilon_{L,t}\right)\text{Var}_{t-1}\left(\epsilon_{L,t}\right)^{-1}\text{Cov}_{t-1}\left(\epsilon_{L,t},\epsilon_{i,t}\right), \\ & = & e_i'H_te_i - \left(e_i'H_te_L\right).\left(e_L'H_te_L\right)^{-1}.\left(e_L'H_te_i\right). \end{array}$$

Above, we have introduced some deterministic matrices (of zeros and ones) e_L s.t. $\epsilon_{L,t} = e_L'\epsilon_t$. The dimension of e_L is $N \times |L|$. More generally, for any $m \times N$-matrix A, Ae_L concatenates the A-columns whose index belongs to L. Using the fact that B is a diagonal matrix and $\epsilon_{i,t} = \sqrt{h_{i,t}}u_{i,t}$, we obtain $\vec{\epsilon}_{i,t} = h_{i,t}u_{i,t}^2$ and

$$\text{Vecd}\left(D_t^2\right) = V + A.\text{Vecd}\left(D_{t-1}^2\right) + B.D_{t-1}^2\vec{u}_{t-1}.$$

where $D_t^2.e = \text{Vecd}\left(D_t^2\right)$.

We first focus on the uniqueness of the conditional variance process. To do so, we consider the difference

$$\begin{array}{rcl} \text{Vecd}\left(D_t^2\right) - \text{Vecd}\left(\tilde{D}_t^2\right) & = & A.\left[\text{Vecd}\left(D_{t-1}^2\right) - \text{Vecd}\left(\tilde{D}_{t-1}^2\right)\right] + B.\left[D_t^2\vec{u}_{t-1} - \tilde{D}_{t-1}^2\vec{\tilde{u}}_{t-1}\right], \\ & = & A.\left[\text{Vecd}\left(D_{t-1}^2\right) - \text{Vecd}\left(\tilde{D}_{t-1}^2\right)\right] + B.\left[D_{t-1}^2 - \tilde{D}_{t-1}^2\right]\vec{u}_{t-1} \\ & + & B.\tilde{D}_{t-1}^2.\left[\vec{u}_{t-1} - \vec{\tilde{u}}_{t-1}\right]. \end{array}$$

Using $D_t^2\vec{u}_t = \vec{u}_t \odot \text{Vecd}\left(D_{t-1}^2\right)$, we obtain

$$\begin{array}{l} \text{Vecd}\left(D_t^2\right) - \text{Vecd}\left(\tilde{D}_t^2\right) = A.\left[\text{Vecd}\left(D_{t-1}^2\right) - \text{Vecd}\left(\tilde{D}_{t-1}^2\right)\right] \\ + \quad B\vec{u}_{t-1} \odot \left[\text{Vecd}\left(D_{t-1}^2\right) - \text{Vecd}\left(\tilde{D}_{t-1}^2\right)\right] + B\left[\vec{u}_{t-1} - \vec{\tilde{u}}_{t-1}\right] \odot \text{Vecd}\left(\tilde{D}_{t-1}^2\right). \end{array}$$

Furthermore

$$\begin{aligned}\vec{u}_t - \vec{\tilde{u}}_t &= (u_t - \tilde{u}_t) \odot (u_t + \tilde{u}_t) \\ &= (u_t + \tilde{u}_t) \odot \left(R_t^{1/2} - \tilde{R}_t^{1/2}\right) \eta_t^*.\end{aligned}$$

Using the spectral norm, the previous quantity can be upper bounded as

$$\|\vec{u}_t - \vec{\tilde{u}}_t\|_s \leq \|u_t + \tilde{u}_t\|_\infty . \| \left(R_t^{1/2} - \tilde{R}_t^{1/2}\right) \eta_t^* \|_s.$$

Since $\|\eta_t^*\|_s = \|\eta_t^*\|_2$ (as for any vector), note that

$$\|u_t\|_\infty = \|R_t^{1/2}\eta_t^*\|_\infty \leq \|R_t^{1/2}\eta_t^*\|_s \leq \|R_t\|_s^{1/2}\|\eta_t^*\|_2 \leq \sqrt{N}\|\eta_t^*\|_2.$$

Using theorem 6.2 of Higham (2008), for any unitarily invariant norm $\|\cdot\|$, we have

$$\|R_t^{1/2} - \tilde{R}_t^{1/2}\| \leq \frac{1}{\lambda_1^{1/2}(R_t) + \lambda_1^{1/2}\left(\tilde{R}_t\right)} \|R_t - \tilde{R}_t\|.$$

Recall that the norm $\|.\|$ is unitarily invariant if $\|UAV\| = \|A\|$ for all matrix A and all unitary matrices U and V, ie $UU' = Id$ and $VV' = Id$. For instance, the spectral norm $\|A\|_s = \rho(A'A)^{1/2} = \lambda_{max}(A)$ satisfies

$$\|UAV\|_s = \rho\left((UAV)'.UAV\right)^{1/2} = \rho(V'A'AV)^{1/2} = \rho(A'A)^{1/2} = \|A\|_s,$$

and is then unitarily invariant. Hence

$$\begin{aligned}\| \left(R_t^{1/2} - \tilde{R}_t^{1/2}\right) \eta_t^* \|_s &\leq \|R_t^{1/2} - \tilde{R}_t^{1/2}\|_s \|\eta_t^*\|_s \\ &\leq \frac{1}{\lambda_1^{1/2}(R_t) + \lambda_1^{1/2}\left(\tilde{R}_t\right)} \|R_t - \tilde{R}_t\|_s \|\eta_t^*\|_2.\end{aligned}$$

Besides,

$$\begin{aligned}\|R_t - \tilde{R}_t\|_s &\leq \sqrt{N}\|R_t - \tilde{R}_t\|_\infty \\ &\leq \sqrt{N}C_F\|Pc_t - \tilde{P}c_t\|_\infty \\ &\leq \sqrt{N}C_F C_{\Psi^{-1}}\|\Psi(Pc_t) - \Psi\left(\tilde{P}c_t\right)\|_\infty.\end{aligned}$$

As B and A are diagonal matrices, their spectral norms are equal to their infinite norm. We obtain the upper bound

$$\begin{aligned}
\|\text{Vecd}\left(D_t^2\right) - \text{Vecd}\left(\tilde{D}_t^2\right)\|_s &\leq \|A\|_s \|\text{Vecd}\left(D_{t-1}^2\right) - \text{Vecd}\left(\tilde{D}_{t-1}^2\right)\|_s \\
&+ \|B\|_s \|\vec{u}_{t-1}\|_s \|\text{Vecd}\left(D_{t-1}^2\right) - \text{Vecd}\left(\tilde{D}_{t-1}^2\right)\|_s \\
&+ \|B\|_s \|\text{Vecd}\left(\tilde{D}_{t-1}^2\right)\|_\infty \|u_t + \tilde{u}_t\|_\infty \|\eta_t^*\|_2 \frac{\|R_t - \tilde{R}_t\|_s}{\lambda_1^{1/2}\left(R_t\right) + \lambda_1^{1/2}\left(\tilde{R}_t\right)} \\
&\leq \Gamma_{1,t} \|\text{Vecd}\left(D_{t-1}^2\right) - \text{Vecd}\left(\tilde{D}_{t-1}^2\right)\|_s + \Gamma_{2,t} \|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty, \quad (1.5.23)
\end{aligned}$$

where

$$\begin{cases}
\Gamma_{1,t} &= \|A\|_\infty + N\|B\|_\infty \|\eta_{t-1}^*\|_2^2, \\
\Gamma_{2,t} &= \|B\|_\infty \|\tilde{D}_{t-1}\|_s^2 \frac{2\|\eta_t^*\|_2^2}{\lambda_1^{1/2}\left(R_t\right) + \lambda_1^{1/2}\left(\tilde{R}_t\right)} N C_F C_{\Psi^{-1}}.
\end{cases}$$

We now focus on the uniqueness of the partial correlation process. We consider the difference

$$\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right) = \Xi\left(\Psi\left(Pc_{t-1}\right) - \Psi\left(\tilde{P}c_{t-1}\right)\right) + \Lambda\left(\zeta_{t-1} - \tilde{\zeta}_{t-1}\right).$$

In this framework, Ξ and Λ are parameterized as diagonal matrices. We have

$$\|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty \leq \|\Xi\|_\infty \|\Psi\left(Pc_{t-1}\right) - \Psi\left(\tilde{P}c_{t-1}\right)\|_\infty + \|\Lambda\|_\infty \|\zeta_{t-1} - \tilde{\zeta}_{t-1}\|_\infty. \quad (1.5.24)$$

The quantity of interest is the vector of innovations, that is

$$v_{ij|L,t} - \tilde{v}_{ij|L,t} = \frac{r_{i|L,t} r_{j|L,t}}{\sqrt{h_{i|L,t}}\sqrt{h_{j|L,t}}} - \frac{\tilde{r}_{i|L,t}\tilde{r}_{j|L,t}}{\sqrt{\tilde{h}_{i|L,t}}\sqrt{\tilde{h}_{j|L,t}}}, \quad (1.5.25)$$

where, using the Gaussian assumption, we have

$$\begin{aligned}
r_{i|L,t} &= \epsilon_{i,t} - \mathbb{E}_{t-1}\left[\epsilon_{i,t}|\epsilon_{L,t}\right] \\
&= \epsilon_{i,t} - \left(e_i' H_t e_L\right).\left(e_L' H_t e_L\right)^{-1} \epsilon_{L,t} \\
&= \left[e_i' - \left(e_i' H_t e_L\right).\left(e_L' H_t e_L\right)^{-1} e_L'\right] \epsilon_t \\
&:= e_i' \mathbf{p_L}\left(\boldsymbol{\epsilon_t}\right).
\end{aligned} \quad (1.5.26)$$

Here, $\mathbf{p_L}(\cdot)$ is the projector on the orthogonal of the subspace $< H_t e_L >$ in $\mathbb{R}^N$, relatively to the H_t^{-1}-euclidian norm, defined by $\|\boldsymbol{x}\|_H = \boldsymbol{x}' H_t^{-1} \boldsymbol{x}$ [4]. By decomposing the projector $\mathbf{p_L}$ in its canonical space, we see that $\|\mathbf{p_L}\|_\mathbf{s} = \mathbf{1}$ obviously. Similarly, $\|\tilde{\mathbf{p}}_L\|_s = 1$.

Recall that $\epsilon_t = D_t R_t^{1/2} \eta_t^*$. Using the same steps as in (1.5.26), we obtain

$$\tilde{r}_{i|L,t} = e_i' \tilde{\mathbf{p}}_L(\tilde{\epsilon}_t), \ \tilde{\epsilon}_t = \tilde{D}_t \tilde{R}_t^{1/2} \eta_t^*.$$

Now we have

$$\|\zeta_{t-1} - \tilde{\zeta}_{t-1}\|_\infty = \sup_{(i,j|L)} |v_{ij|L,t} - \tilde{v}_{ij|L,t}|,$$

which implies we need to control $|r_{i|L,t} - \tilde{r}_{i|L,t}|$ and $|h_{i|L,t} - \tilde{h}_{i|L,t}|$.

Step 1. We have

$$\begin{aligned} r_{i|L,t} - \tilde{r}_{i|L,t} &= e_i' \mathbf{p_L}(\epsilon_\mathbf{t}) - \mathbf{e_i'} \tilde{\mathbf{p}}_\mathbf{L}(\tilde{\epsilon}_\mathbf{t}) \\ &= e_i' [\mathbf{p_L} - \tilde{\mathbf{p}}_\mathbf{L}](\epsilon_t) + e_i' \tilde{\mathbf{p}}_L(\epsilon_t - \tilde{\epsilon}_t) \end{aligned}$$

We obtain

$$\begin{aligned} |r_{i|L,t} - \tilde{r}_{i|L,t}| &\leq \|(\mathbf{p_L} - \tilde{\mathbf{p}}_\mathbf{L})(\epsilon_\mathbf{t})\|_\infty + \|\tilde{\mathbf{p}}_\mathbf{L}(\epsilon_\mathbf{t} - \tilde{\epsilon}_\mathbf{t})\|_\infty \\ &\leq \|(\mathbf{p_L} - \tilde{\mathbf{p}}_\mathbf{L})(\epsilon_\mathbf{t})\|_\mathbf{2} + \|\tilde{\mathbf{p}}_\mathbf{L}(\epsilon_\mathbf{t} - \tilde{\epsilon}_\mathbf{t})\|_\mathbf{2}. \end{aligned}$$

Note that, for any vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_H^2 = \boldsymbol{x} H_t^{-1} \boldsymbol{x} \geq \boldsymbol{x}'\boldsymbol{x}/\rho(H_t)$. Since $\rho(H_t) \leq Tr(H_t) \leq \sum_{j=1}^N h_{j,t} \leq N \|D_t\|_s^2$. Therefore, we get

$$\|\boldsymbol{x}\|_2 \leq \sqrt{N} \|D_t\|_s \|\boldsymbol{x}\|_H.$$

Moreover, for every vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_H^2 = \boldsymbol{x}' H_t^{-1} \boldsymbol{x} \leq \|\boldsymbol{x}\|_2^2 \|H_t^{-1}\|_s$ (diagonalize H_t in an orthonormal basis). This means

$$\|\boldsymbol{x}\|_H \leq C_v^{1/2} \lambda_1(R_t)^{-1/2} \|\boldsymbol{x}\|_2.$$

[4] Indeed, if $\boldsymbol{x}_j = H_t e_L g_j$ for any $|L| \times 1$-vector $g_j = [\delta_{i,j}]_{j=1,\ldots,|L|}$, we check that $\mathbf{p_L}(\boldsymbol{x_j}) = \mathbf{0}$. Moreover, when a vector $\boldsymbol{v}$ belongs to $< H_t e_L >^\perp$, then $\boldsymbol{v}' H_t^{-1} H_t e_L g_j = \boldsymbol{v}' e_L g_j = 0$ for every j, i.e. $\boldsymbol{v}' e_L = 0$. This implies $\mathbf{p_L}(\boldsymbol{v}) = \boldsymbol{v}$.

Since the spectral norm is the matrix norm that is associated to the usual euclidian norm $\|\cdot\|_2$, we have

$$\begin{aligned}|r_{i|L,t} - \tilde{r}_{i|L,t}| &\leq \|(\mathbf{p_L} - \tilde{\mathbf{p}}_\mathbf{L})(\epsilon_\mathbf{t})\|_\mathbf{2} + \|\tilde{\mathbf{p}}_\mathbf{L}(\epsilon_\mathbf{t} - \tilde{\epsilon}_\mathbf{t})\|_\mathbf{2} \\ &\leq \|(\mathbf{p_L} - \tilde{\mathbf{p}}_\mathbf{L})\|_\mathbf{s}\|\epsilon_\mathbf{t}\|_\mathbf{s} + \|\tilde{\mathbf{p}}_\mathbf{L}\|_\mathbf{s}\|\epsilon_\mathbf{t} - \tilde{\epsilon}_\mathbf{t}\|_\mathbf{2} \\ &\leq \|(\mathbf{p_L} - \tilde{\mathbf{p}}_\mathbf{L})\|_\mathbf{s}\|\epsilon_\mathbf{t}\|_\mathbf{2} + \|\epsilon_\mathbf{t} - \tilde{\epsilon}_\mathbf{t}\|_\mathbf{2}.\end{aligned}$$

Furthermore,

$$\begin{aligned}&\tilde{\mathbf{p}}_L - \mathbf{p_L} = -\left(\mathbf{I_N} - \mathbf{H_t e_L}\left(\mathbf{e'_L H_t e_L}\right)^{-\mathbf{1}}\mathbf{e'_L}\right) + \left(\mathbf{I_N} - \tilde{\mathbf{H}}_\mathbf{t}\mathbf{e_L}\left(\mathbf{e'_L}\tilde{\mathbf{H}}_\mathbf{t}\mathbf{e_L}\right)^{-\mathbf{1}}\mathbf{e'_L}\right) \\ &= \left(H_t - \tilde{H}_t\right)e_L\left(e'_L H_t e_L\right)^{-1}e'_L + \tilde{H}_t e_L\left[\left(e'_L H_t e_L\right)^{-1} - \left(e'_L\tilde{H}_t e_L\right)^{-1}\right]e'_L \\ &= \left[\left(D_t - \tilde{D}_t\right)R_t D_t + \tilde{D}_t\left(R_t - \tilde{R}_t\right)D_t + \tilde{D}_t\tilde{R}_t\left(D_t - \tilde{D}_t\right)\right]e_L\left(e'_L H_t e_L\right)^{-1}e'_L \\ &+\tilde{H}_t e_L\left(e'_L H_t e_L\right)^{-1}\left[\left(e'_L\tilde{H}_t e_L\right) - \left(e'_L H_t e_L\right)\right]\left(e'_L\tilde{H}_t e_L\right)^{-1}e'_L.\end{aligned}$$

Note that $\|(e'_L H_t e_L)^{-1}\|_s$ is the inverse of the smallest eigenvalue of $e'_L H_t e_L$. By the Courant-Raleigh theorem, $\lambda_1(e'_L H_t e_L)$ is larger than $\lambda_1(H_t)$. Then, $\|(e'_L H_t e_L)^{-1}\|_s \leq \lambda_1(H_t)^{-1} = \|H_t^{-1}\|_s$. Since $H_t^{-1} = D_t^{-1}R_t^{-1}D_t^{-1}$, we obtain

$$\|(e'_L H_t e_L)^{-1}\|_s \leq \|H_t^{-1}\|_s \leq \|D_t^{-1}\|_s^2\|R_t^{-1}\|_s \leq C_v\lambda_1(R_t)^{-1}.$$

Moreover, it is easy to check that $\|e_L\|_s = \|e'_L\|_s = 1$. Since

$$\|D_t - \tilde{D}_t\|_s \leq \max_i |h_{i,t} - \tilde{h}_{i,t}|/(h_{i,t}^{1/2} + \tilde{h}_{i,t}^{1/2}) \leq C_v^{1/2}\|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s,$$

we have

$$\begin{aligned}\|\mathbf{p_L} - \tilde{\mathbf{p}}_\mathbf{L}\|_\mathbf{s} &\leq \{\|D_t - \tilde{D}_t\|_s\|R_t\|_s\|D_t\|_s + \|\tilde{D}_t\|_s\|R_t - \tilde{R}_t\|_s\|D_t\|_s + \|\tilde{D}_t\|_s\|\tilde{R}_t\|_s\|D_t - \tilde{D}_t\|_s\} \\ &\cdot \|(e'_L H_t e_L)^{-1}\|_s\left(\|1 + \|\tilde{H}_t\|_s\|(e'_L\tilde{H}_t e_L)^{-1}\|_s\right) \\ &\leq \{C_v^{1/2}\|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s N(\|D_t\|_s + \|\tilde{D}_t\|_s) \\ &+ \sqrt{N}C_F C_{\Psi^{-1}}\|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty\|D_t\|_s\|\tilde{D}_t\|_s\} \\ &\cdot C_h\lambda_1(R_t)^{-1}\left(1 + N\|\tilde{D}_t\|_s^2 C_h\lambda_1(\tilde{R}_t)^{-1}\right).\end{aligned}$$

We also have

$$\begin{aligned}\|\epsilon_t\|_2 &\leq \|D_t\|_s\|R_t^{1/2}\|_s\|\eta_t^*\|_2 \\ &\leq \|D_t\|_s\lambda_{max}^{1/2}\left(R_t\right)\|\eta_t^*\|_2 \leq \|D_t\|_s\sqrt{N}\|\eta_t^*\|_2.\end{aligned}$$

Moreover,

$$\begin{aligned}
\|\epsilon_t - \tilde{\epsilon}_t\|_s &\leq \| \left(D_t R_t^{1/2} - \tilde{D}_t \tilde{R}_t^{1/2} \right) \eta_t^* \|_s \\
&\leq \|D_t - \tilde{D}_t\|_s \|R_t^{1/2}\|_s \|\eta_t^*\|_s + \|\tilde{D}_t\|_s \|R_t^{1/2} - \tilde{R}_t^{1/2}\|_s \|\eta_t^*\|_s \\
&\leq \left[C_v^{1/2} \|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s \sqrt{N} + \|\tilde{D}_t\|_s K \|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty \right] \|\eta_t^*\|_s, \\
\text{where } K &= \frac{1}{\lambda_1^{1/2}\left(R_t\right) + \lambda_1^{1/2}\left(\tilde{R}_t\right)} \sqrt{N} C_F C_{\Psi^{-1}}.
\end{aligned}$$

Consequently, for every (i, L) deduced from the vine structure, we obtain

$$|r_{i|L,t} - \tilde{r}_{i|L,t}| \leq \alpha \|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s + \beta \|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty.$$

where

$$\left\{ \begin{array}{rcl}
\alpha &=& \sqrt{N} C_v^{1/2} \|\eta_t^*\|_s \left\{ 1 + \dfrac{N\|D_t\|_s C_h}{\lambda_1(R_t)} \{\|D_t\|_s + \|\tilde{D}_t\|_s\} \left[1 + \dfrac{N\|\tilde{D}_t\|_s^2 C_h}{\lambda_1(\tilde{R}_t)} \right] \right\}, \\
\beta &=& \sqrt{N} C_F C_{\Psi^{-1}} \|\tilde{D}_t\|_s \|\eta_t^*\|_s \left\{ \dfrac{\sqrt{N}\|D_t\|_s^2 C_h}{\lambda_1(R_t)} \left[1 + \dfrac{N\|\tilde{D}_t\|_s^2 C_h}{\lambda_1(\tilde{R}_t)} \right] + \dfrac{1}{\lambda_1^{1/2}\left(R_t\right) + \lambda_1^{1/2}\left(\tilde{R}_t\right)} \right\}.
\end{array} \right.$$

Step 2. We now focus on the discrepancy $|h_{i|L,t} - \tilde{h}_{i|L,t}|$. We have

$$\begin{aligned}
h_{i|L,t} - \tilde{h}_{i|L,t} &= e_i' \left(H_t - \tilde{H}_t \right) e_i - e_i' \left(H_t - \tilde{H}_t \right) e_L \left(e_L' H_t e_L \right)^{-1} \left(e_L' H_t e_i \right) \\
&+ e_i' \tilde{H}_t e_L \left(e_L' H_t e_L \right)^{-1} \left[e_L' \tilde{H}_t e_L - e_L' H_t e_L \right] \left(e_L' \tilde{H}_t e_L \right)^{-1} \left(e_L' H_t e_i \right) \\
&+ e_i' \tilde{H}_t e_L \left(e_L' \tilde{H}_t e_L \right)^{-1} e_L' \left(H_t - \tilde{H}_t \right) e_i,
\end{aligned}$$

which implies

$$\begin{aligned}
|h_{i|L,t} - \tilde{h}_{i|L,t}| &\leq \|H_t - \tilde{H}_t\|_s \Big[1 + C_v \lambda_1(R_t)^{-1} \|H_t\|_s + C_v \lambda_1(\tilde{R}_t)^{-1} \|\tilde{H}_t\|_s \\
&+ C_v^2 \lambda_1(R_t)^{-1} \lambda_1(\tilde{R}_t)^{-1} \|H_t\|_s \|\tilde{H}_t\|_s \Big] \\
&\leq \Big(C_v^{1/2} \|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s N \{\|D_t\|_s + \|\tilde{D}_t\|_s\} \\
&+ \sqrt{N} C_F C_{\Psi^{-1}} \|D_t\|_s \|\tilde{D}_t\|_s \|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty \Big) \\
&\cdot \left[1 + \frac{N C_v \|D_t\|_s^2}{\lambda_1(R_t)} + \frac{N C_v \|\tilde{D}_t\|_s^2}{\lambda_1(\tilde{R}_t)} + \frac{N^2 C_v^2 \|D_t\|_s^2 \|\tilde{D}_t\|_s^2}{\lambda_1(R_t)\lambda_1(\tilde{R}_t)} \right] \\
&\leq \gamma \|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s + \delta \|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty, \qquad (1.5.27)
\end{aligned}$$

where

$$\begin{cases} \gamma &= C_v^{1/2} N\{\|D_t\|_s + \|\tilde{D}_t\|_s\} \left[1 + \dfrac{NC_v\|D_t\|_s^2}{\lambda_1(R_t)} + \dfrac{NC_v\|\tilde{D}_t\|_s^2}{\lambda_1(\tilde{R}_t)} + \dfrac{N^2C_v^2\|D_t\|_s^2\|\tilde{D}_t\|_s^2}{\lambda_1(R_t)\lambda_1(\tilde{R}_t)}\right], \\ \delta &= \sqrt{N}C_F C_{\Psi^{-1}}\|D_t\|_s\|\tilde{D}_t\|_s \left[1 + \dfrac{NC_v\|D_t\|_s^2}{\lambda_1(R_t)} + \dfrac{NC_v\|\tilde{D}_t\|_s^2}{\lambda_1(\tilde{R}_t)} + \dfrac{N^2C_v^2\|D_t\|_s^2\|\tilde{D}_t\|_s^2}{\lambda_1(R_t)\lambda_1(\tilde{R}_t)}\right]. \end{cases}$$

Consequently, we obtain the following relationship for (1.5.25)

$$\begin{aligned} v_{i,j|L,t} - \tilde{v}_{i,j|L,t} &= \frac{\left(r_{i|L,t} - \tilde{r}_{i|L,t}\right) r_{j|L,t}}{\sqrt{h_{i|L,t}}\sqrt{h_{j|L,t}}} + \frac{\tilde{r}_{i|L,t}\left(r_{j|L,t} - \tilde{r}_{j|L,t}\right)}{\sqrt{h_{i|L,t}}\sqrt{h_{j|L,t}}} \\ &+ \tilde{r}_{i|L,t}\tilde{r}_{j|L,t}\{\frac{1}{\sqrt{h_{i|L,t}}\sqrt{h_{j|L,t}}} - \frac{1}{\sqrt{\tilde{h}_{i|L,t}}\sqrt{\tilde{h}_{j|L,t}}}\}. \end{aligned}$$

For any (i, L) we consider, $h_{i|L,t} \leq \|D_t\|_s^2$ everywhere, because the variance of a residual is smaller than the variance of any random variable. Therefore, we get

$$\begin{aligned} |\frac{1}{\sqrt{h_{i|L,t}}\sqrt{h_{j|L,t}}} - \frac{1}{\sqrt{\tilde{h}_{i|L,t}}\sqrt{\tilde{h}_{j|L,t}}}| &\leq \frac{C_h^2}{\sqrt{\tilde{h}_{i|L,t}}\sqrt{\tilde{h}_{j|L,t}} + \sqrt{\tilde{h}_{i|L,t}}\sqrt{\tilde{h}_{j|L,t}}}\{h_{i|L,t}h_{j|L,t} - \tilde{h}_{i|L,t}\tilde{h}_{j|L,t}\} \\ &\leq C_h^3\left[\left(h_{i|L,t} - \tilde{h}_{i|L,t}\right)h_{j|L,t} + \tilde{h}_{i|L,t}\left(h_{j|L,t} - \tilde{h}_{j|L,t}\right)\right] \\ &\leq C_h^3\{\|D_t\|_s^2|h_{i|L,t} - \tilde{h}_{i|L,t}| + \|\tilde{D}_t\|_s^2|h_{j|L,t} - \tilde{h}_{j|L,t}|\}, \end{aligned} \tag{1.5.28}$$

and

$$|r_{i|L,t}| \leq \|\mathbf{p_L}\left(\epsilon_\mathbf{t}\right)\|_\infty \leq \|\mathbf{p_L}\left(\epsilon_\mathbf{t}\right)\|_\mathbf{2} \leq \|\mathbf{p_L}\|_\mathbf{s}.\|\epsilon_\mathbf{t}\|_\mathbf{2} \leq \|\epsilon_\mathbf{t}\|_\mathbf{2} \leq \sqrt{\mathbf{N}}\|\mathbf{D_t}\|_\mathbf{s}\|_\mathbf{s}\|\eta_\mathbf{t}^*\|_\mathbf{2}. \tag{1.5.29}$$

Consequently, using (1.5.27), (1.5.28) and (1.5.29), (1.5.25) can be upper bounded as

$$\begin{aligned} |v_{i,j|L,t} - \tilde{v}_{i,j|L,t}| &\leq C_h\sqrt{N}\left(\|D_t\|_s + \|\tilde{D}_t\|_s\right) \\ \cdot &\left\{\left(\alpha\|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s + \beta\|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty\right)\right. \\ + &\left.\sqrt{N}\|\tilde{D}_t\|_s^2\|\eta_t^*\|_2^2 C_h^2\left(\gamma\|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s + \delta\|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty\right)\right\}. \end{aligned}$$

Hence using the previous inequality, we obtain

$$\|\zeta_t - \tilde{\zeta}_t\|_\infty \leq \Upsilon_{1,t}\|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s + \Upsilon_{2,t}\|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty, \tag{1.5.30}$$

with

$$\begin{cases} \Upsilon_{1,t} &= C_h\sqrt{N}\left(\|D_t\|_s + \|\tilde{D}_t\|_s\right)\{\alpha + \sqrt{N}\|\tilde{D}_t\|_s^2\|\eta_t^*\|_2^2 C_h^2\gamma\}, \\ \Upsilon_{2,t} &= C_h\sqrt{N}\left(\|D_t\|_s + \|\tilde{D}_t\|_s\right)\{\beta + \sqrt{N}\|\tilde{D}_t\|_s^2\|\eta_t^*\|_2^2 C_h^2\delta\}. \end{cases}$$

Using (1.5.30) and (1.5.24), we have

$$\begin{aligned} \|\Psi\left(Pc_t\right) - \Psi\left(\tilde{P}c_t\right)\|_\infty &\leq \{\|\Xi\|_\infty + \|\Lambda\|_\infty\Upsilon_{2,t}\}\|\Psi\left(Pc_{t-1}\right) - \Psi\left(\tilde{P}c_{t-1}\right)\|_\infty \\ &+ \|\Lambda\|_\infty\Upsilon_{1,t}\|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s. \end{aligned} \tag{1.5.31}$$

We denote

$$\|\mu_t\| = \begin{pmatrix} \|\Psi\left(Pc_{t-1}\right) - \Psi\left(\tilde{P}c_{t-1}\right)\|_\infty \\ \|\mathrm{Vecd}\left(D_t^2\right) - \mathrm{Vecd}\left(\tilde{D}_t^2\right)\|_s \end{pmatrix}, \; M_t = \begin{pmatrix} \|\Xi\|_\infty + \|\Lambda\|_\infty\Upsilon_{2,t} & \|\Lambda\|_\infty\Upsilon_{1,t} \\ \Gamma_{2,t} & \Gamma_{1,t} \end{pmatrix}.$$

Using (1.5.23) and (1.5.31), we deduce that

$$\begin{aligned} \|\mu_t\| &\leq M_t\|\mu_{t-1}\| \\ &\leq \{\prod_{k=0}^{t-p} M_{t-k}\}\|\mu_{t-p-1}\|, \end{aligned}$$

for any $p \in \mathbb{N}$. Under assumption 7, $\lim_{p\to\infty}\|M_tM_{t-1}\cdots M_{t-p}\| = 0$ $\mathbb{P}-$a.s., for a fixed t using Lemma 2.1 of Francq and Zakoian (2010). We deduce that $\mu_t \underset{t\to\infty}{\longrightarrow} 0$. This implies that $\Psi(Pc_t) = \Psi(\tilde{P}c_t)$ a.s. and $D_t = \tilde{D}_t$ a.s., which then implies $R_t = \tilde{R}_t$ a.s. and $\epsilon_t = \tilde{\epsilon}_t$ a.s.. This concludes the proof of uniqueness. Furthermore, ergodicity is obtained as a consequence of corollary 7.17 in Douc, Moulines and Stoffer (2014).

A sufficient condition for uniqueness is that the top Lyapunov exponent γ_M is strictly negative. This condition holds if $\mathbb{E}\left[\log\left(\|M_t\|\right)\right] < 0$.

□

1.6 Asymptotic theory

The conditions for the existence and uniqueness of a strictly stationary solution of the vine-GARCH process have been established. We thus can provide a sound asymptotic theory. In this section, we state the asymptotic properties of the two-step quasi-maximum likelihood estimator, but not the estimator obtained by the iterative process,

for which the limiting behavior would be more complex. Our vine-GARCH model is specified by choosing a R-vine, and by Equations (1.3.1)-(1.3.5).

The proofs of consistency and asymptotic normality require some matrix computations, in particular the differentiation of some quantities involving matrices. Recalling some results recorded in Lütkepohl (1996), we have

$$\begin{aligned}
\frac{\partial x'Xx}{\partial X} &= xx',\ X \in \mathcal{M}_{m\times m}(\mathbb{R}), x \in \mathbb{R}^m,\\
\frac{\partial \text{Trace}\,(AX'B)}{\partial X} &= BA,\ X \in \mathcal{M}_{m\times n}(\mathbb{R}), A \in \mathcal{M}_{p\times n}(\mathbb{R}), B \in \mathcal{M}_{m\times p}(\mathbb{R})\\
\frac{\partial \text{Trace}\,(AX^{-1}B)}{\partial X} &= -\left(X^{-1}BAX^{-1}\right)',\ X \in \mathcal{M}_{m\times m}(\mathbb{R}),\ \text{nonsingular}, A, B \in \mathcal{M}_{m\times m}(\mathbb{R}),\\
\frac{\partial \log\left(\det(X)\right)}{\partial X} &= (X')^{-1},\ X \in \mathcal{M}_{m\times m}(\mathbb{R}),\ \text{nonsingular},\\
\frac{\partial X^{-1}}{\partial x} &= -(X')^{-1}(\partial_x X)X^{-1},\ X \in \mathcal{M}_{m\times m}(\mathbb{R}),\ \text{nonsingular}.
\end{aligned}$$

For convenience and to get explicit assumptions, assume hereafter that any conditional variance series follows a univariate GARCH process defined as

$$h_{i,t} = \varsigma_i + \sum_{k=1}^{q_i} \kappa_{i,k}\epsilon_{i,t-k}^2 + \sum_{l=1}^{p_i} \tau_{i,l}h_{i,t-l}, \tag{1.6.1}$$

such that $\theta_v^{(i)} = (\varsigma_i, \boldsymbol{\kappa}_i, \boldsymbol{\tau}_i)' \in \mathbb{R}_+^{p_i+q_i+1}$ for all $i = 1, \ldots, N$. It would be more or less straightforward to obtain similar theoretical results with different volatility dynamics, such as spill-over effects. Nonetheless, this would induce additional technicalities that would digress us from the core of the vine-GARCH models.

Assume we observe a T-path $(\epsilon_t)_{t=1,\ldots,T}$ that corresponds to a realization drawn following a unique, strictly stationary and non-anticipative solution $(\epsilon_t)_{t\in\mathbb{Z}}$ of this model. We will denote by $D_t(\theta)$, $R_t(\theta)$ and $H_t(\theta)$ the t-matrices of conditional volatilities, conditional correlations and conditional covariances respectively, as generated by our model and assuming θ is the underlying parameter. We estimate this model by a Gaussian QMLE and by applying the two-step estimation method of Section 1.4.

To calculate log-likelihoods, a practical issue is the choice of some initial values to generate the sequences (D_t), (R_t) and then (H_t), $t = 1, \ldots, T$. Given some fixed values for ϵ_0, D_0 and R_0, we obtain log-likelihoods. In this Section only, the latter

log-likelihoods will be denoted by $\widetilde{\mathbb{G}}_T l_1(\epsilon;\theta_v)$ and $\widetilde{\mathbb{G}}_T l_2(\epsilon;\theta_v,\theta_c)$. More generally, all quantities with a " ˜ " are deduced from the process with fixed arbitrary starting values at $t=0$. Therefore, they are distinct from the "theoretical" log-likelihoods $\mathbb{G}_T l_1(\epsilon;\theta_v)$ and $\mathbb{G}_T l_2(\epsilon;\theta_v,\theta_c)$, for which the initial values are coming from the stationary laws. Equivalently, they can be seen as coming from a stationary solution $(\epsilon_t)_{t\in\mathbb{Z}}$. Actually, this subtlety has no consequence because we will assume irrelevance of initial values: see assumption 13 and 15.

In the following, we use the sub-multiplicative matrix norm $\|A\| := \sup\{\frac{\|Ax\|}{\|x\|},\ x\neq 0\}$, for any $A\in\mathcal{M}_{n\times m}(\mathbb{R})$, $x\in\mathbb{R}^m$ and $\|x\|$ denotes the Euclidean norm of x. We also need the spectral radius norm of squared non-negative matrices, which is submultiplicative: $\|A\|_s := \max\{|\lambda_i| : \mathrm{Spec}(A)=(\lambda_1,\cdots,\lambda_m)\}$.

1.6.1 Consistency

Assumption 8. The variance parameters θ_v (resp. correlation parameters θ_c) belong to a compact set Θ_v in $\mathbb{R}_+^s$, $s:=\sum_{i=1}^N(p_i+q_i+1)$ (resp. Θ_c in in $\mathbb{R}^{(p+q)N^2(N-1)^2/4+N(N-1)/2}$). The true parameter $\theta_0=(\theta_{0,v},\theta_{0,c})'$ belongs to the interior of the compact set $\Theta := \Theta_v\times\Theta_c$.

Denoting by ρ_Ξ the spectral radius of the companion block-matrix associated to $(\Xi_1,\ldots,\Xi_p)$, a necessary condition is $\rho_\Xi<1$ in particular. When $p=1$, this means simply that all eigenvalues of $\Xi_1:=\Xi$ are smaller than one in absolute value.

Assumption 9. The sequence of innovations (η_t) is strongly stationary. The law of η_t given $\mathcal{F}_{t-1}$ is elliptical s.t. $\mathbb{E}_{t-1}[\eta_t]=0$ and $\mathbb{E}_{t-1}[\eta_{i,t}|\eta_{j,t}]=0$ when $i\neq j$.

In particular, every $\eta_{k,t}^2$, $k=1,\ldots,N$, has a nondegenerate conditional distribution. With an underlying elliptical distribution, the conditional expectation of any η_{it} given $M\eta_t$ is a linear transform of η_t, for an arbitrary $m\times N$ matrix M, $m<N$. This property is necessary to ensure the identifiability of vine-GARCH processes. Considering elliptical random vectors (including Gaussian ones) can be seen as restrictive, but it is convenient and realistic here. This implies that the true DGP can induce fatter tails than conditionally Gaussian processes, for instance by choosing student-distributed noises η_t.

Set the polynomials $\mathcal{A}_{i,\theta}(z)=\sum_{k=1}^{q_i}\kappa_{i,k}z^k$ and $\mathcal{B}_{i,\theta}(z)=1-\sum_{l=1}^{p_i}\tau_{i,l}z^l$.

Assumption 10. For every $i = 1, \dots, N$, when $p_i > 0$, the polynomials $\mathcal{A}_{i,\theta_0}(z)$ and $\mathcal{B}_{i,\theta_0}(z)$ have no common roots, $\mathcal{A}_{i,\theta_0}(1) \neq 0$ and $\kappa_{i,q_i} + \tau_{i,p_i} \neq 0$.

For any $i = 1, \dots, N$, let the random matrix

$$\mathbf{A}_{i,0,t} := \begin{bmatrix} \kappa_{i,1}\eta_{i,t}^2 & \dots & \dots & \dots & \kappa_{i,q_i}\eta_{i,t}^2 & \tau_{i,1}\eta_{i,t}^2 & \dots & \dots & \dots & \tau_{i,p_i}\eta_{i,t}^2 \\ 1 & 0 & \dots & \dots & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ \kappa_{i,1} & \dots & \dots & \dots & \kappa_{i,q_i} & \tau_{i,1} & \dots & \dots & \dots & \tau_{i,p_i} \\ 0 & \dots & \dots & \dots & 0 & 1 & 0 & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 & 0 & \dots & 0 & 1 & 0 \end{bmatrix},$$

that depends on the true parameter values.

Assumption 11. For any $i = 1, \dots, N$, the top Lyapunov exponent $\gamma(\mathbf{A}_{i,0}) := \gamma_i$, defined as

$$\gamma_i := \inf_{t \in \mathbb{N}^\star} \frac{1}{t} \mathbb{E}\left[\log\left(\|\mathbf{A}_{i,0,t}\mathbf{A}_{i,0,t-1} \cdots \mathbf{A}_{i,0,1}\|\right)\right] = \lim_{t \to \infty} \frac{1}{t} \log\left(\|\mathbf{A}_{i,0,t}\mathbf{A}_{i,0,t-1} \cdots \mathbf{A}_{i,0,1}\|\right) \text{ a.s.}$$

is strictly negative. Moreover, for all $\theta_{i,v}$, $\sum_{l=1}^{p_i} \tau_{i,l} < 1$.

Such assumptions on Lyapunov exponents are standard in the GARCH literature. When $p = q = 1$, this is equivalent to $\mathbb{E}[\ln\left(\kappa_{i,1}\eta_{i,t}^2 + \tau_{i,1}\right)] < 0$. More generally, it is sufficient to check that $\mathbb{E}\left[\ln \|A_{i,0,t}A_{i,0,t-1} \dots A_{i,0,1}\|\right] < 0$.

Define $\mathcal{D}_\theta(z) = \sum_{l=1}^{q} \Lambda_l z^l$ and $\mathcal{Q}_\theta(z) = I_N - \sum_{k=1}^{p} \Xi_k z^k$. The following technical assumptions is required to get the identifiability of θ_c (see Section 11.4.1. in Francq and Zakoïan, 2010, for formal definitions of "left coprime" and of the matrix $M(\cdot, \cdot)$).

Assumption 12. For any $\theta \in \Theta$, $\mathcal{Q}_\theta(z)$ is nonsingular, i.e. the roots of $det(\mathcal{Q}_\theta(z)) = 0$ are outside the unit disk. If $p > 0$, $\mathcal{D}_{\theta_0}(z)$ and $\mathcal{Q}_{\theta_0}(z)$ are left coprime and $M(\mathcal{D}_{\theta_0}, \mathcal{Q}_{\theta_0}(z))$ has full rank $N(N-1)/2$.

Assumption 13. The initial values are asymptotically irrelevant, which means

$$\sup_{\theta \in \Theta} |\mathbb{G}_T l_2(\epsilon; \theta) - \widetilde{\mathbb{G}}_T l_2(\epsilon; \theta)| = o_p(1).$$

This assumption is proved as a technical result in Appendix A, due to its technicality.

Assumption 14. There exists a number $a \in (0,1)$ such that, for almost every trajectory and every $\theta \in \Theta$, the partial correlations associated to our R-vine (i.e. the components of the vectors $Pc_t(\theta)$) belong to the fixed interval $[-1+a, 1-a]$.

The latter assumption implies that, for every $\theta \in \Theta$, the determinant of almost every correlation matrices $R_t(\theta)$ is strictly larger than $a^{N(N-1)} > 0$ (apply Kurowicka and Cooke, 2006, Theorem 3.2), and that the norm of $R_t^{-1}(\theta)$ is bounded from above a.e. Indeed, $\|R_t^{-1}\|_s \leq \lambda_{min}(R_t)^{-N} \leq a^{N^2(N-1)}$. Moreover, the function $F_{\text{vine}}(.)$ that maps partial correlations to usual correlations has a bounded derivative, when applied to the trajectories $(Pc_t(\theta))$ generated by the model.

The next assumption allows to control the influence of the first step estimator $\hat{\theta}_{T,v}$ on the second step estimator.

Assumption 15. If $(\tilde{\theta}_{T,v})$ is a sequence in Θ_v that tends to $\theta_{0,v}$ in probability, then

$$\sup_{\theta \in \Theta} |\mathbb{G}_T l_2(\epsilon, \tilde{\theta}_{T,v}; \theta_c) - \mathbb{G}_T l_2(\epsilon, \theta_{0,v}; \theta_c)| = o_P(1).$$

This assumption is proved as a technical result in Appendix B. There, the influence of the correlation-related parameters Ω, Ξ and Λ appears explicitly.

Theorem 1.6.19. *Let $\hat{\theta}_T = (\hat{\theta}_{T,v}, \hat{\theta}_{T,c})'$ be a sequence of QML estimators defined by (1.4.1) and (1.4.2). Then, under assumptions 8-15, $\hat{\theta}_T \xrightarrow{\mathbb{P}} \theta_0$ when $T \to \infty$.*

Set $\theta_{0\backslash c} = (\theta_{0,v}, \theta_c)$. The consistency proof requires some preliminary lemmas. The next three steps will be demonstrated successively.

1. Identifiability of the parameters, which can be expressed in our framework as: $\{\forall t \in \mathbb{Z},\ D_t(\theta_v) = D_t(\theta_{0,v}) \text{ and } R_t(\theta) = R_t(\theta_0)\ \mathbb{P}_{\theta_0} \text{ as}\} \Rightarrow \theta = \theta_0$.

2. The optimum θ_0 is well-separated: if $\|\theta_c - \theta_{0,c}\| > \gamma$ for some $\gamma > 0$, then $l_{2,t}(\epsilon_t; \theta_{0,v}, \theta_{0,c}) \in L^1(\mathbb{R})$ and $\mathbb{E}_{\theta_0}[l_{2,t}(\epsilon_t; \theta_{0,v}, \theta_c)] > \mathbb{E}_{\theta_0}[l_{2,t}(\epsilon_t; \theta_0)]$.

3. Let $\Theta_{0\backslash c} = \{\theta = (\theta_{0,v}, \theta_c) \in \Theta\} = \{\theta_{0,v}\} \times \Theta_c$. For every $\theta^* \in \Theta_{0\backslash c}$ with $\|\theta_c^* - \theta_{0,c}\| > 0$ and every $\pi > 0$, there exists an open ball $V(\theta^*, \pi)$ around θ^* in the space $\Theta_{0\backslash c}$ s.t.

$$\mathbb{E}_{\theta_0}\left[\inf_{\theta \in V(\theta^*, \pi)} l_{2,t}(\epsilon_t; \theta)\right] \geq \mathbb{E}_{\theta_0}[l_{2,t}(\epsilon_t; \theta^*)] - \pi.$$

Proof of Step 1. Our assumptions 9-11 insure the identifiability of the GARCH(p_i,q_i) models: when $D_t(\theta_v) = D_t(\theta_{0,v})$ for every t and almost everywhere, this means that $\theta_v = \theta_{0,v}$ (see Francq and Zakoïan (2010), Theorem 7.1).

Now, let us state the identifiability of the correlation-related parameters. There is a one-to-one relationship between the components of the lower (or upper) triangular part of $R_t(\theta)$, and $Pc_t(\theta)$, the vector of partial correlations, through $F_{\text{vine}}(.)$. Then $R_t(\theta) = R_t(\theta_0)$ $\mathbb{P}_{\theta_0}$ a.s. implies $Pc_t(\theta) = Pc_t(\theta_0)$ $\mathbb{P}_{\theta_0}$ a.s. For a given sequence of innovations (η_t), we write the partial correlation dynamics as

$$\mathcal{Q}_\theta(B)\Psi\left(Pc_t(\theta)\right) = \Omega + \mathcal{D}_\theta(B)\zeta_t(\theta), \text{ or } \Psi\left(Pc_t(\theta)\right) = \mathcal{Q}_\theta^{-1}(B)\mathcal{D}_\theta(B)\zeta_t(\theta) + \mathcal{Q}_\theta^{-1}(1)\Omega,$$

because $\mathcal{Q}_\theta$ is invertible (assumption 12). Set $\mathcal{P}_\theta(z) := \mathcal{Q}_\theta^{-1}(z)\mathcal{D}_\theta(z)$. Since we assume $R_t(\theta) = R_t(\theta_0)$, $D_t(\theta) = D_t(\theta_0)$ for all t and some θ and θ_0 in Θ, then $H_t(\theta) = H_t(\theta_0)$ and the observations ϵ_t are the same under $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta_0}$. This implies that $\zeta_t(\theta) = \zeta_t(\theta_0) := \zeta_t$ and

$$\left(\mathcal{P}_\theta(B) - \mathcal{P}_{\theta_0}(B)\right)\zeta_t(\theta) = \mathcal{Q}_{\theta_0}^{-1}(1)\Omega_0 - \mathcal{Q}_\theta^{-1}(1)\Omega. \tag{1.6.2}$$

There exists a sequence of matrices $(\mathcal{P}_k)$ s.t. $\mathcal{P}_\theta(B) - \mathcal{P}_{\theta_0}(B) =: \sum_{k\geq 0}\mathcal{P}_k B^k$. Note that $\mathcal{P}_0 = 0$. Isolating the terms that are functions of ζ_{t-1}, we see there exists a random variable $\mathcal{V}_{t-2}$ that is $\mathcal{F}_{t-2}-$mesurable s.t. $\mathcal{P}_1\zeta_{t-1} = \mathcal{V}_{t-2}$ a.s. If $\mathcal{P}_1$ is not zero, its kernel is included in an hyperplan in $\mathbb{R}^{N(N-1)/2}$. Therefore, there exists a constant non-zero vector ϖ and an $\mathcal{F}_{t-2}-$measurable variable $\varkappa_{t-2}$ s.t.

$$\varpi'\zeta_{t-1} = \varkappa_{t-2} \text{ a.s.} \tag{1.6.3}$$

Recall that the $N(N-1)/2$ components of the vector ζ_{t-1} are based on cross-products of the returns $\epsilon_{i,t-1}$. To be specific,

$$\zeta_{t-1} = \left[(\epsilon_{i,t-1} - \mathbb{E}_{t-2}[\epsilon_{i,t-1}|\epsilon_{L,t-1}]).(\epsilon_{j,t-1} - \mathbb{E}_{t-2}[\epsilon_{j,t-1}|\epsilon_{L,t-1}])/\sqrt{h_{i,t-1}h_{j,t-1}}\right],$$

denoting by $(ij|L)$ the nodes of the vine. Note that the volatilities $h_{i,t-1}$ are $\mathcal{F}_{t-2}-$measurable, $i = 1,\ldots,N$. Moreover, since η_t is conditionally elliptical by assumption 9, there exists $\mathcal{F}_{t-2}-$measurable vectors $m_{i,L,t-2}$ s.t. $\mathbb{E}_{t-2}[\epsilon_{i,t-1}|\epsilon_{L,t-1}] = m'_{i,L,t-2}\epsilon_{L,t-1}$. Since ϵ_{t-1} is a $\mathcal{F}_{t-2}-$measurable linear transform of η_{t-1}, Equation (1.6.3) becomes

$$\eta'_{t-1}\Upsilon_{t-2}\eta_{t-1} = \varkappa^*_{t-2} \text{ a.s.} \tag{1.6.4}$$

for some $\mathcal{F}_{t-2}-$measurable random matrix (resp. variable) Υ_{t-2} (resp. $\varkappa^*_{t-2}$). Obviously, $\Upsilon_{t-2} = 0$ implies $\varpi = 0$, contradicting $rg(\mathcal{P}_1) > 0$.

Now, let us prove that $\Upsilon_{t-2} = 0$. It is well-known that any standardized elliptical vector, say η_{t-1}, can be decomposed as $\eta_{t-1} = S_{t-1}.Z_{t-1}$, where S_{t-1} is a positive random variable, $Z_{t-1} \sim \mathcal{N}(0_N, I_N)$ and S_{t-1} and Z_{t-1} are independent. By construction, S_{t-1} and Z_{t-1} are functions of η_{t-1} and are then $\mathcal{F}_{t-1}-$measurable, but not $\mathcal{F}_{t-2}-$measurable (if non-degenerate). Then, Equation (1.6.4) may be rewritten

$$S^2_{t-1}.\left(Z'_{t-1}\Upsilon_{t-2}Z_{t-1}\right) = \varkappa^*_{t-2} \text{ a.s.}$$

Given an (arbitrary) realization of $(\eta_{t-2}, \eta_{t-3}, \ldots)$ and invoking the independence between S_{t-1} and Z_{t-1}, we deduce that S_{t-1} and $Z'_{t-1}\Upsilon_{t-2}Z_{t-1}$ are a.e. $\mathcal{F}_{t-2}-$measurable variables. This is possible only if Υ_{t-2} is zero. Therefore, this proves that $\mathcal{P}_1 = 0$.

By a similar reasoning, we prove successively that $\mathcal{P}_k = 0$, for any $k > 0$, and then $\mathcal{P}_\theta(B) = \mathcal{P}_{\theta_0}(B)$. By assumption 12, this is sufficient to insure that $\mathcal{D}_\theta = \mathcal{D}_{\theta_0}$ and $\mathcal{Q}_\theta = \mathcal{Q}_{\theta_0}$ (see the arguments in Francq and Zakoïan, 2010, Section 11.4.1). As a consequence, Ξ_k and Λ_l are uniquely identified from the sequence (η_t) on. And, through Equation (1.6.2), we check easily that $\Omega = \Omega_0$, i.e. that Ω is identified too. □

Proof of Step 2. We now show that the limit criterion is minimized at the true value. It is important to note that the second step is conditional on the first step estimator, i.e. we deal with $l_{2,t}(\epsilon_t; \hat{\theta}_{T,v}, \theta_c)$. For all $\theta \in \Theta$,

$$\mathbb{E}_{\theta_0}\left[l^-_{2,t}(\epsilon_t;\theta)\right] \leq \mathbb{E}_{\theta_0}\left[\log^-(|R_t|)\right] \leq \mathbb{E}\left[\max\left(0, -\log(|R_t|)\right)\right] < \infty,$$

by assumption 14. Consequently, $\mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta)\right]$ belongs to $\mathbb{R} \cup \{+\infty\}$. Actually, $\mathbb{E}_{\theta_0}\left[|l_{2,t}(\epsilon_t;\theta_0)|\right] < \infty$.

Indeed, the determinant of $R_t(\theta_0)$ is bounded from above by $\mathrm{Tr}(R_t)^N = N^N$. Thus, due to the properties of the trace operator, we have

$$\mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta_0)\right] = \mathbb{E}_{\theta_0}\left[\log|R_t(\theta_0)|\right] + \mathrm{Tr}\left(R_t^{-1}\mathbb{E}_{\theta_0}\left[u_t u'_t\right]\right) \leq N\log N + N.$$

Therefore, we obtain that $l_{2,t}(\epsilon_t;\theta_0)$ belongs to L^1.

