Modélisation multivariée de variables météorologiques

Augustin Touron

NNT : 2019SACLS264

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Laboratoire d'accueil : Laboratoire de mathématiques d'Orsay, UMR 8628 CNRS

Spécialité de doctorat : Mathématiques appliquées

Augustin Touron

Modélisation multivariée de variables météorologiques

Date de soutenance : 19 septembre 2019

Après avis des rapporteurs : M. Denis ALLARD (INRA)
M. Axel MUNK (Georg-August-Universität Göttingen)

Jury de soutenance :

M. Denis ALLARD	(INRA) Rapporteur
Mme Élisabeth GASSIAT	(Université Paris-Sud) Directrice de thèse
Mme Thi Thu Huong HOANG	(EDF R&D) Encadrante EDF
M. Axel MUNK	(Georg-August-Universität Göttingen) Rapporteur
M. Gilles STOLTZ	(Université Paris-Sud) Président
M. Mathieu VRAC	(LSCE) Examinateur

Remerciements

La réussite de ma thèse a tenu en partie à la qualité de son encadrement. Je tiens donc à remercier Élisabeth Gassiat, ma directrice de thèse et directrice du LMO, ainsi que Thi Thu Huong Hoang et Sylvie Parey qui m'ont encadré à EDF. Elles m'ont apporté non seulement leur expertise technique, mais aussi leurs encouragements, leur bienveillance, leur confiance, et ont su me guider progressivement dans cet apprentissage de la recherche qu'est le doctorat. Je remercie aussi chaleureusement Yohann de Castro et Sylvain Le Corff pour leur précieuse contribution à l'encadrement de ma thèse, leurs conseils avisés et leur sympathie.

Mes remerciements vont également aux membres de mon comité de suivi de thèse, pour le temps qu'ils m'ont accordé et leurs remarques toujours pertinentes. Outre ceux que j'ai déjà cités, je remercie donc Pierre Ailliot (Université de Brest), Joël Gailhard (EDF DTG), Luc Lehéricy (LMO), Paul-Antoine Michelangeli (EDF R&D), Aurélien Ribes (Météo France) et Mathieu Vrac (LSCE). En particulier, merci à Joël pour les échanges que nous avons eus et le travail que nous avons mené ensemble, qui m'a permis d'écrire le Chapitre 6 de cette thèse, et à Luc pour son implication dans mon travail tout au long de ma thèse et notre fructueuse collaboration parallèlement à sa propre thèse.

I am thankful to Denis Allard and Axel Munk who accepted to read my thesis and to review it in a detailed and thorough manner. Merci aussi à Gilles Stoltz et Mathieu Vrac qui ont gentiment accepté de faire partie de mon jury.

La réalisation de cette thèse a nécessité de nombreuses heures de calcul numérique. Le cluster de calcul du LMO m'a été d'une grande utilité et j'adresse mes remerciements aux ingénieurs de recherche du LMO grâce auxquels ces moyens de calcul sont disponibles : Benjamin Auder, Sylvain Faure, Hugo Leclerc et Suzanne Varet.

Merci à mes collègues doctorants avec qui j'ai eu le plaisir de partager un bureau et de faire des mathématiques : Martin Royer et Luc Lehéricy à Orsay, Margaux Brégère à EDF. Merci aussi à tous les membres du groupe météo d'EDF pour leur accueil en leur sein et leur sympathie au quotidien pendant ces trois ans.

Last but not least, je remercie mes parents pour leur infatigable soutien.

Table des matières

Chapitre 1

Introduction

1.1 Contexte et objectif

L'*aléa climatique* joue un rôle majeur dans les activités d'un producteur et fournisseur d'électricité comme EDF. En premier lieu, la température est un facteur déterminant de la consommation électrique, en France particulièrement, principalement de par l'importance du parc de chauffages électriques : on parle de *thermosensibilité*. D'après RTE, en période hivernale, une chute de 1°C correspond à une augmentation de 2400MW de la consommation d'électricité (ce qui représente 2 ou 3 réacteurs nucléaires). Ainsi, chaque hiver, les vagues de froid correspondent à des pics de consommation électrique. La thermosensibilité est aussi présente en été, dans une moindre mesure : RTE estime qu'une augmentation de 1°C des températures estivales correspond à une augmentation de 500MW de la consommation électrique, cette augmentation étant due à la mise en route des ventilateurs, climatisations et autres systèmes de refroidissement. La thermosensibilité estivale risque d'augmenter dans les années à venir à cause du réchauffement climatique. En outre, la production d'électricité renouvelable dépend largement des variables climatiques :