Denote by $\alpha_{i,t}$ the eigenvalues of $R_t(\theta_0)R_t^{-1}(\theta_{0\backslash c})$, $\theta_{0\backslash c} = (\theta_{0,v}, \theta_c)$, $i = 1, \ldots, N$. They are positive. Setting $u_t = D_t(\theta_{0,v})^{-1}\epsilon_t$, we have

$$\begin{aligned}
&\mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta_{0\backslash c}) - l_{2,t}(\epsilon_t;\theta_0)\right] \\
&= \mathbb{E}_{\theta_0}\left[\log(|R_t(\theta_{0\backslash c})||R_t^{-1}(\theta_0)|)\right] + \mathbb{E}_{\theta_0}\left[u_t'\left(R_t^{-1}(\theta_{0\backslash c}) - R_t^{-1}(\theta_0)\right)u_t\right] \\
&= \mathbb{E}_{\theta_0}\left[\log(|R_t(\theta_{0\backslash c})||R_t^{-1}(\theta_0)|)\right] + \mathbb{E}_{\theta_0}\left[\mathrm{Tr}\left(\left(R_t^{-1}(\theta_{0\backslash c}) - R_t^{-1}(\theta_0)\right)u_t u_t'\right)\right] \\
&= \mathbb{E}_{\theta_0}\left[\log(|R_t(\theta_{0\backslash c})||R_t^{-1}(\theta_0)|)\right] + \mathbb{E}_{\theta_0}\left[\mathrm{Tr}\left(\left(R_t^{-1}(\theta_{0\backslash c}) - R_t^{-1}(\theta_0)\right)\mathbb{E}_{t-1}[u_t u_t']\right)\right] \\
&= \mathbb{E}_{\theta_0}\left[\sum_{i=1}^{N}\left(\alpha_{i,t} - 1 - \log(\alpha_{i,t})\right)\right] \geq 0.
\end{aligned}$$

The inequality $\log(x) \leq x - 1$ holds if and only if $x = 1$. In our case, that means $\alpha_{i,t} = 1$, for all i, i.e. $R_t(\theta_{0\backslash c}) = R_t(\theta_0)$ a.s. By stationarity, this reasoning can be made at time $t-1$, which would give $R_{t-1}(\theta_{0\backslash c}) = R_{t-1}(\theta_0)$ a.s. Hence for any t, the relationship $R_t(\theta_{0\backslash c}) = R_t(\theta_0)$ $\mathbb{P}_{\theta_0}$, a.s. holds by stationarity. By step 1, this means $\theta_0 = \theta_{0\backslash c}$. □

Proof of Step 3. For a given $\theta^* \in \Theta_{0\backslash v}$, $\theta_c^* \neq \theta_{0,c}$, consider a sequence of open balls of radius $1/k$, $k \in \mathbb{N}$ defined by $V_k(\theta^*) := \{\theta \in \Theta_{0\backslash v} |\, \|\theta - \theta^*\| \leq 1/k\}$. Since the sequence of random variable $(\inf_{\theta \in V_k(\theta^*)} l_{2,t}(\epsilon_t;\theta))_k$ is increasing, the Beppo-Levi Theorem applies:

$$\lim_{k\to\infty} \mathbb{E}_{\theta_0}\left[\inf_{\theta \in V_k(\theta^*)} l_{2,t}(\epsilon_t;\theta)\right] = \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta^*)\right],$$

providing the result. □

Proof of Theorem 1.6.19. Under our assumptions, $\hat{\theta}_{T,v}$ converges weakly to $\theta_{0,v}$ (see Theorem 7.1 in Francq and Zakoïan, 2010, e.g.). Now, let us prove the weak convergence of $\hat{\theta}_{T,c}$ to $\theta_{0,c}$, that is, for all $\alpha > 0$, $\lim_{T\to\infty} \mathbb{P}(\|\hat{\theta}_{T,c} - \theta_{0,c}\| > \alpha) = 0$. Invoking Step 3, for any given $\pi > 0$ and for every $\theta^* \in \Theta_{0\backslash c}$, $\theta^* \neq \theta_0$ with $\|\theta_c^* - \theta_{0,c}\| \geq \alpha/2$, we can find an open ball $U(\theta^*) \subset \Theta_{0\backslash c}$ s.t.

$$\mathbb{E}_{\theta_0}\left[\inf_{\theta \in U(\theta^*)} l_{2,t}(\epsilon_t;\theta)\right] \geq \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta^*)\right] - \pi.$$

Since the function $\theta \mapsto \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta_{0,v},\theta_c)\right] - \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta_0)\right]$, defined on $\Theta_{0\backslash c}$, is strictly positive (c.f. Step 2) and continuous on the compact subset $\mathcal{C}_0(\alpha) := \{\theta \in \Theta_{0\backslash c} |\, \|\theta_c - \theta_{0,c}\| \geq \alpha/2\}$, it reaches its minimum $2\mu > 0$. Therefore, for any given $\theta^* \in \mathcal{C}_0(\alpha)$, set $\pi := \pi(\theta^*) = \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta^*)\right] - \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta_0)\right] - \mu > 0$.

Moreover, set $U(\theta_0) := \{\theta \in \Theta_{0\backslash c} : \|\theta - \theta_0\| < \alpha\}$. Then

$$\Theta_{0\backslash c} \subset U(\theta_0) \cup \bigcup_{\theta \in \mathcal{C}_0(\alpha)} U(\theta).$$

Since $\Theta_{0\backslash c}$ can be covered by a finite set of open balls, there is a finite set of points $\theta_1, \ldots, \theta_n$ in $\mathcal{C}_0(\alpha)$ s.t. $\Theta_{0\backslash c} \subset U(\theta_0) \cup \bigcup_{i=1,\ldots,n} U(\theta_i)$. We deduce

$$\mathbb{P}\left(\|\hat{\theta}_{T,c} - \theta_{0,c}\| > \alpha\right) \leq \mathbb{P}\left((\theta_{0,v}, \hat{\theta}_{T,c}) \in \cup_{i=1}^{n} U(\theta_i)\right) \leq \sum_{i=1}^{n} \mathbb{P}\left((\theta_{0,v}, \hat{\theta}_{T,c}) \in U(\theta_i)\right).$$

By definition of $\hat{\theta}_T$ and for all $i = 1, \cdots, n$, we obtain

$$\begin{aligned}
\mathbb{P}\left((\theta_{0,v}, \hat{\theta}_{T,c}) \in U(\theta_i)\right) &\leq \mathbb{P}\left(\inf_{\theta \in U(\theta_i)} \widetilde{\mathbb{G}}_T l_2(\epsilon; \theta) \leq \widetilde{\mathbb{G}}_T l_2(\epsilon; \theta_{0,v}, \hat{\theta}_{T,c})\right) \\
&\leq \mathbb{P}\left(\inf_{\theta \in U(\theta_i)} \mathbb{G}_T l_2(\epsilon; \theta) \leq \mathbb{G}_T l_2(\epsilon; \hat{\theta}_T) + 2 \sup_{\theta \in \Theta} |\mathbb{G}_T l_2(\epsilon; \theta) - \widetilde{\mathbb{G}}_T l_2(\epsilon; \theta)|\right. \\
&+ \left. |\mathbb{G}_T l_2(\epsilon; \theta_{0,v}, \hat{\theta}_{T,c}) - \mathbb{G}_T l_2(\epsilon; \hat{\theta}_T)|\right) \\
&\leq \mathbb{P}\left(\inf_{\theta \in U(\theta_i)} \mathbb{G}_T l_2(\epsilon; \theta) \leq \mathbb{G}_T l_2(\epsilon; \hat{\theta}_{T,v}, \theta_{0,c})| + 2 \sup_{\theta \in \Theta} |\mathbb{G}_T l_2(\epsilon; \theta) - \widetilde{\mathbb{G}}_T l_2(\epsilon; \theta)|\right. \\
&+ \left. |\mathbb{G}_T l_2(\epsilon; \theta_{0,v}, \hat{\theta}_{T,c}) - \mathbb{G}_T l_2(\epsilon; \hat{\theta}_T)|\right) \\
&\leq \mathbb{P}\left(\mathbb{E}_{\theta_0}\left[\inf_{\theta \in U(\theta_i)} l_{2,t}(\epsilon_t; \theta)\right] \leq \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t; \theta_0)\right] + 2 \sup_{\theta \in \Theta} |\mathbb{G}_T l_2(\epsilon; \theta) - \widetilde{\mathbb{G}}_T l_2(\epsilon; \theta)|\right. \\
&+ \left. |\mathbb{G}_T l_2(\epsilon; \theta_0) - \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t; \theta_0)\right]| \quad + |\mathbb{G}_T l_2(\epsilon; \theta_{0,v}, \hat{\theta}_{T,c}) - \mathbb{G}_T l_2(\epsilon; \hat{\theta}_T)| + |\mathcal{R}_{\theta_i}|\right),
\end{aligned}$$

where $\mathcal{R}_{\theta_i} = \frac{1}{T}\sum_{t=1}^{T} \inf_{\theta \in U(\theta_i)} l_{2,t}(\epsilon_t; \theta) - \mathbb{E}_{\theta_0}\left[\inf_{\theta \in U(\theta_i)} l_{2,t}(\epsilon_t; \theta)\right]$. Invoking step 3 and the way the neighborhoods have been built, for any $i = 1, \ldots, n$,

$$\mathbb{E}_{\theta_0}\left[\inf_{\theta \in U(\theta_i)} l_{2,t}(\epsilon_t; \theta)\right] \geq \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t; \theta_0)\right] + \mu.$$

Using the property $\{X+Y \leq a+b\} \subset \{X \leq a\} \cup \{Y \leq b\}$, $a, b \geq 0$ and X, Y any random variables, we obtain

$$\begin{aligned}
\mathbb{P}\left((\theta_{0,v}, \hat{\theta}_{c,T}) \in U(\theta_i)\right) \leq \mathbb{P}\Bigg(\mu \leq 2\sup_{\theta\in\Theta}|\mathbb{G}_T l_2(\epsilon;\theta) - \widetilde{\mathbb{G}}_T l_2(\epsilon;\theta)| + |\mathcal{R}_{\theta_i}| \\
+ \quad |\mathbb{G}_T l_2(\epsilon;\theta_0) - \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta_0)\right]| + |\mathbb{G}_T l_2(\epsilon;\theta_{0,v},\hat{\theta}_{T,c}) - \mathbb{G}_T l_2(\epsilon;\hat{\theta}_T)|\Bigg) \\
\leq \quad \mathbb{P}\left(\frac{\mu}{4} < 2\sup_{\theta\in\Theta}|\mathbb{G}_T l_2(\epsilon;\theta) - \widetilde{\mathbb{G}}_T l_2(\epsilon;\theta)|\right) + \mathbb{P}\left(\frac{\mu}{4} < |\mathbb{G}_T l_2(\epsilon;\theta_0) - \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta_0)\right]|\right) \\
+ \quad \mathbb{P}\left(\frac{\mu}{4} < |\mathcal{R}_{\theta_i}|\right) + \mathbb{P}\left(\frac{\mu}{4} < |\mathbb{G}_T l_2(\epsilon;\theta_{0,v},\hat{\theta}_{T,c}) - \mathbb{G}_T l_2(\epsilon;\hat{\theta}_T)|\right). \qquad (1.6.5)
\end{aligned}$$

Under assumption 13, the initial values generating the process are asymptotically irrelevant. For some $\delta > 0$ and $T > T_1$, this implies

$$\mathbb{P}\left(\frac{\mu}{4} < 2\sup_{\theta\in\Theta}|\mathbb{G}_T l_2(\epsilon;\theta) - \widetilde{\mathbb{G}}_T l_2(\epsilon;\theta)|\right) < \delta/4. \qquad (1.6.6)$$

As for the second probability of the r.h.s. in (1.6.5), we use the ergodic theorem (see Billingsley 1995), and for $T > T_2$, we obtain

$$\mathbb{P}\left(\frac{\mu}{4} < |\mathbb{G}_T l_2(\epsilon;\theta_0) - \mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta_0)\right]|\right) < \delta/4. \qquad (1.6.7)$$

Let us focus on the the third term in the r.h.s. Although the quantity $l_{2,t}(\epsilon_t;\theta)$ is not necessarily integrable, the Ergodic Theorem can still be used as $\mathbb{E}_{\theta_0}\left[l_{2,t}(\epsilon_t;\theta)\right] \in \mathbb{R}\cup\{\infty\}$. Furthermore, $l_{2,t}(\epsilon_t;\theta)$ is a measurable function of an ergodic process, hence, as in Exercise 7.4 in Francq and Zakoïan (2010), the Ergodic Theorem can be applied to $(\inf_{\theta\in U(\theta_i)} l_{2,t}(\epsilon_t;\theta))$:

$$\liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\inf_{\theta\in U(\theta_i)} l_{2,t}(\epsilon_t;\theta) = \mathbb{E}_{\theta_0}\left[\inf_{\theta\in U(\theta_i)} l_{2,t}(\epsilon_t;\theta)\right].$$

Plugging this convergence result into (1.6.5), for $\delta > 0$, $T > T_3$, we obtain

$$\mathbb{P}\left(\mu/4 < |\mathcal{R}_{\theta_i}|\right) < \delta/4. \qquad (1.6.8)$$

Note that the derivative of $\theta_v \mapsto \mathbb{G}_T l_2(\epsilon;\theta_v,\theta_{0,c})$ is uniformly bounded under assumption 14 (recall the arguments in the proof of Step 2). Invoking assumption 15, we can tackle the fourth term of (1.6.5): if $t > T_4$, we have

$$\mathbb{P}\left(\mu/4 < |\mathbb{G}_T l_2(\epsilon;\theta_0) - \mathbb{G}_T l_2(\epsilon;\hat{\theta}_{T,v},\theta_{0,c})|\right) < \delta/4. \qquad (1.6.9)$$

Consequently, with (1.6.6), (1.6.7), (1.6.8) and (1.6.9), for $T > T_1 \vee T_2 \vee T_3 \vee T_4$, (1.6.5) becomes $\mathbb{P}\left(\hat{\theta}_T \in U(\theta_i)\right) \leq \delta$. Since δ can be chosen arbitrarily small, this proves the convergence in probability of $(\hat{\theta}_{T,v}, \hat{\theta}_{T,c})'$ to the true parameter vector θ_0. □

1.6.2 Asymptotic Normality

To define $\hat{\theta}_T$, the first order conditions are

$$\begin{cases} \Delta_T(\hat{\theta}_{T,v}) & = & \frac{1}{T}\sum_{t=1}^{T}\delta_t(\hat{\theta}_{T,v}) = 0, & \text{with } \delta_t(\theta_{T,v}) := \nabla_{\theta_v} l_{1,t}(\epsilon_t;\theta_v), \\ \Psi_T(\hat{\theta}_{T,v}, \hat{\theta}_{T,c}) & = & \frac{1}{T}\sum_{t=1}^{T}\psi_t(\hat{\theta}_{T,v}, \hat{\theta}_{T,c}) = 0, & \text{with } \psi_t(\theta) := \nabla_{\theta_c} l_{2,t}(\epsilon_t;\theta). \end{cases}$$

We stress that $l_{2,t}$ and its derivatives w.r.t. θ cannot be written explicitly in practice, because the functional relationship between a Gaussian likelihood and the underlying partial correlations (through our previous function F_{vine}) is too complex in analytical terms. Therefore, we have to rely on some numerical routines to evaluate numerically such functions: see Brechmann and Schepsmeier (2013). In particular, this is necessary to calculate $\hat{\theta}_{T,c}$ and to approximate the asymptotic variance-covariance matrix in Theorem 1.6.20 below.

Assumption 16. The innovations η_t have finite fourth order moments.

The next regularity conditions are classic and necessary to justify the existence of the asymptotic covariance in the next Theorem.

Assumption 17. The first order moments of $\|\psi_t(\theta_0)\psi_t(\theta_0)'\|$ and $\|\delta_t(\theta_{0,v})\psi_t(\theta_0)'\|$ are finite.

Under the price of additional technicalities, it is possible to establish some sufficient and more explicit conditions on the model parameters to satisfy assumption 17: see assumption 15 in Poignard and Fermanian (2016). Note that the existence of $\mathbb{E}\left[\|\delta_t(\theta_{0,v})\delta_t(\theta_{0,v})'\|\right]$ and $\mathbb{E}\left[\|\nabla_{\theta_v}\delta_t(\theta_{0,v})\|\right]$ has been established by Francq and Zakoïan (2004), as they are related to usual GARCH processes and Gaussian QMLE. Here, we require additional conditions of regularity to manage the correlation part of the likelihood.

Assumption 18. The variables $\nabla_{\theta_v\theta_c'} l_{2,t}(\epsilon_t;\theta_0)$, $\nabla_{\theta_c\theta_c'} l_{2,t}(\epsilon_t;\theta_0)$, $\sup_{\theta:\|\theta-\theta_0\|<\alpha}\|\nabla_{\theta_c\theta_c'}\psi(\theta_0)\|$ and $\sup_{\theta:\|\theta-\theta_0\|<\alpha}\|\nabla_{\theta_c\theta_v'}\psi(\theta_0)\|$ are integrable, for some $\alpha > 0$.

Assumption 19. $\mathbb{E}\left[\nabla_{\theta_c\theta_c'} l_{2,t}(\epsilon_t;\theta_{0,v},\theta_{0,c})\right]$ is nonsingular.

As expected, we need to assume that the initial values of the process are asymptotically irrelevant to evaluate score functions. The multiplication by $\sqrt{T}$ renders this task more difficult than in the proof of consistency.

Assumption 20. $\sqrt{T}\|\Delta_T(\theta_{0,v})-\tilde{\Delta}_T(\theta_{0,v})\| = o_p(1)$ and $\sqrt{T}\|\Psi_T(\theta_{0,v},\theta_{0,c})-\tilde{\Psi}_T(\theta_{0,v},\theta_{0,c})\| = o_p(1)$. For some $\alpha > 0$,

$$\sup_{\theta_v:\|\theta_v-\theta_{0,v}\|<\alpha} \|\nabla_{\theta_v}\Delta_T(\theta_v) - \nabla_{\theta_v}\tilde{\Delta}_T(\theta_v)\| + \sup_{\theta:\|\theta-\theta_0\|<\alpha} \|\nabla_\theta \Psi_T(\theta) - \nabla_\theta\tilde{\Psi}_T(\theta)\| = o_p(1).$$

Theorem 1.6.20. *Assume (8)-(20), then $\hat{\theta}_{T,v}$ and $\hat{\theta}_{T,c}$ are asymptotically normal, and $\sqrt{T}\left(\hat{\theta}_T - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, J^{-1}IJ^{-1}\right)$, where*

$$\begin{aligned}
J &= \mathbb{E}_{\theta_0}\left[\begin{pmatrix} \nabla_{\theta_v\theta_v'} l_{1,t}(\epsilon_t;\theta_{0,v}) & 0 \\ \nabla_{\theta_v\theta_c'} l_{2,t}(\epsilon_t;\theta_{0,v},\theta_{0,c}) & \nabla_{\theta_c\theta_c'} l_{2,t}(\epsilon_t;\theta_0)\end{pmatrix}\right], \\
I &= \mathbb{V}\left(\delta_t(\theta_{0,v}), \psi_t(\theta_{0,v},\theta_{0,c})\right)' \\
&= \mathbb{E}_{\theta_0}\left[\begin{pmatrix} \nabla_{\theta_v} l_{1,t}(\epsilon_t;\theta_{0,v})\nabla_{\theta_v'} l_{1,t}(\epsilon_t;\theta_{0,v}) & \nabla_{\theta_v} l_{1,t}(\epsilon_t;\theta_{0,v})\nabla_{\theta_c'} l_{2,t}(\epsilon_t;\theta_{0,v},\theta_{0,c}) \\ \nabla_{\theta_c} l_{2,t}(\epsilon_t;\theta_0)\nabla_{\theta_v'} l_{1,t}(\epsilon_t;\theta_{0,v}) & \nabla_{\theta_c} l_{2,t}(\epsilon_t;\theta_{0,v},\theta_{0,c})\nabla_{\theta_c'} l_{2,t}(\epsilon_t;\theta_0)\end{pmatrix}\right].
\end{aligned}$$

This usual "sandwich" asymptotic covariance illustrates the two-stage estimation procedure. As we mentioned above, the matrices I and J can be estimated empirically, evaluating the second-order derivatives of the likelihood numerically.

Lemma 1.6.21. *Suppose the assumptions of Theorem 1.6.20 hold. If $\bar{\theta}_T \to \theta_0$ in probability, then*

(i) $\nabla_{\theta_v}\Delta_T(\bar{\theta}_{T,v}) \xrightarrow[T\to\infty]{\mathbb{P}} J_{1,1}$, $\nabla_{\theta_c}\Psi_T(\bar{\theta}_T) \xrightarrow[T\to\infty]{\mathbb{P}} J_{2,2}$, *and* $\nabla_{\theta_v}\Psi_T(\bar{\theta}_T) \xrightarrow[T\to\infty]{\mathbb{P}} J_{2,1}$.

(ii) $\sqrt{T}\begin{pmatrix}\Delta_T(\theta_{0,v}) \\ \Psi_T(\theta_{0,v},\theta_{0,c})\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, I\right)$.

Proof of Lemma 1.6.21. *(i)* The first convergence corresponds to scores of usual GARCH log-likelihoods. This result can be found in Francq and Zakoïan (2004), for instance.

Moreover, applying a Taylor expansion of $\nabla_{\theta_c}\Psi(\bar{\theta}_T)$ around θ_0, we get

$$\begin{aligned}
\nabla_{\theta_c}\Psi_T(\bar{\theta}_T) &= \nabla_{\theta_c}\Psi_T(\theta_0) + \nabla_{\theta_c\theta_v'}\Psi_T(\tilde{\theta}_T).\left(\bar{\theta}_{T,v} - \theta_{0,v}\right) \\
&+ \nabla_{\theta_c\theta_c'}\Psi_T(\tilde{\theta}_T).\left(\bar{\theta}_{T,c} - \theta_{0,c}\right),
\end{aligned} \tag{1.6.10}$$

for some $\tilde{\theta}_T$, $\|\tilde{\theta}_T - \theta_0\| \leq \|\bar{\theta}_T - \theta_0\|$. Furthermore, we can apply an Ergodic Theorem (Billingsley, 1995) to the two sequences $\sup_{\theta:\|\theta-\theta_0\|<\alpha} \|\nabla_{\theta_c\theta_v'}\psi_t(\theta)\|$ and $\sup_{\theta:\|\theta-\theta_0\|<\alpha} \|\nabla_{\theta_c\theta_c'}\psi_t(\epsilon_t;\theta)\|$. Those results imply

$$\begin{aligned} \limsup_{T\to\infty} \|\nabla_{\theta_c\theta_v'}\Psi_T(\theta)\| &\leq \limsup_{T\to\infty} T^{-1}\sum_{t=1}^{T} \sup_{\theta:\|\theta-\theta_0\|<\alpha} \|\nabla_{\theta_c\theta_v'}\psi_t(\theta)\| \\ &= \mathbb{E}\left[\sup_{\theta:\|\theta-\theta_0\|<\alpha} \|\nabla_{\theta_c\theta_v'}\psi_t(\theta)\|\right], \text{ and} \end{aligned} \tag{1.6.11}$$

$$\begin{aligned} \limsup_{T\to\infty} \|\nabla_{\theta_c\theta_c'}\Psi_T(\theta)\| &\leq \limsup_{T\to\infty} T^{-1}\sum_{t=1}^{T} \sup_{\theta:\|\theta-\theta_0\|<\alpha} \|\nabla_{\theta_c\theta_c'}\psi_t(\theta)\| \\ &= \mathbb{E}\left[\sup_{\theta:\|\theta-\theta_0\|<\alpha} \|\nabla_{\theta_c\theta_c'}\psi_t(\theta)\|\right]. \end{aligned} \tag{1.6.12}$$

By assumption 18, both expectations of (1.6.11) and (1.6.12) are finite. Since $\bar{\theta}_T \underset{T\to\infty}{\longrightarrow} \theta_0$ in probability, the two last terms of the r.h.s. of (1.6.10) converge to 0. Finally, the Ergodic Theorem applied to $(\nabla_{\theta_c}\Psi_T(\theta_{0,v},\theta_{0,c}))$ proves the second assertion of *(i)*. The third assertion of *(i)* can be proved similarly.

(ii) To apply a CLT, we prove that $(\delta_t(\theta_{0,v}), \psi_t(\theta_0))'$ is a square integrable martingale difference. Denote by $\delta_t^{(i)}(\theta_v)$ (resp. $\psi_t^{(i)}(\theta_v,\theta_c)$) the i-th component of $\nabla_{\theta_v} l_{1,t}(\theta_v)$ (resp. $\nabla_{\theta_c} l_{2,t}(\theta_v,\theta_c)$). Through usual matrix derivatives (see Lütkepohl, 1996), we get

$$\delta_t^{(i)}(\theta_v) = \mathrm{Tr}\left((I_N - D_t^{-1}\epsilon_t\epsilon_t' D_t^{-1}).(D_t^{-1}(\partial_{\theta_v^i} D_t) + (\partial_{\theta_v^i} D_t)D_t^{-1})\right).$$

Using the $\mathcal{F}_{t-1}$ measurability of D_t, we obtain

$$\begin{aligned} \mathbb{E}\left[\delta_t^{(i)}(\theta_v)|\mathcal{F}_{t-1}\right] &= 2\mathrm{Tr}\left((\partial_{\theta_v^i} D_t)D_t^{-1})\right) - \mathrm{Tr}\left(\mathbb{E}[u_t u_t'|\mathcal{F}_{t-1}](\partial_{\theta_v^i} D_t)D_t^{-1} + D_t^{-1}(\partial_{\theta_v^i} D_t))\right) \\ &= 2\mathrm{Tr}\left((\partial_{\theta_v^i} D_t)D_t^{-1})\right) - 2\mathrm{Tr}\left((\partial_{\theta_v^i} D_t)D_t^{-1})\right) = 0. \end{aligned}$$

Concerning the correlation components, for $i = 1, \cdots, 3N(N-1)/2$, the score is

$$\psi_t^{(i)}(\theta_{0,v},\theta_{0,c}) = \mathrm{Tr}\left((I_N - R_t^{-1}u_t u_t')R_t^{-1}(\partial_{\theta_c^i} R_t)\right).$$

Using the $\mathcal{F}_{t-1}$ measurability of R_t, we obtain

$$\begin{aligned} \mathbb{E}\left[\psi_t^{(i)}(\theta_{0,v},\theta_{0,c})|\mathcal{F}_{t-1}\right] &= \mathrm{Tr}\left(\left(I_N - R_t^{-1}\mathbb{E}\left[u_t u_t'|\mathcal{F}_{t-1}\right]\right)R_t^{-1}(\partial_{\theta_c^i} R_t)\right) \\ &= \mathrm{Tr}\left(\left(I_N - R_t^{-1}R_t\right)R_t^{-1}(\partial_{\theta_c^i} R_t)\right) = 0. \end{aligned}$$

Consequently, $(\delta_t(\theta_{0,v}), \psi_t(\theta_{0,v}, \theta_{0,c}))'$ is a square integrable martingale difference, invoking assumption 17. The process $(\delta_t(\theta_{0,v}), \psi_t(\theta_{0,v}, \theta_{0,c}))$ is stationary, as a measurable function of the stationary process (ϵ_t). Consequently, by a central limit theorem for stationary square integrable martingale differences (see Billinsgley 1995), we obtain the asymptotic normality of $\sqrt{T}(\Delta_T(\theta_{0,v}, \Psi_T(\theta_0))$. □

Proof of Theorem 1.6.20. Through a Taylor expansion around θ_0, we obtain $0 = \Delta_T(\hat{\theta}_{T,v}) = \Delta_T(\theta_{0,v}) + \nabla_{\theta_v}\Delta_T(\bar{\theta}_{T,v}).(\hat{\theta}_{T,v} - \theta_{0,v})$, and

$$0 = \Psi_T(\hat{\theta}_{T,v}, \hat{\theta}_{T,c}) = \Psi_T(\theta_0) + \nabla_{\theta_v}\Psi_T(\bar{\theta}_T)(\hat{\theta}_{T,v} - \theta_{0,v}) + \nabla_{\theta_c}\Psi_T(\bar{\theta}_T)(\hat{\theta}_{T,c} - \theta_{0,c}),$$

where $\|\bar{\theta}_T - \theta_0\| < \|\hat{\theta}_T - \theta_0\|$. Inverting these relationships and multiplying by $\sqrt{T}$, we have $\sqrt{T}\left(\hat{\theta}_{T,v} - \theta_{0,v}\right) = \left(-\nabla_{\theta_v}\Delta_T(\bar{\theta}_{T,v})\right)^{-1}\sqrt{T}\ \Delta_T(\theta_{0,v})$, and

$$\begin{aligned}\sqrt{T}(\hat{\theta}_{T,c} - \theta_{0,c}) &= \left(-\nabla_{\theta_c}\Psi_T(\bar{\theta}_T)\right)^{-1}\nabla_{\theta_v}\Psi_T(\bar{\theta}_T)\left(-\nabla_{\theta_v}\Delta_T(\bar{\theta}_{T,v})\right)^{-1}\sqrt{T}\ \Delta_T(\theta_{0,v}) \\ &+ \left(-\nabla_{\theta_c}\Psi_T(\bar{\theta}_T)\right)^{-1}\sqrt{T}\ \Psi_T(\theta_0).\end{aligned}$$

Therefore, $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is a linear transform of $\sqrt{T}[\Delta_T(\theta_{0,v}), \Psi_T(\theta_0]'$:

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = M_T \cdot \sqrt{T}[\Delta_T(\theta_{0,v}), \Psi_T(\theta_0)]',$$

for some sequence of random matrices (M_T) that tends to J^{-1} in probability. By Lemma 1.6.21 and Slutsky's theorem, we obtain the asymptotic normality of $\sqrt{T}(\hat{\theta}_T - \theta_0)$. □

As a by-product, simple calculations provide the asymptotic variances of $\hat{\theta}_{T,v}$ and $\hat{\theta}_{T,c}$: with obvious notations, $\mathbb{V}_{\text{as}}(\hat{\theta}_{T,v}) = J_{11}^{-1}I_{11}J_{11}^{-1}$, and

$$\mathbb{V}_{\text{as}}(\hat{\theta}_{T,c}) = J_{22}^{-1}I_{22}J_{22}^{-1} - \Gamma I_{12}J_{22}^{-1} - J_{22}^{-1}I_{21}\Gamma' + \Gamma I_{11}\Gamma',\ \Gamma := J_{22}^{-1}J_{21}J_{11}^{-1}.$$

1.7 Empirical applications

To simplify and to lighten notations, we restrict ourselves to one-order models in this section. Moreover, we consider no cross-effects between all the individual partial correlation processes, i.e. the matrices Ξ_k and Λ_l are assumed to be diagonal. Then,

when $p = q = 1$, the $N-1$ first elements of Pc_t correspond to usual correlations, i.e. $\rho_{ij|\emptyset,t} = \rho_{ij,t}$, and they follow the processes

$$\psi\left(\rho_{ij,t}\right) = \omega_{ij} + \xi_{ij}\psi\left(\rho_{ij,t-1}\right) + \lambda_{ij}\hat{\upsilon}_{i,t-1}\hat{\upsilon}_{j,t-1}, \tag{1.7.1}$$

with $\hat{\upsilon}_{k,t} = \epsilon_{k,t}/\sqrt{\hat{h}_{k,t}}$. From the N-th component on, the elements of Pc_t are "true" partial correlations for which $L \neq \emptyset$. Their dynamics are given by

$$\psi(\rho_{ij|L,t}) = \omega_{ij|L} + \xi_{ij|L}\psi(\rho_{ij|L,t-1}) + \lambda_{ij|L}\hat{\upsilon}_{i|L,t-1}\hat{\upsilon}_{j|L,t-1}. \tag{1.7.2}$$

Strictly speaking, the partial correlation dynamics we invoke for inference or simulation purpose is given by (1.7.1) and (1.7.2).

1.7.1 A simulation study

We consider as a data generating process (DGP) multivariate series (ϵ_t) of size $N = 6, 10, 20, 30, 50$. Their innovations η_t are standardized normal white noises. The conditional covariance matrices of these processes are deduced from a MGARCH form $H_t = D_t R_t D_t$. To generate N univariate variance processes along (1.6.1), we choose randomly the corresponding $3N$ parameters such that $\varsigma \sim U\left(10^{-5}, 9.10^{-5}\right)$, $\kappa \sim U\left(0.01, 0.15\right)$ and $\tau \sim U\left(0.95, 0.85\right)$, under the stationarity constraint $\kappa + \tau < 1$. As for the correlation dynamics, we first choose randomly $N(N-1)/2$ deterministic processes among the cosinus, sinus, modulo and constant functions, and then generate some series

$$a_1 + a_2\cos(2\pi t/\alpha),\ b_1 + b_2\sin(2\pi t/\beta),\ c_1 + c_2\bmod(t/\mu),\ d_1 + d_2\text{const},$$

for every $t = 1, \ldots, T$. Our parameters $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2$ are chosen randomly and independently following a $\mathcal{U}\left(-0.4, 0.4\right)$ and α, β, μ are randomly (equally) selected among the fixed subset $\{100, 200, 500, 1000, 1500, 2000\}$. All these series constitute the components of a lower triangular matrix K_t with ones on the main diagonal. Then, we generate symmetric and positive definite matrices $C_t = K_t K_t'$ and $R_t = C_t^{\star -1/2} C_t C_t^{\star -1/2}$. Those processes allow for rapid, gradual changes or constant correlation patterns, and they do not depend on a specific statistical model. Initializing each of the GARCH processes randomly and given ϵ_1, we simulate the successive values of a MGARCH process with conditional covariance matrices (H_t). We do this iterative procedure for $T = 10000$ and we consider 300 different correlation matrix patterns.

Once a series is simulated, we estimate the model under different model assumptions: a C-vine-GARCH, a diagonal QFDCC and a scalar DCC. As a benchmark, we also compute the empirical correlation matrices of our returns through a rolling-window of size 200 observations. The estimated parameters allow the calculation of successive correlation matrices, which are here $\hat{R}_t^{vine}$ (C-vine-GARCH), $\hat{R}_t^{qfdcc}$ (QFDCC model), $\hat{R}_t^{dcc}$ (DCC model), and $\hat{R}_t^{rw}$ (rolling-window) correlations. Moreover, we consider a constrained version of the vine-GARCH. For $N = 6, 10$, the partial correlations of the last two trees are constrained to their unconditional partial correlation values, as estimated over the whole sample. For $N = 20$, the partial correlations from the 11th level are set to their unconditional partial correlation values. The same applies for $N = 30, 50$, where we only consider the constrained vine-GARCH case. Alternatively, we could set zero partial correlations for these two last trees of the C-vine and the results would be comparable. We denote by $\hat{R}_t^{vine\star}$ the correlation matrices obtained with the constrained version of the C-vine. Both vine specifications are estimated by the C-vine iterative process. The first level of the C-vine has been chosen following the procedure of Subsection 1.3.3.

We compare the true correlation process and the estimated correlation processes through the aforementioned models. To do so, we specify a matrix distance, namely the Frobenius norm, defined as $||A - B||_F := \sqrt{\text{Trace}((A-B)'(A-B))}$. We compute the previous norm for each t and for

$$A = R_t, \text{ and } B \in \{\hat{R}_t^{dcc}, \hat{R}_t^{qfdcc}, \hat{R}_t^{rw}, \hat{R}_t^{vine}, \hat{R}_t^{vine\star}\}.$$

We take the average of those quantities over $T = 10000$ periods of time. Since we repeat this experiment 300 times, this provides an average gap for all those simulations. Table 1.1 reports the results.

The C-vine model clearly outperforms the other specifications. The DCC displays a significant gap, which highlights that it is too restrictive to capture complex dynamics with only two parameters. As for the rolling-window correlation, the result emphasizes this empirical measure should be taken with great care. The rolling nature of the samples makes the rolling-window correlation very low to react to a rapid correlation fluctuations. Interestingly, for every N level, both C-vine specifications clearly outperform other usual DCC-type dynamics. The QFDCC specification performs poorly compared to other models. Therefore, this justifies the use of constrained C-vine dynamics, allowing for parsimony.

1.7.2 Application to real portfolios

In this subsection, we estimate by Quasi-Maximum Likelihood the DCC-GARCH and vine-GARCH models for two financial portfolios. They are composed of daily series of stock log-returns related to the Morgan Stanley Capital International (MSCI) Developed Markets indices. In the so-called Portfolio I, we consider Germany, Italy, France, the Netherlands and the United Kingdom. Portfolio II is more diversified geographically because it is composed of Germany, the United-States, Greece, Italy, Japan and Australia. For both portfolios, the samples start in January 1999 and end in August 2013, which amounts to 3669 observations.

First, we have centered the time series by assuming that $\mathbb{E}_{t-1}[r_t] = \mu_t(\theta)$ follows a one-order autoregressive process (estimated by OLS). Second, we estimate the conditional variance processes of the components of $\epsilon_t = r_t - \mu_t$. The GARCH(1,1) specification was chosen a priori for modeling these marginal dynamics. Indeed, this is by far the reference model used in the literature. The estimation results are reported in Table 1.2.

We now turn to the second QML step, i.e. the estimation of the conditional correlation dynamics, knowing the GARCH(1,1) estimates. For portfolios I and II, we select a relevant C-vine, according to the Kendall's tau selection procedure (see Subsection 1.3.3). We associate an index to each country. This number corresponds to the index of the tree for which this country is the "center" (the node with maximal degree). Since Portfolio I is composed of European stocks, it can be considered as relatively homogenous, including the main countries of the Eurozone. The selecting procedure induces the following order: Germany (1), United-Kingdom (2), Italy (3), France (4) and Netherlands (5). In this case, Germany (1) is the root of the first C-vine tree. That means we consider the partial correlations of two countries given Germany on Tree 2. Then, on Tree 3, the conditioning subset is Germany (1) and United-Kingdom (2), etc. The composition of the "heterogenous" portfolio II is given as follows: Germany (1), Greece (2), United-States (3), Italy (4), Japan (5) and Australia (6).

Actually, we consider two cases of C-vine-GARCH models. The first one is the usual unconstrained C-vine tree. The second one is a constrained version of the previous one, where the partial correlations of the last two trees are fixed. Therefore, in portfolio I, $\rho_{45|123}$, $\rho_{35|12}$ and $\rho_{34|12}$ are set to their unconditional values that have been estimated over the whole sample. Thus the size of the parameter space is reduced by 9 parameters for both portfolios. In every case, the parameters are estimated by simulated

annealing. Table 1.3 reports the estimation results of the vine-GARCH model for the unconstrained case. For the sake of comparison, Table 1.4 (resp. Table 1.5) provides the estimation results of the scalar DCC (resp. diagonal QFDCC). The results for the constrained case are very close to those of the unconstrained case: see Tables 1.11 and 1.12.

The same model is implemented for portfolio II, which is heterogenous in terms of geographical areas. Table 1.6 (resp. Table 1.7, Table 1.8) reports the estimation results of the C-vine-GARCH (resp. diagonal QFDCC, scalar DCC).

Concerning Portfolio I, the higher the level of the tree is, the smaller are the partial correlation coefficients ω and λ. We may infer that once we control for the information given by Germany (1) (the core of the Eurozone) and United-Kingdom (2), the dynamics of partial correlations on trees T_3 and T_4 are not very informative. This looks like evaluating a white noise. This is confirmed by the modeling of constrained vines, where the estimation results are close to the unconstrained case. On the contrary, this effect does not appear with the heterogenous portfolio II. Controlling for Germany, Greece and the US in portfolio II is not enough to deduce the whole information about the correlation dynamics between Japan and Australia, due to significant remaining idiosyncratic risks.

1.7.3 Specification testing

Once the model is estimated, we are able to forecast the covariance matrices H_t, at least one-period ahead. There exist several methods to evaluate the absolute and/or relative efficiency of these predictions. See Patton and Sheppard (2009) for a survey. In this study, we focus on direct out-of-sample evaluation methods, which allow for pairwise comparisons. They test whether some of the previous models provide better forecasts in terms of portfolio volatility behavior. Following the methodology of Engle and Colacito (2006), we develop a mean-variance portfolio approach to test the H_t forecasts. Intuitively, if a conditional covariance process is misspecified, then the minimum variance portfolio should emphasize such a shortcoming, compared to other models. Then, consider an investor who allocates a fixed amount between N stocks, according to a minimum-variance strategy and independently at each time t. At each date t, he/she solves

$$\min_{w_t} \; w_t' H_t w_t, \quad \text{s.t.} \quad \iota' w_t = 1, \tag{1.7.3}$$

where w_t is the $N \times 1$ vector of portfolio weights chosen at (the end of) time $t-1$, ι is a $N \times 1$ vector of 1 and H_t is the estimated conditional covariance matrix of the asset returns at time t. They are deduced from some dynamics that have been estimated on the sub-sample January 1999 - October 2011. Once the latter process is estimated in-sample, out-of-sample predictions are plugged into the program (1.7.3) between November 2011 and August 2013. The solution of (1.7.3) is given by the global minimum variance portfolio $w_t = H_t^{-1}\iota/\iota' H_t^{-1}\iota$.

Engle and Colacito (2006) show that the realized portfolio volatility is the smallest one when the model covariance matrices are correctly specified. As a consequence, if wealth is allocated using two different dynamic models i and j, whose predicted covariance matrices are (H_t^i) and (H_t^j), the strategy providing the smallest portfolio variance will be considered as the best one. To do so, we consider a sequence of minimum variance portfolio weights $(w_{i,t})$ and $(w_{j,t})$, depending on the model. Then, we consider a distance based on the difference of the squared returns of the two portfolios, defined as $u_{ij,t} = \left\{w_{i,t}'\epsilon_t\right\}^2 - \left\{w_{j,t}'\epsilon_t\right\}^2$. The portfolio variances are the same if the predicted covariance matrices are the same. Thus we test the null hypothesis $\mathcal{H}_0 : \mathbb{E}[u_{ij,t}] = 0$ by the Diebold and Mariano (1995) test. It consists of a least square regression using HAC standard errors, given by $u_{ij,t} = \alpha + \epsilon_{u,t}$, $\mathbb{E}[\epsilon_{u,t}] = 0$, and we test $\mathcal{H}_0 : \alpha = 0$. If the mean of $u_{ij,t}$ is significantly positive (resp. negative), then the forecasts given by the covariance matrices of model j (resp. i) are preferred.

We run the latter test for portfolios I and II and to compare the scalar DCC, QFDCC, constrained C-vine-GARCH (C-vine-c) and unconstrained C-vine-GARCH (C-vine) models. We also compare these parameterizations to a factor model, the O-GARCH(1,1) [5]. The results are reported in Tables 1.9 and 1.10. Those tables provide the out-of-sample Dielbold-Mariano test statistics that check the equality of a pair of series of covariance matrices using the loss function $u_{ij,t}$ over the period November 2011 - August 2013.

We first note that in the homogenous case, the DCC specifications do not provide better covariance forecasts. Interestingly, the constrained case of the C-vine provides better prediction accuracy than the unconstrained case. For the heterogenous portfolio, we

[5] The O-GARCH assumes the decomposition $H_t = P\Lambda_t P'$, where $\Lambda_t = \text{diag}(\lambda_{1,t}, \cdots, \lambda_{K,t})$, with K the number of factors. Here, we choose $K = N$ factors and each λ_t is supposed to follow a univariate GARCH(1,1) process that is estimated by maximum likelihood. The matrix P is nonsingular and it is estimated by applying a PCA on the empirical variance covariance matrix of ϵ_t. See Alexander (2001), e.g.

obtain the reverse. The C-vine specification outperforms the constrained case in terms of prediction accuracy: the two last levels of the tree should be estimated as, once the dynamics are controlled by Germany, Greece and the US, there remains a significant amount of idiosyncratic risk. Both versions of the C-vine are not outperformed by the scalar DCC, and the C-vine provides better covariance forecasts than the QFDCC. The QFDCC is also slightly outperformed by the scalar DCC specification for the heterogenous portfolio, what is rather surprising. Finally, the O-GARCH model is beaten by all the others distinctly. But all these results are not sufficiently clear-cut to draw any strong conclusion concerning a potential hierarchy between all these models, at least in terms of a "naive" investment strategy.

1.8 Conclusion

We have proposed to rely on vines to define a new family of multivariate GARCH-type models. The main feature of our methodology is the specification/estimation of partial correlation processes "independently" and largely arbitrarily, and their use to generate sequences of correlation matrices. The canonical vine is particularly intuitive to model a hierarchy between asset returns, as reasonings are close to factor models. Our approach does not rely on any normalization stage and we model directly correlation processes. Besides, the vine-GARCH approach allows for building parsimonious models. Indeed, we can assume (theoretically and often empirically) no partial correlation dynamics (or at least, constant, simpler, homogenous, etc., dynamics) at all nodes in the vine from some level on. All these elements foster flexibility and enable to generate high-dimensional matrices.

Therefore, a new framework has been opened in the field of MGARCH models. We have provided sufficient conditions for the consistency and the asymptotic normality of a two-step quasi-maximum estimator. The performances of the vine-GARCH and DCC estimators have been compared by means of applications to simulated and real data. The simulation study confirmed that a more flexible specification (the C-vine-GARCH) provides a better accuracy. The constrained case is particularly adapted to homogenous portfolios and challenges the unconstrained case. The performances calculated from real data support the use of vine dynamics but more empirical work is probably necessary to evaluate all the advantages of such approaches w.r.t. more classic ones, as the standard DCC family.

1.9 Tables and figures

TABLE 1.1: Simulation study: Average distance between true and estimated correlation matrices.

$\|\|R_t - B\|\|_F$	$B = \hat{R}_t^{dcc}$	$B = \hat{R}_t^{qfdcc}$	$B = \hat{R}_t^{rw}$	$B = \hat{R}_t^{vine}$	$B = \hat{R}_t^{vine\star}$
$N = 6$	0.4995	0.4791	0.5275	0.3906	0.4137
$N = 10$	0.8270	0.9237	0.8784	0.6413	0.6825
$N = 20$	1.6931	2.0106	1.7372	1.3250	1.3766
$N = 30$	2.4876	2.6681	2.5151	-	2.0583
$N = 50$	3.2839	3.8662	3.7691	-	2.6800

TABLE 1.2: GARCH(1,1) Models estimated by QML for 9 stock indices. The Bollerslev-Wooldridge standard deviations are in parentheses.

Asset	ς	κ	τ
Australia	0.657e-5 (0.114e-5)	0.124 (0.014)	0.846 (0.011)
France	0.388e-5 (0.076e-5)	0.111 (0.009)	0.876 (0.008)
Germany	0.368e-5 (0.080e-5)	0.100 (0.011)	0.889 (0.010)
Greece	0.191e-5 (0.147e-5)	0.090 (0.010)	0.917 (0.015)
Italy	0.235e-5 (0.052e-5)	0.113 (0.010)	0.883 (0.008)
Japan	0.997e-5 (0.157e-5)	0.103 (0.012)	0.849 (0.013)
Netherlands	0.363e-5 (0.069e-5)	0.110 (0.010)	0.876 (0.009)
United-Kingdom	0.338e-5 (0.067e-5)	0.115 (0.011)	0.868 (0.009)
United-States	0.223e-5 (0.056e-5)	0.102 (0.010)	0.884 (0.008)

TABLE 1.3: C-vine-GARCH estimated by QML for Portfolio I. The Bollerslev-Wooldridge standard deviations are in parentheses.

Ω	Estimate (Std Err)	Ξ	Estimate (Std Err)	Λ	Estimate (Std Err)
ω_{12}	-0.0629 (0.0288)	ξ_{12}	0.9749 (0.0064)	λ_{12}	0.1977 (0.0515)
ω_{13}	-0.0772 (0.0355)	ξ_{13}	0.9748 (0.0053)	λ_{13}	0.2230 (0.0472)
ω_{14}	-0.1388 (0.1928)	ξ_{14}	0.9878 (0.0109)	λ_{14}	0.2594 (0.2994)
ω_{15}	-0.0893 (0.0672)	ξ_{15}	0.9850 (0.0031)	λ_{15}	0.1976 (0.0973)
$\omega_{23\|1}$	0.0191 (0.0071)	$\xi_{23\|1}$	0.9521 (0.0145)	$\lambda_{23\|1}$	0.0097 (0.0100)
$\omega_{24\|1}$	0.0733 (0.0369)	$\xi_{24\|1}$	0.8839 (0.0540)	$\lambda_{24\|1}$	0.0311 (0.0161)
$\omega_{25\|1}$	0.0332 (0.0117)	$\xi_{25\|1}$	0.9375 (0.0162)	$\lambda_{25\|1}$	0.0216 (0.0116)
$\omega_{34\|12}$	0.0181 (0.0068)	$\xi_{34\|12}$	0.9894 (0.0048)	$\lambda_{34\|12}$	-0.0117 (0.0034)
$\omega_{35\|12}$	0.0289 (0.0064)	$\xi_{35\|12}$	0.9619 (0.0090)	$\lambda_{35\|12}$	-0.0136 (0.0077)
$\omega_{45\|123}$	0.0618 (0.0246)	$\xi_{45\|123}$	0.9174 (0.0370)	$\lambda_{45\|123}$	-0.0056 (0.0128)

TABLE 1.4: scalar DCC-GARCH estimated by QML for portfolio I. The Bollerslev-Wooldridge standard deviations are in parentheses.

Model	α	β
DCC	0.0284 (0.0032)	0.9674 (0.0041)

TABLE 1.5: Diagonal QFDCC estimated by QML for Portfolio I. The Bollerslev-Wooldridge standard deviations are in parentheses.

C^2	Estimate (Std Err)	A^2	Estimate (Std Err)	B^2	Estimate (Std Err)
c_{11}^2	0.0068 (0.0255)	a_{11}^2	0.0174 (0.0645)	b_{11}^2	0.9786 (0.0130)
c_{22}^2	0.0111 (0.0584)	a_{22}^2	0.0217 (0.1080)	b_{22}^2	0.9773 (0.0273)
c_{33}^2	0.0087 (0.0380)	a_{33}^2	0.0195 (0.2307)	b_{33}^2	0.9795 (0.0285)
c_{44}^2	0.0082 (0.0147)	a_{44}^2	0.0202 (0.0356)	b_{44}^2	0.9788 (0.0084)
c_{55}^2	0.0025 (0.0021)	a_{55}^2	0.0063 (0.0525)	b_{55}^2	0.9797 (0.0136)

TABLE 1.6: vine-GARCH estimated by QML for Portfolio II. The Bollerslev-Wooldridge standard deviations are in parentheses.

Ω	Estimate (Std Err)	Ξ	Estimate (Std Err)	Λ	Estimate (Std Err)
ω_{12}	0.0009 (0.0363)	ξ_{12}	0.9764 (0.0980)	λ_{12}	0.0473 (0.1015)
ω_{13}	0.0034 (0.0044)	ξ_{13}	0.9787 (0.0044)	λ_{13}	0.0421 (0.0080)
ω_{14}	-0.0637 (0.0258)	ξ_{14}	0.9795 (0.0043)	λ_{14}	0.1884 (0.0414)
ω_{15}	0.0059 (0.0041)	ξ_{15}	0.9714 (0.0127)	λ_{15}	0.0175 (0.0066)
ω_{16}	0.0045 (0.0036)	ξ_{16}	0.9772 (0.0047)	λ_{16}	0.0360 (0.0059)
$\omega_{23\|1}$	-0.0064 (0.0225)	$\xi_{23\|1}$	0.9388 (0.2172)	$\lambda_{23\|1}$	0.0016 (0.0271)
$\omega_{24\|1}$	0.0304 (0.1100)	$\xi_{24\|1}$	0.8828 (0.4267)	$\lambda_{24\|1}$	0.0092 (0.0350)
$\omega_{25\|1}$	0.0080 (0.0074)	$\xi_{25\|1}$	0.9601 (0.0211)	$\lambda_{25\|1}$	0.0034 (0.0191)
$\omega_{26\|1}$	0.0265 (0.0924)	$\xi_{26\|1}$	0.9101 (0.2596)	$\lambda_{26\|1}$	0.0121 (0.0497)
$\omega_{34\|12}$	0.0015 (0.0035)	$\xi_{34\|12}$	0.9551 (0.1663)	$\lambda_{34\|12}$	0.0115 (0.0110)
$\omega_{35\|12}$	-0.0001 (0.0003)	$\xi_{35\|12}$	0.9942 (0.0055)	$\lambda_{35\|12}$	0.0051 (0.0031)
$\omega_{36\|12}$	-0.0008 (0.0016)	$\xi_{36\|12}$	0.9805 (0.0356)	$\lambda_{36\|12}$	0.0094 (0.0101)
$\omega_{45\|123}$	0.0033 (0.0096)	$\xi_{45\|123}$	0.7327 (0.2485)	$\lambda_{45\|123}$	0.0128 (0.0217)
$\omega_{46\|123}$	0.0035 (0.0031)	$\xi_{46\|123}$	0.9512 (0.0191)	$\lambda_{46\|123}$	0.0130 (0.0117)
$\omega_{56\|1234}$	0.0134 (0.0067)	$\xi_{56\|1234}$	0.9660 (0.0062)	$\lambda_{56\|1234}$	0.0334 (0.0124)

TABLE 1.7: Diagonal QFDCC estimated by QML for Portfolio II. The Bollerslev-Wooldridge standard deviations are in parentheses.

C^2	Estimate (Std Err)	A^2	Estimate (Std Err)	B^2	Estimate (Std Err)
c_{11}^2	0.0065 (0.0029)	a_{11}^2	0.0139 (0.0061)	b_{11}^2	0.9851 (0.0025)
c_{22}^2	0.0012 (0.0016)	a_{22}^2	0.0021 (0.0026)	b_{22}^2	0.9931 (0.0026)
c_{33}^2	0.0020 (0.0036)	a_{33}^2	0.0029 (0.0054)	b_{33}^2	0.9876 (0.0029)
c_{44}^2	0.0064 (0.0050)	a_{44}^2	0.0134 (0.0103)	b_{44}^2	0.9856 (0.0028)
c_{55}^2	0.0021 (0.0091)	a_{55}^2	0.0021 (0.0097)	b_{55}^2	0.9925 (0.0041)
c_{66}^2	0.0067 (0.0172)	a_{66}^2	0.0086 (0.0231)	b_{66}^2	0.9904 (0.0030)

TABLE 1.8: scalar DCC GARCH estimated by QML for Portfolio II. The Bollerslev-Wooldridge standard deviations are in parentheses.

Model	α	β
DCC	0.0097 (0.0018)	0.9879 (0.0025)

TABLE 1.9: Diebold Mariano Test of Multivariate GARCH models for Portfolio I.

	DCC	QFDCC	GO-GARCH	C-vine	C-vine-c
DCC		0.6509	-5.9350***	0.7784	0.3551
QFDCC	-0.6509		-6.1426***	0.4237	0.0475
GO-GARCH	5.9350***	6.1426***		5.9438***	5.6779***
C-vine	-0.7784	-0.4237	-5.9498***		-2.1206**
C-vine-c	-0.3551	-0.0475	-5.6779***	2.1206**	

Rejection of the nul hypothesis at: 10% for *, 5% for **, 1% for ***. When the null hypothesis of equal predictive accuracy is rejected, a positive number is evidence in favor of the model in the column.

TABLE 1.10: Diebold Mariano Test of Multivariate GARCH models for Portfolio II.

	DCC	QFDCC	GO-GARCH	C-vine	C-vine-c
DCC		-0.6220	-4.9369***	0.0908	-0.7952
QFDCC	0.6220		-4.9783***	0.2650	-0.5991
GO-GARCH	4.9369***	4.9783***		4.6416***	4.1741***
C-vine	-0.0908	-0.2650	-4.6416***		-3.0709***
C-vine-c	0.7952	0.5991	-4.1741***	3.0709***	

Rejection of the null hypothesis at: 10% for *, 5% for **, 1% for ***. When the null hypothesis of equal predictive accuracy is rejected, a positive number is evidence in favor of the model in the column.

TABLE 1.11: C-vine-GARCH Model estimated by QML for Portfolio I. The Bollerslev-Wooldridge standard deviations are in parentheses.

Ω	Estimate (Std Err)	Ξ	Estimate (Std Err)	Λ	Estimate (Std Err)
ω_{12}	-0.0661 (0.0174)	ξ_{12}	0.9769 (0.0433)	λ_{12}	0.1932 (0.0193)
ω_{13}	-0.0771 (0.0441)	ξ_{13}	0.9804 (0.0659)	λ_{13}	0.1986 (0.0182)
ω_{14}	-0.1665 (0.6173)	ξ_{14}	0.9923 (0.1121)	λ_{14}	0.2590 (0.0638)
ω_{15}	-0.0858 (0.0709)	ξ_{15}	0.9915 (0.0613)	λ_{15}	0.1554 (0.0431)
$\omega_{23\|1}$	0.0081 (0.0047)	$\xi_{23\|1}$	0.9799 (0.1265)	$\lambda_{23\|1}$	0.0013 (0.0165)
$\omega_{24\|1}$	0.0248 (0.0666)	$\xi_{24\|1}$	0.9577 (0.0934)	$\lambda_{24\|1}$	0.0112 (0.0113)
$\omega_{25\|1}$	0.0172 (0.0081)	$\xi_{25\|1}$	0.9641 (0.0329)	$\lambda_{25\|1}$	0.0135 (0.0221)
$\omega_{34\|12}$	1.0821	$\xi_{34\|12}$	-	$\lambda_{34\|12}$	-
$\omega_{35\|12}$	0.6300	$\xi_{35\|12}$	-	$\lambda_{35\|12}$	-
$\omega_{45\|123}$	0.7957	$\xi_{45\|123}$	-	$\lambda_{45\|123}$	-

TABLE 1.12: C-vine-GARCH estimated by QML for Portfolio II. The Bollerslev-Wooldridge standard deviations are in parentheses.

Ω	Estimate (StdE)	Ξ	Estimate (StdE)	Λ	Estimate (StdE)
ω_{12}	-0.0008 (0.0359)	ξ_{12}	0.9823 (0.1935)	λ_{12}	0.0387 (0.0227)
ω_{13}	0.0016 (0.0036)	ξ_{13}	0.9821 (0.0530)	λ_{13}	0.0382 (0.0052)
ω_{14}	-0.0694 (0.0259)	ξ_{14}	0.9801 (0.0180)	λ_{14}	0.1915 (0.0092)
ω_{15}	0.0046 (0.0081)	ξ_{15}	0.9777 (0.0265)	λ_{15}	0.0146 (0.0043)
ω_{16}	0.0017 (0.0058)	ξ_{16}	0.9835 (0.0142)	λ_{16}	0.0288 (0.0012)
$\omega_{23\|1}$	-0.0072 (0.0399)	$\xi_{23\|1}$	0.9334 (0.4570)	$\lambda_{23\|1}$	-0.0007 (0.0037)
$\omega_{24\|1}$	0.0043 (0.0156)	$\xi_{24\|1}$	0.9837 (0.2434)	$\lambda_{24\|1}$	0.0001 (0.0100)
$\omega_{25\|1}$	0.0129 (0.0272)	$\xi_{25\|1}$	0.9384 (0.0790)	$\lambda_{25\|1}$	0.0028 (0.0061)
$\omega_{26\|1}$	0.0022 (0.0076)	$\xi_{26\|1}$	0.9906 (0.0238)	$\lambda_{26\|1}$	0.0049 (0.0175)
$\omega_{34\|12}$	0.0009 (0.0013)	$\xi_{34\|12}$	0.9729 (0.0187)	$\lambda_{34\|12}$	0.0107 (0.0082)
$\omega_{35\|12}$	-0.0001 (0.0002)	$\xi_{35\|12}$	0.9953 (0.0045)	$\lambda_{35\|12}$	0.0047 (0.0035)
$\omega_{36\|12}$	-0.0004 (0019)	$\xi_{36\|12}$	0.9888 (0.0393)	$\lambda_{36\|12}$	0.0053 (0.0118)
$\omega_{45\|123}$	0.0311	$\xi_{45\|123}$	-	$\lambda_{45\|123}$	-
$\omega_{46\|123}$	0.2472	$\xi_{46\|123}$	-	$\lambda_{46\|123}$	-
$\omega_{56\|1234}$	0.8669	$\xi_{56\|1234}$	-	$\lambda_{56\|1234}$	-

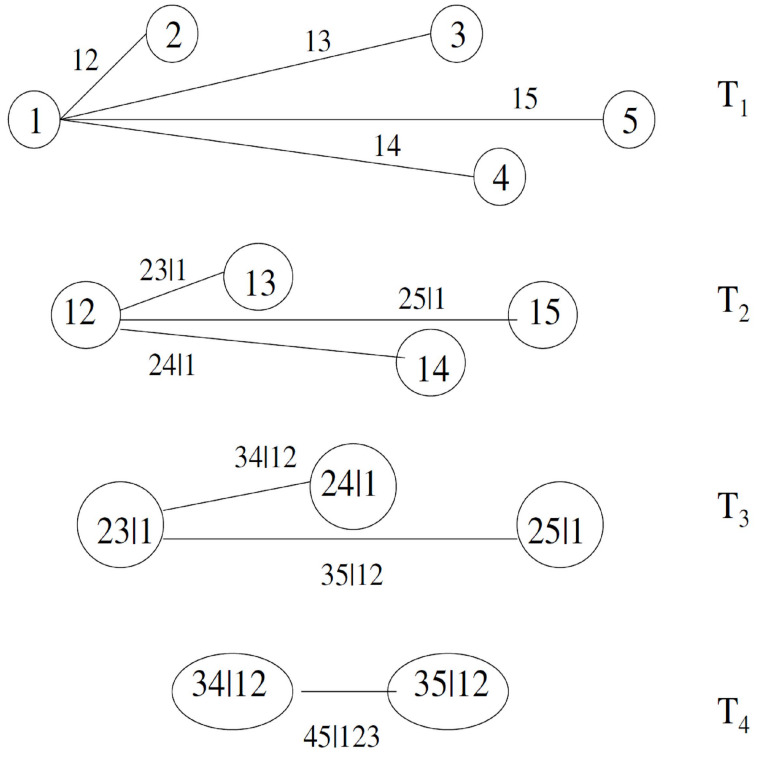

FIGURE 1.1: Example of a C-vine on five variables. Lecture: the two nodes $(1,2)$ and $(1,3)$ in T_2 are connected by the edge $(2,3|1)$, whose constraint set is $\{1,2,3\}$, conditioned set is $\{2,3\}$ and conditioning set is $\{1\}$.

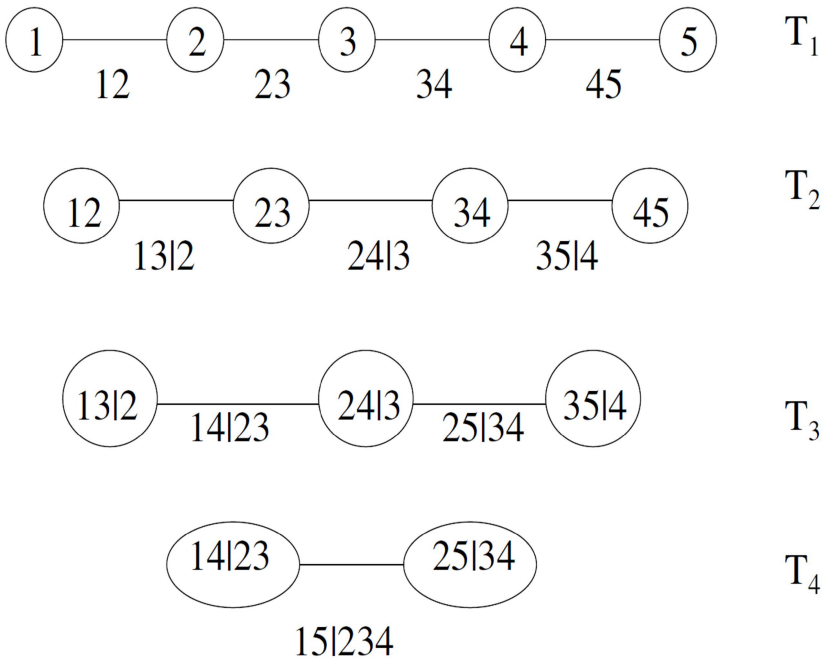

FIGURE 1.2: Example of a D-vine on five variables. Lecture: the two nodes $(1, 3|2)$ and $(2, 4|3)$ in T_3 are connected by the edge $(1, 4|2, 3)$, whose constraint set is $\{1, 2, 3, 4\}$, conditioned set is $\{1, 4\}$ and conditioning set is $\{2, 3\}$.

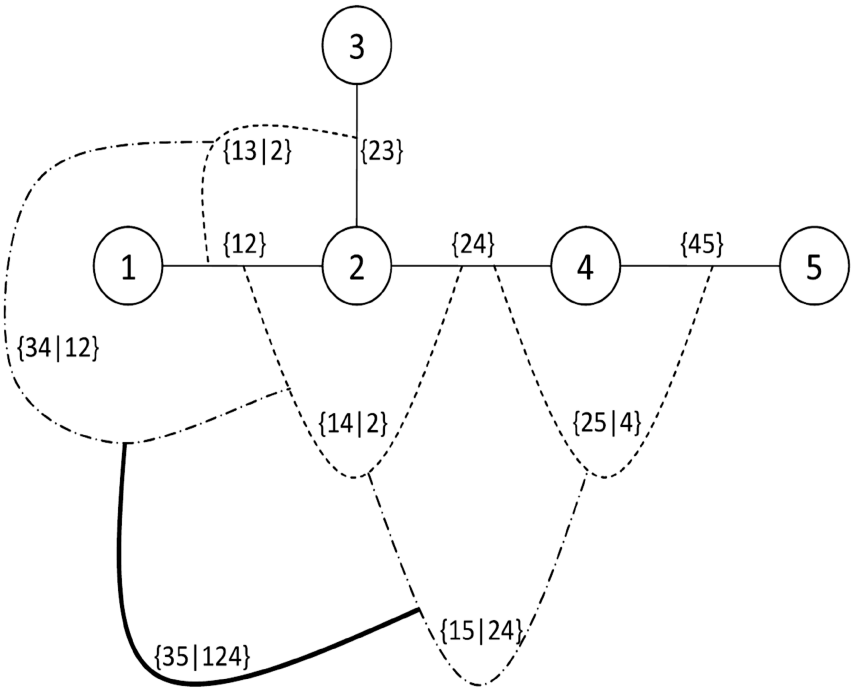

FIGURE 1.3: Example of a R-vine on five variables. The solid, dotted, dashed-dotted and black solid lines correspond to the edges of T_1, T_2, T_3 and T_4 respectively.

Appendix A

Technical result: Proof of assumption 13, Theorem 1.6.19

In this appendix, both technical results are established for the $p = q = 1$ case.

Assumption 13 is proved in this section. It is probably the most difficult part as the nonlinear dynamic of R_t should be controlled. To prove assumption 13, we need a technical assumption.

Assumption 21. Ξ and Λ are diagonal matrices such that $\|\Xi\|_s < 1$, and $\mathbb{E}\left[\log\left(\|B_{t,m}(\chi, \epsilon)\|\right)\right] < 0$, where

$$\mathbb{B}_{t-1,m}(\bar{\chi}, \epsilon) = \begin{pmatrix} \frac{2}{\pi}\|\nabla_1 \zeta_{t-1}\|\|\Lambda\| & \frac{2}{\pi}\|\nabla_1 \zeta_{t-2}\|\|\Lambda\|\|\Xi\| & \cdots & \cdots & \frac{2}{\pi}\|\nabla_1 \zeta_{t-m}\|\|\Lambda\|\|\Xi\|^{m-1} \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix},$$

Above, $\zeta_t = \zeta(\chi_t, \eta_t)$ is the t-innovation of our partial correlation process, where $\chi_t = \left(\bar{P}c_t, \bar{D}_t\right)$ is a $\mathcal{F}_{t-1}$ measurable random vector, denoting by $\bar{P}c_t$ a random set of partial correlations that satisfies 4, and $\bar{D}_t$ is bounded a.e. Moreover, for $i = 1, 2$, $\nabla_i \zeta_t$ is the derivative of ζ_t with respect to its i-th component. Finally, $E[\|\epsilon_t\|^4] < \infty$.

Now assumption 13 becomes

$$\sup_{\theta\in\Theta}|\mathbb{G}_T l_2(\epsilon;\theta) - \widetilde{\mathbb{G}}_T l_2(\epsilon;\theta)| \leq \frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}|\log(|R_t|) - \log(|\tilde{R}_t|)| + \frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}|u_t' R_t^{-1} u_t - \tilde{u}_t'\tilde{R}_t^{-1}\tilde{u}_t|. \tag{A.0.1}$$

We focus on the second sum, which can be written as

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}|u_t' R_t^{-1} u_t - \tilde{u}_t'\tilde{R}_t^{-1}\tilde{u}_t| &= \frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}|u_t'(R_t^{-1}-\tilde{R}_t^{-1})\tilde{u}_t + u_t' R_t^{-1}(u_t-\tilde{u}_t) + (u_t-\tilde{u}_t)'\tilde{R}_t^{-1}\tilde{u}_t| \\ &= \frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}|\text{Trace}\left(u_t'(R_t^{-1}-\tilde{R}_t^{-1})\tilde{u}_t + u_t' R_t^{-1}(u_t-\tilde{u}_t) + (u_t-\tilde{u}_t)'\tilde{R}_t^{-1}\tilde{u}_t\right)|.\end{aligned}$$

By definition, $u_t = D_t^{-1}\epsilon_t$ and $\tilde{u}_t = \tilde{D}_t^{-1}\epsilon_t$. Thus, the previous quantity can be written as

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}|\text{Tr}\left(\epsilon_t'\left[D_t^{-1}(R_t^{-1}-\tilde{R}_t^{-1})\tilde{D}_t^{-1} + D_t^{-1}R_t^{-1}(D_t^{-1}-\tilde{D}_t^{-1}) + (D_t^{-1}-\tilde{D}_t^{-1})\tilde{R}_t^{-1}\tilde{D}_t^{-1}\right]\epsilon_t\right)| \\ &= \frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}|\text{Tr}\left(\left[D_t^{-1}(R_t^{-1}-\tilde{R}_t^{-1})\tilde{D}_t^{-1} + D_t^{-1}R_t^{-1}(D_t^{-1}-\tilde{D}_t^{-1}) + (D_t^{-1}-\tilde{D}_t^{-1})\tilde{R}_t^{-1}\tilde{D}_t^{-1}\right]\epsilon_t\epsilon_t'\right)|\end{aligned}$$

We shall consider a multiplicative norm for matrices. To fix the ideas, this will be the spectral norm. Hence, we can bound the Trace operator as

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}|\text{Tr}\left(\left[D_t^{-1}(R_t^{-1}-\tilde{R}_t^{-1})\tilde{D}_t^{-1} + D_t^{-1}R_t^{-1}(D_t^{-1}-\tilde{D}_t^{-1}) + (D_t^{-1}-\tilde{D}_t^{-1})\tilde{R}_t^{-1}\tilde{D}_t^{-1}\right]\epsilon_t\epsilon_t'\right)| \\ &\leq \frac{N}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}(\|D_t^{-1}\|\|\tilde{R}_t^{-1}\|\|R_t-\tilde{R}_t\|\|R_t^{-1}\|\|\tilde{D}_t^{-1}\| + \|D_t^{-1}\|\|\tilde{D}_t^{-1}\|\|D_t-\tilde{D}_t\|\|D_t^{-1}\|(\|R_t^{-1}\|+\|\tilde{R}_t^{-1}\|))\|\epsilon_t\epsilon_t'\|.\end{aligned}$$

We denote

$$\begin{aligned}\mathbb{T}_t &= \|D_t^{-1}\|\|\tilde{R}_t^{-1}\|\|R_t-\tilde{R}_t\|\|R_t^{-1}\|\|\tilde{D}_t^{-1}\| \\ \mathbb{M}_t &= \|D_t^{-1}\|\|\tilde{D}_t^{-1}\|\|D_t-\tilde{D}_t\|\|D_t^{-1}\|\left(\|R_t^{-1}\|+\|\tilde{R}_t^{-1}\|\right)\end{aligned}$$

The main issue consists of controlling for $(R_t - \tilde{R}_t)$. We focus now on the quantity $\mathbb{T}_t$, and firstly on $\|R_t - \tilde{R}_t\|$.

$$\begin{aligned}R_t - \tilde{R}_t &= \text{vechof}(F_{\text{vine}}(Pc_t)) - \text{vechof}(F_{\text{vine}}(\tilde{Pc}_t)), \\ &= \left[F_{\text{vine}}(Pc_t(i,j|L(i,j))) - F_{\text{vine}}(\tilde{Pc}_t(i,j|L(i,j)))\right]_{1\leq i,j\leq N}.\end{aligned}$$

Let $\epsilon > 0$, and define the compact set $A_\epsilon = [-1+\epsilon, 1-\epsilon]^{N(N-1)/2}$. The one-to-one

mapping $F_{\text{vine}}(.)$ maps A_ϵ to $[-1+\tilde{\epsilon}, 1-\tilde{\epsilon}]^{N(N-1)/2}$, for some $\tilde{\epsilon} > 0$. On A_ϵ, $F_{\text{vine}}(.)$ is C^1, hence $\nabla F_{\text{vine}}(.)$ is bounded. Consequently, $F_{\text{vine}}(.)$ satisfies the Lipschitz condition: there exists $C > 0$ s.t., for all x and $\tilde{x} \in A_\epsilon^2$, we have

$$\|F_{\text{vine}}(x) - F_{\text{vine}}(\tilde{x})\|_\infty \leq C\|x - \tilde{x}\|_\infty. \tag{A.0.2}$$

If we control the dynamics of these partial correlations, then we can ensure to generate trajectories within $[-1+\tilde{\epsilon}, 1-\tilde{\epsilon}]$. The stationary partial correlation processes are defined as

$$\Psi(Pc_t) = \Omega + \Xi\Psi(Pc_{t-1}) + \Lambda\zeta_{t-1}. \tag{A.0.3}$$

When generating the partial correlation dynamics from arbitrarily fixed initial values, they are defined as

$$\Psi(\tilde{Pc}_t) = \Omega + \Xi\Psi(\tilde{Pc}_{t-1}) + \Lambda\zeta_{t-1}.$$

In this process, the matrices are diagonal. Iterating (A.0.3), we get

$$\Psi(Pc_t) = \sum_{k=1}^{t}\Xi^{k-1}\Omega + \Xi^t\Psi(Pc_0) + \sum_{k=1}^{t}\Xi^{k-1}\Lambda\zeta_{t-k},$$

where $\Psi(.)$ is applied to each component of the vector Pc_t and ζ_{t-k} is a function of Pc_{t-k}. The r.h.s. is an element of $\mathbb{R}^{N(N-1)/2}$. We recover Pc_t by inverting $\Psi(.)$ componentwise. (A.0.3) becomes

$$Pc_t = \Psi^{-1}(\sum_{k=1}^{t}\Xi^{k-1}\Omega + \Xi^t\Psi(Pc_0) + \sum_{k=1}^{t}\Xi^{k-1}\Lambda\zeta_{t-k}).$$

The trickiest part of this proof consists of controlling for the difference $Pc_t - \tilde{Pc}_t$. The difficulty comes from the necessary transformation of ϵ_t, D_t and R_t to recover ζ_t. Now we have

$$\begin{aligned} Pc_t - \tilde{Pc}_t &= \Psi^{-1}(\sum_{k=1}^{t}\Xi^{k-1}\Omega + \Xi^t\Psi(Pc_0) + \sum_{k=1}^{t}\Xi^{k-1}\Lambda\zeta_{t-k}) - \Psi^{-1}(\sum_{k=1}^{t}\Xi^{k-1}\Omega + \Xi^t\Psi(\tilde{Pc}_0) \\ &+ \sum_{k=1}^{t}\Xi^{k-1}\Lambda\tilde{\zeta}_{t-k}) \\ &= \nabla\Psi^{-1}(X)\left[\Xi^t(\Psi(Pc_0) - \Psi(\tilde{Pc}_0)) + \sum_{k=1}^{t}\Xi^{k-1}\Lambda(\zeta_{t-k} - \tilde{\zeta}_{t-k})\right], \end{aligned}$$

for some matrix random X. The componentwise derivatives of Ψ^{-1} are the bounded functions $x \mapsto \dfrac{2}{\pi(1+x^2)}$. Hence $\|\nabla\Psi^{-1}\|_\infty \leq 2/\pi$ and we obtain

$$\|Pc_t - \tilde{P}c_t\| \leq \frac{2}{\pi}\|\Xi\|^t\|\Psi(Pc_0) - \Psi(\tilde{P}c_0)\| + \frac{2}{\pi}\|\Lambda\|\sum_{k=1}^{t}\|\Xi\|^{k-1}\|\zeta_{t-k} - \tilde{\zeta}_{t-k}\|,$$

where $\zeta_{t-k} = \zeta(\chi_{t-k}, \epsilon_{t-k})$, with $\chi_{t-k} = (Pc_{t-k}, D_{t-k})$. This gives the expansion

$$\zeta(\chi_{t-k}, \epsilon_{t-k}) - \zeta(\tilde{\chi}_{t-k}, \epsilon_{t-k}) = \nabla_1\zeta(\bar{\chi}_{t-k}, \epsilon_{t-k})(Pc_{t-k} - \tilde{P}c_{t-k}) + \nabla_2\zeta(\bar{\chi}_{t-k}, \epsilon_{t-k})(D_{t-k} - \tilde{D}_{t-k}),$$

where $\bar{\chi}_t$ is located between χ_t and $\tilde{\chi}_t$. Consequently, we deduce

$$\begin{aligned}\tfrac{\pi}{2}\|Pc_t - \tilde{P}c_t\| &\leq A_t + \tfrac{2}{\pi}\|\Lambda\|\sum_{k=1}^{t}\|\Xi\|^{k-1}\Big(\|\nabla_1\zeta(\bar{\chi}_{t-k}, \epsilon_{t-k})\|\|Pc_{t-k} - \tilde{P}c_{t-k}\| \\ &+ \|\nabla_2\zeta(\bar{\chi}_{t-k}, \epsilon_{t-k})\|\|D_{t-k} - \tilde{D}_{t-k}\|\Big),\end{aligned}$$

with $A_t = 2\|\Xi\|^t\|\Psi(Pc_0) - \Psi(\tilde{P}c_0)\|/\pi$. Denote $r_t = \|Pc_t - \tilde{P}c_t\|$ and $d_t = \|D_t - \tilde{D}_t\|$. Note that r_t is uniformly bounded, by a constant that depends on the considered norm. To simplify and wlog, this constant will be one here. We obtain

$$r_t \leq A_t + \frac{2}{\pi}\|\Lambda\|\sum_{k=1}^{t-1}\|\Xi\|^{k-1}\left(\|\nabla_1\zeta(\bar{\chi}_{t-k}, \epsilon_{t-k})\|r_{t-k} + \|\nabla_2\zeta(\bar{\chi}_{t-k}, \epsilon_{t-k})\|d_{t-k}\right). \quad \text{(A.0.4)}$$

Now we rewrite (A.0.4), for all $t \geq T$ and for some $m \leq t$ large enough that will be stated after, as

$$\vec{r}_{t,m} \leq \mathbb{C}_{t,m} + \mathbb{B}_{t-1,m}(\bar{\chi}, \epsilon)\vec{r}_{t-1,m}, \quad \text{(A.0.5)}$$

where $\mathbb{C}_{t,m} = \vec{A}_t + \vec{\mathcal{K}}_{t,m} + \vec{\mathcal{D}}_t$, and the vectors

$$\begin{aligned}&\vec{r}_{t,m} = (r_t, r_{t-1}, \cdots, r_{t-m+1})', \ \vec{A}_t = (A_t, 0, \cdots, 0)', \ \vec{d}_{t,m} = (d_t, d_{t-1}, \cdots, d_{t-m+1})', \\ &\vec{\mathcal{K}}_{t,m} = (\tfrac{2}{\pi}\|\Lambda\|\sum_{k=m+1}^{t}\|\nabla_1\zeta(\bar{\chi}_{t-k}, \epsilon_{t-k})\|\|\Xi\|^{k-1}r_{t-k}, 0, \cdots, 0)', \\ &\vec{\mathcal{D}}_t = (\tfrac{2}{\pi}\|\Lambda\|\sum_{k=1}^{t}\|\nabla_2\zeta(\bar{\chi}_{t-k}, \epsilon_{t-k})\|\|\Xi\|^{k-1}d_{t-k}, 0, \cdots, 0)'.\end{aligned}$$

These quantities are such that $\vec{r}_{t,m} \in \mathbb{R}^m$, $\vec{A}_t \in \mathbb{R}^m$, $\vec{\mathcal{K}}_{t-1,m} \in \mathbb{R}^m$, $\vec{\mathcal{D}}_t \in \mathbb{R}^m$.

We first focus on $\mathbb{C}_{t,m}$. For our matrix norm, we have

$$\|\mathbb{C}_{t,m}\| \leq \|\vec{A}_t\| + \|\vec{\mathcal{K}}_{t,m}\| + \|\vec{\mathcal{D}}_t\|.$$

Now iterating t in (A.0.5), let $0 < q < t$ fixed, we obtain

$$\vec{r}_{t,m} \leq \mathbb{C}_{t,m} + \sum_{k=1}^{q} \mathbb{B}_{t-1,m}(\bar{\chi},\epsilon)\mathbb{B}_{t-2,m}(\bar{\chi},\epsilon)\dots\mathbb{B}_{t-k,m}(\bar{\chi},\epsilon)\mathbb{C}_{t-k,m} + \mathbb{B}_{t-1,m}(\bar{\chi},\epsilon)\cdots\mathbb{B}_{t-q-1,m}(\bar{\chi},\epsilon)\vec{r}_{t-q-1,m}.$$

The sequence of matrices $\mathbb{B}_{t-k,m}(\bar{\chi},\epsilon)$ is stochastic and each of them has a size depending on m. Under our assumptions, the series $\mathcal{B}_{t,m} := \sum_{k=1}^{+\infty} \prod_{j=1}^{k} \mathbb{B}_{t-j,m}(\bar{\chi},\epsilon)$ is converging a.s. In particular, its main term tends to zero.

$$\begin{aligned}\mathbb{P}\left(|\vec{r}_{t,m}| > \epsilon\right) &\leq \mathbb{P}(\|\mathbb{C}_{t,m}\| > \epsilon/3) + \mathbb{P}(\prod_{j=1}^{q+1} \|\mathbb{B}_{t-j,m}(\bar{\chi},\epsilon)\| > \epsilon/3) \\ &+ \quad \mathbb{P}(\sum_{k=1}^{q}\prod_{j=1}^{k} \|\mathbb{B}_{t-j,m}(\bar{\chi},\epsilon)\|.\|\mathbb{C}_{t-k,m}\| > \epsilon/3) := T_1 + T_2 + T_3.\end{aligned}$$

First, let us manage T_1, i.e. the $\mathbb{C}_{t,m}$ term. Since $\|\Psi(Pc_0) - \Psi(\tilde{P}c_0)\|$ is a fixed finite random variable and since $\|\Xi\| < 1$,

$$\mathbb{P}(\|A_t\| > \epsilon/9) < \epsilon,$$

for t sufficiently large (and independently of m and q). Moreover,

$$\mathbb{P}\left(\vec{\mathcal{K}}_{t,m} > \epsilon/9\right) \leq \mathbb{P}\left(\frac{2}{\pi}\|\Lambda\| \sum_{k=m+1}^{t} \|\nabla_1\zeta(\bar{\chi}_{t-k},\epsilon_{t-k})\|.\|\Xi\|^{k-1-m}.\|\Xi\|^{m} > \epsilon/9\right) \leq \epsilon,$$

for m sufficiently large and because the latter series converges a.s.

Denote by ρ the largest parameter among $\tau_1,\dots,\tau_n$. By assumption, $\rho \in [0,1)$. Equation (4.6) in Francq and Zakoian (2004) provides $\sup_\theta \|D_t - \tilde{D}_t\| \leq K\rho^t$ a.s. Therefore,

$$\begin{aligned}\mathbb{P}\left(\|\vec{\mathcal{D}}_t\| > \epsilon/9\right) &\leq \mathbb{P}\left(\frac{2K}{\pi}\|\Lambda\|\sum_{k=1}^{t}\|\nabla_2\zeta(\bar{\chi}_{t-k},\epsilon_{t-k})\|\|\Xi\|^{k-1}\rho^{t-k} > \epsilon/9\right) \\ &\leq \quad \mathbb{P}\left(\frac{2K\|\Lambda\|}{\pi t}\sum_{k=1}^{t}\|\nabla_2\zeta(\bar{\chi}_{t-k},\epsilon_{t-k})\|.t\max(\|\Xi\|,\rho)^{t-1} > \epsilon/9\right) \\ &\leq \quad \epsilon\end{aligned}$$

for t sufficiently large, under our assumptions and the LLN. We deduce $T_1 \leq 3\epsilon$, for a well-chosen (and now fixed) m and for t sufficiently large.

Second, note that the main term of the series $\mathcal{B}_{t,m}$ tends to zero a.s. Therefore, $T_2 < \epsilon$ for the previous fixed m and q sufficiently large.

Third, it remains to deal with T_3. Actually, it is sufficient to use the same arguments as for T_1. Indeed,

$$\begin{aligned}
\mathbb{P}(\sum_{k=1}^{q}\prod_{j=1}^{k}\|\mathbb{B}_{t-j,m}(\bar{\chi},\epsilon)\|.\|\mathbb{C}_{t-k,m}\| > \epsilon/3) &\leq \mathbb{P}(\sum_{k=1}^{q}\prod_{j=1}^{k}\|\mathbb{B}_{t-j,m}(\bar{\chi},\epsilon)\|.\|\vec{A}_{t-k,m}\| > \epsilon/9) \\
+ \quad \mathbb{P}(\sum_{k=1}^{q}\prod_{j=1}^{k}\|\mathbb{B}_{t-j,m}(\bar{\chi},\epsilon)\|.\|\vec{\mathcal{K}}_{t-k,m}\| > \epsilon/9) &+ \mathbb{P}(\sum_{k=1}^{q}\prod_{j=1}^{k}\|\mathbb{B}_{t-j,m}(\bar{\chi},\epsilon)\|.\|\vec{\mathcal{D}}_{t-k,m}\| > \epsilon/9) \\
:= \quad T_{31} + T_{32} + T_{33}.
\end{aligned}$$

To be specific, due to the finiteness of $\mathcal{B}_{t,m}$,

$$T_{31} \leq \frac{2}{\pi}\mathbb{P}(\|\Psi(Pc_0) - \Psi(\tilde{P}c_0)\|.\|\Xi\|^{t-1}.\sum_{k=1}^{+\infty}\prod_{j=1}^{k}\|\mathbb{B}_{t-j,m}(\bar{\chi},\epsilon)\| > \epsilon/9),$$

that is less than ϵ for t sufficiently large (and a fixed m). The terms T_{32} and T_{33} are managed as above, because the multiplication by the (a.e. finite) random variable $\mathcal{B}_{t,m}$ does not change the reasoning.

By grouping the all inequalities above and since the reasonings were uniform wrt θ, we get

$$\mathbb{P}\left(\sup_{\theta\in\Theta}|\vec{r}_{t,m}| > \epsilon\right) \leq 7\epsilon,$$

proving that $\sup_{\theta\in\Theta} r_t = o_P(1)$. Since it is bounded by one and due to the dominated convergence theorem, this convergence to zero is true in L^1 or L^2. This is true for $\|R_t - \tilde{R}_t\|$ too, because of (A.0.2): $\sup_{\theta\in\Theta}\|R_t - \tilde{R}_t\| = o_P(1)$ and $T^{-1}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\|R_t - \tilde{R}_t\|$ tends to zero when $t \to \infty$.

We now focus on the precision matrix $R_t^{-1} := [\rho_t^{ij}]$. Obviously,

$$\rho_t^{ij} = (-1)^{i+j}\frac{det(R_t^{-(i,j)})}{det(R_t)},$$

where $R_t^{-(i,j)}$ is the covmatrix of R_t (the matrix deduced from R_t after having removed line i and column j). But note that Theorem 3.2 in Kurowicka and Cooke (2006) and

assumption 4 implies that there exists a constant a s.t. $det(R_t) > a > 0$ a.s. Since $det(R_t^{-(i,j)})$ is a finite sum of elements in $[-1, 1]$, this term is bounded from above. Therefore, there exists a constant M_1 s.t.

$$\sup_{\theta\in\Theta}\|R_t^{-1}\| \leq M_1, \text{ a.s.}$$

The same argument holds for $\tilde{R}_t$: $\sup_{\theta\in\Theta}\|\tilde{R}_t^{-1}\| \leq M_2$.

Since $\|D_t^{-1}\|$, $\|\tilde{D}_t^{-1}\|$ and $\|R_t^{-1}\|$ are uniformly bounded from above, we deduce

$$\begin{aligned}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\mathbb{T}_t.\|\epsilon_t\epsilon_t'\| > \epsilon\right) &\leq \mathbb{P}\left(\frac{Cte}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\|Pc_t - \tilde{P}c_t\|.\|\epsilon_t\epsilon_t'\| > \epsilon\right)\\ \leq \quad \frac{Cte}{\epsilon}E\left[\sup_{\theta\in\Theta} r_t.\|\epsilon_t\epsilon_t'\|\right] &\leq \frac{Cte}{\epsilon}E\left[\left(\sup_{\theta\in\Theta} r_t\right)^2\right]^{1/2}.E\left[\|\epsilon_t\epsilon_t'\|^2\right]^{1/2},\end{aligned}$$

that is less than ϵ for t sufficiently large.

The second term $\mathbb{M}_t$ can be bounded more straightforwardly. Using the stationarity assumption of the GARCH process, there exists $U > 0$, and $\rho \in]0, 1[$ such that, a.s.,

$$\sup_{\theta\in\Theta}\sup_{i}\left|h_{i,t} - \tilde{h}_{i,t}\right| \leq U\rho^t.$$

Consequently, $\mathbb{M}_t$ can be bounded as

$$\sup_{\theta\in\Theta}\mathbb{M}_t = \sup_{\theta\in\Theta}\|D_t^{-1}\|\|\tilde{D}_t^{-1}\|\|D_t - \tilde{D}_t\|\|D_t^{-1}\|\left(\|R_t^{-1}\| + \|\tilde{R}_t^{-1}\|\right) \leq C\rho^t, \text{ a.s,}$$

for some constant C. Then

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\mathbb{M}_t\|\epsilon_t\epsilon_t'\| > \epsilon\right) \leq \mathbb{P}\left(\frac{C}{T}\sum_{t=1}^{T}\rho^t\|\epsilon_t\epsilon_t'\| > \epsilon\right) \leq \frac{C}{T\epsilon(1-\rho)}E\left[\|\epsilon_t\epsilon_t'\|\right] < \epsilon,$$

for t sufficiently large.

In other words, we have proved that

$$\frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}(\mathbb{T}_t + \mathbb{M}_t).\|\epsilon_t\epsilon_t'\| = o_P(1).$$

For the first sum of (A.0.1) and considering the spectral norm, we have:

$$\begin{aligned}
\log(|R_t|) - \log(|\tilde{R}_t|) &= \log(|I_N + (R_t - \tilde{R}_t)\tilde{R}_t^{-1}|) \\
&\leq N \log(\|I_N + (R_t - \tilde{R}_t)\tilde{R}_t^{-1}\|) \\
&\leq N \log(\|I_N\| + \|(R_t - \tilde{R}_t)\tilde{R}_t^{-1}\|) \\
&\leq N \log(1 + \|(R_t - \tilde{R}_t)\tilde{R}_t^{-1}\|) \\
&\leq N\|R_t - \tilde{R}_t\|\|\tilde{R}_t^{-1}\|.
\end{aligned}$$

By symmetry $\log(|\tilde{R}_t|) - \log(|R_t|) \leq N\|\tilde{R}_t - R_t\|\|R_t^{-1}\|$. Using the previous arguments, the first sum of (A.0.1) converges to 0 when $T \longrightarrow \infty$. We proved that

$$\sup_{\theta \in \Theta} |\mathbb{G}_T l_2(\epsilon; \theta) - \widetilde{\mathbb{G}}_T l_2(\epsilon; \theta)| = o_p(1).$$

Appendix B

Technical result: Proof of assumption 15, Theorem 1.6.19

In this appendix, both technical results are established for the $p = q = 1$ case.

To prove this statement, we need the following assumption.

Assumption 22. Let (A_t, B_t) defined as

$$\begin{aligned} A_t &:= \sup_{\theta:\|\theta_v-\theta_{0,v}\|<\alpha} \| \left(\nabla\Psi(Pc_t)\right)^{-1} \Lambda\nabla_{D_t}\zeta(Pc_t, D_t, \epsilon_t)\nabla_{\theta_v} D_t\|, \\ B_t &:= \sup_{\theta:\|\theta_v-\theta_{0,v}\|<\alpha} \| \left(\nabla\Psi(Pc_t)\right)^{-1} \left[\Xi\nabla\Psi(Pc_t) + \Lambda\nabla_{Pc_t}\zeta(Pc_t, D_t, \epsilon_t)\right] \|. \end{aligned}$$

For some $\alpha > 0$, the stochastic matrix process (A_t, B_t) is stationary, $\mathbb{E}[A_t] < +\infty$ and

$$\sum_{k\geq 1} \mathbb{E}\left[B_{t-1}B_{t-2}\cdots B_{t-k}A_{t-k-1}\right] < \infty.$$

Proof of Lemma 15. Applying a Taylor expansion to $QL_{2,T}(\hat{\theta}_{T,v}, \theta_c; \epsilon)$ around $\theta_{0,v}$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T} l_{2,t}(\epsilon_t; \hat{\theta}_{T,v}, \theta_c) = \frac{1}{T}\sum_{t=1}^{T} l_{2,t}(\epsilon_t; \theta_{0,v}, \theta_c) + (\hat{\theta}_{T,v} - \theta_{0,v})\frac{1}{T}\sum_{t=1}^{T}\nabla_{\theta_v} l_{2,t}(\epsilon_t; \bar{\theta}_v, \theta_c),$$

for some $\bar{\theta}_v$, $\|\bar{\theta}_v - \theta_{0,v}\| < \|\theta_{0,v} - \hat{\theta}_{T,v}\|$. Using the consistency of $\hat{\theta}_{T,v}$, it is sufficient to prove that

$$\frac{1}{T}\sum_{t=1}^{T} \sup_{\{\theta\in\Theta \,|\, \|\theta_v-\theta_{0,v}\|<\alpha\}} \|\nabla_{\theta_v} l_{2,t}(\epsilon_t; \theta_v, \theta_c)\| = O_P(1), \tag{B.0.1}$$

for some (small) $\alpha > 0$. Applying some matrix derivation rules (see Lütkepohl, 1996), the analytical score of the second step likelihood with respect to the i-th element of θ_v is given by

$$\begin{aligned}
\partial_{\theta_v^i} l_{2,t}(\epsilon_t;\theta) &= \partial_{\theta_v^i}\left[\log\left(|R_t|\right) + \epsilon_t' D_t^{-1} R_t^{-1} D_t^{-1}\epsilon_t\right] \\
&= \text{Trace}\left(R_t^{-1}(\partial_{\theta_v^i} R_t)\right) + \text{Trace}\left(\epsilon_t\epsilon_t'\partial_{\theta_v^i}\left[D_t^{-1}R_t^{-1}D_t^{-1}\right]\right) \\
&= \text{Trace}\left(R_t^{-1}(\partial_{\theta_v^i} R_t)\right) - \text{Trace}\left(\epsilon_t\epsilon_t'\left[D_t^{-1}(\partial_{\theta_v^i} D_t)D_t^{-1}R_t^{-1}D_t^{-1}\right]\right) \\
&- \text{Trace}\left(\epsilon_t\epsilon_t'\left[D_t^{-1}R_t^{-1}(\partial_{\theta_v^i} R_t)R_t^{-1}D_t^{-1}\right]\right) - \text{Trace}\left(\epsilon_t\epsilon_t'\left[D_t^{-1}R_t^{-1}D_t^{-1}(\partial_{\theta_v^i} D_t)D_t^{-1}\right]\right).
\end{aligned}$$

Obviously, the matrices D_t^{-1} are bounded from above by positive constants due to the definition of our univariate GARCH dynamics. Concerning correlations, we know that R_t^{-1} is bounded from above, due to assumption 4. As for the derivatives of R_t, note that $\|\nabla_{\theta_v} R_t\| \leq \|\nabla F_{\text{vine}}(Pc_t).\nabla_{\theta_v} Pc_t\|$ and that the derivative of $F_{\text{vine}}(\cdot)$ is bounded a.e. under the latter assumption.

Consequently, there exists some positive constant C such that, for any $\alpha > 0$,

$$\sup_{\theta:\|\theta_v-\theta_{0,v}\|<\alpha} |\nabla_{\theta_v} l_{2,t}(\epsilon_t;\theta_c,\theta_v)| \leq C. \sup_{\theta:\|\theta_v-\theta_{0,v}\|<\alpha} \{(\|\nabla_{\theta_v} D_t\| + \|\nabla_{\theta_v} Pc_t\|)\|\epsilon_t\|^2 + \|\nabla_{\theta_v} Pc_t\|\}.$$

Let us focus on $\nabla_{\theta_v} Pc_t$. By the chain rule, we have

$$\begin{aligned}
\nabla_{\theta_v} Pc_t &= \left(\nabla\Psi(Pc_{t-1})\right)^{-1}\left[\Xi\nabla\Psi(Pc_{t-1}) + \Lambda\nabla_{Pc}\zeta(Pc_{t-1}, D_{t-1}, \epsilon_{t-1})\right]\nabla_{\theta_v} Pc_{t-1} \\
&+ \left(\nabla\Psi(Pc_{t-1})\right)^{-1}\Lambda\nabla_D\zeta(Pc_{t-1}, D_{t-1}, \epsilon_{t-1})\nabla_{\theta_v} D_{t-1},
\end{aligned}$$

and then

$$\begin{aligned}
\sup_{\theta:\|\theta_v-\theta_{0,v}\|<\alpha} \|\nabla_{\theta_v} Pc_t\| &\leq A_{t-1} + B_{t-1} \sup_{\theta:\|\theta_v-\theta_{0,v}\|<\alpha} \|\nabla_{\theta_v} Pc_{t-1}\| \\
&\leq A_{t-1} + \sum_{k=1}^{\infty} B_{t-1}B_{t-2}\cdots B_{t-k}A_{t-k-1}. \qquad \text{(B.0.2)}
\end{aligned}$$

Assumption 22 provides sufficient conditions so that the latter series belongs to L^1. As a consequence, the existence of the series (B.0.2) is ensured a.s. But we need a stronger assumption than in Theorem 1.1. of Bougerol and Picard (1992) typically, because of the integrability requirement. This implies

$$\frac{1}{T}\sum_{t=1}^{T} \sup_{\theta:\|\theta_v-\theta_{0,v}\|<\alpha} \|\nabla_{\theta_v} Pc_t\|.(\|\epsilon_t\|^2 + 1) = O_P(1).$$

We now focus on $\|\nabla_{\theta_v} D_t\|$, which is determined as $\|\partial_{\theta_v^i} D_t\| = \|D_t^{-1}\text{diag}\left(\partial_{\theta_v^i} h_{j,t}\right)\|/2$, $i = 1, \cdots, 3N$. The partial derivative of the j-th component above is zero when $i \neq j$. Otherwise, note that, by iterating the volatility process equation, we have

$$h_{j,t} = \frac{\varsigma_j}{1-\tau_j} + \kappa_j \left(\sum_{k \geq 1} \tau_j^{k-1} \epsilon_{j,t-k}^2 \right),$$

$$\partial_{\varsigma_j} h_{j,t} = \frac{\varsigma_j}{1-\tau_j}, \ \partial_{\kappa_j} h_{j,t} = \sum_{k \geq 1} \tau_j^{k-1} \epsilon_{j,t-k}^2, \text{ and } \partial_{\tau_j} h_{j,t} = \frac{\varsigma_j}{(1-\tau_j)^2} + \sum_{k \geq 1} (k-1) \tau_j^{k-2} \epsilon_{j,t-k}^2.$$

We deduce there exists some constant C s.t.

$$\sup_{\theta : \|\theta_v - \theta_{0,v}\| < \alpha} \|\nabla_{\theta_v} D_t\| . \|\epsilon_t\|^2 \leq C \left(1 + \sum_{k \geq 1} (k-1) \tau_j^{k-1} \epsilon_{j,t-k}^2 \right) \|\epsilon_t\|^2 \text{ a.s.}$$

The latter r.h.s. belongs to L^1 because $E_{t-1}[\epsilon_{j,t}^2] = 1$ for every j and t. Therefore,

$$\frac{1}{T} \sum_{t=1}^{T} \sup_{\theta : \|\theta_v - \theta_{0,v}\| < \alpha} \|\nabla_{\theta_v} D_t\| . \|\epsilon_t\|^2 = O_P(1),$$

proving (B.0.1) and then our lemma. □

Chapter 2

Asymptotic Theory of the Sparse Group Lasso

2.1 Introduction

Model complexity is an obstacle when one models richly parameterized dynamics such as multivariate nonlinear dynamic systems. For instance, dynamic variance correlation processes of size N have an $O(N^2)$ complexity as in the dynamic conditional correlation parametrization (DCC, Ding and Engle, 2001). Another issue arises when the sample size, say T, is comparable to N, which may reduce the estimation performances. This is typically a high-dimensional statistical framework.

A significant literature developed on model penalization, which consists of reducing the number of parameters and performing variable selection. For instance, the Akaike's or Bayesian information criteria aim at selecting the size of a model. However, these methods are unstable, computationally complex and their sampling properties are difficult to study as Fan and Li (2001) pointed out mainly because they are stepwise and subset selection procedures.

The penalization or regularization procedures aim at overcoming these drawbacks. They specify a penalty function (also called regularizer) to the statistical problem, which is singular at zero to foster sparsity and thus performs variable selection and estimation. The choice of the norm depends on the problem at hand and the key quantity is the tuning parameter, also called the regularization parameter, which depends on the sample size and controls for the bias. The Lasso procedure of Tibshirani

(1996) specifies a l^1 norm over the parameters, which fosters sparsity and allows for continuity of the selected models. Other penalties were proposed such as the smoothly clipped absolute deviation (SCAD) of Fan (1997), which modifies the Lasso to shrink large coefficients less severely. The elastic net regularization procedure of Zou and Hastie (2005) was developed to overcome the collinearity between the variables, which hampers the Lasso to perform well. Their idea consists of mixing a l^1 penalty, which performs variable selection, with a l^2 penalty, which stabilizes the solution paths. The Group Lasso of Yuan and Lin (2006) fosters sparsity and variable selection in a group of variables. Simon, Friedman, Hastie and Tibshirani (2013) designed the Sparse Group Lasso (SGL) to foster sparsity both at a a group level and within a group. Their penalization involves a l^1 Lasso type penalty and a mixed l^1/l^2 penalty for group selection.

All these procedures, together with the algorithms designed for performing selection and estimation, were developed within a linear framework. The penalized Ordinary Least Squares (OLS) loss function is typically used for linear models as it is convex, which makes the computation easier, and allows for closed form solutions, such as the soft-thresholding operator for the Lasso penalty. Furthermore, linear modeling allows for deriving non asymptotic oracle inequalities straightforwardly: see Bühlmann and van de Geer (2011) on this non-asymptotic framework.

Knight and Fu (2000) explored the asymptotic properties of the Lasso penalty for OLS loss functions. Fan and Li (2001) proposed a penalization framework for general likelihood functions and studied the asymptotic properties of the SCAD penalty. They proved that the SCAD estimator satisfies the oracle property, that is the sparsity based estimator recovers the true underlying sparse model and is asymptotically normally distributed. The Lasso as proposed by Tibshirani cannot satisfy the oracle property. To fix this drawback, Zou (2006) proposed the adaptive Lasso within an OLS framework, where adaptive weights are used to penalize different coefficients in the penalty. Nardi and Rinaldo (2008) applied the same methodology for the Group Lasso estimator within an OLS framework and studied its oracle property.

These theoretical studies were developed for fixed dimensional models with i.i.d. data, a case where N does not depend on the sample size, and for least squares type loss functions, except Fan and Li (2001). Fan and Peng (2004) considered the general penalized likelihood framework when the number of parameters grows with the sample size and focused on the oracle property for general penalties. Zou and Zhang (2009) also focused on the oracle property of the adaptive elastic-net within the double-asymptotic

framework. Their work highlights that adaptive weights penalizing different coefficients are key quantities to satisfy the oracle property as one can modify the convergence rate of the penalty terms. Nardi and Rinaldo (2008) also proposed within the double-asymptotic setting selection consistency results, which states that asymptotically the right set of relevant variables is selected.

In this paper, we develop the asymptotic theory of penalized M-estimators for convex criteria and dependent variables. Within a time series framework, we specify a generalization of the Sparse Group Lasso estimator of Simon and al. (2013). More precisely, what is new is that we specify two regularization parameters: one for the l^1 Lasso norm and one for the l^1/l^2 Group Lasso norm. This penalty is relevant for problems where one would like to foster sparsity for selecting active groups, that is a group for which some of the corresponding coefficients are non zero, and active coefficients within an active group, a situation where a coefficient is non zero within an active group. Hence this is somehow a two step approach as first the active groups are selected, and then the active variables within an active group are selected. We prove that the SGL as proposed by Simon and al. (2013) does not satisfy the oracle property. Then we propose a new version of the SGL, the adaptive SGL using the same methodology of Zou (2006), which consists of penalizing different coefficients and groups of coefficients using random weights that are positive functions of a first step estimator. This enables to alter the rate of convergence of the penalties such that the adaptive SGL satisfies the oracle property. We provide explicit convergence rate of the regularization parameters and the asymptotic trade-off between the l^1 Lasso and l^1/l^2 Group Lasso regularizations. We also prove that the adaptive SGL satisfies the oracle property in a double-asymptotic framework, a situation where the model complexity grows with the sample size.

The rest of the paper is organized as follows. In Section 2.2, we describe our general framework for penalized convex empirical criteria and the SGL penalty. In Section 2.3, we derive the optimality conditions of the statistical criterion. In Section 2.4, we derive the asymptotic properties of both the SGL and adaptive SGL when the number of parameters is fixed. In Section 2.5, we prove the oracle property of the adaptive SGL in a double-asymptotic setting. In Section 2.6, we use simulations to compare the finite sample performance of the adaptive Sparse Group Lasso with other competitors.

2.2 Framework and notations

We consider a dynamic system in which the criterion is written as an empirical criterion, that is

$$\theta \mapsto \mathbb{G}_T l(\theta) = \frac{1}{T}\sum_{t=1}^{T} l(\epsilon_t;\theta), \tag{2.2.1}$$

such that $l(.)$ is "a general" known loss function on the sample space such that for any process (ϵ_t), $\theta \mapsto l(\epsilon_t;\theta)$ is convex. This framework encompasses for instance the maximum likelihood method, where the $l(.)$ function corresponds to $l(\epsilon_t;\theta) = -\log f(\epsilon_t;\theta)$, where $f(\epsilon_t;\theta)$ is the density of the observation (ϵ_t) under $\mathbb{P}_\theta$. Alternatively, a linear model would imply $l(\epsilon_t;\theta) = \|\epsilon_t^{(1)} - \theta'\epsilon_t^{(2)}\|_p$, where $(\epsilon_t^{(1)}, \epsilon_t^{(2)}) = \epsilon_t$. We denote the empirical score and Hessian of the empirical criterion respectively as

$$\dot{\mathbb{G}}_T l(\theta) = \frac{1}{T}\sum_{t=1}^{T} \nabla_\theta l(\epsilon_t;\theta),\ \ddot{\mathbb{G}}_T l(\theta) = \frac{1}{T}\sum_{t=1}^{T} \nabla^2_{\theta\theta'} l(\epsilon_t;\theta).$$

The dependent nature of our framework requires the use of particular probabilistic tools to study the asymptotic properties of M-estimators. We extensively use the ergodic theorem and central limit theorem (Billingsley, 1961, 1995) to obtain convergence in probability of empirical quantities to their theoretical counterparts and central limit theorems. To do so, we assume the stationarity and the ergodicity of the underlying process (ϵ_t): see assumption 23 in Section 2.4.

In this setting, $\epsilon_t \in \mathbb{R}^N$ and $\theta \in \mathbb{R}^d$, a vector that can be split into m groups $\mathcal{G}_k, k = 1, \cdots, m$, such that $\text{card}(\mathcal{G}_k) = \boldsymbol{c}_k$ and $\sum_{k=1}^{m} \boldsymbol{c}_k = d$. We suppose no overlap between these groups. We use the notation $\theta^{(l)}$ as the subvector of θ, that is the set $\{\theta_k : k \in \mathcal{G}_l\}$. Hence the vector $\theta = (\theta_j, j = 1, \cdots, d)$ can be written as $\theta = (\theta_i^{(k)}, k \in \{1, \cdots, m\}, i = 1, \cdots, \boldsymbol{c}_k)$ [1]. We denote by θ_0 the true parameter vector of interest. Moreover, $\theta \to \mathbb{E}[l(\epsilon_t;\theta)]$ is supposed to be a one-to-one mapping and is minimized uniquely at $\theta = \theta_0$.

[1] Formally, there is a one-to-one mapping between two ways for writing θ:

$$\psi : \{1, \cdots, d\} \to \{(k,i), k = 1, \ldots, m; i = 1, \cdots, \boldsymbol{c}_k\},$$
$$j \mapsto \psi(j) = (k_j, i_j).$$

In the rest of this paper, this mapping is implicit such that we allow such writings as $j = (k,i)$ or $j = i_k$ where k is clear.

We denote by $\mathcal{S} := \{k : \theta^{(k)} \neq 0\}$ the set of indices for which the groups are active. Let $\mathcal{A} := \{j : \theta_{0,j} \neq 0\}$ be the true subset model, which can be decomposed into sub-groups of active sets as $l \in \mathcal{S}$, $\mathcal{A}_l = \{(l,i) : \theta^{(l)}_{0,i} \neq 0\}$. Besides, there are inactive indices $\mathcal{G}_l \setminus \mathcal{A}_l = \mathcal{A}_l^c = \{(l,i) : \theta^{(l)}_{0,i} = 0\}$. We have $\{l \notin \mathcal{S}\} \Leftrightarrow \{\forall i = 1, \cdots, \boldsymbol{c}_l, \theta^{(l)}_{0,i} = 0\}$. In this setting, $\mathcal{A} = \underset{l\in\mathcal{S}}{\cup}\mathcal{A}_l$ such that for $k \neq l$, $\mathcal{A}_k \cap \mathcal{A}_l = \emptyset$. Furthermore, $\mathcal{A}^c = \overset{m}{\underset{l=1}{\cup}}\mathcal{A}_l^c$ such that for $k \neq l$, $\mathcal{A}_k^c \cap \mathcal{A}_l^c = \emptyset$.

Finally, we need the following notations: $\dot{\mathbb{G}}_T l(\theta)_{(k)} \in \mathbb{R}^{\boldsymbol{c}_k}$ is the "score" vector of the empirical criterion taken over group k of size $\boldsymbol{c}_k$, $\dot{\mathbb{G}}_T l(\theta)_{(k),i} \in \mathbb{R}$ is the i-th component of this score, and $\dot{\mathbb{G}}_T l(\theta)_{\mathcal{A}} \in \mathbb{R}^{\text{card}(\mathcal{A})}$ is the score over the set of active indices. $\ddot{\mathbb{G}}_T l(\theta)_{(k)(k)} \in \mathcal{M}_{\boldsymbol{c}_k \times \boldsymbol{c}_k}(\mathbb{R})$ (resp. $\mathbb{H}_{(k)(k)}$) is the empirical (resp. theoretical) Hessian taken over the block representing group k, and $\ddot{\mathbb{G}}_T l(\theta)_{\mathcal{A}\mathcal{A}} \in \mathcal{M}_{\text{card}(\mathcal{A})\times\text{card}(\mathcal{A})}(\mathbb{R})$ is the Hessian over the set of active indices.

The statistical problem consists of minimizing over the parameter space Θ a penalized criterion of the form

$$\hat{\theta} = \underset{\theta\in\Theta}{\arg\min}\{\mathbb{G}_T\varphi(\theta)\}, \tag{2.2.2}$$

where

$$\begin{aligned} \theta \mapsto \mathbb{G}_T\varphi(\theta) &= \frac{1}{T}\sum_{t=1}^{T}\{l(\epsilon_t;\theta) + \boldsymbol{p}_1(\lambda_T,\theta) + \boldsymbol{p}_2(\gamma_T,\theta)\} \\ &= \mathbb{G}_T l(\theta) + \boldsymbol{p}_1(\lambda_T,\theta) + \boldsymbol{p}_2(\gamma_T,\theta). \end{aligned}$$

and both penalties are specified as

$$\begin{cases} \boldsymbol{p}_1 : \mathbb{R}_+ \times \mathbb{R}_+^m \times \Theta \to \mathbb{R}_+, & \boldsymbol{p}_2 : \mathbb{R}_+ \times \mathbb{R}_+^m \times \Theta \to \mathbb{R}_+, \\ (\lambda_T, \alpha, \theta) \mapsto \boldsymbol{p}_1(\lambda_T,\theta) = \lambda_T T^{-1}\sum_{k=1}^{m}\alpha_k\|\theta^{(k)}\|_1, & (\gamma_T,\xi,\theta) \mapsto \boldsymbol{p}_2(\gamma_T,\theta) = \gamma_T T^{-1}\sum_{l=1}^{m}\xi_l\|\theta^{(l)}\|_2. \end{cases}$$

Both α_k and ξ_l are non negative scalar quantities for each group and the regularization parameters (tuning parameters) λ_T and γ_T vary with T.