- La production éolienne dépend entièrement de la force du vent sur les différents lieux de production.
- La production photovoltaïque dépend principalement de l'intensité du rayonnement solaire (qui varie avec la saison, l'heure dans la journée, et la couverture nuageuse), mais aussi de la température (en cas de forte chaleur, le rendement des panneaux photovoltaïques chute).
- La production hydroélectrique dépend des précipitations et de la température (de celle-ci dépendent l'évaporation et l'enneigement, et donc le débit des cours d'eau, voir chapitre 6).

Comme on ne peut pas stocker l'électricité à grande échelle, il est vital pour le réseau électrique que *l'équilibre offre-demande* soit assuré : à chaque instant, la production d'électricité (offre) doit être égale à sa consommation (demande). EDF doit s'assurer que cette condition pourra être remplie à différents horizons de temps (du temps réel à plusieurs décennies). On comprend donc, compte-tenu des éléments que nous venons de donner, qu'il est nécessaire de tenir compte de l'aléa climatique à différents échelles spatiales et temporelles afin d'assurer l'équilibre entre la consommation et la production. Et ce d'autant plus que la sensibilité de cet équilibre aux variables climatiques ira croissante avec l'augmentation de la part des énergies renouvelables dans le mix énergétique. L'aléa climatique, et notamment ses valeurs extrêmes, doit également

être considéré pour le dimensionnement des infrastructures, qu'il s'agisse du réseau électrique ou des moyens de production : refroidissement des centrales nucléaires, dimensionnement des barrages, choix de l'emplacement des parcs éoliens ou photovoltaïques, par exemple. Si l'on veut adopter une approche globale de ces problématiques, il n'est pas envisageable de traiter chaque variable climatique séparément puisqu'elles n'évoluent pas de façon indépendante : elles sont liées entre elles par des relations de dépendance complexes. Enfin, l'augmentation de la part des énergies renouvelables va de pair avec une *décentralisation* de la production électrique, ce qui implique que l'aléa climatique doit être étudié à l'échelle locale.

Cette thèse émane de la nécessité, pour les activités d'EDF, d'envisager les variables climatiques de façon *multivariée* et à *l'échelle locale*.

Concrètement, pour réaliser des études d'impact sur la consommation, la production, et leur équilibre, on a recours à des simulations pour explorer un maximum de scénarios possibles. S'agissant des variables climatiques, une façon de générer ces scénarios est d'avoir recours à un *générateur stochastique de temps* (ici "temps" est à comprendre au sens météorologique !). L'objectif principal de cette thèse est la conception d'un tel outil de simulation. Dans la prochaine section, nous traitons des générateurs stochastiques de temps dans leur généralité, avant de spécifier davantage les caractéristiques de ceux que nous avons étudiés.

1.2 Générateurs stochastiques de temps

Dans cette section, nous répondons de manière synthétique aux questions suivantes : qu'est-ce qu'un générateur de temps ? A quoi cela sert-il ? De quelle manière sont-ils construits ? Comment sont-ils évalués ?

Un générateur stochastique de temps (*stochastic weather generator*) est un modèle statistique destiné à simuler rapidement de longues séries chronologiques de variables météorologiques : température, vitesse/direction du vent, rayonnement solaire, précipitations... etc. Le plus souvent, le pas de temps est journalier. On s'intéresse par exemple à la température moyenne (ou minimale, maximale) sur une journée, ou aux cumuls journaliers des précipitations. Le pas de temps peut aussi être infra-journalier. On peut chercher à simuler des variables météorologiques en un point de l'espace, ou simultanément en plusieurs endroits, en tenant compte des corrélations spatiales. Dans ce cas on parle de modèle *multisite*. On peut s'intéresser à une seule variable, ou chercher à simuler simultanément plusieurs variables de façon cohérente. Nous parlerons alors de modèle *multivarié*. Pour écarter toute confusion, précisons qu'un générateur stochastique de temps n'a pas pour objectif la prédiction de variables climatiques à un certain horizon futur. Il se contente de fournir des simulations réalistes de telles variables. En général, les générateurs stochastiques de temps n'incluent pas les équations physiques qui régissent les grandeurs météorologiques. Ils sont calibrés de manière purement statistique à partir d'un historique d'observations. L'avantage de cette approche est son faible coût en temps de calcul, comparativement à des modèles physiques qui nécessitent de résoudre des équations aux dérivées partielles issues de la mécanique des fluides, et qui par conséquent sont beaucoup plus coûteux numériquement.