The estimator $\hat{\theta}$ obtained in (2.2.2) is not the minimum of the empirical unpenalized criterion $\mathbb{G}_T l(.)$. Our main interest is to analyze the bias generated by the penalties and how the oracle property can be satisfied in the sense of Fan and Li (2001). More precisely, the sparsity based estimator must satisfy

$(i)\hat{\mathcal{A}} = \{i : \hat{\theta}_i \neq 0\} = \mathcal{A}$ asymptotically, that is "model selection consistency".
$(ii)\sqrt{T}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}_0)$ with $\mathbb{V}_0$ a covariance matrix related to the criterion of interest.

We highlight in Proposition 2.4.13, Section 2.4 that actually the SGL as proposed by Simon and al. (2013) cannot perform the oracle property. Hence in Section 2.4, we propose a new estimator based on the same idea as Zou (2006), the adaptive Sparse Group Lasso, for which the oracle property is obtained when the weights are randomized, as proved in Theorem 2.4.16.

This framework can be adapted to a broad range of problem. For instance, one can penalize a subset of groups with a l^1 penalty only, and the other groups with a l^1/l^2 penalty only. This framework encompasses the SGL, the Lasso and the group Lasso for proper choices of α's and ξ's.

Let us motivate the interets of the SGL approach and illustrate our notations through a simple linear example. In finance, finding the right set of explanatory variables to predict future asset returns is a significant issue. For instance, one may use Japanese companies indices, the Japanese GDP or the Japanese aggregated dividend-price ratio to explain the Nikkei index return through a linear projection. But one should also consider some foreign variables, such as the S&P 500 index or the US yield curve. Consequently, some groups of variables naturally arise: group of financial companies, tech companies, and the like; group of foreign components such as American financial companies, and the like. Hence the set $\mathcal{G}_k$ may represent the k-th ($k \leq m$) group of Japanese financial companies, composed (as a shortcoming) with Nomura (index 1), MUFG-Bank of Tokyo (index 2) and Sumitomo (index 3) represented by the parameter vector $\theta^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)})$; then $k \in \mathcal{S}$ if the whole group has a statistically significant effect on the Nikkei index. Suppose the l^1/l^2 penalty selects this group as active. Then $\mathcal{A}_k$ represents the set of active components in $\mathcal{G}_k$ such that $\boldsymbol{c}_{\mathcal{A}_k} = \text{card}(\mathcal{A}_k) \leq \text{card}(\mathcal{G}_k) = \boldsymbol{c}_k$. The l^1 penalty fosters sparsity within this selected group. If Nomura is the only variable that is expelled, then $1 \in \mathcal{A}_k^c = \mathcal{G}_k \setminus \mathcal{A}_k$, whereas $\{2, 3\} \in \mathcal{A}_k$ and $\boldsymbol{c}_{\mathcal{A}_k} = 2$.

2.3 Optimality conditions

The statistical problem consists of solving (2.2.2). Both $\mathbb{G}_T l(.)$, $\boldsymbol{p}_1(\lambda_T, \alpha, .)$ and $\boldsymbol{p}_2(\gamma_T, \xi, .)$ are convex functions and there are no inequality constraints. Consequently, by the Karush-Kuhn-Tucker optimality conditions, which are necessary and sufficient, the

estimator $\hat{\theta}$ satisfies for a group k

$$\dot{\mathbb{G}}_T l(\hat{\theta})_{(k)} + \lambda_T T^{-1} \alpha_k \hat{\boldsymbol{w}}^{(k)} + \gamma_T T^{-1} \xi_k \hat{\boldsymbol{z}}^{(k)} = 0, \tag{2.3.1}$$

for some vectors $\boldsymbol{w}^{(k)}$ and $\boldsymbol{z}^{(k)}$ satisfying

$$\hat{\boldsymbol{w}}^{(k)} = \begin{cases} \text{sgn}(\hat{\theta}_i^{(k)}) & \text{if } \hat{\theta}_i^{(k)} \neq 0, i = 1, \cdots, \boldsymbol{c}_k, \\ \in \{\hat{\boldsymbol{w}}_i^{(k)} : |\hat{\boldsymbol{w}}_i^{(k)}| \leq 1\} & \text{if } \hat{\theta}_i^{(k)} = 0, i = 1, \cdots, \boldsymbol{c}_k. \end{cases} \quad \hat{\boldsymbol{z}}^{(k)} = \begin{cases} \hat{\theta}^{(k)} / \|\hat{\theta}^{(k)}\|_2 & \text{if } \hat{\theta}^{(k)} \neq 0, \\ \in \{\hat{\boldsymbol{z}}^{(k)} : \|\hat{\boldsymbol{z}}^{(k)}\|_2 \leq 1\} & \text{if } \hat{\theta}^{(k)} = 0. \end{cases}$$

If $\hat{\theta}^{(k)} = \mathbf{0}$, we have $\|\hat{\boldsymbol{z}}^{(k)}\|_2 \leq 1$. Then, from (2.3.1), we obtain for such a $k \notin \mathcal{S}$

$$\sum_{i=1}^{\boldsymbol{c}_k} (\dot{\mathbb{G}}_T l(\hat{\theta})_{(k),i} + \lambda_T T^{-1} \alpha_k \hat{\boldsymbol{w}}_i^{(k)})^2 = \sum_{i=1}^{\boldsymbol{c}_k} (\gamma_T T^{-1} \xi_k \hat{\boldsymbol{z}}_i^{(k)})^2 \leq \gamma_T^2 T^{-2} \xi_k^2 \|\boldsymbol{z}^{(k)}\|_2^2.$$

Consequently, if the subgradient equations are satisfied for $\hat{\theta}^{(k)}$, then $\hat{\theta}^{(k)} = \mathbf{0}$ if

$$\|\dot{\mathbb{G}}_T l(\hat{\theta})_{(k)} + \lambda_T T^{-1} \alpha_k \hat{\boldsymbol{w}}^{(k)}\|_2 \leq \gamma_T T^{-1} \xi_k.$$

On the contrary, if this condition is not satisfied, then $\hat{\theta}^{(k)} \neq \mathbf{0}$. In this case, sparsity is fostered by the l^1 penalty as follows: using the optimality condition of (2.3.1), we have for $\hat{\theta}^{(k)} \neq \mathbf{0}$

$$\forall i = 1, \cdots, \boldsymbol{c}_k, -\dot{\mathbb{G}}_T l(\hat{\theta})_{(k),i} = \lambda_T T^{-1} \alpha_k \hat{\boldsymbol{w}}_i^{(k)} + \gamma_T T^{-1} \xi_k \frac{\hat{\theta}_i^{(k)}}{\|\hat{\theta}^{(k)}\|_2}.$$

If $\hat{\theta}_i^{(k)} = 0$, then $|\hat{\boldsymbol{w}}_i^{(k)}| \leq 1$ and we obtain straightforwardly

$$|\dot{\mathbb{G}}_T l(\hat{\theta})_{(k),i}| \leq \lambda_T T^{-1} \alpha_k.$$

Bertsekas (1995) proposed the use of subdifferential calculus to characterize necessary and sufficient solutions for problems such as (2.2.2). The conditions we derived are close to those of Simon and al. (2013) (obtained for a least square loss function). They will be extensively used in the rest of the paper.

2.4 Asymptotic properties

To prove the asymptotic results, we make the following assumptions.

Assumption 23. (ϵ_t) is a strictly stationary and ergodic process.

Assumption 24. The parameter set $\Theta \subset \mathbb{R}^d$ is convex and not necessarily compact.

Assumption 25. For any (ϵ_t), the function $\theta \mapsto l(\epsilon_t;\theta)$ is convex and $C^3(\mathbb{R},\Theta)$.

Assumption 26. $(\nabla l(\epsilon_t;\theta_0))$ is a square integrable martingale difference.

Assumption 27. $\mathbb{H} := \mathbb{E}[\nabla^2_{\theta\theta'} l(\epsilon_t;\theta_0)]$ and $\mathbb{M} := \mathbb{E}[\nabla_\theta l(\epsilon_t;\theta_0)\nabla_{\theta'} l(\epsilon_t;\theta_0)]$ exist and are positive definite.

Assumption 28. Let $\upsilon_t(C) = \sup_{k,l,m=1,\cdots,d}\{\sup_{\theta:\|\theta-\theta_0\|_2\leq\nu_T C}|\partial^3_{\theta_k\theta_l\theta_m} l(\epsilon_t;\theta_0)|\}$, where $C>0$ is a fixed constant and $\nu_T \underset{T\to\infty}{\longrightarrow} 0$, a quantity that will be made explicit. Then

$$\eta(C) := \frac{1}{T^2}\sum_{t,t'=1}^{T}\mathbb{E}[\upsilon_t(C)\upsilon_{t'}(C)] < \infty.$$

Remark 2.4.1. Assumptions 23 and 26 allow for using the central limit theorem of Billingsley (1961). We remind this result stated as a corollary in Billingsley (1961).

Corollary 2.4.2. (Billingsley, 1961)
If $(x_t,\mathcal{F}_t)$ is a stationary and ergodic sequence of square integrable martingal increments such that $\sigma_x^2 = Var(x_t) \neq 0$, then

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t \xrightarrow{d} \mathcal{N}(0,\sigma_x^2).$$

Note that the square martingale difference condition can be relaxed by α-mixing and moment conditions. For instance, Rio (2013) provides a central limit theorem for strongly mixing and stationary sequences.

Theorem 2.4.3. *Under assumptions 23-25, if $\lambda_T/T \to \lambda_0 \geq 0$ and $\gamma_T/T \to \gamma_0 \geq 0$, then for any compact set $\boldsymbol{B} \subset \Theta$ such that $\theta_0 \in \boldsymbol{B}$,*

$$\hat{\theta} \xrightarrow{\mathbb{P}} \underset{\boldsymbol{x}\in\boldsymbol{B}}{\arg\min}\{\mathbb{G}_\infty\varphi(\boldsymbol{x})\},$$

with

$$\mathbb{G}_\infty\varphi(\boldsymbol{x}) = \mathbb{G}_\infty l(\boldsymbol{x}) + \lambda_0\sum_{k=1}^{m}\alpha_k\|\boldsymbol{x}^{(k)}\|_1 + \gamma_0\sum_{l=1}^{m}\xi_l\|\boldsymbol{x}^{(l)}\|_2,$$

where $\theta_0^* = \underset{\boldsymbol{x}\in\boldsymbol{B}}{\arg\min}\{\mathbb{G}_\infty\varphi(\boldsymbol{x})\}$ is supposed to be a unique minimum, and $\mathbb{G}_\infty l(.)$ is the limit in probability of $\mathbb{G}_T l(.)$.

To prove this theorem, we remind of Theorem II.1 of Andersen and Gill (1982) which proves that pointwise convergence in probability of random concave functions implies uniform convergence on compact subspaces.

Lemma 2.4.4. (Andersen and Gill, 1982)
Let E be an open convex subset of $\mathbb{R}^p$, and let $F_1, F_2, \ldots,$ be a sequence of random concave functions on E such that $F_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} f(x)$ for every $x \in E$ where f is some real function on E. Then f is also concave, and for all compact $A \subset E$,

$$\sup_{x\in A}|F_n(x) - f(x)| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

The proof of this theorem is based on a diagonal argument and Theorem 10.8 of Rockafeller (1970), that is the pointwise convergence of concave random functions on a dense and countable subset of an open set implies uniform convergence on any compact subset of the open set. Then the following corollary is stated.

Corollary 2.4.5. (Andersen and Gill, 1982)
Assume $F_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} f(x)$, for every $x \in E$, an open convex subset of $\mathbb{R}^p$. Suppose f has a unique maximum at $x_0 \in E$. Let $\hat{X}_n$ maximize F_n. Then $\hat{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} x_0$.

Newey and Powell (1987) use a similar theorem to prove the consistency of asymmetric least squares estimators without any compacity assumption on Θ. We apply these results in our framework, where the parameter set Θ is supposed to be convex.

Proof of Theorem 2.4.3. By definition, $\hat{\theta} = \arg\min_{\theta\in\Theta}\{\mathbb{G}_T\varphi(\theta)\}$. In a first step, we prove the uniform convergence of $\mathbb{G}_T\varphi(.)$ to the limit quantity $\mathbb{G}_\infty\varphi(.)$ on any compact set $\boldsymbol{B} \subset \Theta$, idest

$$\sup_{\boldsymbol{x}\in\boldsymbol{B}}|\mathbb{G}_T\varphi(\boldsymbol{x}) - \mathbb{G}_\infty\varphi(\boldsymbol{x})| \xrightarrow[T\to\infty]{\mathbb{P}} 0. \tag{2.4.1}$$

We define $\mathcal{C} \subset \Theta$ an open convex set and pick $\boldsymbol{x} \in \mathcal{C}$. Then by assumption 23, the law of large number implies

$$\mathbb{G}_T l(\boldsymbol{x}) \xrightarrow[T\to\infty]{\mathbb{P}} \mathbb{G}_\infty l(\boldsymbol{x}).$$

Consequently, if $\lambda_T/T \to \lambda_0 \geq 0$ and $\gamma_T/T \to \gamma_0 \geq 0$, we obtain the pointwise convergence

$$|\mathbb{G}_T\varphi(\boldsymbol{x}) - \mathbb{G}_\infty\varphi(\boldsymbol{x})| \xrightarrow[T\to\infty]{\mathbb{P}} 0.$$

By Lemma 2.4.4 of Andersen and Gill, $\mathbb{G}_\infty\varphi(.)$ is a convex function and we deduce the desired uniform convergence over any compact subset of Θ, that is (2.4.1).

Now we would like that $\arg\min\{\mathbb{G}_T\varphi(.)\} \xrightarrow[T\to\infty]{\mathbb{P}} \arg\min\{\mathbb{G}_\infty\varphi(.)\}$. By assumption 25, $\varphi(.)$ is convex, which implies

$$|\mathbb{G}_T\varphi(\theta)| \xrightarrow[\|\theta\|\to\infty]{\mathbb{P}} \infty.$$

Consequently, $\arg\min\{\mathbb{G}_T\varphi(\boldsymbol{x})\} = O(1)$, such that $\hat{\theta} \in \mathcal{B}_o(\theta_0, C)$ with probability approaching one for C large enough, with $\mathcal{B}_o(\theta_0, C)$ an open ball centered at θ_0 and of radius C. Furthermore, as $\mathbb{G}_\infty\varphi(.)$ is convex, continuous, then $\arg\min_{\boldsymbol{x}\in B}\{\mathbb{G}_\infty\varphi(\boldsymbol{x})\}$ exists and is unique. Then by Corollary 2.4.5 of Andersen and Gill, we obtain

$$\arg\min_{\boldsymbol{x}\in\boldsymbol{B}}\{\mathbb{G}_T\varphi(\boldsymbol{x})\} \xrightarrow[T\to\infty]{\mathbb{P}} \arg\min_{\boldsymbol{x}\in\boldsymbol{B}}\{\mathbb{G}_\infty\varphi(\boldsymbol{x})\},$$

that is $\hat{\theta} \xrightarrow[T\to\infty]{\mathbb{P}} \theta_0^*$. □

Theorem 2.4.6. *Under assumptions 23-25 and 28, the sequence of penalized estimators $\hat{\theta}$ satisfies*

$$\|\hat{\theta} - \theta_0\| = O_p(T^{-1/2} + \lambda_T T^{-1} a + \gamma_T T^{-1} b),$$

when $\lambda_T = o(T)$ and $\gamma_T = o(T)$, and $a := card(\mathcal{A}).\{\max_k \alpha_k\}$, $b := card(\mathcal{A}).\{\max_l \xi_l\}$ satisfy $\lambda_T T^{-1} a_T \to 0$ and $\gamma_T T^{-1} b_T \to 0$.

Remark 2.4.7. This probability bound shows an explicit convergence rate for the SGL estimator. If $\lambda_T T^{-1} = O(T^{-1/2})$ and $\gamma_T T^{-1} = O(T^{-1/2})$, then we would obtain a $\sqrt{T}$-consistent $\hat{\theta}$.

Proof of Theorem 2.4.6. We denote $\nu_T = T^{-1/2} + \lambda_T T^{-1} a + \gamma_T T^{-1} b$, with $a = \text{card}(\mathcal{A}).\{\max_k \alpha_k\}$ and $b = \text{card}(\mathcal{A}).\{\max_l \xi_l\}$. We would like to prove that for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$\mathbb{P}(\frac{1}{\nu_T}\|\hat{\theta} - \theta_0\| > C_\epsilon) < \epsilon.$$

We have

$$\mathbb{P}(\frac{1}{\nu_T}\|\hat{\theta} - \theta_0\| > C_\epsilon) \leq \mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\|_2 \geq C_\epsilon : \mathbb{G}_T\varphi(\theta_0 + \nu_T\boldsymbol{u}) \leq \mathbb{G}_T\varphi(\theta_0)).$$

Furthermore, $\|\boldsymbol{u}\|_2$ can potentially be large as it represents the discrepancy $\hat{\theta} - \theta_0$ normalized by ν_T. Now based on the convexity of the objective function, we have

$$\{\exists \boldsymbol{u}^*, \|\boldsymbol{u}^*\|_2 \geq C_\epsilon, \mathbb{G}_T\varphi(\theta_0 + \nu_T\boldsymbol{u}^*) \leq \mathbb{G}_T\varphi(\theta_0)\} \subset \{\exists \bar{\boldsymbol{u}}, \|\bar{\boldsymbol{u}}\|_2 = C_\epsilon, \mathbb{G}_T\varphi(\theta_0 + \nu_T\bar{\boldsymbol{u}}) \leq \mathbb{G}_T\varphi(\theta_0)\}, \tag{2.4.2}$$

a relationship that allows us to work with a fixed $\|\boldsymbol{u}\|_2$. Let us define $\theta_1 = \theta_0 + \nu_T \boldsymbol{u}^*$ such that $\mathbb{G}_T\varphi(\theta_1) \leq \mathbb{G}_T\varphi(\theta_0)$. Let $\alpha \in (0,1)$ and $\theta = \alpha\theta_1 + (1-\alpha)\theta_0$. Then by convexity of $\mathbb{G}_T\varphi(.)$, we obtain

$$\begin{aligned}\mathbb{G}_T\varphi(\theta) &\leq \alpha\mathbb{G}_T\varphi(\theta_1) + (1-\alpha)\mathbb{G}_T\varphi(\theta_0)\\ &\leq \mathbb{G}_T\varphi(\theta_0).\end{aligned}$$

We pick α such that $\|\bar{\boldsymbol{u}}\| = C_\epsilon$ with $\bar{\boldsymbol{u}} := \alpha\theta_1 + (1-\alpha)\theta_0$. Hence (2.4.2) holds, which implies

$$\begin{aligned}\mathbb{P}(\|\hat{\theta} - \theta_0\| > C_\epsilon \nu_T) &\leq \mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\|_2 \geq C_\epsilon : \mathbb{G}_T\varphi(\theta_0 + \nu_T\boldsymbol{u}) \leq \mathbb{G}_T\varphi(\theta_0))\\ &\leq \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \mathbb{G}_T\varphi(\theta_0 + \nu_T\bar{\boldsymbol{u}}) \leq \mathbb{G}_T\varphi(\theta_0)).\end{aligned}$$

Hence, we pick a $\boldsymbol{u}$ such that $\|\boldsymbol{u}\|_2 = C_\epsilon$. Using $\boldsymbol{p}_1(\lambda_T, \alpha, 0) = 0$ and $\boldsymbol{p}_2(\gamma_T, \xi, 0) = 0$, by a Taylor expansion to $\mathbb{G}_T l(\theta_0 + \nu_T\boldsymbol{u})$, we obtain

$$\begin{aligned}\mathbb{G}_T\varphi(\theta_0 + \nu_T\boldsymbol{u}) - \mathbb{G}_T\varphi(\theta_0) &= \nu_T\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u} + \frac{\nu_T^3}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}\\ &+ \boldsymbol{p}_1(\lambda_T, \alpha, \theta_T) - \boldsymbol{p}_1(\lambda_T, \alpha, \theta_0) + \boldsymbol{p}_2(\gamma_T, \xi, \theta_T) - \boldsymbol{p}_2(\gamma_T, \xi, \theta_0),\end{aligned}$$

where $\bar{\theta}$ is defined as $\|\bar{\theta} - \theta_0\| \leq \|\theta_T - \theta_0\|$. We want to prove

$$\begin{aligned}&\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T}{2}\mathbb{E}[\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}] + \frac{\nu_T}{2}\mathcal{R}_T(\theta_0) + \frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}\\ &+\nu_T^{-1}\{\boldsymbol{p}_1(\lambda_T, \alpha, \theta_T) - \boldsymbol{p}_1(\lambda_T, \alpha, \theta_0) + \boldsymbol{p}_2(\gamma_T, \xi, \theta_T) - \boldsymbol{p}_2(\gamma_T, \xi, \theta_0)\} \leq 0) < \epsilon,\end{aligned} \tag{2.4.3}$$

where $\mathcal{R}_T(\theta_0) = \sum_{k,l=1}^{d} \boldsymbol{u}_k\boldsymbol{u}_l\{\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\theta_0) - \mathbb{E}[\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\theta_0)]\}$. By assumption 23, (ϵ_t) is a non anticipative stationary solution and is ergodic. As a square integrable martingale difference by assumption 26,

$$\sqrt{T}\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{u}'\mathbb{M}\boldsymbol{u}),$$

by the central limit theorem of Billingsley (1961), which implies $\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} = O_p(T^{-1/2}\boldsymbol{u}'\mathbb{M}\boldsymbol{u})$. By the ergodic theorem of Billingsley (1995), we have

$$\ddot{\mathbb{G}}_T l(\theta_0) \xrightarrow[T\to\infty]{\mathbb{P}} \mathbb{H}.$$

This implies $\mathcal{R}_T(\theta_0) = o_p(1)$.

Furthermore, we have by the Markov inequality and for $b > 0$ that

$$\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \sup_{\bar{\theta}:\|\theta-\theta_0\|_2 \leq \nu_T C_\epsilon} |\frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}| > b) \leq \frac{\nu_T^4 C_\epsilon^6}{36b^2}\eta(C_\epsilon),$$

where $\eta(C_\epsilon)$ is defined in assumption 28. We now focus on the penalty terms. As $\boldsymbol{p}_1(\lambda_T, \alpha, 0) = 0$, for the l^1 norm penalty, we have

$$\begin{aligned}
\boldsymbol{p}_1(\lambda_T, \alpha, \theta_T) - \boldsymbol{p}_1(\lambda_T, \alpha, \theta_0) &= \lambda_T T^{-1} \sum_{k \in \mathcal{S}} \alpha_k \{\|\theta_0^{(k)} + \nu_T \boldsymbol{u}^{(k)}\|_1 - \|\theta_0^{(k)}\|_1\}, \\
\text{and } |\boldsymbol{p}_1(\lambda_T, \alpha, \theta_T) - \boldsymbol{p}_1(\lambda_T, \alpha, \theta_0)| &\leq \text{card}(\mathcal{S})\{\max_{k \in \mathcal{S}} \alpha_k\}\lambda_T T^{-1}\nu_T \|\boldsymbol{u}\|_1.
\end{aligned}$$

As for the l^1/l^2 norm, we obtain

$$\begin{aligned}
\boldsymbol{p}_2(\gamma_T, \xi, \theta_T) - \boldsymbol{p}_2(\gamma_T, \xi, \theta_0) &= \gamma_T T^{-1} \sum_{l \in \mathcal{S}} \xi_l \{\|\theta_T^{(l)}\|_2 - \|\theta_0^{(l)}\|_2\}, \\
\text{and } |\boldsymbol{p}_2(\gamma_T, \xi, \theta_T) - \boldsymbol{p}_2(\gamma_T, \xi, \theta_0)| &\leq \gamma_T T^{-1} \sum_{l \in \mathcal{S}} \xi_l \nu_T \|\boldsymbol{u}^{(l)}\|_2 \\
&\leq \text{card}(\mathcal{S})\{\max_{l \in \mathcal{S}} \xi_l\}\gamma_T T^{-1}\nu_T \|\boldsymbol{u}\|_2.
\end{aligned}$$

Then denoting by $\delta_T = \lambda_{\min}(\mathbb{H})C_\epsilon^2\nu_T$, and using $\frac{\nu_T}{2}\mathbb{E}[\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}] \geq \delta_T$, we deduce that (2.4.3) can be bounded as

$$\begin{aligned}
&\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} \\
&+\nu_T^{-1}\{\boldsymbol{p}_1(\lambda_T, \alpha, \theta_T) - \boldsymbol{p}_1(\lambda_T, \alpha, \theta_0) + \boldsymbol{p}_2(\gamma_T, \xi, \theta_T) - \boldsymbol{p}_2(\gamma_T, \xi, \theta_0)\} \leq 0) \\
&\leq \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}| > \delta_T/8) + \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \frac{\nu_T}{2}|\mathcal{R}_T(\theta_0)| > \delta_T/8) \\
&+\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}| > \delta_T/8) \\
&+\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\boldsymbol{p}_1(\lambda_T, \alpha, \theta_T) - \boldsymbol{p}_1(\lambda_T, \alpha, \theta_0)| > \nu_T\delta_T/8) \\
&+\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\boldsymbol{p}_2(\gamma_T, \xi, \theta_T) - \boldsymbol{p}_2(\gamma_T, \xi, \theta_0)| > \nu_T\delta_T/8).
\end{aligned}$$

We also have for C_ϵ and T large enough, and using norm equivalences that

$$\begin{aligned}
&\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\boldsymbol{p}_1(\lambda_T, \alpha, \theta_T) - \boldsymbol{p}_1(\lambda_T, \alpha, \theta_0)| > \nu_T\delta_T/8) \\
&\leq \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \text{card}(\mathcal{S})\{\max_{k \in \mathcal{S}} \alpha_k\}\lambda_T T^{-1}\nu_T \|\boldsymbol{u}\|_1 > \nu_T\delta_T/8) < \epsilon/5, \\
&\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\boldsymbol{p}_2(\gamma_T, \xi, \theta_T) - \boldsymbol{p}_2(\gamma_T, \xi, \theta_0)| > \nu_T\delta_T/8) \\
&\leq \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \text{card}(\mathcal{S})\{\max_{l \in \mathcal{S}} \xi_l\}\gamma_T T^{-1}\nu_T \|\boldsymbol{u}\|_2 > \nu_T\delta_T/8) < \epsilon/5.
\end{aligned}$$

Moreover, if $\nu_T = T^{-1/2} + \lambda_T T^{-1} a + \gamma_T T^{-1} b$, then for C_ϵ large enough

$$\begin{aligned}\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}| > \delta_T/8) &\leq \frac{C_\epsilon^2 C_{st}}{T\delta_T^2} \\ &\leq \frac{C_{st}}{C_\epsilon^4} < \epsilon/5.\end{aligned}$$

Moreover

$$\begin{aligned}\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \sup_{\bar{\theta}:\|\bar{\theta}-\theta_0\|_2 < \nu_T C_\epsilon} |\frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}| > \delta_T/8) &\leq \frac{C_{st}\nu_T^4 \eta(C_\epsilon)}{\delta_T^2} \\ &\leq C_{st}\nu_T^2 C_\epsilon^2 \eta(C_\epsilon)\end{aligned}$$

where $C_{st} > 0$ is a generic constant. Consequently, we obtain, for T and C_ϵ large enough, we obtain

$$\begin{aligned}&\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}| > \delta_T/8) + \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \frac{\nu_T}{2}|\mathcal{R}_T(\theta_0)| > \delta_T/8) \\ &+\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}| > \delta_T/8) \\ &+\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\boldsymbol{p}_1(\lambda_T, \alpha, \theta_0) - \boldsymbol{p}_1(\lambda_T, \alpha, \theta_T)| > \nu_T\delta_T/8) \\ &+\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\boldsymbol{p}_2(\gamma_T, \xi, \theta_0) - \boldsymbol{p}_2(\gamma_T, \xi, \theta_T)| > \nu_T\delta_T/8) + 0 \\ &\leq \frac{C_{st}}{C_\epsilon^4} + \nu_T^2 C_\epsilon^2 \eta(C_\epsilon) C_{st} + 3\epsilon/5 \\ &\leq \epsilon,\end{aligned}$$

for C_ϵ sufficiently large, and T large enough. We then deduce

$$\|\hat{\theta} - \theta_0\| = O_p(\nu_T) = O_p(\lambda_T T^{-1} a + \gamma_T T^{-1} b + T^{-1/2}).$$

□

Remark 2.4.8. We would like to highlight the use of the convexity property of $\mathbb{G}_T\varphi(.)$. It allowed us to obtain the upper bound (2.4.3). Otherwise, the inequality would have been uniform over $\|\boldsymbol{u}\|_2 \geq C_\epsilon$. A consequence is that $\|\boldsymbol{u}\|_2$ can take significantly large values, which would have made the control of the random part in the Taylor expansion hard. This issue is overcome thanks to the convexity that allows for working with fixed $\|\boldsymbol{u}\|_2$, as Fan and Li (2001), Fan and Peng (2004) or Nardi and Rinaldo (2008) do.

We now focus on the distribution of the SGL estimator. Deriving the asymptotic distribution for M-estimators is standard in the case the objective function is differentiable.

It consists of characterizing the estimator by the orthogonality conditions and derive a linear representation by Taylor expansions of the estimator. But these techniques do not apply when the objective function is not differentiable. In our case, $\varphi(.)$ is not differentiable at 0 due to the penalty terms. In some specific context, it may be possible to treat the non-differentiability of $\mathbb{G}_T\varphi(.)$ by applying the expectation operator $\mathbb{E}[.]$ to $\varphi(.)$, which then becomes differentiable in θ_0. Then Taylor expansions are feasible and one obtains the distribution, provided some regularity conditions of the empirical criterion, such as stochastic equi-continuity: see Andrews (1994, a,b). This approach works for specific loss functions, such as the LAD. But in our setting, the expectation operator fails at regularizing $\varphi(.)$ due to the penalty functionals.

Another approach to obtain the asymptotic distribution relies on the convexity property of $\varphi(.)$, and hence of $\mathbb{G}_T\varphi(.)$, without assuming strong regularity conditions on $\varphi(.)$. The intuition behind this rather strong statement is as follows. Let $\mathbb{F}_T(\boldsymbol{u})$ and $\mathbb{F}_\infty(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbb{R}^d$, be random convex functions such that their minimum are respectively $\boldsymbol{u}_T$ and $\boldsymbol{u}_\infty$. Then if $\mathbb{F}_T(.)$ converges in finite distribution to $\mathbb{F}_\infty(.)$, and $\boldsymbol{u}_\infty$ is the unique minimum of $\mathbb{F}_\infty$ with probability one, then $\boldsymbol{u}_T$ converges weakly to $\boldsymbol{u}_\infty$. This method to prove the convergence of arg min processes is called the *convexity argument*. It was developed by Davis, Knight and Liu (1992), Hjort, Pollard (1993), Geyer (1996a, 1996b) or Kato (2009). Chernozhukov and Huong (2004), Chernozhukhov (2005) use this convexity argument to obtain the asymptotic distribution of quantile regression type estimators. The convexity argument only requires the lower-semicontinuity and convexity of the empirical criterion. The convexity Lemma, as in Chernozhukov (2005), can be stated as follows.

Lemma 2.4.9. *(Chernozhukov, 2005)*
Suppose
(i) a sequence of convex lower-semicontinuous $\mathbb{F}_T : \mathbb{R}^d \to \bar{\mathbb{R}}$ *marginally converges to* $\mathbb{F}_\infty : \mathbb{R}^d \to \bar{\mathbb{R}}$ *over a dense subset of* $\mathbb{R}^d$*;*
(ii) $\mathbb{F}_\infty$ *is finite over a nonempty open set* $E \subset \mathbb{R}^d$*;*
(iii) $\mathbb{F}_\infty$ *is uniquely minimized at a random vector* $\boldsymbol{u}_\infty$*.*
Then

$$\arg\min_{\boldsymbol{z}\in\mathbb{R}^d} \mathbb{F}_T(\boldsymbol{z}) \xrightarrow{d} \arg\min_{\boldsymbol{z}\in\mathbb{R}^d} \mathbb{F}_\infty(\boldsymbol{z}), \text{ that is } \boldsymbol{u}_T \xrightarrow{d} \boldsymbol{u}_\infty.$$

Theorem 2.4.10. *Under assumptions 23-28, if* $\lambda_T T^{-1/2} \to \lambda_0$ *and* $\gamma_T T^{-1/2} \to \gamma_0$*, then*

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \arg\min_{\boldsymbol{u}\in\mathbb{R}^d} \{\mathbb{F}_\infty(\boldsymbol{u})\},$$

provided $\mathbb{F}_\infty$ *is the random function in* $\mathbb{R}^d$*, where*

$$\begin{aligned}\mathbb{F}_\infty(\boldsymbol{u}) &= \frac{1}{2}\boldsymbol{u}'\mathbb{H}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{Z} + \lambda_0\sum_{k=1}^{m}\alpha_k\sum_{i=1}^{\boldsymbol{c_k}}\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})}=\mathbf{0}} + \boldsymbol{u}_{\mathbf{i}}^{(\mathbf{k})}sgn(\theta_{\mathbf{0,i}}^{(\mathbf{k})})\mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})}\neq\mathbf{0}}\}\\ &+ \gamma_0\sum_{l=1}^{m}\xi_l\{\|\boldsymbol{u}^{(l)}\|_2\mathbf{1}_{\theta_{\mathbf{0}}^{(\mathbf{l})}=\mathbf{0}} + \frac{\boldsymbol{u}^{(\mathbf{l})'}\theta_{\mathbf{0}}^{(\mathbf{l})}}{\|\theta_{\mathbf{0}}^{(\mathbf{l})}\|_{\mathbf{2}}}\mathbf{1}_{\theta_{\mathbf{0}}^{(\mathbf{l})}\neq\mathbf{0}}\},\end{aligned}$$

with $\mathbb{H} = \mathbb{H}(\theta_0) := \mathbb{E}[\nabla^2_{\theta\theta'}l(\epsilon_t;\theta_0)]$ *and some random vector* $\boldsymbol{Z} \sim \mathcal{N}(0,\mathbb{M})$*,* $\mathbb{M} = \mathbb{M}(\theta_0) := \mathbb{E}[\nabla_\theta l(\epsilon_t;\theta_0)\nabla_{\theta'}l(\epsilon_t;\theta_0)]$.

Proof of Theorem 2.4.10. Let $\boldsymbol{u} \in \mathbb{R}^d$ such that $\theta = \theta_0 + \boldsymbol{u}/T^{1/2}$ and we define the empirical criterion $\mathbb{F}_T(\boldsymbol{u}) = T\mathbb{G}_T(\varphi(\theta_0 + \boldsymbol{u}/T^{1/2}) - \varphi(\theta_0))$. First, we are going to prove the finite distributional convergence of $\mathbb{F}_T$ to $\mathbb{F}_\infty$. Then we use the convexity of $\mathbb{F}_T(.)$ to obtain the convergence in distribution of the arg min empirical criterion to the arg min process limit. To do so, let $\boldsymbol{u} = \sqrt{T}(\theta - \theta_0)$. We have

$$\begin{aligned}\mathbb{F}_T(\boldsymbol{u}) &= T\{\mathbb{G}_T(l(\theta) - l(\theta_0)) + \boldsymbol{p}_1(\lambda_T,\alpha,\theta) - \boldsymbol{p}_1(\lambda_T,\alpha,\theta_0) + \boldsymbol{p}_2(\gamma_T,\xi,\theta) - \boldsymbol{p}_2(\gamma_T,\xi,\theta_0)\}\\ &= T\mathbb{G}_T(l(\theta_0 + \boldsymbol{u}/T^{1/2}) - l(\theta_0)) + \lambda_T\sum_{k=1}^{m}\alpha_k[\|\theta_0^{(k)} + \boldsymbol{u}^{(k)}/\sqrt{T}\|_1 - \|\theta_0^{(k)}\|_1]\\ &+ \gamma_T\sum_{l=1}^{m}\xi_l[\|\theta_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\theta_0^{(l)}\|_2],\end{aligned}$$

where $\mathbb{F}_T(.)$ is convex and $C^0(\mathbb{R}^d)$. We now prove the finite dimensional distribution of $\mathbb{F}_T$ to $\mathbb{F}_\infty$ to apply Lemma 2.4.9. For the l^1 penalty, for any group k, we have for T sufficiently large

$$\|\theta_0^{(k)} + \boldsymbol{u}^{(k)}/\sqrt{T}\|_1 - \|\theta_0^{(k)}\|_1 = T^{-1/2}\sum_{i=1}^{\boldsymbol{c_k}}\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})}=\mathbf{0}} + \boldsymbol{u}_{\mathbf{i}}^{(\mathbf{k})}\mathrm{sgn}(\theta_{\mathbf{0,i}}^{(\mathbf{k})})\mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})}\neq\mathbf{0}}\},$$

which implies that

$$\lambda_T\sum_{k=1}^{m}\alpha_k[\|\theta_0^{(k)}+\boldsymbol{u}^{(k)}/\sqrt{T}\|_1-\|\theta_0^{(k)}\|_1] \underset{T\to\infty}{\longrightarrow} \lambda_0\sum_{k=1}^{m}\alpha_k\sum_{i=1}^{\boldsymbol{c_k}}\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})}=\mathbf{0}}+\boldsymbol{u}_{\mathbf{i}}^{(\mathbf{k})}\mathrm{sgn}(\theta_{\mathbf{0,i}}^{(\mathbf{k})})\mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})}\neq\mathbf{0}}\},$$

under the condition that $\lambda_T/\sqrt{T} \to \lambda_0$.

As for the l^1/l^2 quantity, for any group l, we have

$$\|\theta_0^{(l)} + u^{(l)}/\sqrt{T}\|_2 - \|\theta_0^{(l)}\|_2 = T^{-1/2}\{\|u^{(l)}\|_2\mathbf{1}_{\theta_{\mathbf{0}}^{(\mathbf{l})}=\mathbf{0}} + \frac{\mathbf{u}^{(\mathbf{l})'}\theta_{\mathbf{0}}^{(\mathbf{l})}}{\|\theta_{\mathbf{0}}^{(\mathbf{l})}\|_{\mathbf{2}}}\mathbf{1}_{\theta_{\mathbf{0}}^{(\mathbf{l})}\neq\mathbf{0}}\} + \mathbf{o}(\mathbf{T^{-1}}).$$

Consequently, if $\gamma_T T^{-1/2} \to \gamma_0 \geq 0$, we obtain

$$\gamma_T \sum_{l=1}^{m} \xi_l [\|\theta_0^{(l)} + u^{(l)}/\sqrt{T}\|_2 - \|\theta_0^{(l)}\|_2] = \gamma_0 \sum_{l=1}^{m} \xi_l \{\|u^{(l)}\|_2 \mathbf{1}_{\theta_{\mathbf{0,k}}^{(\mathbf{l})}=\mathbf{0}} + \frac{\mathbf{u}^{(\mathbf{l})\prime}\theta_{\mathbf{0}}^{(\mathbf{l})}}{\|\theta_{\mathbf{0}}^{(\mathbf{l})}\|_{\mathbf{2}}} \mathbf{1}_{\theta_{\mathbf{0}}^{(\mathbf{l})}\neq\mathbf{0}}\} + \mathbf{o}(\mathbf{T^{-1}})\gamma_{\mathbf{T}}.$$

Now for the unpenalized criterion $\mathbb{G}_T l(.)$, by a Taylor expansion, we have

$$T\mathbb{G}_T(l(\theta_0 + \boldsymbol{u}/T^{1/2}) - l(\theta_0)) = \boldsymbol{u}' T^{1/2} \dot{\mathbb{G}}_T l(\theta_0) + \frac{1}{2}\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{1}{6T^{1/3}} \nabla' \{\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u},$$

where $\bar{\theta}$ is defined as $\|\bar{\theta} - \theta_0\| \leq \|\boldsymbol{u}\|/\sqrt{T}$. Then by assumption 26, we have the central limit theorem of Billingsley (1961)

$$\sqrt{T}\dot{\mathbb{G}}_T l(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbb{M}),$$

and by the ergodic theorem

$$\ddot{\mathbb{G}}_T l(\theta_0) \xrightarrow[T\to\infty]{\mathbb{P}} \mathbb{H}.$$

Furthermore, we have by assumption 28

$$\begin{aligned} |\nabla' \{\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}|^2 &\leq \frac{1}{T^2} \sum_{t,t'=1}^{T} \sum_{k_1,l_1,m_1}^{d} \sum_{k_2,l_2,m_2}^{d} \boldsymbol{u}_{k_1}\boldsymbol{u}_{l_1}\boldsymbol{u}_{m_1}\boldsymbol{u}_{k_2}\boldsymbol{u}_{l_2}\boldsymbol{u}_{m_2} |\partial^3_{\theta_{k_1}\theta_{l_1}\theta_{m_1}} l(\epsilon_t;\bar{\theta}).\partial^3_{\theta_{k_2}\theta_{l_2}\theta_{m_2}} l(\epsilon_{t'};\bar{\theta})| \\ &\leq \frac{1}{T^2} \sum_{t,t'=1}^{T} \sum_{k_1,l_1,m_1}^{d} \sum_{k_2,l_2,m_2}^{d} \boldsymbol{u}_{k_1}\boldsymbol{u}_{l_1}\boldsymbol{u}_{m_1}\boldsymbol{u}_{k_2}\boldsymbol{u}_{l_2}\boldsymbol{u}_{m_2} \upsilon_t(C)\upsilon_{t'}(C), \end{aligned}$$

for C large enough, such that $\upsilon_t(C) = \sup_{k,l,m=1,\cdots,d} \{ \sup_{\theta:\|\theta-\theta_0\|_2 \leq \nu_T C} |\partial^3_{\theta_k\theta_l\theta_m} l(\epsilon_t;\theta)|\}$ with $\nu_T = T^{-1/2} + \lambda_T T^{-1} a_T + \gamma_T T^{-1} b_T$. We deduce

$$\nabla' \{\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} = O_p(\|\boldsymbol{u}\|_2^3 \eta(C)).$$

Consequently, we obtain

$$\frac{1}{6T^{1/3}} \nabla' \{\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} \xrightarrow[T\to\infty]{\mathbb{P}} 0.$$

Then we proved that $\mathbb{F}_T(\boldsymbol{u}) \xrightarrow{d} \mathbb{F}_\infty(\boldsymbol{u})$, for a fixed $\boldsymbol{u}$. Let us observe that

$$\boldsymbol{u}_T^* = \arg\min_{\boldsymbol{u}} \{\mathbb{F}_T(\boldsymbol{u})\},$$

and $\mathbb{F}_T(.)$ admits as a minimizer $\boldsymbol{u}_T^* = \sqrt{T}(\hat{\theta} - \theta_0)$. As $\mathbb{F}_T$ is convex and $\mathbb{F}_\infty$ is continuous, convex and has a unique minimum, then by the convexity Lemma 2.4.9, we obtain

$$\sqrt{T}(\hat{\theta} - \theta_0) = \arg\min_{\boldsymbol{u}}\{\mathbb{F}_T\} \xrightarrow{d} \arg\min_{\boldsymbol{u}}\{\mathbb{F}_\infty\}.$$

□

Theorem 2.4.11. *Under assumptions 23-28, if* $\gamma_T T^{-1} \to 0$ *and* $\gamma_T T^{-1/2} \to \infty$ *such that* $\lambda_T \gamma_T^{-1} \to \mu_0$, *with* $\mu_0 \geq 0$, *then*

$$\frac{T}{\gamma_T}(\hat{\theta} - \theta_0) \xrightarrow{d} \arg\min_{\boldsymbol{u}} \{\mathbb{K}_\infty(\boldsymbol{u})\},$$

provided $\mathbb{K}_\infty$ *is a uniquely defined deterministic function in* $\mathbb{R}^d$, *where*

$$\begin{aligned}\mathbb{K}_\infty(\boldsymbol{u}) &= \frac{1}{2}\boldsymbol{u}'\mathbb{H}\boldsymbol{u} + \mu_0 \sum_{k=1}^{m} \alpha_k \{\|\boldsymbol{u}^{(k)}\|_1 \mathbf{1}_{\theta_0^{(k)}=0} + \boldsymbol{u}^{(k)'} sgn(\theta_0^{(k)}) \mathbf{1}_{\theta_0^{(k)}\neq 0}\} \\ &+ \sum_{l=1}^{m} \xi_l \{\|\boldsymbol{u}^{(l)}\|_2 \mathbf{1}_{\theta_0^{(l)}=0} + \frac{\boldsymbol{u}^{(l)'}\theta_0^{(l)}}{\|\theta_0^{(l)}\|_2} \mathbf{1}_{\theta_0^{(l)}\neq 0}\}.\end{aligned}$$

The limit quantity $\mathbb{K}_\infty(.)$ *is non-random, which implies that the convergence in distribution implies the convergence in probability* $\frac{T}{\gamma_T}(\hat{\theta} - \theta_0) \xrightarrow{\mathbb{P}} \arg\min_{\boldsymbol{u}} \{\mathbb{K}_\infty(\boldsymbol{u})\}$ *by Shiryaev (ex 7, p 259, 1995).*

Remark 2.4.12. The convergence rate of $\hat{\theta}$ is slower than $\sqrt{T}$ and the limit distribution is not random. To obtain an optimal convergence rate, we should take $\lambda_T = O(T^{1/2})$, $\gamma_T = O(T^{1/2})$.

Proof of Theorem 2.4.11. To prove this convergence result, we proceed as in Theorem 2.4.10. To do so, we define $\theta = \theta_0 + \boldsymbol{u}\gamma_T/T$ and we prove that $\tilde{\mathbb{F}}_T(\boldsymbol{u}) = \mathbb{G}_T(\varphi(\theta_0 + \boldsymbol{u}\gamma_T/T) - \varphi(\theta_0))$ converges in finite distribution to $\mathbb{K}_\infty(.)$. We have

$$\begin{aligned}\tilde{\mathbb{F}}_T(\boldsymbol{u}) &= T\{\mathbb{G}_T(l(\theta) - l(\theta_0)) + \boldsymbol{p}_1(\lambda_T, \alpha, \theta) - \boldsymbol{p}_1(\lambda_T, \alpha, \theta_0) + \boldsymbol{p}_2(\gamma_T, \xi, \theta) - \boldsymbol{p}_2(\gamma_T, \xi, \theta_0)\} \\ &= T\mathbb{G}_T(l(\theta_0 + \boldsymbol{u}\gamma_T/T) - l(\theta_0)) + \lambda_T \sum_{k=1}^{m} \alpha_k [\|\theta_0^{(k)} + \boldsymbol{u}^{(k)}\gamma_T/T\|_1 - \|\theta_0^{(k)}\|_1] \\ &+ \gamma_T \sum_{l=1}^{m} \xi_l [\|\theta_0^{(l)} + \boldsymbol{u}^{(l)}\gamma_T/T\|_2 - \|\theta_0^{(l)}\|_2].\end{aligned}$$

For the unpenalized empirical criterion, we have the expansion

$$T\mathbb{G}_T(l(\theta_0+\boldsymbol{u}\gamma_T/T)-l(\theta_0))=\gamma_T\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}+\frac{\gamma_T^2}{2T}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}+\frac{\gamma_T^3}{6T^2}\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u},$$

where $\bar{\theta}$ lies between θ_0 and $\theta_0+\boldsymbol{u}\gamma_T/T$. This implies $\tilde{\mathbb{F}}_T(\boldsymbol{u})=\frac{\gamma_T^2}{T}\mathbb{K}_T(\boldsymbol{u})$, where

$$\begin{aligned}\mathbb{K}_T(\boldsymbol{u}) &= \tfrac{\sqrt{T}}{\gamma_T}(\sqrt{T}\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u})+\tfrac{1}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}+\tfrac{\gamma_T}{6T}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}\\ &+ \tfrac{T}{\gamma_T^2}\lambda_T\sum_{k=1}^m\alpha_k[\|\theta_0^{(k)}+\boldsymbol{u}^{(k)}\gamma_T/T\|_1-\|\theta_0^{(k)}\|_1]+\tfrac{T}{\gamma_T}\sum_{l=1}^m\xi_l[\|\theta_0^{(l)}+u^{(l)}\gamma_T/T\|_2-\|\theta_0^{(l)}\|_2].\end{aligned}$$

We first focus on the penalty terms. For the l^1 part, for any group k, we have

$$\|\theta_0^{(k)}+\boldsymbol{u}^{(k)}\gamma_T/T\|_1-\|\theta_0^{(k)}\|_1=\gamma_T T^{-1}\sum_{i=1}^{\boldsymbol{c_k}}\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\boldsymbol{\theta_{0,i}^{(k)}=0}}+\boldsymbol{u_i^{(k)}}\mathrm{sgn}(\boldsymbol{\theta_{0,i}^{(k)}})\mathbf{1}_{\boldsymbol{\theta_{0,i}^{(k)}\neq 0}}\}.$$

We deduce that

$$\frac{T}{\gamma_T^2}\lambda_T\alpha_k[\|\theta_0^{(k)}+\boldsymbol{u}^{(k)}\gamma_T/T\|_1-\|\theta_0^{(k)}\|_1]\to\mu_0\sum_{i=1}^{\boldsymbol{c_k}}\alpha_k\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\boldsymbol{\theta_{0,i}^{(k)}=0}}+\boldsymbol{u_i^{(k)}}\mathrm{sgn}(\boldsymbol{\theta_{0,i}^{(k)}})\mathbf{1}_{\boldsymbol{\theta_{0,i}^{(k)}\neq 0}}\},$$

under the condition $\lambda_T\gamma_T^{-1}\to\mu_0$.

As for the l^1/l^2 quantity, for any group l, we have

$$\|\theta_0^{(l)}+u^{(l)}\gamma_T/T\|_2-\|\theta_0^{(l)}\|_2=\gamma_T T^{-1}\{\|u^{(l)}\|_2\mathbf{1}_{\boldsymbol{\theta_0^{(l)}=0}}+\frac{\mathbf{u^{(l)'}\theta_0^{(l)}}}{\boldsymbol{\|\theta_0^{(l)}\|_2}}\mathbf{1}_{\boldsymbol{\theta_0^{(l)}\neq 0}}\}+\mathbf{o(T^{-1})}.$$

Consequently, we obtain

$$\frac{T}{\gamma_T}\xi_l[\|\theta_0^{(l)}+u^{(l)}\gamma_T/T\|_2-\|\theta_0^{(l)}\|_2]\to\xi_l\{\|u^{(l)}\|_2\mathbf{1}_{\boldsymbol{\theta_0^{(l)}=0}}+\frac{\mathbf{u^{(l)'}\theta_0^{(l)}}}{\boldsymbol{\|\theta_0^{(l)}\|_2}}\mathbf{1}_{\boldsymbol{\theta_0^{(l)}\neq 0}}\}.$$

Now for the unpenalized part, by the central limit theorem of Billingsley (1961), $\sqrt{T}\dot{\mathbb{G}}_T l(\theta_0)$ is asymptotically normal, then $\gamma_T T^{-1/2}\to\infty$ implies by the Slutsky theorem

$$\frac{\sqrt{T}}{\gamma_T}(\sqrt{T}\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u})\xrightarrow[T\to\infty]{\mathbb{P}}0.$$

Furthermore, by the ergodic theorem of Billingsley (1961), we have

$$\ddot{\mathbb{G}}_T l(\theta_0)\xrightarrow[T\to\infty]{\mathbb{P}}\mathbb{H}.$$

As for the third order term, by assumption 28 and using the same reasoning as the proof of Theorem 2.4.10, we have

$$\frac{\gamma_T}{6T}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} \xrightarrow[T\to\infty]{\mathbb{P}} 0,$$

using $\gamma_T = o(T)$. Then we proved that $\mathbb{K}_T(\boldsymbol{u}) \xrightarrow{d} \mathbb{K}_\infty(\boldsymbol{u})$, for a fixed $\boldsymbol{u} \in \mathbb{R}^d$. We have

$$\boldsymbol{u}_T^* = \arg\min_{\boldsymbol{u}} \{\mathbb{K}_T(\boldsymbol{u})\},$$

and $\mathbb{K}_T(.)$ admits as a minimizer $\boldsymbol{u}_T^* = \frac{T}{\gamma_T}(\hat{\theta} - \theta_0)$. $\mathbb{K}_T(.)$ is convex and $\mathbb{K}_\infty(.)$ is continuous, then by the convexity Lemma, we deduce

$$\frac{T}{\gamma_T}(\hat{\theta} - \theta_0) = \arg\min\{\mathbb{K}_T\} \xrightarrow{d} \arg\min\{\mathbb{K}_\infty\}.$$

□

We now turn to the oracle property of the SGL. Model selection consistency consists of evaluating the probability that $\{\hat{\mathcal{A}} = \mathcal{A}\}$, for T large enough. That means we check that the regularization asymptotically allows for identifying the right model.

Proposition 2.4.13. *Under assumption 23-28, if* $\lambda_T T^{-1/2} \to \lambda_0$ *and* $\gamma_T T^{-1/2} \to \gamma_0$, *then*

$$\limsup_{T\to\infty} \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \leq c < 1,$$

where c is a constant depending on the true model.

Proof of Proposition 2.4.13. In Theorem 2.4.10, we proved

$$\sqrt{T}(\hat{\theta} - \theta_0) := \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_T\} \xrightarrow{d} \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_\infty\},$$

under the assumption $\lambda_T/\sqrt{T} \to \lambda_0$ and $\gamma_T/\sqrt{T} \to \gamma_0$. The limit random function is

$$\begin{aligned}\mathbb{F}_\infty(\boldsymbol{u}) &= \frac{1}{2}\boldsymbol{u}'\mathbb{H}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{Z} + \lambda_0\sum_{k=1}^{m}\alpha_k\sum_{i=1}^{\boldsymbol{c_k}}\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})}=\mathbf{0}} + \boldsymbol{u}_{\mathbf{i}}^{(\mathbf{k})}\mathrm{sgn}(\theta_{\mathbf{0,i}}^{(\mathbf{k})})\mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})}\neq\mathbf{0}}\} \\ &+ \gamma_0\sum_{l=1}^{m}\xi_l\{\|\boldsymbol{u}^{(l)}\|_2\mathbf{1}_{\theta_{\mathbf{0}}^{(\mathbf{l})}=\mathbf{0}} + \frac{\boldsymbol{u}^{(\mathbf{l})'}\theta_{\mathbf{0}}^{(\mathbf{l})}}{\|\theta_{\mathbf{0}}^{(\mathbf{l})}\|_{\mathbf{2}}}\mathbf{1}_{\theta_{\mathbf{0}}^{(\mathbf{l})}\neq\mathbf{0}}\}.\end{aligned}$$

First, let us observe that

$$\{\hat{\mathcal{A}} = \mathcal{A}\} = \{\forall k = 1, \cdots, m,\ i \in \mathcal{A}_k^c, \hat{\theta}_i^{(k)} = 0\} \cap \{\forall k = 1, \cdots, m,\ i \in \hat{\mathcal{A}}_k^c, \theta_{0,i}^{(k)} = 0\}.$$

Both sets describing $\{\hat{\mathcal{A}} = \mathcal{A}\}$ are symmetric, and thus we can focus on

$$\{\hat{\mathcal{A}} = \mathcal{A}\} \Rightarrow \{\forall k = 1, \cdots, m,\ i \in \mathcal{A}_k^c, T^{1/2}\hat{\theta}_i^{(k)} = 0\}.$$

Hence

$$\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \leq \mathbb{P}(\forall k = 1, \cdots, m, \forall i \in \mathcal{A}_k^c, T^{1/2}\hat{\theta}_i^{(k)} = 0).$$

Denoting by $\boldsymbol{u}^* := \underset{\boldsymbol{u} \in \mathbb{R}^d}{\arg\min}\{\mathbb{F}_\infty(\boldsymbol{u})\}$, Theorem 2.4.10 corresponds to $\sqrt{T}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) \xrightarrow{d} \boldsymbol{u}_{\mathcal{A}}^*$. By the Portmanteau Theorem (see Wellner and van der Vaart, 1996), we have

$$\limsup_{T \to \infty} \mathbb{P}(\forall k = 1, \cdots, m, \forall i \in \mathcal{A}_k^c, T^{1/2}\hat{\theta}_i^{(k)} = 0) \leq \mathbb{P}(\forall k = 1, \cdots, m, \forall i \in \mathcal{A}_k^c, \boldsymbol{u}_i^{(k)*} = 0),$$

as $\theta_{0,\mathcal{A}^c} = \mathbf{0}$. Consequently, we need to prove that the probability of the right hand side is strictly inferior to 1, which is upper-bounded by

$$\begin{aligned} &\mathbb{P}(\forall k = 1, \cdots, m, \forall i \in \mathcal{A}_k^c, \boldsymbol{u}_i^{(k)*} = 0) \leq \\ &\min(\mathbb{P}(k \notin \mathcal{S}, \boldsymbol{u}^{(k)*} = 0), \mathbb{P}(k \in \mathcal{S}, \forall i \in \mathcal{A}_k^c, \boldsymbol{u}_i^{(k)*} = 0)). \end{aligned} \tag{2.4.4}$$

If $\lambda_0 = \gamma_0 = 0$, then $\boldsymbol{u}^* = -\mathbb{H}^{-1}\boldsymbol{Z}$, such that $\mathbb{P}_{\boldsymbol{u}^*} = \mathcal{N}(0, \mathbb{H}^{-1}\mathbb{M}\mathbb{H}^{-1})$. Hence , $c = 0$.

If $\lambda_0 \neq 0$ or $\gamma_0 \neq 0$, the necessary and sufficient optimality conditions for a group k tell us that $\boldsymbol{u}^*$ satisfies

$$\begin{cases} (\mathbb{H}\boldsymbol{u}^* + \boldsymbol{Z})_{(k)} + \lambda_0 \alpha_k \boldsymbol{p}^{(k)} + \gamma_0 \xi_k \dfrac{\theta_0^{(k)}}{\|\theta_0^{(k)}\|_2} = 0, & k \in \mathcal{S}, \\ (\mathbb{H}\boldsymbol{u}^* + \boldsymbol{Z})_{(k)} + \lambda_0 \alpha_k \boldsymbol{w}^{(k)} + \gamma_0 \xi_k \boldsymbol{z}^{(k)} = 0, & \text{otherwise}, \end{cases} \tag{2.4.5}$$

where $\boldsymbol{w}^{(k)}$ and $\boldsymbol{z}^{(k)}$ are the subgradients of $\|\boldsymbol{u}^{(k)}\|_1$ and $\|\boldsymbol{u}^{(k)}\|_2$ given by

$$\boldsymbol{w}_i^{(k)} = \begin{cases} \operatorname{sgn}(\boldsymbol{u}_i^{(k)}) \text{ if } \boldsymbol{u}_i^{(k)} \neq 0, \\ \in \{\boldsymbol{w}_i^{(k)} : |\boldsymbol{w}_i^{(k)}| \leq 1\} \text{ if } \boldsymbol{u}_i^{(k)} = 0, \end{cases} \qquad \boldsymbol{z}^{(k)} = \begin{cases} \dfrac{\boldsymbol{u}^{(k)}}{\|\boldsymbol{u}^{(k)}\|_2} \text{ if } \boldsymbol{u}^{(k)} \neq 0, \\ \in \{\boldsymbol{z}^{(k)} : \|\boldsymbol{z}^{(k)}\|_2 \leq 1\} \text{ if } \boldsymbol{u}^{(k)} = 0, \end{cases}$$

and $\boldsymbol{p}_i^{(k)} = \partial_{\boldsymbol{u}_i}\{|\boldsymbol{u}_i^{(k)}| \mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})} = \mathbf{0}} + \boldsymbol{u}_{\mathbf{i}}^{(\mathbf{k})} \operatorname{sgn}(\theta_{\mathbf{0,i}}^{(\mathbf{k})}) \mathbf{1}_{\theta_{\mathbf{0,i}}^{(\mathbf{k})} \neq \mathbf{0}}\}$.

If $\boldsymbol{u}^{(m)*} = 0, \forall m \notin \mathcal{S}$, then the optimality conditions (2.4.5) become

$$\begin{cases} \mathbb{H}_{\mathcal{S}\mathcal{S}}\boldsymbol{u}^*_{\mathcal{S}} + \boldsymbol{Z}_{\mathcal{S}} + \lambda_0\tau_{\mathcal{S}} + \gamma_0\zeta_{\mathcal{S}} = 0, \\ \| - \mathbb{H}_{(l)\mathcal{S}}\boldsymbol{u}^*_{\mathcal{S}} - \boldsymbol{Z}_{(l)} - \lambda_0\alpha_l\boldsymbol{w}^{(l)}\|_2 \leq \gamma_0\xi_l, \text{ as } \|\boldsymbol{z}^{(l)}\|_2 \leq 1,\ l \in \mathcal{S}^c, \end{cases} \tag{2.4.6}$$

with $\tau_{\mathcal{S}} = \text{vec}(k \in \mathcal{S}, \alpha_k\boldsymbol{p}^{(k)})$ and $\zeta_{\mathcal{S}} = \text{vec}(k \in \mathcal{S}, \xi_k\dfrac{\theta_0^{(k)}}{\|\theta_0^{(k)}\|_2})$, which are vectors of $\mathbb{R}^{\text{card}(\mathcal{S})}$.

For $k \in \mathcal{S}$, that is the vector $\theta_0^{(k)}$ is at least non-zero, then

$$\begin{cases} (\mathbb{H}\boldsymbol{u}^* + \boldsymbol{Z})_i + \lambda_0\alpha_k\text{sgn}(\theta_{0,i}^{(k)}) + \gamma_0\xi_k\dfrac{\theta_{0,i}^{(k)}}{\|\theta_0^{(k)}\|_2} = 0, \text{ if } k \in \mathcal{S}, i \in \mathcal{A}_k, \\ (\mathbb{H}\boldsymbol{u}^* + \boldsymbol{Z})_i + \lambda_0\alpha_k\boldsymbol{w}_i^{(k)} = 0,\ i \in \mathcal{A}_k^c. \end{cases} \tag{2.4.7}$$

Consequently, if $\boldsymbol{u}_i^{(k)*} = 0, \forall i \in \mathcal{A}_k^c$, with $k \in \mathcal{S}$, then the conditions (2.4.7) become

$$\begin{cases} \mathbb{H}_{\mathcal{A}_k\mathcal{A}_k}\boldsymbol{u}^*_{\mathcal{A}_k} + \boldsymbol{Z}_{\mathcal{A}_k} + \lambda_0\alpha_k\text{sgn}(\theta_{0,\mathcal{A}_k}) + \gamma_0\xi_k\dfrac{\theta_{0,\mathcal{A}_k}}{\|\theta_{0,\mathcal{A}_k}\|_2} = 0, \\ |-(\mathbb{H}_{\mathcal{A}_k^c\mathcal{A}_k}\boldsymbol{u}^*_{\mathcal{A}_k} + \boldsymbol{Z}_{\mathcal{A}_k^c})_i| \leq \lambda_0\alpha_k. \end{cases}$$

Combining relationships in (2.4.6), we obtain

$$\|\mathbb{H}_{(l)\mathcal{S}}\mathbb{H}_{\mathcal{S}\mathcal{S}}^{-1}(\boldsymbol{Z}_{\mathcal{S}} + \lambda_0\tau_{\mathcal{S}} + \gamma_0\zeta_{\mathcal{S}}) - \boldsymbol{Z}_{(l)} - \lambda_0\alpha_l\boldsymbol{w}^{(l)}\|_2 \leq \gamma_0\xi_l, l \in \mathcal{S}^c.$$

The same reasoning applies for active groups with inactive components, such that combining relationships in (2.4.7), we obtain

$$|(\mathbb{H}_{\mathcal{A}_k^c\mathcal{A}_k}\mathbb{H}_{\mathcal{A}_k\mathcal{A}_k}^{-1}(\boldsymbol{Z}_{\mathcal{A}_k} + \lambda_0\alpha_k\text{sgn}(\theta_{0,\mathcal{A}_k}) + \gamma_0\xi_k\frac{\theta_{0,\mathcal{A}_k}}{\|\theta_{0,\mathcal{A}_k}\|_2}) - \boldsymbol{Z}_{\mathcal{A}_k^c})_i| \leq \lambda_0\alpha_k.$$

Hence we deduce

$$\begin{aligned} &\mathbb{P}(\forall k = 1, \cdots, m, \forall \in \mathcal{A}_k^c, \boldsymbol{u}_i^{(k)*} = 0) \leq \\ &\min(\mathbb{P}(k \notin \mathcal{S}, \boldsymbol{u}^{(k)*} = 0), \mathbb{P}(k \in \mathcal{S}, \forall i \in \mathcal{A}_k^c, \boldsymbol{u}_i^{(k)*} = 0)) := \min(a_1, a_2). \end{aligned}$$

Under the assumption that $\lambda_0 < \infty$ and $\gamma_0 < \infty$, we obtain

$$a_1 = \mathbb{P}(l \in \mathcal{S}^c, \|\mathbb{H}_{(l)\mathcal{S}}\mathbb{H}_{\mathcal{S}\mathcal{S}}^{-1}(\boldsymbol{Z}_{\mathcal{S}} + \lambda_0\tau_{\mathcal{S}} + \gamma_0\zeta_{\mathcal{S}}) - \boldsymbol{Z}_{(l)} - \lambda_0\alpha_l\boldsymbol{w}^{(l)}\|_2 \leq \gamma_0\xi_l) < 1,$$

$$a_2 = \mathbb{P}(k \in \mathcal{S}, i \in \mathcal{A}_k^c, |(\mathbb{H}_{\mathcal{A}_k^c\mathcal{A}_k}\mathbb{H}_{\mathcal{A}_k\mathcal{A}_k}^{-1}(\boldsymbol{Z}_{\mathcal{A}_k} + \lambda_0\alpha_k\text{sgn}(\theta_{0,\mathcal{A}_k}) + \gamma_0\xi_k\frac{\theta_{0,\mathcal{A}_k}}{\|\theta_{0,\mathcal{A}_k}\|_2}) - \boldsymbol{Z}_{\mathcal{A}_k^c})_i| \leq \lambda_0\alpha_k) < 1.$$

Thus $c < 1$, which proves (2.4.4), that is proposition 2.4.13. □

Remark 2.4.14. The result in Proposition 2.4.13 highlights that the SGL as proposed by Simon and al. (2013) cannot satisfy the oracle property since the penalties cannot recover the unknown set of active indices $\mathcal{A}$, which is called model selection consistency. To fix this drawback in an ordinary least square framework, Zou (2006) proposed the adaptive Lasso, where random weights are used to penalize different coefficients and proves that the adaptive Lasso estimator satisfies the oracle property in the sense of Fan and Li (2001), that is asymptotic normality and selection consistency for a proper choice of λ_T and $\alpha_i^{(k)}$. That is also the case for the adaptive Group Lasso model proposed by Nardi and Rinaldo (2008), where adaptive weights are used to penalize grouped coefficients differently. We propose the same approach than Zou (2006) and use adaptive weights in the penalties such that the adaptive SGL satisfies the oracle property in the sense of Fan and Li (2001) as proved in Theorem 2.4.16.

The adaptive specification of the proposed estimator now becomes

$$\hat{\theta} = \arg\min_{\theta\in\Theta}\{\mathbb{G}_T\psi(\theta)\}, \tag{2.4.8}$$

where

$$\begin{aligned}\theta \mapsto \mathbb{G}_T\psi(\theta) &= \frac{1}{T}\sum_{t=1}^{T} l(\epsilon_t;\theta) + \boldsymbol{p}_1(\lambda_T,\tilde{\theta},\theta) + \boldsymbol{p}_2(\gamma_T,\tilde{\theta},\theta)\\ &= \mathbb{G}_T l(\theta) + \boldsymbol{p}_1(\lambda_T,\tilde{\theta},\theta) + \boldsymbol{p}_2(\gamma_T,\tilde{\theta},\theta),\end{aligned}$$

such that both penalties are specified as

$$\boldsymbol{p}_1(\lambda_T,\tilde{\theta},\theta) = \lambda_T T^{-1}\sum_{k=1}^{m}\sum_{i=1}^{\boldsymbol{c}_k}\alpha(\tilde{\theta}_i^{(k)})|\theta_i^{(k)}|, \boldsymbol{p}_2(\gamma_T,\tilde{\theta},\theta) = \gamma_T T^{-1}\sum_{l=1}^{m}\xi(\tilde{\theta}^{(l)})\|\theta^{(l)}\|_2.$$

These penalties are now randomized through the $\tilde{\theta}$ argument in the weights α's and ξ's. This first step estimator $\tilde{\theta}$ is supposed to be a $T^{1/2}$-consistent estimator of θ_0. For instance, it can be defined as an M-estimator of the unpenalized empirical criterion $\mathbb{G}_T l(.)$, that is

$$\tilde{\theta} = \arg\min_{\theta\in\Theta}\mathbb{G}_T l(\theta).$$

Adaptive weights are also used by Zou and Zhang (2009), who plug the elastic-net estimator in the adaptive weight and then estimate a new elastic net model using these weights, that is the adaptive elastic net.

The weights we use are now random and for any group k or l, $\alpha(\tilde{\theta}^{(k)}) \in \mathbb{R}_+^{\boldsymbol{c}_k}$, $\xi(\tilde{\theta}^{(l)}) \in \mathbb{R}_+$ are specified as

$$\alpha_T^{(k)} := \alpha(\tilde{\theta}^{(k)}) = (|\tilde{\theta}_i^{(k)}|^{-\eta}, i = 1, \cdots, \boldsymbol{c}_k),\ \xi_{T,l} := \xi(\tilde{\theta}^{(l)}) = \|\tilde{\theta}^{(l)}\|_2^{-\mu},$$

for some constants $\eta > 0$ and $\mu > 0$ (to be specified).

Theorem 2.4.15. *Under assumptions 23-25 and 28, the sequence of penalized estimators $\hat{\theta}$ satisfies*

$$\|\hat{\theta} - \theta_0\| = O_p(T^{-1/2} + \lambda_T T^{-1} a_T + \gamma_T T^{-1} b_T),$$

with $a_T = card(\mathcal{A}).\{\max_{k\in\mathcal{S}}(\max_{i\in\mathcal{A}_k}\alpha_{T,i}^{(k)})\}$, $b_T = card(\mathcal{A}).\{\max_{l\in\mathcal{S}}\xi_{T,l}\}$ stochastic quantities, such that $\lambda_T T^{-1} a_T \xrightarrow{\mathbb{P}} 0$ and $\gamma_T T^{-1} b_T \xrightarrow{\mathbb{P}} 0$.

Proof of Theorem 2.4.15. The proof follows exactly the same steps as for Theorem (2.4.6), except a_T and b_T are random quantities. □

Theorem 2.4.16. *Under assumptions 23-28, if $\lambda_T T^{-1/2} \to 0$, $\gamma_T T^{-1/2} \to 0$, $T^{(\eta-1)/2}\lambda_T \to \infty$, $T^{(\mu-1)/2}\gamma_T \to \infty$ and $T^{(\mu-\eta)/2}\gamma_T\lambda_T^{-1} \to \infty$, then $\hat{\theta}$ obtained in (2.4.8) satisfies*

$$\begin{aligned} \lim_{T\to\infty} \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) &= 1, \text{ and} \\ \sqrt{T}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) &\xrightarrow{d} \mathcal{N}(0, \mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1}\mathbb{M}_{\mathcal{A}\mathcal{A}}\mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1}). \end{aligned}$$

Proof of Theorem 2.4.16. We start with the asymptotic distribution and proceed as in the proof of Theorem 2.4.10, where we used Lemma 2.4.9. To do so, we prove the finite dimensional convergence in distribution of the empirical criterion $\mathbb{F}_T(\boldsymbol{u})$ to $\mathbb{F}_\infty(\boldsymbol{u})$ with $\boldsymbol{u} \in \mathbb{R}^d$, where these quantities are respectively defined as

$$\begin{aligned} \mathbb{F}_T(\boldsymbol{u}) &= T\mathbb{G}_T(\psi(\theta_0 + \boldsymbol{u}/\sqrt{T}) - \psi(\theta_0)) \\ &= T\mathbb{G}_T(l(\theta_0 + \boldsymbol{u}/\sqrt{T}) - l(\theta_0)) + \lambda_T \sum_{k=1}^{m}\sum_{i=1}^{\boldsymbol{c}_k} \alpha_{T,i}^{(k)}[|\theta_{0,i}^{(k)} + \boldsymbol{u}_i^{(k)}/\sqrt{T}| - |\theta_{0,i}^{(k)}|] \\ &+ \gamma_T \sum_{l=1}^{m} \xi_{T,l}[\|\theta_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\theta_0^{(l)}\|_2], \end{aligned}$$

and

$$\mathbb{F}_\infty(\boldsymbol{u}) = \begin{cases} \frac{1}{2}\boldsymbol{u}'_{\mathcal{A}}\mathbb{H}_{\mathcal{A}\mathcal{A}}\boldsymbol{u}_{\mathcal{A}} + \boldsymbol{u}'_{\mathcal{A}}\boldsymbol{Z}_{\mathcal{A}} & \text{if } \boldsymbol{u}_i = 0, \text{ when } i \notin \mathcal{A}, \text{ and} \\ \infty & \text{otherwise,} \end{cases} \tag{2.4.9}$$

with $\boldsymbol{Z}_{\mathcal{A}} \sim \mathcal{N}(0, \mathbb{M}_{\mathcal{A}\mathcal{A}})$. By Lemma 2.4.9, the finite dimensional convergence in distribution implies $\arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_T(\boldsymbol{u})\} \xrightarrow{d} \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_\infty(\boldsymbol{u})\}$. We first consider the unpenalized empirical criterion of $\mathbb{F}_T(.)$, which can be expanded as

$$T\mathbb{G}_T(\psi(\theta_0 + \boldsymbol{u}/\sqrt{T}) - \psi(\theta_0)) = T^{1/2}\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u} + \frac{1}{6T^{1/3}}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\}\boldsymbol{u},$$

where $\bar{\theta}$ lies between θ_0 and $\theta_0 + \boldsymbol{u}/\sqrt{T}$. First, using the same reasoning on the third order term, we obtain

$$\frac{1}{6T^{1/3}}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\}\boldsymbol{u} \xrightarrow[T\to\infty]{\mathbb{P}} 0.$$

By the ergodic theorem, we deduce $\ddot{\mathbb{G}}_T l(\theta_0) \xrightarrow[T\to\infty]{\mathbb{P}} \mathbb{H}$ and by assumption 26, $\sqrt{T}\dot{\mathbb{G}}_T l(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbb{M})$.

We now focus on the penalty terms of (2.4.8), we remind that $\alpha_{T,i}^{(k)} = |\tilde{\theta}_i^{(k)}|^{-\eta}$, such that for $i \in \mathcal{A}_k, k \in \mathcal{S}$, $\tilde{\theta}_i^{(k)} \xrightarrow[T\to\infty]{\mathbb{P}} \theta_{0,i}^{(k)} \neq 0$. Note that

$$\sqrt{T}(|\theta_0^{(k)} + \boldsymbol{u}^{(k)}/\sqrt{T}| - |\theta_0^{(k)}|] \xrightarrow[T\to\infty]{\mathbb{P}} \boldsymbol{u}_i^{(k)}\text{sgn}(\theta_{0,i}^{(k)})\mathbf{1}_{\boldsymbol{\theta}_{\mathbf{0,i}}^{(\mathbf{k})}\neq\mathbf{0}}.$$

This implies that, for $i \in \mathcal{A}_k$, $k \in \mathcal{S}$, we have

$$\lambda_T T^{-1/2}\sum_{i=1}^{\boldsymbol{c}_k}\alpha_{T,i}^{(k)}\sqrt{T}(|\theta_{0,i}^{(k)} + \boldsymbol{u}_i^{(k)}/\sqrt{T}| - |\theta_{0,i}^{(k)}|) \xrightarrow[T\to\infty]{\mathbb{P}} 0,$$

under the condition $\lambda_T T^{-1/2} \to 0$. For $i \in \mathcal{A}_k^c$, $\theta_{0,i}^{(k)} = 0$, then $T^{\eta/2}(|\tilde{\theta}_i^{(k)}|)^\eta = O_p(1)$. Hence under the assumption $\lambda_T T^{(\eta-1)/2} \to \infty$, we obtain

$$\lambda_T T^{-1/2}\alpha_{T,i}^{(k)}\sqrt{T}(|\theta_{0,i}^{(k)} + \boldsymbol{u}_i^{(k)}/\sqrt{T}| - |\theta_{0,i}^{(k)}|) = \lambda_T T^{-1/2}|\boldsymbol{u}_i^{(k)}|\frac{T^{\eta/2}}{(T^{1/2}|\tilde{\theta}_i^{(k)}|)^\eta} \xrightarrow[T\to\infty]{\mathbb{P}} \infty. \tag{2.4.10}$$

As for the l^1/l^2 quantity, we remind that $\xi_{T,l} = \|\tilde{\theta}^{(l)}\|_2^{-\mu}$, such that for $l \in \mathcal{S}$, $\tilde{\theta}^{(l)} \xrightarrow[T\to\infty]{\mathbb{P}} \theta_0^{(l)}$, and in this case

$$\sqrt{T}\{\|\theta_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\theta_0^{(l)}\|_2\} = \frac{\boldsymbol{u}^{(l)'}\theta_0^{(l)}}{\|\theta_0^{(l)}\|_2} + o(T^{-1/2}).$$

Consequently, using $\gamma_T T^{-1/2} \to 0$, and for $l \in \mathcal{S}$, we obtain

$$\gamma_T T^{-1/2}\sqrt{T}\xi_{T,l}(\|\theta_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\theta_0^{(l)}\|_2) \xrightarrow[T\to\infty]{\mathbb{P}} 0.$$

Combining the fact $k \in \mathcal{S}$ and $\theta_0^{(k)}$ is partially zero, that is $i \in \mathcal{A}_k^c$, we obtain the divergence given in (2.4.10). Furthermore, if $l \notin \mathcal{S}$, that is $\theta_0^{(l)} = 0$, then

$$\sqrt{T}\{\|\theta_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\theta_0^{(l)}\|_2\} = \|\boldsymbol{u}^{(l)}\|_2,$$

and $T^{\mu/2}(\|\tilde{\theta}^{(l)}\|_2)^{\mu} = O_p(1)$, then under the assumption $\gamma_T T^{(\mu-1)/2} \to \infty$, we obtain

$$\gamma_T T^{-1/2}\xi_{T,l}\sqrt{T}[\|\theta_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\theta_0^{(l)}\|_2] = \gamma_T T^{-1/2}\|\boldsymbol{u}^{(l)}\|_2 \frac{T^{\mu/2}}{(T^{1/2}\|\tilde{\theta}^{(l)}\|_2)^{\mu}} \xrightarrow[T\to\infty]{\mathbb{P}} \infty.$$

We deduce the pointwise convergence $\mathbb{F}_T(\boldsymbol{u}) \xrightarrow{d} \mathbb{F}_\infty(\boldsymbol{u})$, where $\mathbb{F}_\infty(.)$ is given in (2.4.9). As $\mathbb{F}_T(.)$ is convex and $\mathbb{F}_\infty(.)$ is convex and has a unique minimum $(\mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{Z}_{\mathcal{A}}, \boldsymbol{0}_{\mathcal{A}^c})$, by Lemma 2.4.9, we obtain

$$\sqrt{T}(\hat{\theta} - \theta_0) = \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_T(\boldsymbol{u})\} \xrightarrow{d} \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_\infty(\boldsymbol{u})\},$$

that is to say

$$\sqrt{T}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) \xrightarrow{d} \mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{Z}_{\mathcal{A}}, \text{ and } \sqrt{T}(\hat{\theta}_{\mathcal{A}^c} - \theta_{0,\mathcal{A}^c}) \xrightarrow{d} \boldsymbol{0}_{\mathcal{A}^c}.$$

We now prove the model selection consistency. Let $i \in \mathcal{A}_k$, then by the asymptotic normality result, $\hat{\theta}_i^{(k)} \xrightarrow[T\to\infty]{\mathbb{P}} \theta_0^{(k)}$, which implies $\mathbb{P}(i \in \hat{\mathcal{A}}_k) \to 1$. Thus the proof consists of proving

$$\forall k = 1, \cdots, m, \forall i \in \mathcal{A}_k^c, \mathbb{P}(i \in \hat{\mathcal{A}}_k) \to 0.$$

This problem can be split into two parts as

$$\forall k \notin \mathcal{S}, \mathbb{P}(k \in \hat{\mathcal{S}}) \to 0, \text{ and } \forall k \in \mathcal{S}, \forall i \in \mathcal{A}_k^c, \mathbb{P}(i \in \hat{A}_k) \to 0. \tag{2.4.11}$$

Let us start with the case $k \notin \mathcal{S}$. If $k \in \hat{\mathcal{S}}$, by the optimality conditions given by the Karush-Kuhn-Tucker theorem applied on $\mathbb{G}_T\psi(\hat{\theta})$, we have

$$\dot{\mathbb{G}}_T l(\hat{\theta})_{(k)} + \frac{\lambda_T}{T}\alpha_T^{(k)} \odot \hat{\boldsymbol{w}}^{(k)} + \frac{\gamma_T}{T}\xi_{T,k}\frac{\hat{\theta}^{(k)}}{\|\hat{\theta}^{(k)}\|_2} = 0,$$

$\odot$ is the Hadamard product and

$$\hat{\boldsymbol{w}}_i^{(k)} = \begin{cases} \text{sgn}(\hat{\theta}_i^{(k)}) \text{ if } \hat{\theta}_i^{(k)} \neq 0, \\ \in \{\hat{\boldsymbol{w}}_i^{(k)} : |\hat{\boldsymbol{w}}_i^{(k)}| \leq 1\} \text{ if } \hat{\theta}_i^{(k)} = 0. \end{cases}$$

Multiplying the unpenalized part by $T^{1/2}$, we have the expansion

$$\begin{aligned} T^{1/2}\dot{\mathbb{G}}_T l(\hat{\theta})_{(k)} &= T^{1/2}\dot{\mathbb{G}}_T l(\theta_0)_{(k)} + T^{1/2}\ddot{\mathbb{G}}_T l(\theta_0)_{(k)(k)}(\hat{\theta}-\theta_0)_{(k)} \\ &+ T^{1/2}\nabla'\{(\hat{\theta}-\theta_0)'_{(k)}\ddot{\mathbb{G}}_T l(\bar{\theta})_{(k)(k)}(\hat{\theta}-\theta_0)_{(k)}\}, \end{aligned}$$

which is asymptotically normal by consistency, assumption 28 regarding the bound on the third order term, the Slutsky theorem and the central limit theorem of Billingsley (1961). Furthermore, we have

$$\gamma_T T^{-1/2}\xi_{T,k}\frac{\hat{\theta}^{(k)}}{\|\hat{\theta}^{(k)}\|_2} = \gamma_T T^{(\mu-1)/2}(T^{1/2}\|\tilde{\theta}^{(k)}\|_2)^{-\mu}\frac{\hat{\theta}^{(k)}}{\|\hat{\theta}^{(k)}\|_2} \xrightarrow[T\to\infty]{\mathbb{P}} \infty,$$

using $T^{(\mu-\eta)/2}\gamma_T\lambda_T^{-1} \to \infty$. Therefore, we have

$$\forall k \notin \mathcal{S}, \mathbb{P}(k \in \hat{\mathcal{S}}) \leq \mathbb{P}(-\dot{\mathbb{G}}_T l(\hat{\theta})_{(k)} = \frac{\lambda_T}{T}\alpha_T^{(k)} \odot \hat{\boldsymbol{w}}_i^{(k)} + \frac{\gamma_T}{T}\xi_{T,k}\frac{\hat{\theta}^{(k)}}{\|\hat{\theta}^{(k)}\|_2}) \to 0.$$

We now pick $k \in \mathcal{S}$ and consider the event $\{i \in \hat{\mathcal{A}}_k\}$. Then the Karush-Kuhn-Tucker conditions for $\mathbb{G}_T\psi(\hat{\theta})$ are given by

$$(\dot{\mathbb{G}}_T l(\hat{\theta}))_{(k),i} + \frac{\lambda_T}{T}\alpha_{T,i}^{(k)}\text{sgn}(\hat{\theta}_{T,i}^{(k)}) + \frac{\gamma_T}{T}\xi_{T,k}\frac{\hat{\theta}_i^{(k)}}{\|\hat{\theta}^{(k)}\|_2} = 0.$$

Using the same reasoning as previously, $T^{1/2}(\dot{\mathbb{G}}_T l(\hat{\theta}))_{(k),i}$ is also asymptotically normal, and $\tilde{\theta}^{(k)} \xrightarrow[T\to\infty]{\mathbb{P}} \theta_0^{(k)}$ for $k \in \mathcal{S}$, and besides

$$\lambda_T T^{-1/2} \alpha_{T,i}^{(k)} \text{sgn}(\hat{\theta}_i^{(k)}) = \lambda_T \frac{T^{(\eta-1)/2}}{(T^{1/2}|\tilde{\theta}_i^{(k)}|)^\eta} \xrightarrow[T\to\infty]{\mathbb{P}} \infty,$$

such that we obtain the same when adding $\gamma_T T^{-1/2} \xi_{T,k} \dfrac{\hat{\theta}_i^{(k)}}{\|\hat{\theta}^{(k)}\|_2}$. Therefore, we have

$$\forall k \in \mathcal{S}, \forall i \notin \mathcal{A}_k, \mathbb{P}(i \in \hat{\mathcal{A}}_k) \leq \mathbb{P}(-(\dot{\mathbb{G}}_T l(\hat{\theta}))_{(k),i} = \frac{\lambda_T}{T} \alpha_{T,i}^{(k)} \text{sgn}(\hat{\theta}_i^{(k)}) + \frac{\gamma_T}{T} \xi_{T,k} \frac{\hat{\theta}_i^{(k)}}{\|\hat{\theta}^{(k)}\|_2}) \to 0.$$

We have proved (2.4.11). □

2.5 Double-asymptotic

In the previous sections, we worked with a fixed dimension d, where $d = \sum_{k=1}^{m} \boldsymbol{c}_k$. From now on, let us consider the case where $d = d_T$, such that $d_T \to \infty$ as $T \to \infty$. Note that $\text{card}(\mathcal{S}) = O(\text{card}(\mathcal{A})) = O(d_T)$. The speed of growth of the dimension is supposed to be $d_T = O(T^c)$ for some $q_2 < c < q_1$. In this section, we prove that the adaptive SGL satisfies the oracle property, that is model selection consistency and optimal rate of convergence for proper choices of $0 \leq q_1 < q_2 < 1$. We highlight that our general framework unfortunatly hampers a high degree of flexibility on the behavior of d_T, that is c cannot be set in $(0, 1)$. This issue was encountered by Fan and Peng (2004) in an i.i.d. and non-adaptive framework. This lack of flexibility is a necessary cost to cope with the random remainder of the Taylor expansions as we should take the third order term into account. This problem is moved aside when considering the simple linear model, where the third order derivative is zero. For instance, Zou and Zhang (2009) proved the oracle property of the adaptive elastic-net in a double-asymptotic framework for linear models where $0 \leq c < 1$.

For the asymptotic normality, we use the method of Fan and Peng (2004) and Zou and Zhang (2009), where we derive the asymptotic distribution of the discrepancy $\sqrt{T}(\hat{\theta} - \theta_0)_{\mathcal{A}}$ times a matrix sequence (Q_T) of size $r \times \text{card}(\mathcal{A})$, r being arbitrary but finite. This allows for switching from infinite dimensional distribution to finite dimensional distribution, where we can apply the usual tools of the asymptotic analysis.

In this section, we provide the conditions to satisfy the oracle property as in Fan and Peng (2004) or Zou and Zhang (2009). In this double-asymptotic framework, the quantities depend on d_T, hence on T. They should be indexed by T, which expresses that the dimension depend on the sample size. In the rest of the paper, we denote $\mathbb{H}_T := \mathbb{E}[\nabla^2_{\theta\theta} l(\epsilon_t;\theta_0)]$ and $\mathbb{M}_T := \mathbb{E}[\nabla_\theta l(\epsilon_t;\theta_0)\nabla_{\theta'} l(\epsilon_t;\theta_0)]$. To make the reading easier, we do not index other quantities by T, which will be implicit. We remind that the criterion is

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta\in\Theta}\{\mathbb{G}_T l(\theta) + \boldsymbol{p}_1(\lambda_T,\tilde{\theta},\theta) + \boldsymbol{p}_2(\gamma_T,\tilde{\theta},\theta)\} \\ &= \arg\min_{\theta\in\Theta}\{\frac{1}{T}\sum_{t=1}^{T} l(\epsilon_t;\theta) + \frac{\lambda_T}{T}\sum_{k=1}^{m}\sum_{i=1}^{\boldsymbol{c}_k}\alpha_{T,i}^{(k)}|\theta_i^{(k)}| + \frac{\gamma_T}{T}\sum_{l=1}^{m}\xi_{T,l}\|\theta^{(l)}\|_2\},\end{aligned} \tag{2.5.1}$$

with $\alpha_{T,i}^{(k)} = |\tilde{\theta}_i^{(k)}|^{-\eta}$ and $\xi_{T,l} = \|\tilde{\theta}^{(l)}\|_2^{-\mu}$, where $\eta > 0, \mu > 0$, and $\tilde{\theta}$ is a first step estimator satisfying

$$\tilde{\theta} = \arg\min_{\theta\in\Theta}\{\mathbb{G}_T l(\theta)\}.$$

The double-asymptotic framework implies that the empirical criterion can be viewed as a sequence of dependent random variables for which we need refined asymptotic theorems for dependent sequence of arrays. Shiryaev (1991) proposed a version of the central limit theorem for dependent sequence of arrays, provided this sequence is a square integrable martingale difference satisfying the so-called Lindeberg condition. A similar theorem can be found in Billingsley (1995, theorem 35.12, p.476). We provide here the theorem of Shiryaev (see Theorem 4, p.543 of Shiryaev, 1991) that we will use to derive the asymptotic distribution of the adaptive SGL estimator.

Theorem 2.5.17. (Shiryaev, 1991)
Let a sequence of square-integrable martingale differences $\xi^T = (\xi_{T,t}, \mathcal{F}_t^T), T \geq 0$, *with* $\mathcal{F}_t^T = \sigma(\xi_{T,s}, s \leq t)$, *satisfy the Lindeberg condition, for* $\epsilon > 0$, *given by*

$$\sum_{t=0}^{T}\mathbb{E}[\xi_{T,t}^2 \mathbf{1}_{|\xi_{\mathbf{T},\mathbf{t}}|>\epsilon}|\mathcal{F}_{\mathbf{t-1}}^{\mathbf{T}}] \xrightarrow[\mathbf{T}\to\infty]{\mathbb{P}} \mathbf{0},$$

then if $\sum_{t=0}^{T}\mathbb{E}[\xi_{T,t}^2|\mathcal{F}_{t-1}^T] \xrightarrow[T\to\infty]{\mathbb{P}} \sigma_t^2$, *or* $\sum_{t=0}^{T}\xi_{T,t}^2 \xrightarrow[T\to\infty]{\mathbb{P}} \sigma_t^2$, *then* $\sum_{t=0}^{T}\xi_{T,t} \xrightarrow{d} \mathcal{N}(0,\sigma_t^2)$.

Remark 2.5.18. Note that central limit theorems relaxing the stationarity and martingale difference assumptions for sequences of arrays exist. Neumann (2013) proposed such a central limit theorem for weakly dependent sequences of arrays. Such sequences

should also satisfy a Lindeberg condition and conditions on covariances. In the rest of the paper, we use Shiryaev's result.

We consider problem (2.5.1), which is the adaptive SGL estimator. In the first step, we study the convergence rate of the first step unpenalized estimator, which is plugged in the adaptive specification. The convergence rate of a classic M-estimator is $T^{1/2}$, for d fixed. For d diverging, we need some additional assumptions.

The two next assumptions are similar to condition (F) of Fan and Peng (2004) and allow for controlling the minimum and maximum eigenvalues of the limits of the empirical Hessian and the score cross-product. We denote by $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ the minimum and maximum eigenvalues of any positive definite square matrix $\boldsymbol{M}$.

Assumption 29. $\mathbb{H}_T$ and $\mathbb{M}_T$ exist. $\mathbb{H}_T$ is nonsingular, and there exist b_1, b_2 with $0 < b_1 < b_2 < \infty$ and c_1, c_2 with $0 < c_1 < c_2 < \infty$ such that, for all T,

$$b_1 < \lambda_{\min}(\mathbb{M}_T) < \lambda_{\max}(\mathbb{M}_T) < b_2,\ c_1 < \lambda_{\min}(\mathbb{H}_T) < \lambda_{\max}(\mathbb{H}_T) < c_2.$$

Let $\mathbb{V}_T = \mathbb{H}_T^{-1}\mathbb{M}_T\mathbb{H}_T^{-1}$, we deduce there exist a_1, a_2 with $0 < a_1 < a_2 < \infty$ such that, for all T,

$$a_1 < \lambda_{\min}(\mathbb{V}_T) < \lambda_{\max}(\mathbb{V}_T) < a_2.$$

Assumption 30. $\mathbb{E}[\{\nabla_\theta l(\epsilon_t;\theta_0)\nabla_{\theta'} l(\epsilon_t;\theta_0)\}^2] < \infty$, for every d_T (and then of T).

Assumption 31. There exist some functions $\Psi(.)$ such that, for all T,

$$\sup_{k=1,\cdots,d_T} \mathbb{E}[\partial_{\theta_k} l(\epsilon_t;\theta)\partial_{\theta_k} l(\epsilon_{t'};\theta)] \leq \Psi(|t-t'|),$$

and

$$\sup_T \frac{1}{T}\sum_{t,t'=1}^{T} \Psi(|t-t'|) < \infty.$$

Assumption 32. Let $\zeta_{kl,t} := \partial^2_{\theta_k\theta_l} l(\epsilon_t;\theta_0) - \mathbb{E}[\partial^2_{\theta_k\theta_l} l(\epsilon_t;\theta_0)]$. There exist some functions $\chi(.)$ such that

$$|\mathbb{E}[\zeta_{kl,t}\zeta_{k'l',t'}]| \leq \chi(|t-t'|),$$

and

$$\sup_T \frac{1}{T}\sum_{t,t'=1}^{T} \chi(|t-t'|) < \infty.$$

Assumption 33. Let $\upsilon_t(C) := \sup_{k,l,m=1,\cdots,d_T} \{ \sup_{\theta:\|\theta-\theta_0\|_2 \leq \nu_T C} |\partial^3_{\theta_k\theta_l\theta_m} l(\epsilon_t;\theta)| \}$, where $C > 0$ is a fixed constant and $\nu_T = (d_T/T)^{1/2}$. Then

$$\eta(C) := \frac{1}{T^2} \sum_{t,t'=1}^{T} \mathbb{E}[\upsilon_t(C)\upsilon_{t'}(C)] < \infty.$$

Theorem 2.5.19. *Under assumptions 23-25, 29-33 and if $d_T^4 = o(T)$, the sequence of unpenalized M-estimators solving $\tilde{\theta} = \arg\min_{\theta\in\Theta} \{\mathbb{G}_T l(\theta)\}$ satisfies*

$$\|\tilde{\theta} - \theta_0\|_2 = O_p((\frac{d_T}{T})^{\frac{1}{2}}).$$

Both vectors $\tilde{\theta}$ and θ_0 depend on T such that $\tilde{\theta} = \tilde{\theta}_T$ and $\theta_0 = \theta_{0,T} := \theta_{0,\infty}.e_T$.

Remark 2.5.20. Note that this consistency result requires at most $d_T^4 = o(T)$, as Theorem 1 of Fan and Peng (2004).

Proof of Theorem 2.5.19. We proceed as in the proof of Theorem 2.4.6. We denote $\nu_T = (d_T/T)^{1/2}$ and we would like to prove that, for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$\mathbb{P}(\|\tilde{\theta} - \theta_0\|_2/\tilde{\nu}_T > C_\epsilon) < \epsilon. \tag{2.5.2}$$

To prove (2.5.2), it is sufficient to show that for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$\begin{aligned} \mathbb{P}(\|\tilde{\theta} - \theta_0\|_2 > C_\epsilon \nu_T) &\leq \mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 \geq C_\epsilon : \mathbb{G}_T l(\theta_0 + \nu_T \boldsymbol{u}) \leq \mathbb{G}_T l(\theta_0)) \\ &= \mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : \mathbb{G}_T l(\theta_0 + \nu_T \boldsymbol{u}) \leq \mathbb{G}_T l(\theta_0)), \end{aligned}$$

by convexity. By a Taylor expansion of $\mathbb{G}_T l(\theta_0 + \nu_T \boldsymbol{u})$, we obtain

$$\mathbb{G}_T l(\theta_0 + \nu_T \boldsymbol{u}) = \mathbb{G}_T l(\theta_0) + \nu_T \dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^2}{2} \boldsymbol{u}' \ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^3}{6} \nabla' \{\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u},$$

where $\bar{\theta} \in \Theta$ such that $\|\bar{\theta} - \theta_0\|_2 \leq C_\epsilon \nu_T$. We would like to prove

$$\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : \nu_T \dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^2}{2} \boldsymbol{u}' \ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^3}{6} \nabla' \{\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} \leq 0) < \epsilon. \tag{2.5.3}$$

To do so, we focus on each quantity of the Taylor expansion to extract the dominant term. First, for $a > 0$ and the Markov inequality, we have for the score term

$$\begin{aligned}
\mathbb{P}(\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=C_\epsilon} |\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}| > a) &\leq \mathbb{P}(\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=C_\epsilon} \|\dot{\mathbb{G}}_T l(\theta_0)\|_2 \|\boldsymbol{u}\|_2 > a) \\
&\leq \mathbb{P}(\|\dot{\mathbb{G}}_T l(\theta_0)\|_2 > \tfrac{a}{C_\epsilon}) \\
&\leq (\tfrac{C_\epsilon}{a})^2 \mathbb{E}[\|\dot{\mathbb{G}}_T l(\theta_0)\|_2^2] \\
&\leq (\tfrac{C_\epsilon}{a})^2 \sum_{k=1}^{d_T} \mathbb{E}[(\partial_{\theta_k} \mathbb{G}_T l(\theta_0))^2] \\
&= (\tfrac{C_\epsilon}{a})^2 \tfrac{1}{T^2} \sum_{t,t'=1}^{T} \sum_{k=1}^{d_T} \mathbb{E}[\partial_{\theta_k} l(\epsilon_t;\theta_0) \partial_{\theta_k} l(\epsilon_{t'};\theta_0)] \\
&\leq (\tfrac{C_\epsilon}{a})^2 \{\tfrac{1}{T^2} \sum_{t,t'=1}^{T} \Psi(|t-t'|)\}.d_T.
\end{aligned}$$

By assumption 31, $\sup_{k=1,\cdots,d_T} \mathbb{E}[\partial_{\theta_k} l(\epsilon_t;\theta_0)\partial_{\theta_k} l(\epsilon_{t'};\theta_0)] \leq \Psi(|t-t'|)$ and $\frac{1}{T}\sum_{t,t'=1}^{T} \Psi(|t-t'|) < \infty$. This implies

$$\mathbb{P}(\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=C_\epsilon} |\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}| > a) \leq \frac{C_\epsilon^2 d_T}{Ta^2} K_1,$$

for some constant $K_1 > 0$.

We now focus on the hessian quantity that can be rewritten as

$$\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\theta_0) \boldsymbol{u} = \boldsymbol{u}' \mathbb{E}[\ddot{\mathbb{G}}_T l(\theta_0)] \boldsymbol{u} + \mathcal{R}_T(\theta_0),$$

where $\mathcal{R}_T(\theta_0) = \sum_{k,l=1}^{d_T} \boldsymbol{u}_k \boldsymbol{u}_l \{\partial^2_{\theta_k\theta_l} \mathbb{G}_T l(\theta_0) - \mathbb{E}[\partial^2_{\theta_k\theta_l} \mathbb{G}_T l(\theta_0)]\}$. We have

$$\mathbb{E}[\mathcal{R}_T(\theta_0)] = 0, \ \mathrm{Var}(\mathcal{R}_T(\theta_0)) = \frac{1}{T^2} \sum_{t,t'=1}^{T} \sum_{k,k',l,l'=1}^{d_T} \boldsymbol{u}_k \boldsymbol{u}_{k'} \boldsymbol{u}_l \boldsymbol{u}_{l'} \mathbb{E}[\zeta_{kl,t}.\zeta_{k'l',t'}],$$

where $\zeta_{kl,t} = \partial^2_{\theta_k\theta_l} l(\epsilon_t;\theta_0) - \mathbb{E}[\partial^2_{\theta_k\theta_l} l(\epsilon_t;\theta_0)]$. Let $b > 0$, we deduce by the Markov inequality and assumption 32,

$$\mathbb{P}(|\mathcal{R}_T(\theta_0)| > b) \leq \frac{1}{b^2}\mathbb{E}[\mathcal{R}_T^2(\theta_0)] \leq \frac{K_2}{b^2}\frac{\|\boldsymbol{u}\|_2^4 d_T^2}{T} \leq \frac{K_2 C_\epsilon^4 d_T^2}{b^2 T},$$

where $K_2 > 0$. Furthermore, by assumption 29,

$$\boldsymbol{u}' \mathbb{E}[\ddot{\mathbb{G}}_T l(\theta_0)] \boldsymbol{u} \geq \lambda_{\min}(\mathbb{H}_T) \boldsymbol{u}' \boldsymbol{u}.$$

As for the third order term, we have

$$\begin{aligned}|\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}|^2 &\leq \frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k_1,k_2,k_3}\sum_{l_1,l_2,l_3}|u_{k_1}u_{k_2}u_{k_3}u_{l_1}u_{l_2}u_{l_3}||\partial^3_{\theta_{k_1}\theta_{k_2}\theta_{k_3}} l(\epsilon_t;\bar{\theta}).\partial^3_{\theta_{l_1}\theta_{l_2}\theta_{l_3}} l(\epsilon_{t'};\bar{\theta})| \\ &\leq \|\boldsymbol{u}\|_2^6 d_T^3 \frac{1}{T^2}\sum_{t,t'=1}^{T}\upsilon_t(C_\epsilon)\upsilon_{t'}(C_\epsilon),\end{aligned}$$

where

$$\upsilon_t(C_0) = \sup_{k_1k_2k_3}\{\sup_{\theta:\|\theta-\theta_0\|_2\leq\nu_T C_0}|\partial^3_{\theta_{k_1}\theta_{k_2}\theta_{k_3}} l(\epsilon_t;\theta)|\}.$$

Note that $\upsilon_t(C_0)$ depends on d_T and C_0. By assumption 33, we have

$$\eta(C_0) := \frac{1}{T^2}\sum_{t,t'=1}^{T}\mathbb{E}[\upsilon_t(C_0)\upsilon_{t'}(C_0)] < \infty.$$

By the Markov inequality, for $c > 0$, we conclude that

$$\mathbb{P}(\exists\boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \frac{\nu_T^2}{6}\sup_{\|\bar{\theta}-\theta_0\|_2\leq\nu_T C_\epsilon}|\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}| > c) \leq \frac{\nu_T^4 d_T^3 C_\epsilon^6}{36c^2}\eta(C_\epsilon).$$

We can now bound (2.5.3) thanks to proper choices of a, b, c and C_ϵ. We denote by $\delta_T = \lambda_{\min}(\mathbb{H}_T)C_\epsilon^2\nu_T$, and using $\frac{\nu_T}{2}\mathbb{E}[\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}] \geq \delta_T$, we have

$$\begin{aligned}&\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : \dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^2}{6}\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} \leq 0) \\ &\leq \mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}| > \delta_T/4) + \mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : \tfrac{\nu_T}{2}|\mathcal{R}_T(\theta_0)| > \delta_T/4) \\ &+\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : \tfrac{\nu_T^2}{6}\sup_{\bar{\theta}:\|\bar{\theta}-\theta_0\|_2<\nu_T C_\epsilon}|\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}| > \delta_T/4) \\ &\leq \frac{16C_\epsilon^2 d_T K_1}{T\delta_T^2} + \frac{4\nu_T^2 d_T^2 C_\epsilon^4}{T\delta_T^2} + \frac{16\nu_T^4 d_T^3 C_\epsilon^6}{36\delta_T^2}\eta(C_\epsilon) \\ &\leq C_1\frac{d_T}{TC_\epsilon^2\nu_T^2} + C_2\frac{d_T^2}{T} + C_3\nu_T^2 d_T^3 C_\epsilon^2\eta(C_\epsilon),\end{aligned}$$

where C_1, C_2, C_3 are strictly positive constants. We chose $\nu_T = (\frac{d_T}{T})^{\frac{1}{2}}$, we then deduce

$$\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : \nu_T \dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^3}{6}\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} \leq 0)$$
$$\leq \frac{C_1}{C_\epsilon^2} + C_2\frac{d_T^2}{T} + \frac{d_T^4 C_\epsilon^2}{T}\eta(C_\epsilon).$$

Now we fix C_ϵ sufficiently large enough, such that $C_1/C_\epsilon^2 < \epsilon/3$. Once this constant is fixed, there exists a T_0 such that for $T > T_0$ we have $C_2\frac{d_T^2}{T} < \epsilon/3$ and $C_3\frac{d_T^4 C_\epsilon^2}{T}\eta(C_\epsilon) < \epsilon/3$ under the assumption that $d_T^4 = o(T)$. Consequently, we obtain

$$\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : \mathbb{G}_T l(\theta_0) + \nu_T \dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^3}{6}\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} \leq 0) < \epsilon.$$

This proves (2.5.2), that is $\|\tilde{\theta} - \theta_0\|_2 = O_p((\frac{d_T}{T})^{\frac{1}{2}})$. □

The first step estimator used for the adaptive weights is $(T/d_T)^{1/2}$-consistent. However, the estimated quantities on $\mathcal{A}^c$ converge to zero by consistency. We then propose a slight modification of the first step estimator, denoted $\tilde{\tilde{\theta}}$, which disappears asymptotically as follows

$$\tilde{\tilde{\theta}} = \tilde{\theta} + e_T,$$

such that $e_T \to 0$ is a strictly positive quantity. We choose $e_T = T^{-\kappa}$ with $\kappa > 0$. This means we add in the adaptive weights a power of T to the first step estimator, that is

$$\alpha_{T,i}^{(k)} = |\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta} = |\tilde{\theta} + T^{-\kappa}|^{-\eta}, \ \xi_{T,l} = \|\tilde{\tilde{\theta}}^{(l)}\|_2^{-\mu} = \|\tilde{\theta}^{(l)} + T^{-\kappa}\|_2^{-\mu}.$$

Theorem 2.5.21. *Under assumptions 23-25, 29-33, if* $d_T^4 = o(T)$, *and if* $\frac{\gamma_T}{\sqrt{T}}T^{\frac{c}{2}+\kappa\mu} \underset{T\to\infty}{\longrightarrow} 0$, $\frac{\lambda_T}{\sqrt{T}}T^{\kappa\eta} \underset{T\to\infty}{\longrightarrow} 0$, *then the sequence of penalized estimators* $\hat{\theta}$ *solving 2.5.1 satisfies*

$$\|\hat{\theta} - \theta_0\|_2 = O_p((\frac{d_T}{T})^{\frac{1}{2}}).$$

Remark 2.5.22. Note that $d_T^4 = o(T)$ is as in Fan and Peng (2004), Theorem 1. Thanks to proper choices of the regularization terms, we obtain a $(T/d_T)^{1/2}$-consistent adaptive SGL estimator.

Proof of Theorem 2.5.21. We proceed as we did for proving Theorem 2.5.19. Let $\nu_T = (d_T/T)^{1/2}$. We would like to prove that for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$\mathbb{P}(\|\hat{\theta} - \theta_0\|_2/\nu_T > C_\epsilon) < \epsilon. \tag{2.5.4}$$

To prove (2.5.4), we show

$$\begin{aligned}\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : \dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} + \frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} \\ +\nu_T^{-1}\{\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\theta}}, \theta_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\theta}}, \theta_0) + \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\theta}}, \theta_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\theta}}, \theta_0\}) \leq 0) < \epsilon.\end{aligned} \tag{2.5.5}$$

a relationship obtained by convexity and a Taylor expansion.

The score quantity can be upper bounded as

$$|\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}| \leq \|\dot{\mathbb{G}}_T l(\theta_0)\|_2\|\boldsymbol{u}\|_2 = O_p((\frac{d_T}{T})^{\frac{1}{2}})\|\boldsymbol{u}\|_2 = O_p(\nu_T)\|\boldsymbol{u}\|_2,$$

where we used assumption 31 to obtain the bound in probability of the score.

As for the third order term, we have by the Cauchy-Schwartz inequality

$$\begin{aligned}|\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}|^2 &\leq \|\boldsymbol{u}\|_2^6 d_T^3 \frac{1}{T^2}\sum_{t,t'=1}^{T^2}\{\sum_{k_1,l_1,m_1=1}^{d_T}\sum_{k_2,l_2,m_2=1}^{d_T}\partial^3_{\theta_{k_1}\theta_{l_1}\theta_{m_1}} l(\epsilon_t;\bar{\theta})\partial^3_{\theta_{k_2}\theta_{l_2}\theta_{m_2}} l(\epsilon_{t'};\bar{\theta})\} \\ &= \|\boldsymbol{u}\|_2^6 d_T^3 \eta(C_\epsilon).\end{aligned}$$

This implies

$$\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u} = O_p(d_T^{3/2}\|\boldsymbol{u}\|_2^3).$$

Hence by the Markov inequality

$$\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\nu_T^2\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}| > a) \leq \frac{\nu_T^4 C_\epsilon^6 d_T^3}{a^2}\eta(C_\epsilon).$$

where we used assumption 33.