Outre la production d'énergie, qui nous intéresse dans cette thèse, les générateurs stochastiques de temps ont diverses applications. En particulier, ils sont utilisés dans le domaine agronomique pour produire des variables d'entrée dans des modèles de simulation de cultures, qui modélisent par exemple le rendement d'une culture en fonction de différents facteurs, dont la température, les précipitations et le rayonnement (voir par exemple Brisson et al. (2003)). Ils peuvent aussi être utilisés comme méthode d'imputation de données manquantes, dont la fiabilité est douteuse, ou encore dont la résolution temporelle est trop grossière. Dans le même esprit, un générateur

de temps peut avoir pour objectif la *descente d'échelle* statistique (Wilby and Wigley, 1997). En modélisant le lien entre les variables grande échelle (par exemple les champs géopotentiels) et les variables locales (température, précipitations...), un générateur stochastique peut produire des séries de variables locales compatibles avec les conditions globales (grande échelle). Une telle procédure peut être utilisée, par exemple, pour étudier l'impact du changement climatique sur certaines variables à l'échelle locale, à partir de scénarios climatiques à plus grand échelle. L'exemple des précipitations est détaillé dans Wilks (2010). Une autre utilisation peut-être moins connue des générateurs de temps concerne la finance, en particulier pour la valorisation des dérivés climatiques. Ces instruments financiers dont la valeur est indexée sur des phénomènes météorologiques sont utilisés pour se prémunir du risque climatique. Voir par exemple Campbell and Diebold (2005) ou Benth and Šaltytė Benth (2011) pour des modèles de températures dans ce cadre.

Historiquement, les générateurs stochastiques de temps ont d'abord concerné le processus des précipitations, en commençant par le processus des occurrences, c'est-à-dire le processus à valeurs dans $\{0, 1\}$ tel que 0 correspond à l'absence de pluie (jour sec) et 1 correspond à un jour pluvieux. On s'est rapidement aperçu (Newnham, 1916) que ce processus n'était pas indépendant : s'il a plu hier, la probabilité pour qu'il pleuve aujourd'hui est plus grande que s'il n'a pas plu hier. Ainsi les jours pluvieux et les jours secs ont tendance à s'agréger pour former des épisodes secs ou pluvieux. Le modèle le plus simple pour un tel processus, une chaîne de Markov d'ordre 1 à deux états (sec et pluvieux), a été proposé par Gabriel and Neumann (1962). Dans un tel modèle, l'état à un l'instant t (0 ou 1) ne dépend que de l'état à l'instant $t-1$, et les durées des épisodes pluvieux/secs suivent des lois géométriques. Dans Katz (1977), ce modèle est étendu de façon à modéliser le processus des occurrences mais aussi le processus des intensités, c'est-à-dire des cumuls journaliers de précipitations. Le processus des occurrences $(J_t)_{t\geq 1}$ est d'abord simulé, puis, lorsque $J_t = 1$ (jour t pluvieux), l'intensité X_t est simulée selon la loi F_0 si $J_{t-1} = 0$ et F_1 si $J_{t-1} = 1$. Les distributions F_0 et F_1 sont choisies parmi la famille des lois gamma. Par la suite, l'idée de générer d'abord le processus des occurrences de pluie, par une chaîne de Markov ou par un autre moyen (voir par exemple Racsko et al. (1991)), puis de générer les autres variables conditionnellement à l'occurrence de pluie, a été abondamment reprise. Le lecteur intéressé pourra consulter Wilks and Wilby (1999) ou Srikanthan and McMahon (2001) pour une revue de ces modèles. L'un des modèles les plus cités dans la littérature est le WGEN (Weather GENerator) de Richardson (1981). Dans Richardson (1981), les autres variables (températures minimale et maximale, rayonnement) sont générées conditionnellement au processus pluie/non pluie par un processus auto-régressif vectoriel gaussien d'ordre 1. Ce modèle a ensuite été étendu pour inclure d'autres variables (voir par exemple Parlange and Katz (2000)).