Finally, the hessian quantity can be treated as in the proof of Theorem 2.5.19. We denote by $\mathcal{R}_T(\theta_0) = \sum_{k,l=1}^{d_T}\boldsymbol{u}_k\boldsymbol{u}_l\{\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\theta_0) - \mathbb{E}[\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\theta_0)]\}$. We have

$$\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} = \boldsymbol{u}'\mathbb{E}[\ddot{\mathbb{G}}_T l(\theta_0)]\boldsymbol{u} + \mathcal{R}_T(\theta_0).$$

By assumption 32 and the Markov inequality, for any $\kappa > 0$, we obtain

$$\mathbb{P}(|\mathcal{R}_T(\theta_0)| > \kappa) \leq \frac{1}{\kappa^2}\mathbb{E}[\mathcal{R}_T^2(\theta_0)] \leq \frac{K_2}{\kappa^2}\frac{\|\boldsymbol{u}\|_2^4 d_T^2}{T} \leq \frac{K_2 C_\epsilon^4 d_T^2}{\kappa^2 T},$$

with $K_2 > 0$. This relationship holds for any $\kappa > 0$. Then for T large enough, we deduce that $|\mathcal{R}_T(\theta_0)| = o_p(1)$. Consequently

$$\frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u} \geq \frac{\nu_T^2}{2}\lambda_{\min}(\mathbb{H}_T)\|\boldsymbol{u}\|_2^2 + o_p(1)\nu_T^2\|\boldsymbol{u}\|_2^2.$$

We focus on the penalty terms. We have

$$\begin{aligned}
\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\theta}}, \theta_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\theta}}, \theta_0) &= \lambda_T T^{-1}\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k}\alpha_{T,i}^{(k)}\{|\theta_{0,i}^{(k)} + \nu_T\boldsymbol{u}_i^{(k)}| - |\theta_{0,i}^{(k)}|\}, \\
\text{and } |\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\theta}}, \theta_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\theta}}, \theta_0)| &\leq \lambda_T T^{-1}\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k}\alpha_{T,i}^{(k)}\nu_T|\boldsymbol{u}_i^{(k)}|.
\end{aligned}$$

As for the l^1/l^2 norm, we obtain

$$\begin{aligned}
\boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\theta}}, \theta_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\theta}}, \theta_0) &= \gamma_T T^{-1}\sum_{l\in\mathcal{S}1}\xi_{T,l}\{\|\theta_0^{(l)} + \nu_T\boldsymbol{u}\|_2 - \|\theta_0^{(l)}\|_2\} \\
\text{and } |\boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\theta}}, \theta_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\theta}}, \theta_0)| &\leq \gamma_T T^{-1}\sum_{l\in\mathcal{S}}\xi_{T,l}\nu_T\|\boldsymbol{u}^{(l)}\|_2.
\end{aligned}$$

For the l^1 norm penalty, using $\{\min_{k\in\mathcal{S},i\in\mathcal{A}_k}|\tilde{\tilde{\theta}}_i^{(k)}|\}^{-\eta} \leq T^{\kappa\eta}$, then

$$\begin{aligned}
\lambda_T T^{-1}\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k}\alpha_{T,i}^{(k)}\nu_T|\boldsymbol{u}_i^{(k)}| &\leq \lambda_T T^{-1}\nu_T\{\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k}|\tilde{\tilde{\theta}}_i^{(k)}|^{-2\eta}\}^{1/2}\|\boldsymbol{u}\|_2 \\
&\leq \lambda_T T^{-1}\nu_T\frac{\sqrt{d_T}}{\{\min_{k\in\mathcal{S},i\in\mathcal{A}_k}|\tilde{\tilde{\theta}}_i^{(k)}|\}^{\eta}}\|\boldsymbol{u}\|_2 \\
&\leq \lambda_T T^{-1}\nu_T\sqrt{d_T}T^{\kappa\eta}\|\boldsymbol{u}\|_2,
\end{aligned}$$

by the Cauchy-Schwartz inequality. Then if $\lambda_T T^{\frac{c}{2}-1+\kappa\eta}$ is bounded, we obtain

$$\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\theta}}, \theta_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\theta}}, \theta_0) = O(\nu_T^2)\|\boldsymbol{u}\|_2.$$

As for the l^1/l^2 term, using $\{\min_{l\in\mathcal{S}}\|\tilde{\tilde{\theta}}^{(l)}\|_2\}^{-\mu} \leq T^{\kappa\mu}$, we obtain

$$\begin{aligned}\gamma_T T^{-1}\sum_{l=1}^{m}\xi_{T,l}\nu_T\|\boldsymbol{u}^{(l)}\|_2 &\leq \gamma_T T^{-1}\nu_T\{\sum_{l\in\mathcal{S}}\|\tilde{\tilde{\theta}}^{(l)}\|_2^{-2\mu}\}^{1/2}\|\boldsymbol{u}\|_2\\ &\leq \gamma_T T^{-1}\nu_T\frac{\sqrt{d_T}}{\{\min_{l\in\mathcal{S}}\|\tilde{\tilde{\theta}}^{(l)}\|_2\}^{\mu}}\|\boldsymbol{u}\|_2\\ &\leq \gamma_T T^{-1}\nu_T\sqrt{d_T}T^{\kappa\mu}\|\boldsymbol{u}\|_2,\end{aligned}$$

by the Cauchy-Schwartz inequality. Then if $\gamma_T T^{\frac{c}{2}-1+\kappa\mu}$ is bounded, we obtain

$$\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0+\nu_T\boldsymbol{u})-\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0)=O(\nu_T^2)\|\boldsymbol{u}\|_2.$$

We now can prove (2.5.5). Let $\delta_T=\lambda_{\min}(\mathbb{H}_T)C_\epsilon^2\nu_T$ and using $\frac{\nu_T}{2}\mathbb{E}[\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}]\geq\delta_T$, we have

$$\begin{aligned}&\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T},\|\boldsymbol{u}\|_2=C_\epsilon:\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}+\nu_T\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}/2+\nu_T^2\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}/6\\ &+\nu_T^{-1}\{\boldsymbol{p}_1(\lambda_T,\tilde{\tilde{\theta}},\theta_0+\nu_T\boldsymbol{u})-\boldsymbol{p}_1(\lambda_T,\tilde{\tilde{\theta}},\theta_0)+\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0+\nu_T\boldsymbol{u})-\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0)\}\leq 0)\\ &\leq\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T},\|\boldsymbol{u}\|_2=C_\epsilon:|\nu_T\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}/2|\leq|\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}|+|\nu_T^2\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}/6|\\ &+\nu_T^{-1}\{|\boldsymbol{p}_1(\lambda_T,\tilde{\tilde{\theta}},\theta_0)-\boldsymbol{p}_1(\lambda_T,\tilde{\tilde{\theta}},\theta_0+\nu_T\boldsymbol{u})|+|\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0)-\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0+\nu_T\boldsymbol{u})|\})\\ &\leq\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T},\|\boldsymbol{u}\|_2=C_\epsilon:|\dot{\mathbb{G}}_T l(\theta_0)\boldsymbol{u}|>\delta_T/8)+\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T},\|\boldsymbol{u}\|_2=C_\epsilon:\tfrac{\nu_T}{2}|\mathcal{R}_T(\theta_0)|>\delta_T/8)\\ &+\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T},\|\boldsymbol{u}\|_2=C_\epsilon:|\tfrac{\nu_T^2}{6}\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\theta})\boldsymbol{u}\}\boldsymbol{u}|>\delta_T/8)\\ &+\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T},\|\boldsymbol{u}\|_2=C_\epsilon:|\boldsymbol{p}_1(\lambda_T,\tilde{\tilde{\theta}},\theta_0)-\boldsymbol{p}_1(\lambda_T,\tilde{\tilde{\theta}},\theta_0+\nu_T\boldsymbol{u})|>\nu_T\delta_T/8)\\ &+\mathbb{P}((\exists\boldsymbol{u}\in\mathbb{R}^{d_T},\|\boldsymbol{u}\|_2=C_\epsilon:|\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0)-\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0+\nu_T\boldsymbol{u})|>\nu_T\delta_T/8)\\ &\leq\frac{C_{st}}{C_\epsilon^2}+C_{st}\{\nu_T^2C_\epsilon^6d_T^3\eta(C_\epsilon)\}+\frac{C_{st}\nu_T^2d_T^2C_\epsilon^4}{T\delta_T^2}+\epsilon/5+\epsilon/5\\ &<\epsilon,\end{aligned}$$

with $C_{st}>0$ a generic constant. We used $d_T^4=o(T)$ and for C_ϵ large enough

$$\begin{aligned}\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T},\|\boldsymbol{u}\|_2=C_\epsilon:|\boldsymbol{p}_1(\lambda_T,\tilde{\tilde{\theta}},\theta_0)-\boldsymbol{p}_1(\lambda_T,\tilde{\tilde{\theta}},\theta_0+\nu_T\boldsymbol{u})|>\nu_T\delta_T/8)&\leq\epsilon/5,\\ \mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_T},\|\boldsymbol{u}\|_2=C_\epsilon:|\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0)-\boldsymbol{p}_2(\gamma_T,\tilde{\tilde{\theta}},\theta_0+\nu_T\boldsymbol{u})|>\nu_T\delta_T/8)&\leq\epsilon/5.\end{aligned}$$

Thus we obtain for C_ϵ and T large enough, with the conditions $\gamma_T T^{\frac{c}{2}-1+\kappa\mu}\to 0$ and $\lambda_T T^{\frac{c}{2}-1+\kappa\eta}\to 0$ that

$$\|\hat{\theta}-\theta_0\|_2=O_p(\nu_T)=O_p((\frac{d_T}{T})^{\frac{1}{2}}).$$

□

To satisfy the oracle property, we need some additional assumptions regarding the adaptive penalty components.

Assumption 34. For any T, there exists β such that $0 < \beta < \min_{i \in \mathcal{A}_k} \theta_{0,i,\mathcal{A}_k}, k \in \mathcal{S}$. Moreover,

$$\beta^{-1} T^{-1} \{ \lambda_T d_T^{1/2} \mathbb{E}[\max_{k \in \mathcal{S}, i \in \mathcal{A}_k} \alpha_{T,\mathcal{A}_k,i}] + \gamma_T \mathbb{E}[\max_{k \in \mathcal{S}} \xi_{T,k}] \} \underset{T \to \infty}{\longrightarrow} 0.$$

Assumption 35. The model complexity is assumed to behave as $d_T^5 = o(T)$, which implies that $0 < c < \frac{1}{5}$. The regularization parameters are chosen such that they satisfy

$$\frac{\gamma_T}{\sqrt{T}} T^{\frac{c}{2}+\kappa\mu} \underset{T \to \infty}{\longrightarrow} 0, \qquad \frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(1-c)-1]} \underset{T \to \infty}{\longrightarrow} \infty,$$
$$\frac{\lambda_T}{\sqrt{T}} T^{\kappa\eta} \underset{T \to \infty}{\longrightarrow} 0, \qquad \frac{\lambda_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\eta)(1-c)-1]} \underset{T \to \infty}{\longrightarrow} \infty,$$
$$\frac{\gamma_T}{\lambda_T^{1+\mu}} T^{(1+\mu)(1-\frac{c}{2}-\kappa\eta)-1} \underset{T \to \infty}{\longrightarrow} \infty.$$

Remark 2.5.23. The main condition is $d_T^5 = o(T)$, which is the same as Fan and Peng (2004). This condition comes from the control for the third order derivative of the empirical criterion. Note that simple cases allow for a framework where $0 \leq c < 1$. Moreover, these asymptotic behaviors are closely related to condition (A5) of Zou and Zhang (2009). In Section 2.6, we provide further details about the calibration of the adaptive weights and κ.

Assumption 36. Let $\mathcal{F}_t^T = \sigma(X_{T,s}, s \leq t)$ with $X_{T,t} = \sqrt{T} Q_T \mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2} \mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1} \dot{\mathbb{G}}_T l_t(\theta_0)_{\mathcal{A}}$, (Q_T) is a sequence of $r \times \text{card}(\mathcal{A})$ matrices such that $Q_T \times Q_T' \xrightarrow{\mathbb{P}} \mathbb{C}$, for some $r \times r$ nonnegative symmetric matrix $\mathbb{C}$, $\mathbb{V}_{T,\mathcal{A}\mathcal{A}} = (\mathbb{H}_T^{-1} \mathbb{M}_T \mathbb{H}_T^{-1})_{\mathcal{A}\mathcal{A}}$ and $\dot{\mathbb{G}}_T l_t(\theta_0)_{\mathcal{A}} = \frac{1}{T} \nabla_{\mathcal{A}} l(\epsilon_t; \theta_0)$. Then $X_{T,t}$ is a martingale difference and we have

$$\mathbb{E}[\sup_{i,j=1,\cdots,d_T} \mathbb{E}[\{\partial_{\theta_i} l(\epsilon_t; \theta_0) \partial_{\theta_j} l(\epsilon_t; \theta_0)\}^2 | \mathcal{F}_{t-1}^T] \lambda_{\max,t-1}(\mathbb{H}_{t-1}^T)] \leq \bar{B} < \infty,$$

with

$$\mathbb{H}_{t-1}^T := \mathbb{E}[\nabla l(\epsilon_t; \theta_0) \nabla' l(\epsilon_t; \theta_0) | \mathcal{F}_{t-1}^T] \leq \lambda_{\max}(\mathbb{H}_{t-1}^T) < \infty.$$

Theorem 2.5.24. *Under assumptions 23-25, and assumptions 29-36, the sequence of adaptive estimator $\hat{\theta}$ solving (2.5.1) satisfies*

$$\lim_{T \to \infty} \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) = 1, \text{ and}$$
$$\sqrt{T} Q_T \mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2} (\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0, \mathbb{C}),$$

where (Q_T) *is a sequence of* $r \times card(\mathcal{A})$ *matrices such that* $Q_T \times Q_T' \xrightarrow{\mathbb{P}} \mathbb{C}$, *for some* $r \times r$ *nonnegative symmetric matrix* $\mathbb{C}$ *and* $\mathbb{V}_{T,\mathcal{A}\mathcal{A}} = (\mathbb{H}_T^{-1}\mathbb{M}_T\mathbb{H}_T^{-1})_{\mathcal{A}\mathcal{A}}$.

Proof of Theorem 2.5.24. Model selection consistency consists of proving that the probability of the event $\{\hat{\mathcal{A}} = \mathcal{A}\}$ tends to one asymptotically. This event is

$$\{\hat{\mathcal{A}} = \mathcal{A}\} = \{\forall k \in \mathcal{S}, \forall i \in \mathcal{A}_k, |\hat{\theta}_i^{(k)}| > 0\} \cap \{\forall k = 1, \cdots, m, \forall i \in \mathcal{A}_k^c, \hat{\theta}_i^{(k)} = 0\}.$$

Hence we prove

$$\mathbb{P}(\{\forall k \in \mathcal{S}, \forall i \in \mathcal{A}_k, |\hat{\theta}_i^{(k)}| > 0\} \cap \{\forall k = 1, \cdots, m, \forall i \in \mathcal{A}_k^c, \hat{\theta}_i^{(k)} = 0\}) \underset{T\to\infty}{\longrightarrow} 1. \quad (2.5.6)$$

Model selection consistency can be decomposed into two parts: recovering the active indices by estimating nonzero coefficients; discarding the inactive indices by shrinking to zero the related coefficients. Now (2.5.6) can be proved by first showing that for any T, there exists β such that $0 < \beta < \min_{i\in\mathcal{A}_k} \theta_{0,i,\mathcal{A}_k}$, with $k \in \mathcal{S}$ and

$$\mathbb{P}(\|\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}\|_2 < \beta) \underset{T\to\infty}{\longrightarrow} 1. \quad (2.5.7)$$

The second part regarding nonactive indices can be proved as

$$\begin{cases} \mathbb{P}(\bigcap_{k\in\mathcal{S}^c} \{\|\hat{\boldsymbol{z}}^{(k)}\|_2 < 1\}) \underset{T\to\infty}{\longrightarrow} 1, \\ \mathbb{P}(\bigcap_{k\in\mathcal{S}} \bigcap_{i\in\mathcal{A}_k^c} \{|\hat{\boldsymbol{w}}_i^{(k)}| < 1\}) \underset{T\to\infty}{\longrightarrow} 1, \end{cases} \quad (2.5.8)$$

where $\hat{\boldsymbol{z}}^{(k)}$ (resp. $\hat{\boldsymbol{w}}^{(k)}$) is the subgradient of $\|\hat{\theta}^{(k)}\|_2$ (resp. $\|\hat{\theta}^{(k)}\|_1$) given in (2.3.1). Hence (2.5.7) and (2.5.8) prove (2.5.6).

We first focus on (2.5.7), which is equivalent to

$$\mathbb{P}(\|\hat{\theta}_{\mathcal{A}_k} - \theta_{0,\mathcal{A}_k}\|_2 > \beta) \underset{T\to\infty}{\longrightarrow} 0.$$

By the Karush-Kuhn-Tucker optimality conditions, we have

$$\dot{\mathbb{G}}_T l(\hat{\theta})_{\mathcal{A}} + \lambda_T T^{-1} \alpha_{T,\mathcal{A}} \odot \operatorname{sgn}(\hat{\theta}_{\mathcal{A}}) + \gamma_T T^{-1} \varsigma_T = 0,$$

where $\varsigma_T = \text{vec}(\xi_{T,k}\frac{\hat{\theta}_{\mathcal{A}_k}}{\|\hat{\theta}_{\mathcal{A}_k}\|_2}, k \in \mathcal{S})$. We denote by $\alpha_{T,\mathcal{A}_k} = (\alpha_{T,i}, i \in \mathcal{A}_k)$, a vector of size $\mathbb{R}^{\boldsymbol{c}_{\mathcal{A}_k}}$. By a Taylor expansion of the gradient component around $\theta_{0,\mathcal{A}}$, we have

$$
\begin{aligned}
&\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}} + \mathbb{H}_{T,\mathcal{A}\mathcal{A}}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) + \mathcal{P}_T(\theta_0)(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) + \frac{1}{2}\nabla'_{\mathcal{A}}\{(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\theta})(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})\} \\
&+\lambda_T T^{-1}\alpha_{T,\mathcal{A}} \odot \text{sgn}(\hat{\theta}_{\mathcal{A}}) + \gamma_T T^{-1}\varsigma_T = 0 \\
&\Leftrightarrow \hat{\theta}_{\mathcal{A}} = \theta_{0,\mathcal{A}} - \mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}(\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}} + \lambda_T T^{-1}\alpha_{T,\mathcal{A}} \odot \text{sgn}(\hat{\theta}_{\mathcal{A}}) + \gamma_T T^{-1}\varsigma_T \\
&-\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\frac{1}{2}\nabla'_{\mathcal{A}}\{(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})\} - \mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\mathcal{P}_T(\theta_0)(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}),
\end{aligned}
$$

where $\|\bar{\theta}-\theta_0\|_2 \leq \|\hat{\theta}-\theta_0\|_2$, $\mathcal{P}_T(\theta_0) = \ddot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}\mathcal{A}} - \mathbb{H}_{T,\mathcal{A}\mathcal{A}}$ and $\mathbb{H}_{T,\mathcal{A}\mathcal{A}} = \mathbb{E}[\nabla^2_{\theta\theta'} l(\epsilon_t;\theta_0)]_{\mathcal{A}\mathcal{A}}$. Then using $\|\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}\|_2 = O_p((\frac{d_T}{T})^{\frac{1}{2}})$, we obtain

$$
\begin{aligned}
\mathbb{P}(\|\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}\|_2 > \beta) \leq\ & \mathbb{P}(\|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}}\|_2 + \|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\|_2\|\lambda_T T^{-1}\alpha_{T,\mathcal{A}} \odot \text{sgn}(\hat{\theta}_{\mathcal{A}})\|_2 \\
&+ \|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\|_2\|\gamma_T T^{-1}\varsigma_T\|_2 + \|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\|_2\|\nabla'_{\mathcal{A}}\{(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})\}/2\|_2 \\
&+ \|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\|_2\|\mathcal{P}_T(\theta_0)(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})\|_2 > \beta) \\
\leq\ & \mathbb{P}(\lambda^{-1}_{\min}(\mathbb{H}_T)\|\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}}\|_2 + \lambda^{-1}_{\min}(\mathbb{H}_T)\lambda_T T^{-1}\|\alpha_{T,\mathcal{A}}\|_2 \\
&+ \lambda^{-1}_{\min}(\mathbb{H}_T)\gamma_T T^{-1}\|\varsigma_T\|_2 + \lambda^{-1}_{\min}(\mathbb{H}_T)C_0^2(d_T/2T)\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}\}\|_2 \\
&+ \lambda^{-1}_{\min}(\mathbb{H}_T)C_0(d_T/T)^{1/2}\|\mathcal{P}_T(\theta_0)\|_2 > \beta) + \mathbb{P}(\|\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}\|_2 > (d_T/T)^{1/2}C_0),
\end{aligned}
$$

for $C_0 > 0$ large enough, and we used $\|\mathbb{H}_T^{-1}\boldsymbol{x}\|_2 \leq \lambda^{-1}_{\min}(\mathbb{H}_T)\|\boldsymbol{x}\|_2$ for any vector $\boldsymbol{x} \in \mathbb{R}^{d_T}$.

Let us proceed element-by-element. We have by the Markov inequality

$$
\mathbb{P}(\lambda^{-1}_{\min}(\mathbb{H}_T)C_0\sqrt{\frac{d_T}{T}}\|\mathcal{P}_T(\theta_0)\|_2 > \frac{\beta}{6}) \leq \frac{36\lambda^{-2}_{\min}(\mathbb{H}_T)C_0^2 d_T}{T\beta^2}\mathbb{E}[\|\mathcal{P}_T(\theta_0)\|_2^2].
$$

We have

$$
\mathbb{E}[\|\mathcal{P}_T\|_2^2] = \frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k,k'\in\mathcal{A}}\sum_{l,l'\in\mathcal{A}}\mathbb{E}[\zeta_{kl,t}\zeta_{k'l',t'}],
$$

where $\zeta_{kl,t} = \partial^2_{\theta_k\theta_l} l(\epsilon_t;\theta_0) - \mathbb{E}[\partial^2_{\theta_k\theta_l} l(\epsilon_t;\theta_0)]$. By assumption 32, we obtain

$$
\mathbb{P}(\lambda^{-1}_{\min}(\mathbb{H}_T)C_0\sqrt{\frac{d_T}{T}}\|\mathcal{P}_T(\theta_0)\|_2 > \frac{\beta}{6}) \leq \frac{36\lambda^{-2}_{\min}(\mathbb{H}_T)C_0^2}{\beta^2}\frac{d_T^3}{T^2}.
$$

As for the third order term, by the Markov inequality

$$\mathbb{P}(\frac{1}{2}\lambda_{\min}^{-1}(\mathbb{H}_T)C_0^2\frac{d_T}{T}\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}\}\|_2 > \frac{\beta}{6}) \leq \frac{9\lambda_{\min}^{-2}(\mathbb{H}_T)C_0^4 d_T^2}{T^2}\mathbb{E}[\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}\}\|_2^2].$$

We obtain

$$\begin{aligned}\mathbb{E}[\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}\}\|_2^2] &\leq \frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k_1,k_2,k_3\in\mathcal{A}}\sum_{l_1,l_2,l_3\in\mathcal{A}}\mathbb{E}[|\partial^3_{\theta_{k_1}\theta_{k_2}\theta_{k_3}} l(\epsilon_t;\theta_0).\partial^3_{\theta_{l_1}\theta_{l_2}\theta_{l_3}} l(\epsilon_t;\theta_0) l(\epsilon_{t'};\theta_0)|] \\ &\leq \frac{1}{T^2}d_T^3\sum_{t,t'=1}^{T}\mathbb{E}[\upsilon_t(C_0)\upsilon_{t'}(C_0)] = \eta(C_0)d_T^3,\end{aligned}$$

by assumption 33, where $\upsilon_t(C_0) = \sup_{k_1k_2k_3}\sup_{\theta:\|\theta-\theta_0\|_2\leq\sqrt{\frac{d_T}{T}}C_0}|\partial^3_{\theta_{k_1}\theta_{k_2}\theta_{k_3}} l(\epsilon_t;\theta_0)|$. We deduce that

$$\mathbb{P}(\frac{1}{2}\lambda_{\min}^{-1}(\mathbb{H}_T)C_0^2\frac{d_T}{T}\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}\}\|_2 > \frac{\beta}{6}) \leq \frac{9\lambda_{\min}^{-2}(\mathbb{H}_T)C_0^4}{4}\frac{d_T^5}{T^2}\eta(C_0).$$

We now turn to the score quantity. By the Markov inequality and assumption 31, we have

$$\begin{aligned}\mathbb{P}(\lambda_{\min}^{-1}(\mathbb{H}_T)\|\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}}\|_2 > \beta/6) &\leq \frac{\lambda_{\min}^{-2}(\mathbb{H}_T)36}{\beta^2}\mathbb{E}[\|\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}}\|_2] \\ &\leq \frac{\lambda_{\min}^{-2}(\mathbb{H}_T)36}{\beta^2}\frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k\in\mathcal{A}}\mathbb{E}[\partial_{\theta_k} l(\epsilon_t;\theta_0)\partial_{\theta_k} l(\epsilon_{t'};\theta_0)] \\ &\leq \frac{\lambda_{\min}^{-2}(\mathbb{H}_T)36}{\beta^2}\frac{1}{T}\{\frac{1}{T}\sum_{t,t'=1}^{T}\Psi(|t-t'|)\}d_T \\ &\leq \frac{\lambda_{\min}^{-2}(\mathbb{H}_T)36Kd_T}{T\beta^2},\end{aligned}$$

with $K > 0$. Hence we deduce

$$\begin{aligned}
\mathbb{P}(\|\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}\|_2 > \beta) &\leq \mathbb{P}(\lambda_{\min}^{-1}(\mathbb{H}_T)\lambda_T T^{-1}\|\alpha_{T,\mathcal{A}}\|_2 + \lambda_{\min}^{-1}(\mathbb{H}_T)\gamma_T T^{-1}\|\varsigma_T\|_2 > \beta/2) \\
&+ \mathbb{P}(\|\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}\|_2 > (d_T/T)^{1/2}C_0) + \mathbb{P}(\lambda_{\min}^{-1}(\mathbb{H}_T)C_0(d_T/T)^{1/2}\|\mathcal{P}_T(\theta_0)\|_2 > \beta/6) \\
&+ \mathbb{P}(\lambda_{\min}^{-1}(\mathbb{H}_T)C_0^2(d_T/2T)\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}\}\|_2 > \beta/6) \\
&+ \mathbb{P}(\lambda_{\min}^{-1}(\mathbb{H}_T)\|\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}}\|_2 > \beta/6) \\
&\leq \frac{2\lambda_{\min}^{-1}(\mathbb{H}_T)}{\beta}\{\lambda_T T^{-1} d_T^{1/2}\mathbb{E}[\max_{k\in\mathcal{S}, i\in\mathcal{A}_k}\alpha_{T,\mathcal{A}_k,i}] + \gamma_T T^{-1}\mathbb{E}[\max_{k\in\mathcal{S}}\xi_{T,k}]\} \\
&+ \frac{36\lambda_{\min}^{-2}(\mathbb{H}_T)Kd_T}{T\beta^2} + \frac{9\lambda_{\min}^{-2}(\mathbb{H}_T)C_0^4 d_T^5}{4T^2}\eta(C_0) + \frac{36\lambda_{\min}^{-2}(\mathbb{H}_T)C_0^2 d_T^3}{\beta^2 T^2} + \epsilon.
\end{aligned}$$

For T and C_0 large enough, if $d_T^5 = o(T)$, by assumption 34, that is if

$$\beta^{-1}T^{-1}\{\lambda_T d_T^{1/2}\mathbb{E}[\max_{k\in\mathcal{S}, i\in\mathcal{A}_k}\alpha_{T,\mathcal{A}_k,i}] + \gamma_T\mathbb{E}[\max_{k\in\mathcal{S}}\xi_{T,k}]\} \underset{T\to\infty}{\longrightarrow} 0,$$

then

$$\mathbb{P}(\|\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}\|_2 > \beta) \underset{T\to\infty}{\longrightarrow} 0.$$

We now turn to the second step of model selection consistency. First we prove

$$\mathbb{P}(\bigcap_{k\in\mathcal{S}^c}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 < 1\}) \underset{T\to\infty}{\longrightarrow} 1 \Leftrightarrow \mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \underset{T\to\infty}{\longrightarrow} 0. \tag{2.5.9}$$

This is equivalent to proving

$$\mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\dot{\mathbb{G}}_T l(\hat{\theta})_{(k)} + \lambda_T T^{-1}\alpha_T^{(k)} \odot \hat{\boldsymbol{w}}^{(k)}\|_2 \geq \gamma_T T^{-1}\xi_{T,k}\}) \underset{T\to\infty}{\longrightarrow} 0.$$

We have for $k \in \mathcal{S}^c$ that $\|\hat{\boldsymbol{w}}^{(k)}\|_\infty \leq 1$, which implies by the optimality conditions of Karush-Kuhn-Tucker that

$$\begin{aligned}
&\mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\dot{\mathbb{G}}_T l(\hat{\theta})_{(k)} + \lambda_T T^{-1}\alpha_T^{(k)} \odot \hat{\boldsymbol{w}}^{(k)}\|_2 \geq \gamma_T T^{-1}\xi_{T,k}\}) \\
&\leq \mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\dot{\mathbb{G}}_T l(\hat{\theta})_{(k)}\|_2 \geq \gamma_T T^{-1}\xi_{T,k} - \lambda_T T^{-1}\|\alpha_T^{(k)}\|_2\}).
\end{aligned}$$

By a Taylor expansion around θ_0, let $\bar{\theta}$ such that $\|\bar{\theta}-\theta_0\| \leq \|\hat{\theta}-\theta_0\|$, we have

$$\begin{aligned}
\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \leq \mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{&\|\dot{\mathbb{G}}_T l(\theta_0)_{(k)}\|_2 \geq \gamma_T T^{-1}\xi_{T,k} - \lambda_T T^{-1}\|\alpha_T^{(k)}\|_2 \\
- \ &\|\ddot{\mathbb{G}}_T l(\theta_0)_{(k)(k)}\|_2\|\hat{\theta}-\theta_0\|_2 - \|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{(k)(k)}\}_{(k)}\|_2\|\hat{\theta}-\theta_0\|_2^2\}) \\
\leq \ \mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{&\|\dot{\mathbb{G}}_T l(\theta_0)_{(k)}\|_2 \geq \gamma_T T^{-1}\|\tilde{\tilde{\theta}}^{(k)}\|_2^{-\mu} - \lambda_T T^{-1} d_T^{1/2} \max_{k\in\mathcal{S}^c, i\in\mathcal{G}_k}(|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta}) \\
- \ &\|\ddot{\mathbb{G}}_T l(\theta_0)_{(k)(k)}\|_2\|\hat{\theta}-\theta_0\|_2 - \|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{(k)(k)}\}_{(k)}\|_2\|\hat{\theta}-\theta_0\|_2^2\}),
\end{aligned}$$

where we used $\|\ddot{\mathbb{G}}_T l(\theta_0)_{(k)(k)}(\hat{\theta}-\theta_0)\|_2 \leq \|\ddot{\mathbb{G}}_T l(\bar{\theta})_{(k)(k)}\|_2\|\hat{\theta}-\theta_0\|_2$ and $\|\ddot{\mathbb{G}}_T l(\theta_0)_{(k)(k)}\|_2 = \|\ddot{\mathbb{G}}_T l(\theta_0)_{(k)(k)}\|_s$. Let $\epsilon > 0$, and K_ϵ strictly positive constants, we proved for T large enough that

$$\mathbb{P}(\|\hat{\theta}-\theta_0\|_2 > K_\epsilon (d_T/T)^{1/2}) < \epsilon/6.$$

We deduce that

$$\begin{aligned}
\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \leq \mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{&\|\dot{\mathbb{G}}_T l(\theta_0)_{(k)}\|_2 \geq \gamma_T T^{-1}\|\tilde{\tilde{\theta}}^{(k)}\|_2^{-\mu} - \lambda_T T^{-1} d_T^{1/2} \max_{k\in\mathcal{S}^c, i\in\mathcal{G}_k}(|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta}) \\
- \ &\|\ddot{\mathbb{G}}_T l(\theta_0)_{(k)(k)}\|_2 (d_T/T)^{1/2} K_\epsilon - \|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{(k)(k)}\}_{(k)}\|_2 (\frac{d_T}{T})^2 K_\epsilon^2\}) + \epsilon/6.
\end{aligned}$$

Let $M_{1,T} = (\frac{\gamma_T}{T})^{\frac{1}{1+\mu}}$, then we obtain

$$\begin{aligned}
\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \leq \sum_{k\in\mathcal{S}^c}\{&\mathbb{P}(\|\dot{\mathbb{G}}_T l(\theta_0)_{(k)}\|_2 \geq \gamma_T T^{-1}\|\tilde{\tilde{\theta}}^{(k)}\|_2^{-\mu} - \lambda_T T^{-1} d_T^{1/2} \max_{k\in\mathcal{S}^c, i\in\mathcal{G}_k}(|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta}) \\
- \ &\|\ddot{\mathbb{G}}_T l(\theta_0)_{(k)(k)}\|_2 (d_T/T)^{\frac{1}{2}} K_\epsilon - \|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{(k)(k)}\}_{(k)}\|_2 (\frac{d_T}{T})^2 K_\epsilon^2, \|\tilde{\tilde{\theta}}^{(k)}\|_2 \leq M_{1,T}) \\
+ \ &\mathbb{P}(\|\tilde{\tilde{\theta}}^{(k)}\|_2 > M_{1,T})\} + \epsilon/6.
\end{aligned}$$

Consequently, we have the relationship

$$\begin{aligned}
\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \ \leq \ &\sum_{k\in\mathcal{S}^c}\{\mathbb{P}(\|\dot{\mathbb{G}}_T l(\theta_0)_{(k)}\|_2 \geq \gamma_T T^{-1} M_{1,T}^{-\mu}/4) \\
+ \ &\mathbb{P}(\lambda_T T^{-1} d_T^{1/2} \max_{k\in\mathcal{S}^c, i\in\mathcal{G}_k}(|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta}) > \gamma_T T^{-1} M_{1,T}^{-\mu}/4) \\
+ \ &\mathbb{P}(\|\ddot{\mathbb{G}}_T l(\theta_0)_{(k)(k)}\|_2 (d_T/T)^{1/2} K_\epsilon > \gamma_T T^{-1} M_{1,T}^{-\mu}/4) \\
+ \ &\mathbb{P}(\|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{(k)(k)}\}_{(k)}\|_2 (\frac{d_T}{T})^2 K_\epsilon^2 > \gamma_T T^{-1} M_{1,T}^{-\mu}/4) \\
+ \ &\mathbb{P}(\|\tilde{\tilde{\theta}}^{(k)}\|_2 > M_{1,T})\} + \epsilon/6 := \sum_{i=1}^5 T_i + \epsilon/6.
\end{aligned}$$

We then focus on each T_i. We have by the Markov inequality

$$\begin{aligned} T_1 := \sum_{k \in \mathcal{S}^c} \mathbb{P}(\|\dot{\mathbb{G}}_T l(\theta_0)_{(k)}\|_2 > \gamma_T T^{-1} M_{1,T}^{-\mu}/4) &\leq \sum_{k \in \mathcal{S}^c} \frac{16\mathbb{E}[\|\dot{\mathbb{G}}_T l(\theta_0)_{(k)}\|_2^2]}{\{\gamma_T T^{-1} M_{1,T}^{-\mu}\}^2} \\ &\leq \frac{16\mathbb{E}[\|\dot{\mathbb{G}}_T l(\theta_0)\|_2^2]}{\{\gamma_T T^{-1} M_{1,T}^{-\mu}\}^2} \\ &\leq \frac{16 d_T}{T\{\gamma_T T^{-1} M_{1,T}^{-\mu}\}^2} \\ &= O((\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(1-c)-1]})^{-\frac{2}{1+\mu}}). \end{aligned}$$

Furthermore, using $|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta} \leq T^{\kappa\eta}$, we have for T_2 that

$$\begin{aligned} \mathbb{P}(\lambda_T T^{-1} d_T^{1/2} \max_{k \in \mathcal{S}^c, i \in \mathcal{G}_k} (|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta}) > \gamma_T T^{-1} M_{1,T}^{-\mu}/4) &\leq \mathbb{P}(\lambda_T T^{-1} d_T^{1/2} T^{\kappa\eta} > \gamma_T T^{-1} M_{1,T}^{-\mu}/4) \\ &\leq \mathbb{P}(\gamma_T T^{-1} M_{1,T}^{-\mu}/4\{1 - 4\lambda_T \gamma_T^{-1} d_T^{1/2} M_{1,T}^{\mu} T^{\kappa\eta}\} \quad (2.5.10) \end{aligned}$$

The quantity of interest is $\gamma_T \lambda_T^{-1} d_T^{-1/2} M_{1,T}^{-\mu} T^{-\kappa\eta}$ that has to converge to ∞ such that (2.5.10) converge to zero for T sufficiently large enough. We have

$$\gamma_T \lambda_T^{-1} d_T^{-1/2} M_{1,T}^{-\mu} T^{-\kappa\eta} \to \infty \Leftrightarrow \frac{\gamma_T}{\lambda_T^{1+\mu}} d_T^{-\frac{1+\mu}{2}} T^{-\kappa\eta(1+\mu)+\mu} \to \infty.$$

As for T_3, we have by the Markov inequality

$$
\begin{aligned}
&T_3 := \sum_{k\in\mathcal{S}^c} \mathbb{P}((\|\mathbb{H}_{T,(k)(k)}\|_2 + \|\mathcal{R}_{T,(k)}(\theta_0)\|_2)(d_T/T)^{1/2}K_\epsilon > \gamma_T T^{-1} M_{1,T}^{-\mu}/4) \\
&\leq \sum_{k\in\mathcal{S}^c} \mathbb{P}((\|\mathcal{R}_{T,(k)}(\theta_0)\|_2 (d_T/T)^{1/2}K_\epsilon > \gamma_T T^{-1} M_{1,T}^{-\mu}/4 - \|\mathbb{H}_{T,(k),(k)}\|_2 (d_T/T)^{1/2}K_\epsilon) \\
&\leq \sum_{k\in\mathcal{S}^c} \{\mathbb{P}((\|\mathcal{R}_{T,(k)}(\theta_0)\|_2 (d_T/T)^{1/2}K_\epsilon > \gamma_T T^{-1} M_{1,T}^{-\mu}/8) \\
&+\mathbb{P}(\|\mathbb{H}_{T,(k),(k)}\|_2 (d_T/T)^{1/2}K_\epsilon > \gamma_T T^{-1} M_{1,T}^{-\mu}/8)\} \\
&\leq \sum_{k\in\mathcal{S}^c} \{\frac{64K_\epsilon^2 d_T \mathbb{E}[\|\mathcal{R}_{T,(k)}(\theta_0)\|_2^2]}{T\gamma_T^2 T^{-2} M_{1,T}^{-2\mu}} + \frac{64K_\epsilon^2 d_T \|\mathbb{H}_{T,(k)(k)}\|_2^2}{T\gamma_T^2 T^{-2} M_{1,T}^{-2\mu}}\} \\
&\leq \frac{64K_\epsilon^2 d_T \|\mathbb{H}_T\|_2^2}{\gamma_T^2 T^{-1} M_{1,T}^{-2\mu}} + \frac{64K_\epsilon^2 \mathbb{E}[\|\mathcal{R}_T(\theta_0)\|_2^2]}{\gamma_T^2 M_{1,T}^{-2\mu}} \\
&\leq \frac{64K_\epsilon^2 d_T \lambda_{\max}^2(\mathbb{H}_T)}{\gamma_T^2 T^{-1} M_{1,T}^{-2\mu}} + \frac{64K_\epsilon^2 d_T^3}{\gamma_T^2 M_{1,T}^{-2\mu}} \\
&\leq \frac{64K_\epsilon^2 \lambda_{\max}^2(\mathbb{H}_T)}{\{\gamma_T T^{-1/2} d_T^{-1/2} M_{1,T}^{-\mu}\}^2} + \frac{64K_\epsilon^2}{\{\gamma_T d_T^{-3/2} M_{1,T}^{-\mu}\}^2} \\
&= O((\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(1-c)-1]})^{-\frac{2}{1+\mu}}) + O((\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(2-3c)-1]})^{-\frac{2}{1+\mu}}).
\end{aligned}
$$

We obtain for T_4 by the Markov inequality

$$
\begin{aligned}
&T_4 := \sum_{k\in\mathcal{S}^c} \mathbb{P}(\|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{(k)(k)}\}_{(k)}\|_2 (\frac{d_T}{T})^2 K_\epsilon^2 > \gamma_T T^{-1} M_{1,T}^{-\mu}/4) \\
&\leq \sum_{k\in\mathcal{S}^c} \frac{16K_\epsilon^4 d_T^2 \mathbb{E}[\|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\theta})_{(k)(k)}\}_{(k)}\|_2^2]}{T^2 \gamma_T T^{-2} M_{1,T}^{-2\mu}} \\
&\leq \frac{16K_\epsilon^4 d_T^5 \mathbb{E}[\|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\theta})\}\|_2^2]}{\gamma_T^2 M_{1,T}^{-2\mu}} \\
&\leq \frac{16K_\epsilon^4 d_T^5 \eta(K_\epsilon)}{\gamma_T^2 M_{1,T}^{-2\mu}} = \frac{16K_\epsilon^4 \eta(K_\epsilon)}{\{\gamma_T d_T^{-5/2} M_{1,T}^{-\mu}\}^2} = O((\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(2-5c)-1]})^{-\frac{2}{1+\mu}}).
\end{aligned}
$$

Finally, we have for T_5 that

$$
\begin{aligned}
T_5 := \sum_{k\in\mathcal{S}^c} \mathbb{P}(\|\tilde{\tilde{\theta}}^{(k)}\|_2 > M_{1,T}) &\leq \sum_{k\in\mathcal{S}^c} \frac{\mathbb{E}[\|\tilde{\tilde{\theta}}^{(k)}\|_2^2]}{M_{1,T}^2} \\
&\leq \frac{\mathbb{E}[\|\tilde{\tilde{\theta}} - \theta_0\|_2^2]}{M_{1,T}^2} \\
&= O((\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(1-c)-1]})^{-\frac{2}{1+\mu}}).
\end{aligned}
$$

Hence we obtain from these relationships and using assumption 35

$$\frac{\gamma_T}{\lambda_T^{1+\mu}} T^{\mu-(\frac{c}{2}+\kappa\eta)(1+\mu)} \underset{T\to\infty}{\longrightarrow} \infty,$$
$$\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(1-c)-1]} \underset{T\to\infty}{\longrightarrow} \infty,$$

such that the latter implies

$$\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(2-3c)-1]} \underset{T\to\infty}{\longrightarrow} \infty,$$
$$\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(2-5c)-1]} \underset{T\to\infty}{\longrightarrow} \infty.$$

Consequently each T_i converges to zero for T large enough. Hence

$$\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \leq \sum_{i=1}^{5} T_i + \epsilon/6 \underset{T\to\infty}{\longrightarrow} \epsilon.$$

For $\epsilon \to 0$, we prove $\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \to 0$ for T large enough.

As for the second part of the model selection procedure, we prove that

$$\mathbb{P}(\underset{k\in\mathcal{S}}{\cap}\underset{i\in\mathcal{A}_k^c}{\cap}\{|\hat{\boldsymbol{w}}_i^{(k)}| < 1\}) \underset{T\to\infty}{\longrightarrow} 1 \Leftrightarrow \mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\hat{\boldsymbol{w}}_i^{(k)}| \geq 1\}) \underset{T\to\infty}{\longrightarrow} 0.$$

By the optimality conditions, we have

$$\mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\hat{\boldsymbol{w}}_i^{(k)}| \geq 1\}) = \mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\dot{\mathbb{G}}_T l(\hat{\theta})_{(k),i}| \geq \lambda_T T^{-1}\alpha_{T,i}^{(k)}\}).$$

Then by a Taylor expansion around θ_0, with $\bar{\theta}$ between $\hat{\theta}$ and θ_0, we have

$$\begin{aligned}
\mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\hat{\boldsymbol{w}}_i^{(k)}| \geq 1\}) &= \mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\dot{\mathbb{G}}_T l(\theta_0)_{(k),i} + [\sum_j \partial_{ij}^2 \mathbb{G}_T l(\theta_0)(\hat{\theta}_j - \theta_{0,j})]_i \\
&+ [\sum_{j,k} T^{-1}\sum_{t=1}^{T} \partial_{ijk}^3 l(\epsilon_t;\bar{\theta})(\hat{\theta}_j - \theta_{0,j})^2/2]_i| \geq \lambda_T T^{-1}\alpha_{T,i}^{(k)}\}) \\
&\leq \mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\dot{\mathbb{G}}_T l(\theta_0)_{(k),i}| \geq \lambda_T T^{-1}\alpha_{T,i}^{(k)} - [\sum_j \partial_{ij}^2 \mathbb{G}_T l(\theta_0)(\hat{\theta}_j - \theta_{0,j})]_i \\
&- [\sum_{j,k} T^{-1}\sum_{t=1}^{T} \partial_{ijk}^3 l(\epsilon_t;\bar{\theta})(\hat{\theta}_j - \theta_{0,j})^2/2]_i|\}).
\end{aligned}$$

Let $M_{2,T} = (\frac{\lambda_T}{T})^{\frac{1}{1+\eta}}$. Then using $\|\hat{\theta} - \theta_0\|_2 = O_p((\frac{d_T}{T})^{\frac{1}{2}})$ and the Cauchy-Schwartz inequality, we obtain

$$
\begin{aligned}
\mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\hat{\boldsymbol{w}}_i^{(k)}| \geq 1\}) &\leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\{\mathbb{P}(|\dot{\mathbb{G}}_T l(\theta_0)_{(k),i}| \geq \lambda_T T^{-1}\alpha_{T,i}^{(k)} - [\sum_j \partial_{ij}^2 \mathbb{G}_T l(\theta_0)(\hat{\theta}_j - \theta_{0,j})]_i \\
&- [\sum_{j,k} T^{-1}\sum_{t=1}^T \partial_{ijk}^3 l(\epsilon_t;\bar{\theta})(\hat{\theta}_j - \theta_{0,j})^2/2]_i|, |\tilde{\tilde{\theta}}_i^{(k)}| \leq M_{2,T}) + \mathbb{P}(|\tilde{\tilde{\theta}}_i^{(k)}| > M_{2,T})\} \\
&\leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\{\mathbb{P}(|\dot{\mathbb{G}}_T l(\theta_0)_{(k),i}| \geq \lambda_T T^{-1} M_{2,T}^{-\eta} - \{\sum_j (\partial_{ij}^2 \mathbb{G}_T l(\theta_0))^2\}^{1/2} K_\epsilon (d_T/T)^{1/2} \\
&- \{\sum_{j,k,l,m} T^{-2}\sum_{t,t'=1}^T \partial_{ijk}^3 l(\epsilon_t;\bar{\theta})\partial_{ilm}^3 l(\epsilon_{t'};\bar{\theta})\}^{1/2} K_\epsilon^2 (d_T/T)) \\
&+ \mathbb{P}(|\tilde{\tilde{\theta}}_i^{(k)}| > M_{2,T})\} + \epsilon/5 \\
&\leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\{\mathbb{P}(|\dot{\mathbb{G}}_T l(\theta_0)_{(k),i}| \geq \lambda_T T^{-1} M_{2,T}^{-\eta}/3) \\
&+ \mathbb{P}(\{\sum_j (\partial_{ij}^2 \mathbb{G}_T l(\theta_0))^2\}^{1/2} K_\epsilon (d_T/T)^{1/2} > \lambda_T T^{-1} M_{2,T}^{-\eta}/3) \\
&+ \mathbb{P}(\{\sum_{j,k,l,m} T^{-2}\sum_{t,t'=1}^T \partial_{ijk}^3 l(\epsilon_t;\bar{\theta})\partial_{ilm}^3 l(\epsilon_{t'};\bar{\theta})\}^{1/2} K_\epsilon^2 (d_T/T) > \lambda_T T^{-1} M_{2,T}^{-\eta}/3) \\
&+ \mathbb{P}(|\tilde{\tilde{\theta}}_i^{(k)}| > M_{2,T})\} + \epsilon/5 := \sum_{i=1}^4 T_i + \epsilon/5.
\end{aligned}
$$

We proceed as for inactive groups. For T_1, we have by the Markov inequality

$$
\begin{aligned}
T_1 := \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\mathbb{P}(|\dot{\mathbb{G}}_T l(\theta_0)_{(k),i}| \geq \lambda_T T^{-1} M_{2,T}^{-\eta}/3) &\leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\frac{9\mathbb{E}[|\dot{\mathbb{G}}_T l(\theta_0)_{(k),i}|^2]}{\{\lambda_T T^{-1} M_{2,T}^{-\eta}\}^2} \\
&\leq \frac{9\mathbb{E}[\|\dot{\mathbb{G}}_T l(\theta_0)\|_2^2]}{\{\lambda_T T^{-1} M_{2,T}^{-\eta}\}^2} \\
&= O((\frac{\lambda_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\eta)(1-c)-1]})^{-\frac{2}{1+\eta}}).
\end{aligned}
$$

As for T_2, we have

$$
\begin{aligned}
T_2 &:= \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\mathbb{P}(\{\sum_j(\partial^2_{ij}\mathbb{G}_T l(\theta_0))^2\}^{1/2}K_\epsilon(d_T/T)^{1/2} > \lambda_T T^{-1}M_{2,T}^{-\eta}/3)\\
&= \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\mathbb{P}((\{\sum_j\mathcal{P}_{T,(k),j}(\theta_0)\}^{1/2}+\{\sum_j\mathbb{H}^2_{T,(k),j}\}^{1/2})K_\epsilon(d_T/T)^{1/2} > \lambda_T T^{-1}M_{2,T}^{-\eta}/3)\\
&\leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\sum_j\{\frac{36d_T\mathbb{E}[\mathcal{P}^2_{T,(k),j}(\theta_0)]}{T\{\lambda_T T^{-1}M_{2,T}^{-\eta}\}^2}\}+\frac{36d_T\|\mathbb{H}_T\|_2^2}{T\{\lambda_T T^{-1}M_{2,T}^{-\eta}\}^2}\\
&\leq \frac{36d_T\lambda^2_{\max}(\mathbb{H}_T)}{T\{\lambda_T T^{-1}M_{2,T}^{-\eta}\}^2}+\frac{36d_T\mathbb{E}[\|\mathcal{P}_T(\theta_0)\|_2^2]}{T\{\lambda_T T^{-1}M_{2,T}^{-\eta}\}^2}\\
&= O((\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(1-c)-1]})^{-\frac{2}{1+\eta}})+O((\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(2-3c)-1]})^{-\frac{2}{1+\eta}}).
\end{aligned}
$$

Furthermore, for the third order term in T_3, we have

$$
\begin{aligned}
T_3 &:= \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\mathbb{P}(\{\sum_{j,k,l,m}T^{-2}\sum_{t,t'=1}^T\partial^3_{ijk}l(\epsilon_t;\bar\theta)\partial^3_{ilm}l(\epsilon_{t'};\bar\theta)\}^{1/2}K^2_\epsilon(d_T/T) > \lambda_T T^{-1}M_{2,T}^{-\eta}/3)\\
&\leq \frac{9d_T^2\mathbb{E}[\|\nabla'\{\ddot{\mathbb{G}}_T l(\bar\theta\}\|_2^2]}{T^2\{\lambda_T T^{-1}M_{2,T}^{-\eta}\}^2} = O((\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(2-5c)-1]})^{-\frac{2}{1+\eta}}).
\end{aligned}
$$

Finally, we have for T_4 that

$$
\begin{aligned}
T_4 := \sum_{i\in\mathcal{A}_k^c}\mathbb{P}(|\tilde{\tilde\theta}_i^{(k)}| > M_{2,T}) &\leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\frac{\mathbb{E}[|\tilde{\tilde\theta}_i^{(k)}|^2]}{M^2_{2,T}}\\
&\leq \frac{\mathbb{E}[\|\tilde{\tilde\theta}-\theta_0\|_2^2]}{M^2_{2,T}} = O((\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(1-c)-1]})^{-\frac{2}{1+\eta}}).
\end{aligned}
$$

We have from these relationships and by assumption 35, $\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(1-c)-1]}\underset{T\to\infty}{\longrightarrow}\infty$ implies

$$
\begin{aligned}
&\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(2-3c)-1]}\underset{T\to\infty}{\longrightarrow}\infty,\\
&\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(2-5c)-1]}\underset{T\to\infty}{\longrightarrow}\infty.
\end{aligned}
$$

We deduce

$$
\mathbb{P}(\bigcup_{k\in\mathcal{S}}\bigcup_{i\in\mathcal{A}_k^c}\{|\hat{\boldsymbol{w}}_i^{(k)}|\geq 1\})\underset{T\to\infty}{\longrightarrow}\epsilon,
$$

for T sufficiently large enough. We have then concluded the model selection consistency.

We now focus on the asymptotic normality. Model selection implies that

$$\mathbb{P}(\{k \in \mathcal{S}, i \in \mathcal{A}_k, : \hat{\theta}_i^{(k)} \neq 0\} = \mathcal{A}) \underset{T\to\infty}{\longrightarrow} 1.$$

As a consequence, the next relationship holds

$$\mathbb{P}(\forall k \in \mathcal{S}, \dot{\mathbb{G}}_T l(\hat{\theta})_{\mathcal{A}_k} + \lambda_T T^{-1} \alpha_{T,\mathcal{A}_k} \odot \operatorname{sgn}(\hat{\theta}_{\mathcal{A}_k}) + \gamma_T T^{-1} \xi_{T,k} \frac{\hat{\theta}_{\mathcal{A}_k}}{\|\hat{\theta}_{\mathcal{A}_k}\|_2} = 0) \underset{T\to\infty}{\longrightarrow} 1.$$

By a Taylor expansion of the gradient term around $\theta_{0,\mathcal{A}}$, we obtain

$$\begin{aligned}\mathbb{P}(\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}} + \ddot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}\mathcal{A}}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) &+ \frac{1}{2}\nabla'\{(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})\} \\ &+ \lambda_T T^{-1}\alpha_{T,\mathcal{A}} \odot \operatorname{sgn}(\hat{\theta}_{\mathcal{A}}) + \gamma_T T^{-1}\eta_T = 0) \underset{T\to\infty}{\longrightarrow} 1,\end{aligned}$$

where $\eta_T = \operatorname{vec}(\xi_{T,k}\frac{\hat{\theta}_{\mathcal{A}_k}}{\|\hat{\theta}_{\mathcal{A}_k}\|_2}, k \in \mathcal{S})$ and $\|\bar{\theta} - \theta_0\|_2 \leq \|\hat{\theta} - \theta_0\|_2$. As a consequence, we have

$$\begin{aligned}\mathcal{P}(\theta_0)(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) + \mathbb{H}_{T,\mathcal{A}\mathcal{A}}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) &= -\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}} - \frac{1}{2}\nabla'\{(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})\} \\ &- \lambda_T T^{-1}\alpha_{T,\mathcal{A}} \odot \operatorname{sgn}(\hat{\theta}_{\mathcal{A}}) - \gamma_T T^{-1}\eta_T + o_p(1),\end{aligned}$$

where $\mathcal{P}(\theta_0) = \ddot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}\mathcal{A}} - \mathbb{H}_{T,\mathcal{A}\mathcal{A}}$ and $\mathbb{H}_{T,\mathcal{A}\mathcal{A}} = \mathbb{E}[\nabla^2_{\theta\theta'} l(\epsilon_t;\theta_0)]_{\mathcal{A}\mathcal{A}}$. Then multiplying by $\sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}$, we obtain

$$\begin{aligned}\sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) &= -\sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}(\lambda_T T^{-1}\alpha_{T,\mathcal{A}} \odot \operatorname{sgn}(\hat{\theta}_{\mathcal{A}}) + \gamma_T T^{-1}\eta_T) \\ &- \sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}} \\ &- \sqrt{T}/2Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}\nabla'\{(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})\} \\ &- \sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}\mathcal{P}(\theta_0)(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) + o_p(1).\end{aligned}$$

We focus on the l^1 penalty term, which can be upper bounded as

$$\begin{aligned}N_{1,T} := |\sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}(\frac{\lambda_T}{T}\alpha_{T,\mathcal{A}} \odot \operatorname{sgn}(\hat{\theta}_{\mathcal{A}}))| &\leq |Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}||\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}|\lambda_T T^{-1/2}\max_{k\in\mathcal{S},i\in\mathcal{A}_k}\alpha_{T,i,\mathcal{A}} \\ &\leq |Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}|\lambda_{\min}^{-1}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}})\lambda_T T^{-1/2}\{\min_{k\in\mathcal{S},i\in\mathcal{A}_k}|\tilde{\tilde{\theta}}_i^{(k)}|\}^{-\eta} \\ &\leq |Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}|\lambda_{\min}^{-1}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}})\lambda_T T^{\kappa\eta-\frac{1}{2}}.\end{aligned}$$

If $\lambda_T T^{\kappa\eta} \to 0$, then $N_{1,T} = o_p(1)$.

As for the l^1/l^2 penalty, it can be upper bounded as

$$\begin{aligned} N_{2,T} := |\sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}\frac{\gamma_T}{T}\eta_T| &\leq |Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}||\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}|\gamma_T T^{-1/2}\|\eta_T\|_2 \\ &\leq |Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}||\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}|\gamma_T T^{-1/2}\sqrt{\sum_{k\in\mathcal{S}}\|\tilde{\tilde{\theta}}^{(k)}\|_2^{-2\mu}} \\ &\leq |Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}|\lambda_{\min}^{-1}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}})\gamma_T T^{-1/2}d_T^{1/2}\{\min_{k\in\mathcal{S}}\|\tilde{\tilde{\theta}}^{(k)}\|_2\}^{-\mu} \\ &\leq |Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}|\lambda_{\min}^{-1}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}})\gamma_T T^{-1/2}d_T^{1/2}T^{\kappa\mu}. \end{aligned}$$

Using $d_T = O(T^c)$, if $\gamma_T T^{\frac{c-1}{2}+\kappa\mu} \to 0$, then $N_{2,T} = o_p(1)$. Consequently, we have $N_{1,T} + N_{2,T} = o_p(1)$.

We now turn to the hessian quantity of the Taylor expansion and prove the discrepancy $\mathcal{P}(\theta_0)$ converges uniformly to zero in probability. For any $\epsilon > 0$, by the Markov's inequality, we have

$$\begin{aligned} \mathbb{P}(\|\ddot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}\mathcal{A}} - \mathbb{H}_{T,\mathcal{A}\mathcal{A}}\|_2^2 > (\epsilon/d_T)^2) &\leq \frac{d_T^2}{\epsilon^2 T^2}\mathbb{E}[\sum_{(k,l)\in\mathcal{A}}\{\partial^2_{\theta_k\theta_l}l(\epsilon_t;\theta_0) - \mathbb{E}[\nabla^2_{\theta_k\theta_l}l(\epsilon_t;\theta_0)]\}^2] \\ &\leq \frac{d_T^4}{\epsilon^2 T^2}\lambda^2_{\max}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}}). \end{aligned}$$

As for the third order term, by the Cauchy-Schwartz inequality

$$\begin{aligned} \|\nabla'\{(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\theta})_{\mathcal{A}\mathcal{A}}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}})\}\|_2^2 &\leq \frac{1}{T^2}\sum_{t=1}^T\{\sum_{(k,l,m)\in\mathcal{A}}\partial^3_{\theta_k\theta_l\theta_m}l_T^2(\epsilon_t;\bar{\theta})\}\|\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}\|_2^4 \\ &\leq \frac{1}{T^2}\sum_{t=1}^T\{\sum_{(k,l,m)\in\mathcal{A}}\psi_T^2(\epsilon_t)\}\|\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}\|_2^4 \\ &= O_p(\frac{d_T^5}{T^2}) = o_p(\frac{1}{T}). \end{aligned}$$

We now prove $X_{T,t} = \sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}\dot{\mathbb{G}}_T l_t(\theta_0)_{\mathcal{A}}, t = 1, \cdots, T$, is asymptotically normal by checking the Lindeberg-Feller's condition for applying Shiryaev's Theorem 2.5.17. We remind that $\dot{\mathbb{G}}_T l_{T,t}(\theta_0)$ is the t-th point of the score of the empirical criterion. Let $\beta > 0$, and to use Shiryaev's Theorem, we need to prove that for any $\epsilon > 0$, we have

$$\mathbb{P}(\sum_{t=0}^T\mathbb{E}[\|X_{T,t}\|_2^2\mathbf{1}_{\|\mathbf{X_{T,t}}\|_\mathbf{2}>\beta}|\mathcal{F}^{\mathbf{T}}_{\mathbf{t-1}}] > \epsilon) \underset{\mathbf{T}\to\infty}{\longrightarrow} \mathbf{0}.$$

By the Markov inequality, we obtain

$$\begin{aligned}\mathbb{P}(\sum_{t=0}^{T}\mathbb{E}[\|X_{T,t}\|_2^2\mathbf{1}_{\|\mathbf{X_{T,t}}\|_\mathbf{2}>\beta}|\mathcal{F}_{\mathbf{t-1}}^{\mathbf{T}}]>\epsilon) &\leq \tfrac{1}{\epsilon}\sum_{t=0}^{T}\mathbb{E}[\|X_{T,t}\|_2^2\mathbf{1}_{\|\mathbf{X_{T,t}}\|_\mathbf{2}>\beta}|\mathcal{F}_{\mathbf{t-1}}^{\mathbf{T}}]\\ &\leq \tfrac{1}{\epsilon}\sum_{t=0}^{T}\mathbb{E}[\mathbb{E}[\|X_{T,t}\|_2^4|\mathcal{F}_{t-1}^{T}]^{1/2}\mathbb{P}(\|X_{T,t}\|_2>\beta|\mathcal{F}_{t-1}^{T})^{1/2}]\\ &\leq \tfrac{1}{\epsilon}\sum_{t=0}^{T}\mathbb{E}[\{\tfrac{C_{st}}{T^2}\mathbb{E}[\|\nabla l(\epsilon_t;\theta_0)\nabla' l(\epsilon_t;\theta_0)\|_2^2|\mathcal{F}_{t-1}^{T}]\}^{1/2}\\ &\quad .\tfrac{1}{\beta}\mathbb{E}[\|\sqrt{T}Q_T\mathbb{V}_{T,\mathcal{AA}}^{-1/2}\mathbb{H}_{T,\mathcal{AA}}^{-1}\dot{\mathbb{G}}_T l_t(\theta_0)_{\mathcal{A}}\|_2^2|\mathcal{F}_{t-1}^{T}]^{1/2}],\end{aligned}$$

with $C_{st}>0$. First, let $\mathbb{K}_T=Q_T\mathbb{V}_{T,\mathcal{AA}}^{-1/2}\mathbb{H}_{T,\mathcal{AA}}^{-1}$, we have

$$\begin{aligned}\mathbb{E}[\|\sqrt{T}\mathbb{K}_T\dot{\mathbb{G}}_T l_t(\theta_0)_{\mathcal{A}}\|_2^2|\mathcal{F}_{t-1}^{T}] &= \frac{1}{T}\mathbb{E}[\nabla' l(\epsilon_t;\theta_0)\mathbb{K}_T'\mathbb{K}_T\nabla l(\epsilon_t;\theta_0)|\mathcal{F}_{t-1}^{T}]\\ &= \frac{1}{T}\mathbb{E}[\mathrm{Trace}(\nabla' l(\epsilon_t;\theta_0)\mathbb{K}_T'\mathbb{K}_T\nabla l(\epsilon_t;\theta_0))|\mathcal{F}_{t-1}^{T}]\\ &= \frac{1}{T}\mathrm{Trace}(\mathbb{E}[\nabla l(\epsilon_t;\theta_0)\nabla' l(\epsilon_t;\theta_0)|\mathcal{F}_{t-1}^{T}]\mathbb{K}_T'\mathbb{K}_T)\leq\frac{1}{T}\lambda_{\max}(\mathbb{H}_{t-1}^{T})\tilde{C}_{st},\end{aligned}$$

where $\tilde{C}_{st}>0$. Furthermore, we have

$$\begin{aligned}\mathbb{E}[\|\nabla l(\epsilon_t;\theta_0)\nabla' l(\epsilon_t;\theta_0)\|_2^2|\mathcal{F}_{t-1}^{T}] &= \mathbb{E}[\sum_{i,j=0}^{d_T}\{\partial_{\theta_i}l(\epsilon_t;\theta_0)\partial_{\theta_j}l(\epsilon_t;\theta_0)\}^2|\mathcal{F}_{t-1}^{T}]\\ &\leq d_T^2\sup_{i,j=1,\cdots,d_T}\mathbb{E}[\{\partial_{\theta_i}l(\epsilon_t;\theta_0)\partial_{\theta_j}l(\epsilon_t;\theta_0)\}^2|\mathcal{F}_{t-1}^{T}].\end{aligned}$$

By assumption 36, we have

$$\begin{aligned}&\mathbb{P}(\sum_{t=0}^{T}\mathbb{E}[\|X_{T,t}\|_2^2\mathbf{1}_{\|\mathbf{X_{T,t}}\|_\mathbf{2}>\beta}|\mathcal{F}_{\mathbf{t-1}}^{\mathbf{T}}]>\epsilon)\\ &\leq\frac{C_{st}^{\frac{1}{2}}\tilde{C}_{st}^{\frac{1}{2}}d_T}{T^{\frac{3}{2}}}\sum_{t=0}^{T}\mathbb{E}[\sup_{i,j=1,\cdots,d_T}\mathbb{E}[\{\partial_{\theta_i}l(\epsilon_t;\theta_0)\partial_{\theta_j}l(\epsilon_t;\theta_0)\}^2|\mathcal{F}_{t-1}^{T}]\lambda_{\max}(\mathbb{H}_{t-1}^{T})]\leq\frac{C_{st}^{\frac{1}{2}}\tilde{C}_{st}^{\frac{1}{2}}\bar{B}Td_T}{T^{\frac{3}{2}}}.\end{aligned}$$

Consequently, we obtain

$$\sum_{t=0}^{T}\mathbb{E}[\|X_{T,t}\|_2^2\mathbf{1}_{\|\mathbf{X_{T,t}}\|_\mathbf{2}>\beta}|\mathcal{F}_{\mathbf{t-1}}^{\mathbf{T}}]=\mathbf{o_p(1)}.$$

We deduce that $X_{T,t}$ satisfies the Lindeberg-Feller condition, and by Theorem 2.5.17, $\sqrt{T}Q_T\mathbb{V}_{T,\mathcal{AA}}^{-1/2}\mathbb{H}_{T,\mathcal{AA}}^{-1}\dot{\mathbb{G}}_T l(\theta_0)_{\mathcal{A}}$ is asymptotically normally distributed. The asymptotic distribution of Theorem 2.5.24 follows.

□

2.6 Simulation experiments

In this section, we carry out a simulation study to explore the finite sample performance of the adaptive Sparse Group Lasso. We first focus on the calibration of the adaptive weights entering the penalties. The regularization parameters must satisfy conditions to satisfy the oracle property in the double-asymptotic case. To do so, we suppose $\lambda_T = T^\beta$ and $\gamma_T = T^\alpha$, where β and α are both strictly positive constant. Regarding assumption 35, we obtain the conditions

$$\begin{cases} \alpha + \frac{c}{2} + \kappa\mu - \frac{1}{2} < 0, \\ \alpha - \frac{1}{2} + \frac{1}{2}[(1+\mu)(1-c) - 1] > 0, \\ \beta + \kappa\eta - \frac{1}{2} < 0, \\ \beta - \frac{1}{2} + \frac{1}{2}[(1+\eta)(1-c) - 1] > 0, \\ (1+\mu)[1 - \frac{c}{2} - \kappa\eta - \beta] + \alpha - 1 > 0. \end{cases}$$

This system allows for flexibility when choosing μ and η once κ, c, α and β are fixed. For instance, for $c = 1/6$, $\kappa = 0.05$, $\alpha = 1/10$ and $\beta = 1/10$, then $\mu \in [0.4, 6.3]$ and $\eta \in [0.6, 7.9]$. If $\alpha = \beta = 1/5$ and for $c = 1/6$ and $\kappa = 0.05$, then $\mu \in [0.4, 4.3]$ and $\eta \in [0.3, 5.9]$.

We consider 6 methods in the experiment: the Lasso (L), the Adaptive Lasso (AL), the Group Lasso (GL), the Adaptive Group Lasso (AGL), the Sparse Group Lasso (SGL) and the Adaptive Sparse Group Lasso (ASGL).

There are several methods to numerically solve the non-differentiable statistical problem (2.5.1). Fan and Li (2001) proposed a local quadratic approximation (LQA) of the first order derivative of the penalty function and a Newton-Raphson type algorithm. To circumvent numerical instability, they suggest to shrunk to zero coefficients that are close to zero, that is a coefficient $|\theta_j| < \epsilon$, with $\epsilon > 0$ to be calibrated. The drawback is that once it is set to zero, it will be excluded at any step of the LQA algorithm. Hunter and Li (2005) proposed a more sophisticated version of the LQA algorithm to avoid the drawback of the stepwise selection and numerical instability. They also studied the convergence properties of the LQA method. Zou and Li (2008) proposed a local linear approximation (LLA) of the penalty function such that the estimated coefficients have naturally a sparse representation, under the condition that the penalty function satisfies the continuity condition. Zou (2006) or Zou and Zhang (2009) used

the LQA algorithm for their empirical study. Other approaches are also possible such as gradient descent methods.

When one consider the OLS loss function, closed form algorithms can be applied to our problem. Bühlmann and van de Geer (2011) compiled these methodologies for solving the Lasso and the Group Lasso using gradient descent methods for general penalized convex empirical function. We used these algorithms in our study for solving the group Lasso. As for the Lasso, we applied the shooting algorithm developed by Fu (1998), which is a particular case of the gradient descent method. Simon and al. (2013) proposed an algorithm for solving the SGL that can accommodate likelihood criteria. This is a "two-step" method, where we first check whether the group is active, and then, if active, check if the coefficient within this group is active. In this simulation study, we used the alternative direction method of multipliers provided by Li, Mo, Yuan and Zhang (2014) since it provides better convergence performances.

We used a cross-validation procedure to select both parameters λ_T and γ_T such that both terms are defined by $\lambda_T = T^\beta$ and $\gamma_T = T^\alpha$, and $\beta = \alpha = 1/8$. The adaptive weights are computed as follows: we first compute an OLS estimator $\tilde{\theta}$ such that the adaptive weights entering the penalties correspond to $\tilde{\tilde{\theta}} = \tilde{\theta} + T^{-\kappa}$, with $\kappa = 0.2$. As for the adaptive weights, they are chosen such that the above system is satisfied: we set $\eta = 3.5$ and $\mu = 2.5$.

We report the variable selection performance through the number of zero coefficients correctly estimated, denoted as C and, the number of nonzero coefficients incorrectly estimated, denoted IC. Besides, the mean squared error is reported as an estimation accuracy measure.

Simulated experiment 1. We consider a data generating process

$$y = \sum_l \beta_0^{(l)} \mathbf{X^{(l)}} + \sigma\eta,$$

where η is a strong white noise, normally distributed, centered with unit variance and $\sigma = 0.3$. The matrices $\mathbf{X^{(l)}}$ follow $\boldsymbol{c}_l$- dimensional multivariate normal distributions, centered and with variance covariance $\Sigma^{(l)}$ such that the entries are defined as $\Sigma_{ij}^{(l)} = \rho_{(l)}^{|i-j|}, 1 \leq j, i \leq \boldsymbol{c}_l$ with $\rho_{(l)} \in \mathcal{U}([0.5, 0.9])$ for each group. Moreover, the dimension $d_T = [x \times T^{1/6}]$ with $T = 500, 2000, 4000$ and $x = 10, 30, 50$ respectively for the values of T. As $d_T = O(T^c)$ with $c = 1/6$, we can multiply by x to consider more realistic

settings. The number of groups is defined as $N_g = 4$ (resp. $N_g = 8$, resp. $N_g = 18$) for $T = 500$ (resp. for $T = 2000$, resp. for $T = 4000$) and the size of each of them is randomly chosen among $\{5, \cdots, 30\}$. The number of active groups is defined as $|\mathcal{S}| = 2a_T$ with $a_T = [N_g/3]$. Moreover, zero coefficients are randomly chosen among the whole vector β for active groups, such that the total number of zeros -both the zero subvectors for inactive groups and zero components for active groups - matches the total number of inactive indices. The total number of active indices is defined as $|\mathcal{A}| = 3b_T$ with $b_T = [d_T/9]$. Finally, we generate the active indices among a uniform law $\mathcal{U}([0.1, 0.99])$. Zou and Zhang (2009) experiment influenced our framework.

TABLE 2.1: Simulated experiment 1: Model selection and precision accuracy based on 100 replications.

T	d_T	N_g	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$	Model	MSE	C	IC
500	28	4	2	9	Truth		19	0
					Lasso	0.0178	13.13	0
					aLasso	0.0118	17.98	0
					GLasso	0.0146	12.77	0
					AGLasso	0.0129	13.57	0
					SGL	0.0183	12.97	0
					ASGL	0.0101	18.83	0
2000	106	8	4	33	Truth		73	0
					Lasso	0.0118	49.65	0
					aLasso	0.0103	70.95	0
					GLasso	0.0150	57.48	0
					AGLasso	0.0160	60.78	0
					SGL	0.0125	58.88	0
					ASGL	0.0095	72.70	0
4000	199	18	12	66	Truth		133	0
					Lasso	0.0105	87.17	0
					aLasso	0.0093	131.33	0
					GLasso	0.0140	113.42	0
					AGLasso	0.0150	113.17	0
					SGL	0.0102	98.92	0
					ASGL	0.0094	133	0

We can highlight some interesting remarks from this simulation study. First, the adaptive versions of the Lasso, the Group Lasso or the SGL outperfom their non adaptive versions. The difference is significant for the adaptive Lasso and the adaptive SGL. This is in line with the asymptotic theory. The adaptive SGL performs well as it

can discard inactive groups and inactive indices among active groups and outperform other adaptive penalization methods.

Simulated experiment 2. We consider a data generating process

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + u_t,$$

with $y_t = (y_{1,t}, \cdots, y_{N,t})'$, $u_t \sim \mathcal{N}_{\mathbb{R}^N}(0, \Sigma)$ such that $\Sigma = D^{\frac{1}{2}} R D^{\frac{1}{2}}$, with $R_{ij} = \rho^{|i-j|}, 1 \leq j, i \leq N$, $D = diag(\sigma_1^2, \cdots, \sigma_N^2)$, $\forall i, \sigma_i \in \mathcal{U}([0.01; 0.03])$ and $\rho \in \mathcal{U}([0.5, 0.9])$. We set $N = 5$ and $T = 5000$. It corresponds to a VAR(p) dynamic, with $p = 2$ such that we generate Φ_1 and Φ_2 under the usual stationarity constraints together with an ordering constraint, idest $\forall i, j$, $\Phi_{2,ij} \leq \Phi_{1,ij}$. We also set zero coefficients among Φ_1 and Φ_2: the number of zeros is 30 such that the number of nonzero coefficients is 20. Each of these active coefficients is simulated in $\mathcal{U}([0.05, 0.9])$.

Then we estimate a VAR(p) model, with $p = 4$. The total number of estimated parameters would be $d = p \times N^2 = 100$ and the total number of zero to recover is 80. In this setting d is not indexed by T. We define the group as the lags for the Group Lasso and the SGL procedures, which implies there are 4 groups in total, with 2 active groups.

TABLE 2.2: Simulated experiment 2: Model selection and precision accuracy based on 100 replications.

Model	MSE	C	IC
Truth		80	0
Lasso	0.1130	60.10	1.01
aLasso	0.0917	75.00	1.42
GLasso	0.1512	67.07	3.38
AGLasso	0.1545	67.07	3.38
SGL	0.1062	67.73	1.54
ASGL	0.0709	78.27	0.95

These results illustrate the abbility of the adaptive SGL procedure to properly perform for variable selection. The adaptive Lasso also provide proper performance results regarding both estimation precision and variable selection.

2.7 Conclusion

We explored the asymptotic properties of the Sparse Group Lasso estimator within the M-estimator framework for dependent variables. We showed that the non-adaptive estimator does not satisfy the oracle property in the sense of Fan and Li (2001). We then proposed the adaptive Sparse Group Lasso estimator using the approach of Zou (2006) and proved that this estimator satisfies the oracle property both in a fixed and double-asymptotic framework. Our asymptotic oracle theorems provide the proper choices of the regularization parameters.

Our simulation experiment illustrated the asymptotic results as the adaptive Sparse Group Lasso estimator provides better performance results than other oracle-like methods for model selection and estimation precision.

Chapter 3

Sparse dynamic variance-covariance matrix processes

3.1 Introduction

The multivariate modeling has gained a significant relevance for both practitioners and academics. The main challenge consists in developing a framework that is flexible enough, idest sufficiently parameterized to capture complex patterns, and parsimonious, where the parameters are constrained to avoid overfitting. In a discrete time framework, the usual key quantity in such multivariate processes is the variance covariance matrix of the joint distribution. The curse of dimensionality is an inherent hurdle as general dynamics imply an explosive number of parameters, even when some two-step optimization procedures would be feasible. Furthermore, the corresponding (quasi-)likelihood functions are highly nonlinear - multivariate Gaussian or Student - with a significant number of free parameters that necessitates fast solving optimization procedures.

Scalar versions are often considered: see the scalar Dynamic Conditional Correlation (DCC, Engle, 2002) when modeling correlation processes, the scalar BEKK (Engle and Kroner, 1995), for instance. However, it would be unrealistic to capture heterogenous patterns with scalar dynamic models. Indeed, in such models, the influence of past returns is similar for all components of the variance covariance matrix. But heterogeneity typically occurs when considering high-dimensional vectors. Another approach

is the factor modeling, which aims at reducing the model complexity. Fan, Fan and Lv (2008) emphasized the relevance of factor models for high-dimensional precision matrix estimation. They proved that there is a statistical gain in terms of precision. However, this modeling requires the identification of the relevant factors. An "expert" approach is based on some priors regarding the leading underlying factors. Otherwise, latent unobserved factors induce particular estimation issues and their number is questionable.

In this paper, we propose to tackle both the curse of dimensionality within the multivariate GARCH framework. The objective of this paper consists in modeling high-dimensional variance covariance matrices in a flexible way and breaking the curse of dimensionality. To do so, we propose extensions of the univariate ARCH model to multivariate ones and estimate such models through a penalized ordinary least squares (OLS) procedure. Indeed, multivariate ARCH models admit a linear representation with respect to the parameters, which is a clear advantage wrt GARCH ones as the related loss function can be easily handled. Besides, our multivariate ARCH specification can approximately recover the autoregressive feature of a general GARCH process by using a large number of lags. The idea is to set to zeros the model coefficients from a particular lag on using a regularization procedure. The OLS objective function is particularly adapted for regularization procedures and fast closed form algorithms can be applied. The natural regularization procedure is the Sparse Group Lasso of Simon, Friedman, Hastie and Tibshirani (2013), as it fosters sparsity at a group level and within a group, where the groups would be the lagged variables. The penalized loss function satisfies the convex property such that the adaptive SGL satisfies the oracle property (Fan and Li, 2001). We thus propose a general penalized OLS objective function for a wide range of multivariate ARCH processes.

The main challenge is the positive-definiteness constraint for generating conditional variance covariance matrices. Indeed, the model parameters must then satisfy eigenvalue-type constraints such that the estimation problem is not convex. This prevents from using fast solving algorithms. Besides, the oracle property of Fan and Li (2001), which ensures the right identification of the underlying sparse set, can not be satisfied as it heavily relies on the convex property of the criterion and parameter set. To fix this issue, we propose new multivariate ARCH parameterizations that ensure linear dynamics with linear constraints, if any, imposed over the parameters. Our main objective is

to devise processes that can be estimated thanks to a penalized OLS criterion, where the regularizer is meant to select the relevant lag.

The rest of the paper is organized as follows. In Section 3.2, we describe the multivariate ARCH framework and the penalized ordinary least squares criterion. In Section 3.3, we propose several ARCH-type parameterizations. In Section 3.4, we describe a Cholesky-GARCH model. In Section 3.5, we use simulations to compare the performance of the penalized multivariate ARCH process with other competitors.

3.2 Framework

3.2.1 Dynamic processes of variance covariance

We consider a N-dimensional vectorial stochastic process $(r_t)_{t=1,\cdots,T}$ and denote by θ the vector of the model parameters. Decompose the stochastic process $(r_t)_{t=1,\cdots,T}$ as the sum of conditional expected returns and a residual

$$\begin{aligned} r_t &= \mu_t(\theta) + \epsilon_t, \\ \epsilon_t &= H_t^{1/2}(\theta)\eta_t. \end{aligned}$$

The expected return given the past is $\mu_t(\theta) = \mathbb{E}[r_t|\mathcal{F}_{t-1}] := \mathbb{E}_{t-1}[r_t]$, where $\mathcal{F}_t$ denotes the market information until (and including) time t. We suppose $H_t(\theta) = \mathrm{Var}(r_t|\mathcal{F}_{t-1}) := \mathrm{Var}_{t-1}(r_t) = \mathrm{Var}_{t-1}(\epsilon_t)$ is a $N \times N$ positive definite matrix. The series (η_t) is supposed to be a strong white noise, i.e. a sequence of independent and identically distributed random variables s.t. $\mathbb{E}[\eta_t] = 0$ and $\mathrm{Var}(\eta_t) = I_N$.

The model will be semi-parametric. Its specification is complete when the law of η_t is specified and when the functional form of both $\mu_t(\theta)$ and $H_t(\theta)$ are given. In this paper, we focus on the latter point. For convenience, we will denote $\mu_t(\theta) = \mu_t$ and $H_t(\theta) = H_t = [h_{k,l,t}]_{1\leq k,l\leq N}$.

Actually, we will focus on the centered dynamics (ϵ_t) after removing the first conditional moment. Typically, most authors suppose that the conditional expected returns are modeled as an $ARMA(p,q)$. Since we are interested in ϵ_t only in this paper, we simply assume that (μ_t) follows an $AR(1)$ process. Then, we estimate μ_t by OLS and subtract

it from r_t. Now, these estimated residuals will be considered as our observations (still denoted by ϵ_t). The information set is defined by $\mathcal{F}_t = \sigma(r_s, s \leq t) = \sigma(\epsilon_s, s \leq t)$.

The quantity of interest is H_t and we would like to specify directly its dynamics. A significant stream of the literature has been developed in this direction. A general formulation of H_t-dynamics has been proposed by Bollerslev et al. (1988). In the general VEC model, each element of H_t is a linear function of the lagged squared errors, cross-products of errors and the components of lagged H_t matrices. The most general formulation of a VEC(p, q) model is

$$h_{i,j,t} = a_{i,j} + \sum_{k=1}^{q} \epsilon'_{t-k} B_{ijk} \epsilon_{t-k} + \sum_{l=1}^{p} C_{ij,l} \text{vec}(H_{t-l}), \tag{3.2.1}$$

for every t and every indices i, j in $\{1, \ldots, N\}$. The model parameters are the unknown $N \times N$ matrices B_{ijk}, $i, j \in \{1, \ldots, N\}$, $k = 1, \ldots, q$, $C_{ij,l}$ for $l = 1, \ldots, p$ and $A := [a_{ij}]$ are $N(N+1)/2$ vectors. Some tedious constraints have to be fulfilled to ensure the definite positiveness of H_t. In this paper [1], we will not consider the auto-regressive part in (3.2.1). Then, the model can be rewritten

$$H_t = A + \sum_{k=1}^{q} (I_N \otimes \epsilon'_{t-k}) B_k (I_N \otimes \epsilon_{t-k}), \tag{3.2.2}$$

where B_k is the $N^2 \times N^2$ block matrix given by $B_k := [B_{ijk}]_{1 \leq i,j \leq N}$, I_N is the identity matrix in $\mathbb{R}^N$ and $\otimes$ is the usual Kronecker product. In Gouriéroux (1997), it is noticed that sufficient conditions for obtaining nonnegative covariance matrices H_t are

(i) A and B_k, $k = 1, \ldots, q$, are symmetric, and

(ii) A and B_k, $k = 1, \ldots, q$, are non-negative.

Clearly, (i) can be imposed easily, but (ii) is a lot more tricky. Indeed, in general, the latter condition imposes complex non-linear constraints on the model parameters. Moreover, it is not realistic to estimate general non-negative matrices B, due to their sizes ($qN^2(N^2+1)/2$ unknown parameters!), under the tedious nonlinear constraints imposed by non-negativeness (particularly at the optimization stage). Therefore, we have to exhibit flexible (but realistic) sub-families of models as (3.2.2). This will be done hereafter.

[1] And for some reasons that will appear hereafter.

Note that (3.2.2) can be rewritten as a linear model

$$\epsilon_t\epsilon_t' = A + \sum_{k=1}^{q}(I_N \otimes \epsilon_{t-k}')B_k(I_N \otimes \epsilon_{t-k}) + \zeta_t, \;\; \mathbb{E}[\zeta_t|\mathcal{F}_{t-1}] = 0. \tag{3.2.3}$$

Introducing the usual operator Vech(.) that transforms any $m \times m$ symmetric matrix M into the $m(m+1)/2$ vector of its component, this is equivalent to

$$\text{Vech}(\epsilon_t\epsilon_t') = \text{Vech}(A) + \sum_{k=1}^{q}\text{Vech}\left((I_N \otimes \epsilon_{t-k}')B_k(I_N \otimes \epsilon_{t-k})\right) + \text{Vech}(\zeta_t).$$

More explicitly, this can be rewritten: for every couple $(i,j) \in \{1,\ldots,N\}^2$ such that $i \leq j$,

$$\epsilon_{i,t}\epsilon_{j,t} = a_{i,j} + \sum_{k=1}^{q}\sum_{r,s=1}^{N} b_{ijk,rs}\epsilon_{r,t-k}\epsilon_{s,t-k} + \zeta_{i,j,t}, \;\; \mathbb{E}[\zeta_{i,j,t}|\mathcal{F}_{t-1}] = 0, \tag{3.2.4}$$

where $B_{ijk} = [b_{ijk,rs}]_{1\leq r,s\leq N}$. Note that the elements of the N^2-squared matrix B_k will be indexed by quadruplets (i,j,r,s). The latter elements are related to the coefficients of B_k that define the dynamics of $\epsilon_{i,t}\epsilon_{j,t}$. Moreover, note that $B_{ijk} = B_{jik}$ and $\zeta_{i,j,t} = \zeta_{j,i,t}$ for every couple (i,j) and every k. Hereafter, the couples of indices (i,j) and (r,s) will be sorted in the lexicographical order

$$(1,1),(1,2),\ldots,(1,N),(2,1),(2,2),\ldots,(N,N-1),(N,N),$$

even when we restrict ourselves to the couples (i,j) s.t. $i \leq j$.

The previous linear model will be estimated by a penalized least squares procedure. In terms of inference, this is a dramatic advantage wrt the usual QML estimation procedure of GARCH models. Therefore, in practical terms, it is easier to estimate ARCH-type models with a lot of assets and lags ($N >> 1$, $q >> 1$) than a GARCH model with the same N and $q = 1$.

3.2.2 Statistical criterion

Contrary to GARCH-type dynamics that require the optimization of a nonlinear objective function - typically Gaussian or Student type likelihoods -, the multivariate ARCH process has the advantage of a direct linear estimation by specifying an ordinary least

squares objective function. Assuming that the true model is (3.2.4), a regularization procedure with q sufficiently large would likely set to zero the parameters after the true q_0. Now if the true model is a GARCH process, then its autoregressive component [2] can be written as in (3.2.4) with $q = \infty$. In such a case, a regularization procedure performed over a large q would produce a relevant approximation of the GARCH process. For the sake of parsimony, the parameters need to be constrained to avoid overfitting. That is the key idea of this paper: specifying a regularization procedure to perform variable selection and estimation. The OLS objective function is particularly adapted to the penalization procedures and the asymptotic properties of the oracle-like penalties can be used such as the oracle property of Fan and Li (2001). The regularization procedure aims at identifying this relevant subset to described the instantaneous covariance. It belongs to a bigger set formed by the specified lagged variables (typically a large number a priori). This means that the regularizer performs both estimation and variable selection.

To illustrate this idea, consider a univariate ARCH(1) process, which is defined as

$$h_t = \omega + \alpha\epsilon_{t-1}^2,\ \omega > 0, \alpha \in [0, 1).$$

This dynamics can be rewritten as a linear model

$$\epsilon_t^2 = \omega + \alpha\epsilon_{t-1}^2 + u_t,\ u_t = \epsilon_t^2 - h_t,$$

by noting that $\mathbb{E}[u_t|\mathcal{F}_{t-1}] = 0$. Then, it is natural to consider the corresponding OLS estimator of $\theta := (\omega, \alpha)$: $\hat{\theta}$ is defined by

$$\hat{\theta} = \arg\min_{\theta} \|Y - X\theta\|_2^2 = (X'X)^{-1}X'Y,$$

where

$$X = \begin{pmatrix} 1 & \epsilon_1^2 \\ 1 & \epsilon_2^2 \\ \vdots & \vdots \\ 1 & \epsilon_{T-1}^2 \end{pmatrix}, \quad Y = \begin{pmatrix} \epsilon_2^2 \\ \epsilon_3^2 \\ \vdots \\ \epsilon_T^2 \end{pmatrix}.$$

The previous criterion can be extended to the multivariate case, provided that the estimated dynamics generate positive definite covariance matrices. Then our least

[2] think of the invertibility of the autoregressive matrix component

squares objective function can be specified as

$$\begin{cases} \mathbb{G}_T l(\theta) &= \frac{1}{T}\sum_{t=1}^{T} l(\epsilon_t;\theta), \\ l(\epsilon_t;\theta) &= \|\text{Vech}(\epsilon_t\epsilon_t') - \Psi(\underline{\epsilon_{t-1}})\theta\|_2^2, \end{cases} \tag{3.2.5}$$

where $\Psi(\epsilon_{t-1})$ is a $\mathcal{F}_{t-1}$-measurable random matrix, whose particular analytic form depends on the model specification. For instance, for the process (3.2.4) and without any additional constraint on the parameters, the parameter vector can be decomposed as

$$\theta = (\theta^{(ij)}, 1 \leq i \leq j \leq N),$$

such that the ij-th sub-vector is

$$\theta^{(ij)} := (a_{ij}, \theta^{(ij1)}, \ldots, \theta^{(ijq)}),$$

$$\theta^{(ijk)} := (b_{ijk,11}, 2b_{ijk,12}, \cdots, 2b_{ijk,1N}, b_{ijk,22}, 2b_{ijk,23}, \ldots, 2b_{ijk,(N-1)N}, b_{ijk,NN})'.$$

This means that the number of unknown parameters is $d(1+qd)$, with $d = N(N+1)/2$. Then, in such a case, $\Psi(\underline{\epsilon_t})$ is the $d \times d(1+qd)$ matrix

$$\Psi(\underline{\epsilon_t}) = \begin{pmatrix} \psi(\underline{\epsilon_t}) & 0_{1+qd} & 0_{1+qd} & 0_{1+qd} & \cdots & 0_{1+qd} \\ 0_{1+qd} & \psi(\underline{\epsilon_t}) & 0_{1+qd} & 0_{1+qd} & \cdots & 0_{1+qd} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{1+qd} & 0_{1+qd} & 0_{1+qd} & \cdots & 0_{1+qd} & \psi(\underline{\epsilon_t}) \end{pmatrix},$$

where 0_{1+qd} is a $1+qd$-row vector of zeros and

$$\psi(\underline{\epsilon_t}) = (1, \text{Vech}(\epsilon_{t-1}\epsilon_{t-1}')', \ldots, \text{Vech}(\epsilon_{t-q}\epsilon_{t-q}')').$$

Note that the latter criterion has most often to be rewritten as long as some constraints on the model parameters are included. Indeed, in such a case, the number of free parameters is typically reduced, and/or some parameters are shared by several univariate linear equations of the type (3.2.4). See for instance the so-called "homogeneous model" below.

The autoregressive feature of some MGARCH models should be reproduced by specifying a sufficiently large number of lags q in the model (3.2.4). Moreover, in a lot of situations, it is likely that the most recent observations should have a higher level

effect on the current covariance matrix than older observations. In this setting it is natural to assume that the coefficients decay as we move farther away from the current observation. We could consider a procedure that would impose inequality constraints among the coefficients to recover such ordering effect. Tibshirani and Suo (2016) proposed an order-constrained version of the Lasso. This framework is left for further extensions. At least, it makes sense that the coefficients go to zero from a certain rank, which is a minimal assumption we make.

We propose a penalization approach to constrain the parameters and foster parsimony. The intuition is as follows: we specify a large number of lags a priori to approximate an autoregressive pattern. We assume that only a subset of potential features (the lagged variances and covariances) has a statistically significant effect on the output: that is the sparsity assumption. As this subset is unknown, the penalization procedure enables to recover it since it provides an estimation of the set of indices for which the corresponding coefficients are non-zero. To achieve this subset identification, the Sparse Group Lasso is the most relevant regularizer as it fosters sparsity both at a group level and within a group. Intuitively, the natural groups should be all the parameters that are associated to a given lagged observed vector ϵ_{t-k} (i.e. all quantities $b_{ijk,r,s}$ for every quadruplet (i,j,r,s)), but other choices are possible, obviously.

The statistical problem consists in minimizing over the parameter space $\Theta \subset \mathbb{R}^m$ a penalized criterion of the form

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} \{\mathbb{G}_T \varphi(\theta)\}, \tag{3.2.6}$$

where

$$\begin{aligned} \theta \mapsto \mathbb{G}_T \varphi(\theta) &= \frac{1}{T}\sum_{t=1}^{T}\{l(\epsilon_t;\theta) + \boldsymbol{p}_1(\lambda_T, \tilde{\theta}, \theta) + \boldsymbol{p}_2(\gamma_T, \tilde{\theta}, \theta)\} \\ &= \mathbb{G}_T l(\theta) + \boldsymbol{p}_1(\lambda_T, \tilde{\theta}, \theta) + \boldsymbol{p}_2(\gamma_T, \tilde{\theta}, \theta), \end{aligned}$$

and both penalties are specified as

$$\begin{cases} \boldsymbol{p}_1 : \mathbb{R}_+ \times \Theta \times \Theta \to \mathbb{R}_+, & \boldsymbol{p}_2 : \mathbb{R}_+ \times \Theta \times \Theta \to \mathbb{R}_+, \\ (\lambda_T, \tilde{\theta}, \theta) \mapsto \boldsymbol{p}_1(\lambda_T, \tilde{\theta}, \theta) = \dfrac{\lambda_T}{T}\displaystyle\sum_{k=1}^{m}\sum_{i=1}^{\boldsymbol{c}_k}\alpha_{T,i}^{(k)}|\theta_i^{(k)}|, & (\gamma_T, \tilde{\theta}, \theta) \mapsto \boldsymbol{p}_2(\gamma_T, \tilde{\theta}, \theta) = \dfrac{\gamma_T}{T}\displaystyle\sum_{l=1}^{m}\xi_{T,l}\|\theta^{(l)}\|_2, \end{cases}$$

with $\alpha_{T,i}^{(k)} = |\tilde{\theta}_i^{(k)}|^{-\eta}$ and $\xi_{T,l} = \|\tilde{\theta}^{(l)}\|_2^{-\mu}$, where $\eta > 0, \mu > 0$, and $\tilde{\theta}$ is a first step estimator, which is supposed to be a $\sqrt{T}$-consistent estimator [3].

This reduces to the classic OLS estimator when there is no penalization. The proposed penalization framework includes the usual Lasso criterion when $\gamma_T = 0$, the Group Lasso when $\lambda_T = 0$ and the Sparse Group Lasso when λ_T and γ_T are non zero.

Obtaining the positive definiteness of the conditional covariance matrices induced by (3.2.4) is the main technical challenge in practice. To ensure this constraint, the parameters in (3.2.4) must satisfy eigenvalue-type constraints such that Θ will not be convex. This is a drawback from both an empirical and theoretical point of views: empirically, it hampers fast solving algorithms; theoretically, the non-convexity prevents the Sparse Group Lasso estimator from satisfying the oracle property of Fan and Li (2001). Thus, in the next section we aim at devising parameterizations that allow for generating positive definite matrices while remaining linear with respect to the parameters. This would discard processes that require a normalization step or non convex constraint sets for the parameters.

3.3 ARCH Parameterizations

In this section, we propose parameterizations of (3.2.2) to ensure the positive definiteness of H_t. Our main objective is to obtain a linear process with linear constraints that must be satisfied by the parameters. These are sufficient conditions to obtain a convex objective function for a convex parameter set.

3.3.1 Evaluation of A

We first focus on a covariance targeting procedure for the estimation of A. Although this parameter could be estimated with B simultaneously, the covariance targeting step fosters dimension reduction as it splits the problem. This will allow to satisfy the non-negativeness of the (estimated) A matrix more easily. To do so, note that taking

[3] For instance, $\tilde{\theta}$ can be an unpenalized OLS estimator. The $\sqrt{T}$-consistency is a necessary condition to satisfy the oracle property.

the unconditional expectation of (3.2.4), we have

$$\mathbb{E}[\epsilon_{i,t}\epsilon_{j,t}] = a_{i,j} + \sum_{k=1}^{q}\sum_{r,s=1}^{N} b_{ijk,rs}\mathbb{E}[\epsilon_{r,t-k}\epsilon_{s,t-k}],$$

for every couple (i,j). If the coefficients $b_{ijk,rs}$ were known, and assuming we have estimated consistently $\mathbb{E}[\epsilon_{i,t}\epsilon_{j,t}]$ by $\widehat{\text{cov}}_{i,j}$, then the coefficients $a_{i,j}$ could be estimated as

$$\hat{a}_{i,j} = \widehat{\text{cov}}_{i,j} - \sum_{k=1}^{q}\sum_{r,s=1}^{N} b_{ijk,rs}\widehat{\text{cov}}_{r,s}.$$

When T is large and assuming the model is well specified, $\hat{a}_{i,j}$ will converge towards $a_{i,j}$ and we would observe that the estimated matrix $\hat{A} := [\hat{a}_{i,j}]$ is definite positive if this is the case for A. Nonetheless, at finite distance, it is likely the latter condition will not be the satisfied. Fortunately, our OLS estimation procedure does not require per se that we manipulate nonnegative matrices A and B. This is required only for prediction and likelihood-based methods. Therefore, to estimate (3.2.2) (and then (3.2.4)), we propose to replace $a_{i,j}$ by $\hat{a}_{i,j}$, and the model is then parameterized by B only. Once B is estimated (see below) by $\hat{B}$, the matrix A will be approximated by $\tilde{A}$ whose components are

$$\tilde{a}_{i,j} = \widehat{\text{cov}}_{i,j} - \sum_{k=1}^{q}\sum_{r,s=1}^{N} \hat{b}_{ijk,rs}\widehat{\text{cov}}_{r,s}.$$

Afterwards, a projection of $\tilde{A}$ on the cone of nonnegative matrices would provide the final estimate of A.

As an alternative strategy, we can invoke a parametrization of A in the cone of nonnegative matrices directly. The natural basis would be provided by the spectral decomposition of $\mathbb{E}[\epsilon_t\epsilon_t']$ (or its empirical approximation $[\widehat{\text{cov}}_{i,j}]$ instead). Indeed, there exists an orthonormal basis $(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N)$ in $\mathbb{R}^N$ s.t.

$$\mathbb{E}[\epsilon_t\epsilon_t'] \simeq [\widehat{\text{cov}}_{i,j}]_{1\le i,j\le N} = \sum_{l=1}^{N}\lambda_l\boldsymbol{v}_l\boldsymbol{v}_l',$$

where $(\lambda_1,\ldots,\lambda_N)$ is the associated spectrum, $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N \ge 0$. Then, we would assume that there exist nonnegative real numbers μ_l, $l = 1,\ldots,N$ s.t.

$$A = \sum_{l=1}^{N}\mu_l\boldsymbol{v}_l\boldsymbol{v}_l'.$$

We have replaced the $N(N+1)/2$ unknown coefficients of A by only N parameters $(\lambda_1, \ldots, \lambda_N)$. And such a matrix A will be nonnegative by construction.

We now focus on the evaluation of B-type matrices in (3.2.4). To do so, we propose three ARCH-type parameterizations that aim at reducing the dimensionality while generating positive definite processes. These models share the fact that they are linear with respect to the parameters and shall satisfy, if any, convex constraints. First, we propose a constraint free multivariate ARCH dynamic, where the B-parameters are unconstrained and the corresponding process is projected onto the space of positive definite matrices to generate a variance covariance matrix sequence. The second case is called "homogeneous" and is relevant for random vectors with positively correlated components. Finally we propose a "heterogenous" parameterization that it is adapted to random vectors with discordant patterns.

3.3.2 Constraint free and matrix projection

This approach consists in projecting a matrix process, which may not be necessarily positive definite, onto the space of positive definite matrices. This method allows flexibility because one can independently specify/estimate the processes that are associated to each component of $\text{vec}(\epsilon_t \epsilon_t')$. We rewrite the general dynamics given by (3.2.4) for each component of the $\epsilon_t \epsilon_t'$ matrix as

$$\epsilon_{i,t}\epsilon_{j,t} = a_{i,j} + \sum_{k=1}^{q}\sum_{r=1}^{N} b_{ijk,rr}\epsilon_{r,t-k}^2 + \sum_{k=1}^{q}\sum_{r,s=1,r<s}^{N} 2b_{ijk,rs}\epsilon_{r,t-k}\epsilon_{s,t-k} + \zeta_{i,j,t},\ \mathbb{E}[\zeta_{i,j,t}|\mathcal{F}_{t-1}] = 0, \tag{3.3.1}$$

if $i \leq j$. The OLS is a natural estimator but the symmetric matrix coefficients A and B are not necessarily positive definite. Nonetheless, these matrix can be approximated by positive definite ones. Here is a loss we need to accept as we eventually obtain an approximation of (3.3.1) that would generate true conditional covariance matrices (H_t).

To this goal, we propose two methods: consider the singular value decomposition of a symmetric matrix M as $M = P'\text{diag}(\lambda_1, \cdots, \lambda_N)P$, where P is an orthogonal matrix composed with N eigenvectors. We define two projections $f_k : \mathcal{M}_{N\times N}(\mathbb{R}) \to \mathcal{M}^+_{N\times N}(\mathbb{R})$ with $k = 1, 2$. A first projection would be

$$f_1(M) = P'\text{diag}(\lambda_1^+, \cdots, \lambda_N^+)P,$$

with λ_k^+ the positive part of λ_k. A second projection would be

$$f_2(M) = (M + \lambda_{\min}^- I_d)/(1 + \lambda_{\min}^-),$$

with $\lambda_{\min}^-$ the negative part of the minimum eigenvalue of M. The eigenvectors remain the same as M.

The first stage estimated matrix is denoted by $\tilde{H}_t = [\tilde{h}_{ij,t}]$, given by

$$h_{ij,t} = \hat{a}_{i,j} + \sum_{k=1}^{q}\sum_{r=1}^{N} \hat{b}_{ijk,rr}\epsilon_{r,t-k}^2 + \sum_{k=1}^{q}\sum_{r,s=1,r<s}^{N} 2\hat{b}_{ijk,rs}\epsilon_{r,t-k}\epsilon_{s,t-k},$$

for any couple (i,j). For and projection method $k = 1 \in \{1,2\}$, the final estimated covariance matrix of ϵ_t given $\mathcal{F}_{t-1}$ would be $H_t = f_k(\tilde{H}_t)$.

This method allows for an equation-by-equation estimation procedure, where each equation corresponds to a couple, which is particularly adapted for high-dimensional regression settings. Such dynamics are linear with respect to the parameters so that the estimation can be carried out by the ordinary least squares objective function or by penalized OLS.

3.3.3 The homogeneous case

First, we need some matrix notations.

- For any subset J of indices in $I := \{1, \ldots, m\}$, the m-column vector $e_{m,J}$ of zeros and ones is defined by $e_{m,J} := [\mathbf{1}(i \in J)]_{1 \le i \le m}$. When its size is obvious, it is written e_J simply. Moreover, set $e_{m,I} = e_m$ the m-vector of ones.
- For any vector $\boldsymbol{x} \in \mathbb{R}^m$, $D(\boldsymbol{x})$ denotes the $m \times m$ diagonal matrix given by $D(\boldsymbol{x}) = [\mathbf{1}(i = j)x_i]_{1 \le i,j \le m}$.

Set $\mathcal{J} = \{1, N{+}2, 2N{+}3, \ldots, (N{-}2)N{+}N{-}1, (N{-}1)N{+}N\}$, a subset of $\{1, \ldots, N^2\}$. Let us consider the parametric family $\mathcal{B}$ of matrices given by

$$\mathcal{B} = \{M \in \mathcal{M}_{N^2 \times N^2}(\mathbb{R}) \,|\, M = \alpha e_{N^2} e_{N^2}' + \beta e_{\mathcal{J}} e_{\mathcal{J}}' + \gamma D(e_{\mathcal{J}}), (\alpha, \beta, \gamma) \in [0,1]^3\}.$$

Clearly, all matrices in $\mathcal{B}$ are non-negative. By assumption, we will choose our matrices B_k, $k = 1, \ldots, q$, inside $\mathcal{B}$. More explicitly, the model becomes: for every indices i, j and time t, then

$$\epsilon_{it}\epsilon_{jt} = a_{ij} + \sum_{k=1}^{q} \left((\alpha_k + \beta_k + \gamma_k \mathbf{1}(i=j))\epsilon_{i,t-k}\epsilon_{j,t-k} + \alpha_k \sum_{(r,s)\neq(i,j)} \epsilon_{r,t-k}\epsilon_{s,t-k} \right) + \zeta_{ij,t},$$

where $\zeta_{ij,t} = \epsilon_{it}\epsilon_{jt} - h_{ij,t} = \{\eta_{it}\eta_{jt} - 1\}h_{ij,t}$. Note that the matrix $e_{\mathcal{J}}e'_{\mathcal{J}}$ can be rewritten as a block-matrix $[E_{ij}]_{1\leq i,j\leq N}$, where $E_{ij} = [\mathbf{1}((i,j) = (r,s))]_{1\leq r,s\leq 1}$. In other words, this model tries to capture three effects on the dynamics of $\epsilon_{i,t}\epsilon_{j,t}$:

(i) a uniform effect of all past cross-product among the components of $\epsilon_t\epsilon'_t$ through the α_k coefficients;

(ii) a more important bump caused by the past values of $\epsilon_{i,t}\epsilon_{j,t}$ on itself through β_k;

(iii) an additional bump when variances are managed (ie when $i = j$) through the parameters γ_k.

As for the estimation step, the (non penalized) OLS objective function in (3.2.5) corresponds to

$$\theta = (\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_q, \gamma_1, \ldots, \gamma_q,),$$

when the constant $a_{i,j}$ has been removed as explained in Subsection 3.3.1. In this case, the matrix $\Psi(\underline{\epsilon_{t-1}})$ of regressors is

$$\Psi(\underline{\epsilon_{t-1}}) = \begin{pmatrix} s_{t-1} & \ldots & s_{t-q} & \vec{\epsilon}_{11,t,q} & \vec{\epsilon}_{11,t,q} \\ s_{t-1} & \ldots & s_{t-q} & \vec{\epsilon}_{12,t,q} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{t-1} & \ldots & s_{t-q} & \vec{\epsilon}_{NN,t,q} & \vec{\epsilon}_{NN,t,q} \end{pmatrix}$$

with $s_{t-k} := \sum_{r,s=1}^{N} \epsilon_{r,t-k}\epsilon_{s,t-k}$, for $k = 1, \ldots, q$ and $\vec{\epsilon}_{ij,t,q} := (\epsilon_{i,t}\epsilon_{j,t-1}, \cdots, \epsilon_{i,t-q}\epsilon_{j,t-q})$. Note that the size of $\Psi(\underline{\epsilon_{t-1}})$ is $N(N+1)/2 \times 3q$. Moreover, the regressors in the last column of are $\Psi(\underline{\epsilon_{t-1}})$ are zero, except when $i = j$ (lexicographical order).

3.3.4 The heterogenous case

In this case, we have identified two homogeneous sub-portfolios but whose dynamics behave differently. The first (resp. second) portfolio corresponds to the assets that are numbered $\{1,\ldots,p\}$ (resp. $\{p+1,\ldots,N\}$). This necessitates to extend the previous model and to introduce more parameters. Let us introduce additional notations:

- For any real numbers $\alpha_1,\alpha_2,\alpha_3$, and two integers n and m, $n<m$. Set the $m\times m$ matrix
$$M(\alpha_1,\alpha_2,\alpha_3,m,n) := \left[\begin{array}{cc} \alpha_1 e_n e_n' & \alpha_2 e_n e_{m-n}' \\ \alpha_2 e_{m-n} e_n' & \alpha_3 e_{m-n} e_{m-n}' \end{array}\right].$$
By some standard algebraic calculations, we can prove that the characteristic polynomial of the symmetrical matrix $M(\alpha_1,\alpha_2,\alpha_3,m,n)$ is
$$x \mapsto (-1)^m x^{m-2}\left[(x-n\alpha_1)(x-(m-n)\alpha_3)-n(m-n)\alpha_2^2\right].$$
Therefore, the associated spectrum is $\{\delta_+,\delta_-,0\}$, $x_\pm := (n\alpha_1+(m-n)\alpha_3 \pm \sqrt{\Delta})/2$, where
$$\Delta := (n\alpha_1+(m-n)\alpha_3)^2 - 4n(m-n)(\alpha_1\alpha_3-\alpha_2^2).$$
If $\alpha_1\alpha_3 \geq \alpha_2^2$, then x_+ and x_- are nonnegative and the matrix $M(\alpha_1,\alpha_2,\alpha_3,m,n)$ is nonnegative. Note that this can be achieved in an optimization program with linear constraints by assuming that $\alpha_2 \leq \min(\alpha_1,\alpha_3)$.
- Set the partitioned matrix $\tilde{M}(\beta_1,\beta_2,\beta_3,p) = [\tilde{M}_{i,j}]_{1\leq i,j\leq N}$, where
$$\begin{aligned}\tilde{M}_{i,j} = [\mathbf{1}((r,s)=(i,j)).\{&\beta_1\mathbf{1}(r\leq p, s\leq p)+\beta_3\mathbf{1}(r>p,s>p)\\ +\quad &\beta_2\mathbf{1}(r\leq p,s>p)+\beta_2\mathbf{1}(r>p,s\leq p)\}]_{1\leq r,s\leq N}.\end{aligned}$$
By a similar reasoning as previously, it can be proved that the matrix $\tilde{M}(\beta_1,\beta_2,\beta_3,p)$ is nonnegative if $\beta_1\beta_3 \geq \beta_2^2$. Again, in the optimization stage, we will assume that $\beta_2 \leq \min(\beta_1,\beta_3)$.
- Let γ_1 and γ_2 be two arbitrary nonnegative real numbers, and an integer $p\leq N$. Let $J := \{1, N+2, 2N+3,\ldots,(p-1)N+p\}$ and $\tilde{J} := \{pN+p+1,(p+1)N+$

$p+2,\ldots,(N-1)N+N\}$. Set the diagonal matrix

$$\begin{aligned} N(\gamma_1,\gamma_2,p) &:= D\left(\gamma_1 e_{N^2,J}+\gamma_2 e_{N^2,\tilde{J}}\right) \\ &= \left[\mathbf{1}((r,s)=(i,j)).\left\{\gamma_1\mathbf{1}(i=j\in J)+\gamma_2\mathbf{1}(i=j\in\tilde{J})\right\}\right]. \end{aligned}$$

Obviously, $N(\gamma_1,\gamma_2,p)$ is nonnegative when γ_1 and γ_2 are nonnegative.

With the notations above, we will choose the matrices B_k of (3.2.2) in the following parametric family:

$$\begin{aligned} \tilde{\mathcal{B}} = \{B\in\mathcal{M}_{N^2\times N^2}(\mathbb{R})\,|B = M(\alpha_1,\alpha_2,\alpha_3,N^2,Np)+\tilde{M}(\beta_1,\beta_2,\beta_3,p)+N(\gamma_1,\gamma_2,p), \\ \alpha_1\geq 0,\alpha_3\geq 0,\alpha_1\alpha_3\geq\alpha_2^2,\beta_1\geq 0,\beta_3\geq 0,\beta_1\beta_3\geq\beta_2^2,\gamma_1\geq 0,\gamma_2\geq 0\}. \quad (3.3.2) \end{aligned}$$

To be more explicit, for any $k=1,\ldots,q$,

$$\epsilon_{it}\epsilon_{jt} = a_{ij}+\sum_{k=1}^{q}\left((\alpha_{ij}^{(k)}+\beta_{ij}^{(k)}+\gamma_i^{(k)}\mathbf{1}(i=j))\epsilon_{i,t-k}\epsilon_{j,t-k}+\alpha_{ij}^{(k)}\sum_{(r,s)\neq(i,j)}\epsilon_{r,t-k}\epsilon_{s,t-k}\right)+\zeta_{ij,t},$$

$$\alpha_{i,j}^{(k)} = \alpha_1^{(k)}\mathbf{1}((i,j)\in J^2)+\alpha_3^{(k)}\mathbf{1}((i,j)\in\tilde{J}^2)+\alpha_2^{(k)}\mathbf{1}((i,j)\in J\times\tilde{J}\text{ or }(i,j)\in\tilde{J}\times J),$$

$$\beta_{i,j}^{(k)} = \beta_1^{(k)}\mathbf{1}((i,j)\in J^2)+\beta_3^{(k)}\mathbf{1}((i,j)\in\tilde{J}^2)+\beta_2^{(k)}\mathbf{1}((i,j)\in J\times\tilde{J}\text{ or }(i,j)\in\tilde{J}\times J),$$

$$\gamma_i^{(k)} = \gamma_1^{(k)}\mathbf{1}(i\in J)+\gamma_2^{(k)}\mathbf{1}(i\in\tilde{J}).$$

This parametric model tries to capture three effects on the dynamics of $\epsilon_{i,t}\epsilon_{j,t}$:

(i) a uniform effect of all past cross-products on the $\epsilon_{i,t}\epsilon_{j,t}$ through the coefficients $\alpha_.$; when i and j belong to the first (resp. second) group of assets, we use α_1 (resp. α_3). When i and j do not belong to the same group, we invoke α_3.

(ii) a more important bump caused by the past values of $\epsilon_{i,t}\epsilon_{j,t}$ on itself, through the $\beta_.$; as above, such effects depend on the group of i and j.

(iii) an additional bump when variances are managed (ie when $i=j$) through the parameters $\gamma_.$; if i belongs to the first or the second group of assets, we apply γ_1 or γ_2 respectively.

Actually, the latter model specification can be criticized because the effect of $\epsilon_{r,t-k}\epsilon_{s,t-k}$ on $\epsilon_{i,t-k}\epsilon_{j,t-k}$, $(r,s)\neq(i,j)$, is transmitted through the same coefficient $\alpha_{ij}^{(k)}$, independently of the identify of the (r,s)-group. For instance, it is likely that this effect should

be stronger when (r, s) and (i, j) belong to the same subset, typically. Therefore, a more general parametric model could be considered, where there are different cross-effects on the dynamics of $\epsilon_{i,t}\epsilon_{j,t}$, depending on the considered couples of indices (r, s), with our previous notations.

It makes sense to introduce the family of block matrices $\bar{\mathcal{M}} := \{\bar{M} = [\bar{M}_{i,j}]_{1\leq i,j\leq N}\}$, where the $N \times N$ matrices $\bar{M}_{i,j}$ are defined as

$$\bar{M}_{i,j} = M(\alpha_1^{(1)}, \alpha_2^{(1)}, \alpha_3^{(1)}, N, p) \text{ if } i \text{ and } j \text{ belong to the first group,}$$

$$\bar{M}_{i,j} = M(\alpha_1^{(2)}, \alpha_2^{(2)}, \alpha_3^{(2)}, N, p) \text{ if } i \text{ and } j \text{ belong to the second group, and}$$

$$\bar{M}_{i,j} = M(\delta_1, \delta_2, \delta_3, N, p) \text{ if } i \text{ and } j \text{ do not belong to the same group.}$$

This would enrich the flexibility and the realism of the model. But the calculation of the spectrum of matrices $\bar{M} \in \bar{\mathcal{M}}$ is difficult. And only highly nonlinear conditions will be able to guarantee that such matrices will be nonnegative.

Nonetheless, we are convinced that it is valuable to study the impact of cross-effects on any product dynamics $\epsilon_{i,t}\epsilon_{j,t}$ differently. To stay tractable and to keep the same notations as above, we will simplify the framework by assuming that $\delta_1 = \delta_2 = \delta_3 := \delta$. This means that the effect of all cross products on the dynamics of $\epsilon_{i,t}\epsilon_{j,t}$ is uniform when i and j do not belong to the same portfolio [4] Therefore, under this simplifying assumption, any matrix $\bar{M}$ in $\bar{\mathcal{M}}$ is written

$$\bar{M} = \begin{bmatrix} M(\alpha^{(1)}) & \cdots & M(\alpha^{(1)}) & M(\delta) & \cdots & M(\delta) \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ M(\alpha^{(1)}) & \cdots & M(\alpha^{(1)}) & M(\delta) & \cdots & M(\delta) \\ M(\delta) & \cdots & M(\delta) & M(\alpha^{(2)}) & \cdots & M(\alpha^{(2)}) \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ M(\delta) & \cdots & M(\delta) & M(\alpha^{(2)}) & \cdots & M(\alpha^{(2)}) \end{bmatrix}, \tag{3.3.3}$$

where

$$M(\alpha^{(1)}) := M(\alpha_1^{(1)}, \alpha_2^{(1)}, \alpha_3^{(1)}, N, p) \text{ appears } p^2 \text{ times,}$$

$$M(\alpha^{(2)}) := M(\alpha_1^{(2)}, \alpha_2^{(2)}, \alpha_3^{(2)}, N, p) \text{ appears } (N-p)^2 \text{ times, and}$$

$$M(\delta) := \delta e_N e_N', \ \delta \in \mathbb{R}^+, \text{ appears } 2p(N-p) \text{ times.}$$

[4] This is reasonable, because the dynamics of $\epsilon_{i,t}\epsilon_{j,t}$, when i and j do not belong to the same group, is "poorer" than when i and j belong to the same group.

Proposition 3.3.1. *A matrix $\bar{M}$ defined as in (3.3.3) is definite positive iff*

$$(\alpha_1^{(1)}, \alpha_2^{(1)}, \alpha_3^{(1)}, \alpha_1^{(2)}, \alpha_2^{(2)}, \alpha_3^{(2)}, \delta) \in \mathbb{R}_+^7,$$

$$\Delta^{(1)} := \alpha_1^{(1)}\alpha_3^{(1)} - (\alpha_2^{(1)})^2 > 0,\ \alpha_1^{(2)}\alpha_3^{(2)} > (\alpha_2^{(2)})^2,\ \textit{and}$$

$$\left(\alpha_1^{(2)} - \frac{\delta^2}{\alpha_1^{(1)}} - \frac{(\alpha_2^{(2)})^2\delta^2}{\alpha_1^{(1)}\Delta^{(1)}}\right) \cdot \left(\alpha_3^{(2)} - \frac{\delta^2}{\alpha_1^{(1)}} - \frac{(\alpha_2^{(2)})^2\delta^2}{\alpha_1^{(1)}\Delta^{(1)}}\right) > \left(\alpha_2^{(2)} - \frac{\delta^2}{\alpha_1^{(1)}} - \frac{(\alpha_2^{(2)})^2\delta^2}{\alpha_1^{(1)}\Delta^{(1)}}\right)^2. \tag{3.3.4}$$

As a consequence, the latter condition (3.3.4) is satisfied if $\alpha_2^{(2)} < \min(\alpha_1^{(2)}, \alpha_3^{(2)})$.

Proof of Proposition 3.3.1. First let us study the positiveness of the quadratic form q_0 that is associated to the $pN \times pN$ symmetrical matrix

$$B_0 = \begin{bmatrix} M(\alpha) & \cdots & M(\alpha) \\ \vdots & \cdots & \vdots \\ M(\alpha) & \cdots & M(\alpha) \end{bmatrix}, \tag{3.3.5}$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$. Let the two sets of indices

$$\mathcal{I} := \{1, \ldots, p, N+1, \ldots, N+p, 2N+1, \ldots, 2N+p, \ldots, (p-1)N+1, \ldots, (p-1)N+p\},$$

$$\mathcal{J} := \{p+1, \ldots, N, N+p+1, \ldots, 2N, 2N+p+1, \ldots, 3N, \ldots, (p-1)N+p+1, \ldots, pN\}.$$

Obviously, $\{1, \ldots, pN\} = \mathcal{I} \cup \mathcal{J}$. Then, for any $\boldsymbol{x} \in \mathbb{R}^{pN}$,

$$\begin{aligned} q_0(\boldsymbol{x}) &= \alpha_1 \sum_{(i,j)\in\mathcal{I}^2} x_i x_j + \alpha_3 \sum_{(i,j)\in\mathcal{J}^2} x_i x_j + 2\alpha_2 \left(\sum_{i\in\mathcal{I}} x_i\right) \cdot \left(\sum_{j\in\mathcal{J}} x_j\right) \\ &= \alpha_1 \left(\sum_{i\in\mathcal{I}} x_i + \frac{\alpha_2}{\alpha_1} \sum_{j\in\mathcal{J}} x_j\right)^2 + \frac{\alpha_1\alpha_3 - \alpha_2^2}{\alpha_1} \left(\sum_{j\in\mathcal{J}} x_j\right)^2. \end{aligned}$$

Therefore, the positiveness of q_0 (or B_0) is equivalent to $\alpha > 0$ and $\alpha_1\alpha_3 > \alpha_2^2$.

Now, we consider the quadratic form q that is associated to $\bar{M} \in \bar{\mathcal{M}}$. Introduce $\tilde{\mathcal{I}} = \mathcal{I} + Np$ and $\tilde{\mathcal{J}} = \mathcal{J} + Np$. Set $y_1 := \sum_{i\in\mathcal{I}} x_i$, $y_2 = \sum_{i\in\mathcal{J}} x_i$, $y_3 := \sum_{i\in\tilde{\mathcal{I}}} x_i$ and

$y_4 = \sum_{i\in\tilde{\mathcal{J}}} x_i$. By simple calculations, we get

$$\begin{aligned} q(\boldsymbol{x}) &= \alpha_1^{(1)} y_1^2 + \alpha_3^{(1)} y_2^2 + 2\alpha_2^{(1)} y_1 y_2 + \alpha_1^{(2)} y_3^2 + \alpha_3^{(2)} y_4^2 + 2\alpha_2^{(2)} y_3 y_4 + 2\delta(y_1+y_2)(y_3+y_4) \\ &= \alpha_1^{(1)} \left(y_1 + \frac{\alpha_2^{(1)}}{\alpha_1^{(1)}} y_2 + \frac{\delta}{\alpha_1^{(1)}}(y_3+y_4) \right)^2 + \frac{\Delta^{(1)}}{\alpha_1^{(1)}} \left(y_2 - \frac{\alpha_2^{(1)}\delta}{\Delta^{(1)}}(y_3+y_4) \right)^2 \\ &+ y_3^2 \left(\alpha_1^{(2)} - \frac{\delta^2}{\alpha_1^{(1)}} - \frac{(\alpha_2^{(1)})^2\delta^2}{\alpha_1^{(1)}\Delta^{(1)}} \right) + y_4^2 \left(\alpha_3^{(2)} - \frac{\delta^2}{\alpha_1^{(1)}} - \frac{(\alpha_2^{(1)})^2\delta^2}{\alpha_1^{(1)}\Delta^{(1)}} \right) \\ &+ 2y_3 y_4 \left(\alpha_2^{(2)} - \frac{\delta^2}{\alpha_1^{(1)}} - \frac{(\alpha_2^{(1)})^2\delta^2}{\alpha_1^{(1)}\Delta^{(1)}} \right), \end{aligned}$$

providing the result.

□

Therefore, we propose a second family of parametric matrices B_k in the case of heterogenous portfolios (with two groups):

$$\begin{aligned} \bar{\mathcal{B}} = \{ & B \in \mathcal{M}_{N^2\times N^2}(\mathbb{R}) \,|\, B = \bar{M}(\alpha^{(1)}, \alpha^{(2)}, \delta) + \tilde{M}(\beta_1, \beta_2, \beta_3, p) + N(\gamma_1, \gamma_2, p), \\ & \alpha^{(j)} \in \mathbb{R}_+^3,\ j = 1, 2, (\alpha^{(1)}, \alpha^{(2)}, \delta) \in \mathbb{R}_+^7 \text{ satisfies the conditions of Proposition 3.3.1}, \\ & \beta_1 \geq 0, \beta_3 \geq 0, \beta_1\beta_3 \geq \beta_2^2, \gamma_1 \geq 0, \gamma_2 \geq 0\}. \end{aligned}$$

3.3.5 Stationarity conditions

The model dynamics are specified by the N^2 equations (3.2.4). Strictly speaking, they define a Vectorial Autoregressive model of order p and dimension N^2 (or $N(N+1)/2$ to avoid redundant equations). The vector of noises $(\vec{\zeta}_t)$ is a difference martingale. In other words, setting the N^2 vector $\vec{\boldsymbol{v}}_t = [\epsilon_{it}\epsilon_{jt}]_{(i,j)\in N^2}$, its dynamics is

$$\vec{\boldsymbol{v}}_t = A + \sum_{k=1}^{q} C_k \boldsymbol{v}_{t-k} + \vec{\zeta}_t, \tag{3.3.6}$$

where $C_k := [b_{ijk,rs}]_{\{(i,j),(r,s)\in N^2\}}$, with the previous notations. Obviously, there is a one-to-one mapping between $C_1, \ldots, C_q$ and $(B_1, \ldots, B_q)$. For instance, in the case of an homogeneous portfolio, the parametrization that we proposed in Subsection (3.3.2) induces the matrices $C_k := [\alpha_k + \beta_k \mathbf{1}((i,j) = (r,s)) + \gamma_k \mathbf{1}(i = j = r = s)]_{(i,j),(,s)}$, $k = 1, \ldots, q$.

It is well-known that the system given by (3.3.6) has a unique strongly stationary solution when all complex number λ s.t.

$$det(\lambda^q I_{N^2} - \lambda^{q-1} C_1 - \ldots - \lambda C_{q-1} - C_q) = 0$$

satisfies $|\lambda| < 1$. See Hamilton (1994), for instance. Those λ are the eigenvalues of the $qN^2 \times qN^2$ matrix

$$M_C := \begin{bmatrix} 0_{N^2} & I_{N^2} & 0_{N^2} & \ldots & \ldots & 0_{N^2} \\ \vdots & 0_{N^2} & I_{N^2} & \ddots & \ldots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0_{N^2} \\ \vdots & & & & 0_{N^2} & I_{N^2} \\ C_q & C_{q-1} & \ldots & \ldots & \ldots & C_1 \end{bmatrix}.$$

Unfortunately, calculations in some simple cases show that the stationarity conditions are nonlinear functional of the model parameters. For instance, when $q = 1$ and in the case of an homogeneous portfolio, the stationarity condition is equivalent to the following: the modulus of the eigenvalues is C_1 are strictly smaller than one. In this case, simple algebraic calculations show that the characteristic polynomial of M_C is

$$\chi(x) = (\beta+\gamma-x)^{N-1}(\beta-x)^{N^2-N-1}\left(x^2 - (N^2\alpha + 2\beta + \gamma)x + (N^2\alpha + \beta + \gamma)\beta + \alpha\gamma\right).$$

Its roots are strictly smaller than one iff

$$\beta + \gamma < 1, \text{ and } (N^2\alpha + \beta + \gamma)(1 - \beta) < 1 - \beta + \alpha\gamma. \tag{3.3.7}$$

The latter condition is nonlinear. Note that it is fulfilled if $N^2\alpha + \beta + \gamma < 1$. Note that, when $N \to \infty$, (3.3.7) can be satisfied only if $\alpha(N)$ tends to zero as $O(1/N^2)$.

When $p = 2$, similar calculations allow the calculation of the characteristic polynomial of M_C, but its roots cannot be calculated analytically easily due to a four-order factor.

Remark 3.3.2. Despite that lack of explicitly written eigenvalues of M_C, some (strong) sufficient conditions for stationarity can be obtained. For instance, following Higham and Tisseur (2003) (Equation (2.12)), any eigenvalue λ of M_C satisfies

$$|\lambda| \leq \max\left(\frac{\|C_p\|_1}{\|C_{p-1}\|_1}, 2\frac{\|C_{k+1}\|_1}{\|C_k\|_1}, k = 1, \ldots, p-2\right).$$

In the case of our "homogeneous portfolio" model, $\|C_k\|_1 = N^2\alpha_k + \beta_k + \gamma_k$, and the latter sufficient condition means

$$N^2\alpha_{k+1} + \beta_{k+1} + \gamma_{k+1} \leq \frac{1}{2}(N^2\alpha_k + \beta_k + \gamma_k),$$

for any $k = 1, \ldots, p-1$. In other words, we get stationarity when the autoregressive coefficients of successive lags should decrease to zero exponentially fast (with the lag index k).

The positive definite constraint is a key hurdle since it requires particular constraint sets for the parameters. This constraint is nonlinear - space of positive definite matrices -, which hampers any flexible parameterization. Although the constraint free model is flexible, the variance covariance matrix is not directly evaluated. As for the homogeneous and heterogenous evaluations, the parameters are still constrained to obtain positive definite matrices. In the next section, we present an alternative dynamic, where the driving parameters are not constrained since the generated variance covariance matrix is positive definite by construction.

3.4 Cholesky-GARCH

Let the N-dimensional random vector ε_t s.t. $\varepsilon_t = H_t^{1/2}\eta_t$ where (η_t) is a white noise and H_t is $\mathcal{F}_t$-measurable. We observe the series $(\varepsilon_t)_{t=1,\ldots,T}$. As in Darolles et al. (2017), we propose to use the Cholesky decomposition of H_t, i.e. $H_t = L_t G_t L_t'$, where L_t is lower triangular with ones on the diagonal, and G_t is diagonal. Set $G_t = \text{diag}(g_{i,t})$ and $L_t = [\ell_{ij,t}]$, where $\ell_{ij,t} = 0$ when $j > i$.

We want to define a process for (H_t), by specifying the dynamics of (G_t) and (L_t). Set the random vectors $\boldsymbol{v}_t$ s.t. $\varepsilon_t := L_t\boldsymbol{v}_t$. Then, given $\mathcal{F}_{t-1}$, the components of $\boldsymbol{v}_t$ are uncorrelated: $Cov_{t-1}(\boldsymbol{v}_t) = G_t$. Note that $v_{1t} = \varepsilon_{1t}$ is observable.

First, we assume a dynamics for the conditional volatility of ε_{1t}:

$$\mathbb{E}[\varepsilon_{1t}^2|\mathcal{F}_{t-1}] = \mathbb{E}[v_{1t}^2|\mathcal{F}_{t-1}] = g_{1t},$$

and we assume an ARCH-type model

$$g_{1,t} = a_{1,0} + \sum_{k=1}^{m} a_{11,k} f_{k,t},$$

where every random factor $f_{k,t}$ is $\mathcal{F}_{t-1}$-measurable and for some nonnegative constants $a_{1,0}, a_{11,k}$, $k = 1, \ldots, m$. Typically, the factors f_{kt} are functions of $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ and of some of their relevant crossproducts. For instance,

$$g_{1,t} = a_{1,0} + \sum_{k=1}^{m} \sum_{j=1}^{N} a_{11,jk} \varepsilon_{j,t-k}^2, \tag{3.4.1}$$

for some nonnegative constants $a_{1,0}, a_{11,jk}$. We can estimate the latter linear equation by penalized OLS, as

$$\varepsilon_{1,t}^2 = a_{1,0} + \sum_{k=1}^{m} \sum_{j=1}^{N} a_{11,jk} \varepsilon_{j,t-k}^2 + \zeta_{11,t},$$

with $\mathbb{E}[\zeta_{11,t}|\mathcal{F}_{t-1}] = 0$. This means we can consider we "know" the process (g_{1t}).

Moreover, for every $i > 1$, we have by definition

$$\varepsilon_{it} = \sum_{j=1}^{i-1} \ell_{ijt} v_{jt} + v_{it}, \text{ or } v_{it} = -\sum_{j=1}^{i-1} \beta_{ijt} \varepsilon_{jt} + \varepsilon_{it},$$

by introducing $L_t^{-1} := [-\beta_{ij,t}]$. Then, if $i > j$, we will assume

$$\beta_{ij,t} = a_{ij,0} + \sum_{k=1}^{m} a_{ij,k} f_{k,t}, \; i > j.$$

We can estimate the latter coefficients thanks to the ordinary least squares objective funtion. For instance, we would have

$$\varepsilon_{2t} = \beta_{21t} \varepsilon_{1t} + v_{2t} = (a_{21,0} + \sum_{k=1}^{m} a_{21,k} f_{k,t}) \varepsilon_{1t} + v_{2t},$$

with $\mathbb{E}[v_{2t}|\varepsilon_{1t}] = 0$. This gives us the dynamics of $(\beta_{12,t})$.

This can be done for every couple (i, j), $i > j$, and provides the dynamics of the processes $(\beta_{ij,t})$ and then $(\ell_{ij,t})$, $i > j$, that are "known". Note that we can estimate any vector $v_{i,t}$ because we "know" L_t and we observe ε_t.

Now, we evaluate the process (g_{2t}) by noting that $v_{2t} = \varepsilon_{2t} - \ell_{12,t}\varepsilon_{1t}$ is "observed". Then, as above , we can assume a process as

$$g_{2,t} = a_{2,0} + \sum_{k=1}^{m} a_{22,k} f_{k,t}.$$

The corresponding linear regression is here

$$v_{2t}^2 = a_{2,0} + \sum_{k=1}^{m} a_{22,k} f_{k,t} + \zeta_{22,t}, \ \mathbb{E}[\zeta_{22,t}|\mathcal{F}_{t-1}] = 0.$$

And so on. Iteratively, we estimate the processes (g_{it}).

This procedure automatically generates non negative covariance matrices by construction. Moreover, the necessary and sufficient conditions to get stationary solutions of (3.4.1) are provided by Darolles, Francq and Laurent (2017). But it seems impossible to explicitly take such conditions into account during the estimation stage.

To be able to compare the size of all these coefficients, it may be useful to normalize the vector of returns. For instance, by centering and normalizing any component of ε_t, but by the unconditional volatility of every component and not by their conditional volatilities. Indeed, otherwise, this would induce some annoying constraints as

$$\sum_{j=1}^{i-1} \ell_{ij,t}^2 g_{j,t} + g_{i,t} = E_{t-1}[\epsilon_{i,t}^2] = 1,$$

for every i.

The Cholesky-GARCH process can be iteratively estimated over the index levels such that we would consider "local" OLS objective functions. Each components are observable such that the ordinary least squares objective function can be used to derive the OLS estimator.

3.5 Simulation experiments

In this section, we carry out a simulation study to explore the accuracy performance of the sparse ARCH. To do so, we consider three simulation settings, where we will compare the estimated variance covariance process to the true variance covariance process. Based on the DGP (3.2.4) and given initial values, we simulate the successive values of a MGARCH process with conditional covariance matrices (H_t) of size $N = 4$. We do this iterative procedure for $T = 10000$ and we consider 100 different variance covariance matrix patterns. Once a series is simulated, we estimate the model under different model assumptions: a scalar DCC, a homogeneous ARCH, a constraint free ARCH, a Cholesky ARCH and their penalized versions. The estimated parameters allow the calculation of successive variance covariance matrices, which are here $\hat{H}_t^{dcc}$ for the DCC model, $\hat{H}_t^{hom}$ (resp. $\hat{H}_t^{hom\star}$) for the homogeneous ARCH (resp. penalized homogeneous ARCH), $\hat{H}_t^{cf}$ (resp. $\hat{H}_t^{cf\star}$) for the constraint free ARCH (resp. penalized constraint free ARCH), and $\hat{H}_t^{cho}$ (resp. $\hat{H}_t^{cho\star}$) for the Cholesky ARCH (resp. penalized Cholesky ARCH).

The adaptive version of the Sparse Group Lasso estimator is implemented, where the first step estimator is the unpenalized OLS estimator. In Chapter 2, we described the cross-validation procedure to select the regularization parameter together with the system that determines the convergence rate of the regularization parameters to satisfy the oracle property. The lags in the homogeneous, constraint free and Cholesky models are defined a priori a follows: in the experiments 1 and 2, $q = 10$ (resp. $q = 8$) for the homogeneous model (resp. for the constraint free and Cholesky models). As for the experiment 3, $q = 20$ (resp. $q = 10$) for the homogeneous model (resp. for the constraint free and Cholesky models).

We compare the true variance covariance process and the estimated correlation processes through the aforementioned models. To do so, we specify a matrix distance, namely the Frobenius norm, defined as $||A - B||_F := \sqrt{\text{Trace}((A-B)'(A-B))}$. We compute the previous norm for each t and for

$$A = R_t, \text{ and } B \in \{\hat{H}_t^{dcc}, \hat{H}_t^{hom}, \hat{H}_t^{hom\star}, \hat{H}_t^{cf}, \hat{H}_t^{cf\star}, \hat{H}_t^{cho}, \hat{H}_t^{cho\star}\}.$$

We take the average of those quantities over T = 10000 periods of time. We obtain an average gap for all those simulations as this procedure is repeated 100 times.

Simulated experiment 1. As a particular case of (3.2.4), we consider a data generating process

$$\epsilon_{it}\epsilon_{jt} = a_{ij} + \sum_{k=1}^{q}\left((\alpha_k + \beta_k + \gamma_k \mathbf{1}(i=j))\epsilon_{i,t-k}\epsilon_{j,t-k} + \alpha_k \sum_{(r,s)\neq(i,j)} \epsilon_{r,t-k}\epsilon_{s,t-k}\right) + \zeta_{ij,t},$$

for any couple (i,j). All coefficients $(\alpha_k, \beta_k, \gamma_k)$ are set to zero except $(\alpha_4, \beta_4, \gamma_4)$, a case for which we consider different grid values. The symmetric and positive definite matrix A is simulated as $A_{ij} \sim \mathcal{U}([-0.02, 0.02])$, $i \neq j$ and $A_{ii} \sim \mathcal{U}([0.1, 0.2])$. Denoting $\omega = (\alpha_4, \beta_4, \gamma_4)$, we consider the grids

$$\begin{aligned}
\omega^{(1)} &= (0.001, 0.1, 0.2),\\
\omega^{(2)} &= (0.005, 0.3, 0.1),\\
\omega^{(3)} &= (0.01, 0.5, 0.1),\\
\omega^{(4)} &= (0.01, 0.3, 0.2).
\end{aligned}$$

For each of these 100 patterns, $\omega^{(j)}$ remains fixed for $j = 1, 2, 3, 4$ and A is simulated as described above.

We remind that $q = 10$ for the homogeneous model and $q = 8$ for both the constraint free and Cholesky processes.

TABLE 3.1: Simulated experiment 1: Average distance between true and estimated variance covariance matrices

ω	$B = \hat{H}_t^{dcc}$	$B = \hat{H}_t^{hom}$	$B = \hat{H}_t^{hom\star}$	$B = \hat{H}_t^{cf}$	$B = \hat{H}_t^{cf\star}$	$B = \hat{H}_t^{cho}$	$B = \hat{H}_t^{cho\star}$
$\omega^{(1)}$	0.2015	0.0776	0.1540	0.1042	0.0816	0.1516	0.1657
$\omega^{(2)}$	0.4647	0.1497	0.3117	0.1514	0.1401	0.3219	0.3346
$\omega^{(3)}$	1.2292	0.5341	0.7675	0.5063	0.3983	0.8386	0.8486
$\omega^{(4)}$	0.7353	0.2782	0.3545	0.2378	0.2047	0.4157	0.4350

We can highlight some interesting remarks from this simulation study. First, the DCC specification is outperformed by the competing models, especially by the homogeneous model, which is not surprising. Moreover, there is a gain in precision when applying a regularization procedure: the penalized version of the constraint free model outperforms the unpenalized version. This support the need of constraining the parameters when considering a large number of parameters, even when $N = 4$.

Simulated experiment 2. We consider a data generating process

$$\epsilon_{it}\epsilon_{jt} = a_{ij} + \sum_{k=1}^{q} \left((\alpha_k + \beta_k + \gamma_k \mathbf{1}(i=j))\epsilon_{i,t-k}\epsilon_{j,t-k} + \alpha_k \sum_{(r,s)\neq(i,j)} \epsilon_{r,t-k}\epsilon_{s,t-k} \right) + \zeta_{ij,t},$$

for any couple (i,j). We set all coefficients $(\alpha_k, \beta_k, \gamma_k)$ to zero except for $k = 4,5,6$, where we consider different grid values. A is parameterized as in simulated experiment 1. We denote $\omega = (\alpha_2, \beta_2, \gamma_2, \alpha_3, \beta_3, \gamma_3, \alpha_4, \beta_4, \gamma_4)$ and consider different grids, with

$$\begin{aligned} \omega^{(1)} &= (0.02, 0.2, 0.02, 0.01, 0.1, 0.01, 0.001, 0.01, 0.01), \\ \omega^{(2)} &= (0.001, 0.3, 0.05, 0.0005, 0.2, 0.02, 0.00001, 0.1, 0.01). \end{aligned}$$

For each of these 100 patterns, $\omega^{(j)}$ remains fixed for $j = 1,2$ and A is simulated.

We remind that $q = 10$ for the homogeneous model and $q = 8$ for both the constraint free and Cholesky processes.

TABLE 3.2: Simulated experiment 2: Average distance between true and estimated variance covariance matrices

ω	$B = \hat{H}_t^{dcc}$	$B = \hat{H}_t^{hom}$	$B = \hat{H}_t^{hom\star}$	$B = \hat{H}_t^{cf}$	$B = \hat{H}_t^{cf\star}$	$B = \hat{H}_t^{cho}$	$B = \hat{H}_t^{cho\star}$
$\omega^{(1)}$	0.4914	0.2512	0.4095	0.2079	0.1537	0.4488	0.4503
$\omega^{(2)}$	0.9787	0.5209	0.7895	0.3669	0.3364	0.7658	0.7812

The same remarks hold here as in simulated experiment 1.

Simulated experiment 3. In this experiment setting, we simulate (3.2.4) with

$$H_t = A + \sum_{k=1}^{q} (I_N \otimes \epsilon'_{t-k}) B_k (I_N \otimes \epsilon_{t-k}),$$

where we select $q = 5$. The $N^2 \times N^2$ matrices B_k are selected as $B_{ij,k} \sim \mathcal{U}([-0.2, 0.2])$ and $B_{ii,k} \sim \mathcal{U}([0.1, 0.15])$ such that they satisfy the positive definite, stationarity and "ordering" constraints. This ordering constraint, idest $\forall i,j$, $|B_{ij,k}| \leq |B_{ij,k-1}|$ for $k = 2, \cdots, 5$. As for the symmetric and positive definite matrix A, we define $A_{ij} \sim \mathcal{U}([-0.02, 0.02])$, $i \neq j$ and $A_{ii} \sim \mathcal{U}([0.1, 0.2])$. We consider two settings: setting 1, where the B_k's are not null matrices for each k; setting 2, where B_1 and B_2 are null matrices and the B_k's are not null matrices for $k = 3,4,5$. For each of these 100 patterns, the B_k and A matrices are simulated.

We remind that $q = 20$ for the homogeneous model and $q = 10$ for both the constraint free and Cholesky processes.

TABLE 3.3: Simulated experiment 3: Average distance between true and estimated variance covariance matrices

	$B = \hat{H}_t^{dcc}$	$B = \hat{H}_t^{hom}$	$B = \hat{H}_t^{hom\star}$	$B = \hat{H}_t^{cf}$	$B = \hat{H}_t^{cf\star}$	$B = \hat{H}_t^{cho}$	$B = \hat{H}_t^{cho\star}$
Setting 1	0.4044	0.4181	0.4457	0.3780	0.2024	0.2833	0.2251
Setting 2	0.2870	0.2875	0.2952	0.1979	0.1121	0.1688	0.1440

These results emphasize the good performances of the constraint free and the Cholesky processes when the observed patterns are heterogeneous. The gain in precision is significant once the adaptive SGL regularization is applied. Not surprisingly, the DCC and the homogeneous are outperformed in this simulated framework.

3.6 Conclusion

We proposed several parameterizations for multivariate ARCH models that are linear with respect to the parameters. These models can be estimated thanks to an Ordinary Least Squares procedure. Then we considered a large number of lagged values to approximate a multivariate GARCH pattern such that the optimal lag is selected thanks to a regularization procedure. To do so, the Sparse Group Lasso penalty is relevant as it fosters sparsity both at a group level and within a group. Besides, our multivariate ARCH framework is devised such that the penalized objective function is convex with convex constraints. The regularization procedure thus satisfies the oracle property and identifies the right underlying sparse model.

Our simulated experiments emphasized the abbility of the ARCH-type dynamics to outperform the scalar DCC process. More interestingly, there is a gain in regularizing the estimates once the parameter vector size becomes significant, even for small vector sizes.

Conclusion générale

Le présent manuscrit a traité du problème de la grande dimension, en particulier dans le cadre de la modélisation multivariée en temps discret. Son ambition était de proposer un nouveau processus de matrices de corrélation flexible et parcimonieux au sein des modèles MGARCH et d'en faire l'étude théorique et empirique. Il visait également à apporter des éléments théoriques aux outils de réduction de dimension de type estimateurs pénalisés.

Pour modéliser les dynamiques matricielles de corrélation, le choix a été porté sur la famille des GARCH multivariés. Le processus généralement usité est le Dynamic Conditional Correlation (DCC, Engle, 2002) en version scalaire, cas dans lequel la dynamique de corrélation nécessite l'estimation de deux paramètres si la procédure de "correlation targeting" est appliquée. Le premier chapitre a proposé une nouvelle dynamique dite vine-GARCH et dont la paramétrisation reposait sur un graphe non dirigé appelé "vine". Celui-ci décrit la structure des corrélations partielles au travers des niveaux du graphe contrôlant le degré d'information par la taille des ensembles conditionnant. Cette approche présente les avantages de générer des dynamiques définies-positives et de spécifier des processus de corrélations partielles univariés ouvrant la voie à des approches parcimonieuses. En effet, une des propriétés théoriques de la "vine" mise en évidence est la possibilité de spécifier des corrélations partielles nulles à partir d'un certain niveau du graphe et de telle sorte que la structure de celui-ci aux niveaux suivants n'a aucune influence sur la matrice de corrélation "usuelle".

Ce chapitre a proposé une étude théorique approfondie du modèle vine-GARCH. Dans un premier temps, les propriétés probabilistes d'existence et d'unicité de solutions stationnaires ont été mises en évidence. Puis les propriétés de consistance faible et normalité asymptotique de l'estimateur en deux étapes ont été démontrées. Par ailleurs,

les performances empiriques du vine-GARCH soulignent sa capacité à capturer des dynamiques complexes, en particulier lorsque les tailles des vecteurs sont significatives.

La problématique de la modélisation en grande dimension est naturellement apparue et la spécification à base de graphe peut potentiellement fournir des dynamiques parcimonieuses. De façon plus générale, les approches à base d'estimateurs pénalisés sont une approche au sein de laquelle le choix de conserver certaines variables pour prédire la variable de sortie n'est pas réalisé a priori. Le second chapitre de ce présent document propose ainsi un cadre général de M-estimateurs pénalisés dans lequel la fonction de régularisation étudiée est le Sparse Group Lasso, proposé initialement par Simon et al. (2013). Ce chapitre effectue une analyse asymptotique approfondie, non réalisée jusqu'à présent pour le Sparse Group Lasso, et propose un nouvel estimateur sparse, l'"adaptive Sparse Group Lasso" en utilisant l'idée de Zou et son adaptive Lasso (2006). Le problème de la grande dimension est traité en considérant le cadre dans lequel la taille du vecteur des paramètres diverge avec l'échantillon. Dans cet asymptotique double, le principal résultat démontré est la propriété oracle, idest la capacité de l'estimateur "adaptive Sparse Group Lasso" à identifier le support sparse théorique et sa propriété de normalité asymptotique. Pour ce faire, les vitesses de convergence des paramètres de régularisation sont explicitement données, notamment le compromis entre la pénalisation l^1 Lasso et pénalisation l^1/l^2 Group Lasso. Les résultats de simulations obtenus illustrent la capacité de l'adaptive Sparse Group Lasso à retrouver le vrai support sparse.

Dans ce cadre général M-estimateurs pénalisés, le dernier chapitre a proposé une application de ces procédures de pénalisations pour des dynamiques ARCH multivariées. Celles-ci peuvent être estimées grâce aux moindres carrés ordinaires à l'instar de l'ARCH univarié. Cette représentation linéaire permet des estimations en forme fermées et procure des gains de temps significatifs. En outre, les estimateurs pénalisés par l'"adaptive Sparse Group Lasso" vérifient la propriété oracle du fait de la convexité de la fonction objectif régularisée. Le caractère autoregressif du GARCH peut-être approximé en spécifiant un nombre significatif de retards de telle sorte que le retard optimal soit selectionné par la procédure de régularisation. L'"adaptive Sparse Group Lasso" est particulièrement adapté dans ce cas dans la mesure où les retards sont traités comme des groupes dans la composante l^1/l^2 et la réduction de dimension est

encouragée pour les groupes conservés par la composante l^1.

Pour résumer, ce projet de thèse s'est situé à la charnière de la modélisation multivariée non linéaire et de la statistique en grande dimension. Les objectifs de ce projet de thèse ont été de fournir une méthode innovante pour générer des processus de corrélation de manière flexible et parcimonieuse ainsi qu'un travail théorique portant sur des M-estimateurs pénalisés et d'en proposer des applications. L'idée directrice a été de considérer des processus pouvant accueillir des dynamiques jointes potentiellement grandes. Dans le cadre de la grande dimension, l'analyse théorique menée sur le Sparse Group Lasso a souligné sa capacité à identifier le vrai support sparse, idest à identifier les "facteurs" pertinents pour décrire la dynamique observable.

Trois principaux axes de recherche pertinents ont été identifiés à l'issue de ces travaux. D'une part, les propriétés asymptotiques d'estimateurs pénalisés seront étudiés dans le cadre où les marges sont estimées non-paramétriquement. L'idée est de considérer des combinaisons linéaires de densité de copule pour approximer la distribution jointe et d'appliquer une procédure de régularisation qui doit sélectionner le bon sous-ensemble de copules. Le second axe porterait sur les propriétés asymptotiques des M-estimateurs pénalisés pour des fonctions objectives non convexes. Pour ce faire, cela nécessite de faire appel à des résultats de consistence ne reposant pas sur des hypothèses de régularité trop fortes, idest différentiabilité et convexité. En travaillant avec une fonction objectif explicite - mais non nécessairement convexe -, les inégalités de types oracles en échantillon fini pourraient être obtenues en utilisant les inégalités de concentration données dans Massart (2003). Par exemple Chesneau et Hebiri (2008) utilisent ces outils afin d'obtenir des inégalités dites de sparsité pour des critères quadratiques. Enfin, le troisième axe de recherche porte sur la modélisation dynamique de la sparsité pour des modèles de réseaux stochastiques. Sur la base des travaux de Bühlmann et Meinshausen (2006), dans le cadre gaussien, les branches d'un graphe représentant des corrélations partielles pouvent être estimées par régressions linéaires. Dans le contexte de variables dépendantes, l'idée principale est que la sparsité peut connaître des changements de telle sorte que le graphe sous-jacent serait pénalisé différemment. Pour ce faire, l'introduction de variables latentes de type chaînes de Markov serait une approche naturelle, le paramètre de régularisation variant selon les régimes.

Bibliography

[1] Aas, K., Czado, C., Frigessi, A. and Bakken, H. (2006): *Pair-copula constructions of multiple dependence.* Working paper 487, Munich University.

[2] Aielli, G.P. (2013): *Dynamic Conditional Correlation: on Properties and Estimation.* Journal of Business & Economic Statistics, 31, 282-299.

[3] Alexander, C. (2001):*Orthogonal GARCH.* Alexander, C. (Ed.), Mastering Risk, Financial Times-Prentice Hall, London, pp. 21-28.

[4] Andersen, P.K. and Gill, R.D. (1994,a): *Cox's Regression Model for Counting Processes: A Large Sample Study.* The Annals of Statistics, Vol. 10, No. 4, 1100-1120.

[5] Andrews, D.W.K. (1994,a): *Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity.* Econometrica, Vol. 62, No. 1, 43-72.

[6] Andrews, D.W.K. (1994,b): *Empirical Process Methods in Econometrics.* Handbook of econometrics, Elsevier.

[7] Asai, M., McAleer, M. and Yu, J. (2006): *Multivariate Stochastic Volatility: a Review.* Econometric Reviews, 25, 145-175.

[8] Bauwens, L., Laurent, S. and Rombouts, J. (2006): *Multivariate GARCH models: a Survey.* Journal of Applied Econometrics, 21, 79-109.

[9] Bedford, T. and Cooke, R. (2002): *Vines - a New Graphical Model for Dependent Random Variables.* Annals of Statistics, 30(4), 1031-1068.

[10] Bertsekas, D. (1995): *Nonlinear Programming.* Athena Scientific, Belmont, MA.

[11] Billingsley, P. (1961): *The Lindeberg-Levy theorem for martingales.* Proceedings of the American Mathematical Society 12, 788-792.

[12] Billingsley, P. (1995): *Probability and Measure.* New York: John Wiley and Sons, Inc.

[13] Billio, M. and Caporin, M. (2006): *A Generalized Dynamic Conditional Correlation Model for Portfolio Risk Evaluation.* Working Paper Department of Economics, University of Venice.

[14] Bollerslev, T. (1986): *Generalized Autoregressive Conditional Heteroskedasticity.* Journal of Econometrics, 31, 307-327.

[15] Bollerslev, T., Engle, R.F. and Wooldridge, J.M. (1988): *A Capital Asset Pricing Model with Time-Varying Covariances.* Journal of Political Economy, Vol. 96, No. 1, pp. 116-131.

[16] Bollerslev, T. and Wooldridge, J.M. (1994): *Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances.* Econometric Reviews, 11, 143-172.

[17] Boswijk, H.P. and van der Weide, R. (2011): *Method of Moments Estimation of GO-GARCH Models.* Journal of Econometrics, 163, 118-126.

[18] Bougerol, P. and Picard, N. (1992): *Stationarity of GARCH processes and of some nonnegative time series.* Journal of Econometrics 52, 1714-1729.

[19] Boussama, F., Fuchs, F. and Stelzer (2011), R. (2011): *Stationarity and Geometric Ergodicity of BEKK Multivariate GARCH Models.* Stochastic Processes and their Applications 121, 2331-2360.

[20] Brechmann, E. C. and Joe, H. (2015): *Truncation of Vine Copulas Using Fit Indices.* Working paper TUM.

[21] Brechmann, E. C. and Shepsmeier, U. (2013): *Modeling Dependence with C- and D-vine Copulas: The R Package CDVine.* Journal of Statistical Software, 52, 1-27.

[22] Bühlmann, P. and van de Geer, S. (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer Series in Statistics.

[23] Cambanis, S., Huang, S. and Simons, G. (1981): *On the Theory of Elliptically Contoured Distributions.* Journal of Multivariate Analysis, 11, 368-385.

[24] Caporin, M. and McAleer, Ms. (2013): *Ten Things You Should Know about the Dynamic Conditional Correlation Representation.* Econometrics, 1, 115-126.

[25] Chernozhukov, V. (2005): *Extremal Quantile Regression.* The Annals of Statistics, Vol. 33, No. 2, 806-839.

[26] Chernozhukov, V. and Hong, H. (2004): *Likelihood Estimation and Inference in a Class of Nonregular Econometric Models.* Econometrica, Vol. 72, No. 5, 1445-1480.

[27] Chesneau, C. and Hebiri, M. (2008): *Some Theoretical Results on the Grouped Variables Lasso.* Mathematical Methods of Statistics, Vol. 17, 317326.

[28] Darolles, S., Francq, C. and Laurent, S. (2017): *Cholesky-GARCH, Theory and Application to Conditional Beta.* Working Paper CREST.

[29] Davis, R.A., Knight, K., Liu, J. (1992): *M-Estimation for Autoregressions with Infinite Variance.* Stochastic Processes and their Applications, 40, 145-180.

[30] Diebold, F.X. and Mariano, R.S. (1995): *Comparing Predictive Accuracy.* Journal of Business and Economic Statistics, 13, 253-263.

[31] Ding, Z. and Engle, R.F. (2001): *Large Scale Conditional Covariance Matrix Modeling, Estimation and Testing.* Academia Economic Papers, 29, 157-184.

[32] Dissmann, J., Brechmann, E.C., Czado, C. and Kurowicka, D. (2012): *Selecting and Estimating Regular Vine Copulae and Application to Financial Returns.* ArXiv 1202.2002.

[33] Douc, R., Moulines, E. and Stoffer, D. (2014): *Nonlinear Time Series.* Chapman and Hall.

[34] Engle R.F. (2002): *Dynamic Conditional Correlation: a Simple Class of Multivariate GARCH Models.* Journal of Business and Economic Statistics, 20, 339-350.

[35] Engle, R.F. and Colacito, R. (2006): *Testing and Valuing Dynamic Correlations for Asset Allocation.* Journal of Business and Economic Statistics, 24, 238-253.

[36] Engle, R.F. and Kroner, K. (1995): *Multivariate Simultaneous GARCH.* Econometric Theory, 11, 122-150.

[37] Engle, R.F., Ng, V. and Rothschild, M. (1990): *Asset Pricing with a Factor-ARCH Covariance Structure: Empirical Estimates for Treasury Bills.* Journal of Econometrics, 45, 213-237.

[38] Engle, R.F. and Sheppard, K. (2001): *Theoretical and Empirical Properties of Dynamic Conditional Correlation Multivariate GARCH.* Working Paper No. 8554. National Bureau of Economic Research.

[39] Fan, J. (1997): *Comments on wavelets in statistics: A review, by A. Antoniadis.* Journal of Italian Statisitcal Society, 6,131-138.

[40] Fan, J., Fan, Y. and Lv, J. (2008) *Large dimensional covariance matrix estimation using a factor model.* Journal of Econometrics, 147, 186197.

[41] Fan, J. and Li, R. (2001): *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties.* Journal of the American Statistical Association, 96:456, 1348-1360.

[42] Fan, J. and Peng, H. (2004): *Nonconcave Penalized Likelihood with a Diverging Number of Parameters.* The Annals of Statistics, Vol. 32, No. 3, 928-961.

[43] Fermanian, J.-D. and Malongo, H. (2016): *On the Stationarity of Dynamic Conditional Correlation Models.* Econometric Theory, in press.

[44] Francq, C. and Zakoïan, J.M. (2004): *Maximum Likelihood Estimation of Pure GARCH and AscA-GARCH Processes.* Bernoulli, 10(4), 605-637.

[45] Francq, C. and Zakoïan, J.M. (2010): *GARCH models.* Wiley.

[46] Fu, W.J. (1998):*Penalized Regressions: The Bridge versus the Lasso.* Journal of Computational and Graphical Statistics, 7:3, 397-416.

[47] Geyer, C.J. (1994): *On the Asymptotics of Constrained M-Estimation.* The Annals of Statistics, Vol. 22, No. 4, 1993-2010.

[48] Geyer, C.J. (1996): *On the Asymptotics of Convex Stochastic Optimization.* Unpublished manuscript.

[49] Gouriéroux, C. (1997), *ARCH Models and Financial Applications.* Springer.

[50] Gourieroux, C., Monfort, A. and Trognon, A. (1984): *Pseudo Maximum Likelihood: Theory.* Econometrica, 52, 681-700.

[51] Hamao, Y., R. Masulis and Ng, V. (1990): *Correlations in Price Changes and Volatility Across International Stock Markets.* Review of Financial Studies. 3, 281-307.

[52] Hamilton, J.D. (1994): *Time Series Analysis.* Pinceton U.P.

[53] Hastie, T., Tibshirani, R. and Wainwright, M. (2015): *Statistical Learning with Sparsity: The Lasso and Generalizations.* Chapman and Hall/CRC.

[54] Higham, N.J. (2008): *Functions of Matrices.* SIAM.

[55] N.J. Higham and Tisseur, F. (2001):*Bounds for Eigenvalues of Matrix Polynomials.* Linear Algebra and its Applications, **358**, 5-22.

[56] Hjort, N.L. and Pollard, D. (1993): *Asymptotics for Minimisers of Convex Processes.* Unpublished manuscript.

[57] Hunter, D.R and Li, R. (2005): *Variable Selection Using MM Algorithms.* The Annals of Statistics, Vol. 33, No. 4, 1617-1642.

[58] Joe, H. (2006): *Generating Random Correlation Matrices Based on Partial Correlations.* Journal of Multivariate Analysis, 97, 2177-2189.

[59] Kato, K. (2009): *Asymptotics for Argmin Processes: Convexity Arguments.* Journal of Multivariate Analysis, 1816-1829.

[60] Knight, K. and Fu, W. (2000): *Asymptotics for LASSO-type Estimators.* The Annals of Statistics, Vol. 28, No. 5, 1356-1378.

[61] Koutmos, G. and Booth, G.G. (1995): *Asymmetric Volatility Transmission in International Stock.* Journal of International Money and Finance, 14, 747-762.

[62] Kurowicka, D. (2011): *Optimal Truncation of Vines.* In Dependence Modeling: Handbook on Vine Copulae. World Scientific Publishing Co., 233-248.

[63] Kurowicka, D. and Cooke, R.M. (2003): *A Parametrization of Positive Definite Matrices in Tescs of Partial Correlation Vines.* Linear Algebra and its Applications, 372, 225-251.

[64] Kurowicka, D. and Cooke, R.M. (2006): *Completion Problem With Partial Correlation Vines.* Linear Algebra and its Applications, 418, 188-200.

[65] Kurowicka, D. and Joe, H. (2010): *Dependence Modelling, Vine Copula Handbook.* World Scientific.

[66] Lewandowski, D., Kurowicka, D. and Joe, H. (2009): *Generating Random Correlation Matrices Based on Vines and Extended Onion Method.* Journal of Multivariate Analysis, 100, 1989-2001.

[67] Li, X., Mo, L., Yuan, X., Zhang, J. (2014): *Linearized Alternating Direction Method of Multipliers for Sparse Group and Fused LASSO models.* Computational Statistics and Data Analysis, 79, 203-221.

[68] Liao, A. and Williams, J. (2004): *Volatility Transmission and Changes in Stock Market Interdependence in the European Community.* European Review of Economics and Finance, 3, 203-231.

[69] Ling, S. and McAleer, M. (2003): *Asymptotic theory for a vector AscA-GARCH model.* Econometric Theory 19, 280-310.

[70] Lütkepohl, H. (1996): *Handbook of Matrices.* Wiley.

[71] Massart, P. (2003): *Concentration Inequalities and Model Selection.* Springer.

[72] Meinshausen, N. and Bühlmann, P. (2006): *High-Dimensional Graphs and Variable Selection with the Lasso.* The Annals of Statistics, Vol. 34, No. 3, 1436-1462.

[73] Nardi, Y. and Rinaldo, A. (2008): *On the Asymptotic Properties of the Group LASSO Estimator for Linear Models.* Electronic Journal of Statistics, Vol. 2, 605-633.

[74] Neumann, M.H. (2013): *A Central Limit Theorem for Triangular Arrays of Weakly Dependent Random Variables, with Applications in Statistics.* Probability and Statistics, Vol. 17, 120-134.

[75] Newey, W.K. and Powell, J.L. (2008): *Asymmetric Least Squares Estimation and Testing.* Econometrica, Vol. 55, No. 4, 819-847.

[76] Patton, A.J. and Sheppard, K. (2009): *Evaluating Volatility and Correlation Forecasts.* in Mikosch, T., Kreib, J.-P., Davis, R.A., Andersen, T.G. (Eds.), Handbook of Financial Time Series. Springer Berlin Heidelberg, 801-838.

[77] Poignard, B. and Fermanian, J.-D. (2016): *Vine-GARCH Process: Stationarity and Asymptotic Properties.* Working paper CREST 2016-03.

[78] Rio, E. (2013): *Asymptotic Theory of Weakly Dependent Random Processes.* Springer.

[79] Rockafeller, R.T. (1970): *Convex analysis.* Princeton University Press, Princeton.

[80] Shiryaev, A.N. (1991): *Probability.* Second Edition, Springer.

[81] Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2013): *A sparse group lasso.* Journal of Computational and Graphical Statistics, 22:2, 231-245.

[82] Tibshirani, R. (1996): *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society. Series B, Vol. 58, No. 1, pp. 267-288.

[83] Tibshirani, R. and Suo, X. (2016): *An Ordered Lasso and Sparse Time-Lagged Regression.* Technometrics, Vol. 58, 415-423.

[84] Tse, Y.K. and Tsui, A.K.C. (2002): *A multivariate GARCH Model with Time-Varying Correlations.* Journal of Business and Economic Statistics, 20, 351-362.

[85] Tweedie, R.L. (1988): *Invariant Measure for Markov Chains with no Irreductibility Assumptions.* Journal of Applied Probability 25A, 275-285.

[86] van der Weide, R. (2002): *GO-GARCH: a Multivariate Generalized Orthogonal GARCH Model.* Journal of Applied Econometrics, 17, 549-564.

[87] Veraverbeke, N., Omelka, M. and Gijbels, I. (2011): *Estimation of a Conditional Copula and Association Measures.* Scandinavian Journal of Statistics, 38, 766-780.

[88] White, H. (1994): *Estimation Inference and Specification Analysis.* Cambridge University Press.

[89] Yuan, M. and Lin, Y. (2006): *Model Selection and Estimation in Regression with Grouped Variables.* Journal of the Royal Statistical Society. Series B, Vol. 68, No. 1, pp. 49-67

[90] Zou, H. (2006): *The Adaptive LASSO and its Oracle Properties.* Journal of the American Statistical Association, 101:476, 1418-1429.

[91] Zou, H. and Li, R. (2008): *One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models.* The Annals of Statistics, Vol. 36, No. 4, 1509-1533.

[92] Zou, H. and Zhang, H.H. (2009): *On the Adaptive Elastic-Net with a Diverging Number of Parameters.* The Annals of Statistics, Vol. 37, No. 4, 1733-1751.

Résumé

Ce document traite du problème de la grande dimension dans des processus GARCH multivariés. L'auteur propose une nouvelle dynamique vine-GARCH pour des processus de corrélation paramétrisés par un graphe non dirigé appelé "vine". Cette approche génère directement des matrices définies-positives et encourage la parcimonie. Après avoir établi des résultats d'existence et d'unicité pour les solutions stationnaires du modèle vine-GARCH, l'auteur analyse les propriétés asymptotiques du modèle. Il propose ensuite un cadre général de M-estimateurs pénalisés pour des processus dépendants et se concentre sur les propriétés asymptotiques de l'estimateur "adaptive Sparse Group Lasso". La grande dimension est traitée en considérant le cas où le nombre de paramètres diverge avec la taille de l'échantillon. Les résultats asymptotiques sont illustrés par des expériences simulées. Enfin dans ce cadre l'auteur propose de générer la sparsité pour des dynamiques de matrices de variance covariance. Pour ce faire, la classe des modèles ARCH multivariés est utilisée et les processus correspondants à celle-ci sont estimés par moindres carrés ordinaires pénalisés.

Abstract

This document contributes to high-dimensional statistics for multivariate GARCH processes. First, the author proposes a new dynamic called vine-GARCH for correlation processes parameterized by an undirected graph called vine. The proposed approach directly specifies positive definite matrices and fosters parsimony. The author provides results for the existence and uniqueness of stationary solution of the vine-GARCH model and studies its asymptotic properties. He then proposes a general framework for penalized M-estimators with dependent processes and focuses on the asymptotic properties of the adaptive Sparse Group Lasso regularizer. The high-dimensionality setting is studied when considering a diverging number of parameters with the sample size. The asymptotic properties are illustrated through simulation experiments. Finally, the author proposes to foster sparsity for multivariate variance covariance matrix processes within the latter framework. To do so, the multivariate ARCH family is considered and the corresponding parameterizations are estimated thanks to penalized ordinary least square procedures.

Mots Clés

Corrélations partielles, Estimateur du quasi-maximum de vraisemblance, M-estimateurs pénalisés, Propriété oracle, Stationnarité, Vine régulière.

Keywords

Oracle property, Partial correlations, Penalized M-estimators, Quasi-maximum likelihood estimator, Regular vine, Stationarity.