Modèles à espace d'états

Ces modèles où les variables sont simulées conditionnellement à l'occurrence de pluie peuvent être vus comme un cas particulier de modèles à espace d'état. Ici on a deux états : *sec* et *pluvieux*. Ce concept peut être généralisé pour engendrer toute une famille de générateurs de temps. Pour une revue récente de tels modèles, le lecteur pourra se reporter à Ailliot et al. (2015a). Ils ont tous en commun de faire intervenir une variable discrète, l'*état*, qui détermine en grande partie la dynamique des variables à simuler. Cette définition, volontairement vague, peut recouvrir un grand nombre de situations.

Les états peuvent être définis à partir de descripteurs de conditions météorologiques à grande échelle. Typiquement, on réalise une classification des champs de pression à l'échelle continentale pour déterminer des *régimes de temps* desquels vont dépendre, via des lois conditionnelles, les variables à simuler à l'échelle locale. Un tel choix offre l'avantage d'être directement physiquement interprétable. Par exemple, dans Wilson et al. (1992), les auteurs obtiennent une classification

en 4 classes à partir de l'analyse en composantes principales d'une base de données constituée de plusieurs champs géopotentiels, de vitesse de vent et de température le long de différentes surfaces isobares, en retenant les deux premières composantes principales, suivie par une classification k-means. Cette classification est ensuite utilisée pour définir un générateur stochastique de précipitations sur plusieurs sites. Dans Garavaglia et al. (2010), 8 types de temps sont définis sur la France à partir de 4 champs géopotentiels et de chroniques de précipitations dans le Sud-Est de la France, dans l'optique d'une simulation spatio-temporelle des précipitations extrêmes. Dans Boé and Terray (2008), les auteurs montrent que la variabilité interannuelle des précipitations hivernales en France est en grande partie expliquée par les fréquences d'occurrence de types de temps à grande échelle déterminé à partir du champ de pression au niveau de la mer (SLP).

Dans certains cas, les variables atmosphériques à grande échelle peuvent ne pas être de bons descripteurs des variables que l'on cherche à simuler à l'échelle locale. Les états, ou types de temps, peuvent alors être définis directement par une classification sur les variables cibles (celles que l'on cherche à modéliser). Ainsi Flecher et al. (2010) définissent des types de temps saison par saison en subdivisant la dichotomie classique *jour pluvieux/jour sec* à l'aide d'un algorithme de clustering hiérarchique sur leurs variables cibles : températures minimales et maximales, rayonnement, vitesse du vent et précipitations. Les transitions entre les états sont ensuite modélisées par une chaîne de Markov. Le clustering hiérarchique est aussi utilisé par Vrac et al. (2007) qui définissent des types de temps à partir des précipitations observée sur un réseau de stations. Qu'elles soient basées sur des variables à grande échelle ou sur des variables locales, de telles classifications *a priori* sont efficaces à condition d'avoir identifié les bonnes variables explicatives, la bonne méthode de classification et le bon nombre d'états.

Générateurs de temps et modèles de Markov cachés

Pour gagner en flexibilité et pouvoir capturer une plus grande variété de phénomènes tout en conservant l'approche par types de temps, on peut considérer ces derniers comme des variables latentes, non observées. Nous parlerons d'*états cachés*. Dans ce cas, la détermination des types de temps est guidée directement par les données au lieu de dépendre d'une classification a priori ou d'une variable exogène. Elle est donc adaptée aux variables locales que nous cherchons à modéliser. Il n'est pas nécessaire que les états cachés possèdent une interprétation physique évidente, ni qu'ils correspondent à une situation météorologique identifiable. Cependant, nous verrons que dans certains cas, l'analyse *a posteriori* des paramètres associés aux lois conditionnelles permet d'interpréter les états cachés. L'exemple le plus simple d'une telle modélisation est le *modèle de Markov caché* (Hidden Markov Model, HMM). Dans ce modèle, sur lequel nous reviendrons abondamment, la séquence des états cachés est une chaîne de Markov, et conditionnellement aux états, les observations sont générées de façon indépendante. Zucchini and Guttorp (1991) ont introduit un HMM pour modéliser le processus des occurrences de précipitations simultanément sur un réseau de N sites. En raison de leur simplicité et de leur flexibilité, les modèles de Markov cachés sont devenus un outil très populaire pour construire des générateurs de temps. Voir par exemple Ailliot et al. (2009) pour une modélisation spatio-temporelle des précipitations (occurrence et intensité) via un HMM. L'hypothèse de stationnarité de la chaîne de Markov associée au modèle de Markov caché standard étant souvent trop restrictive, des extensions des HMM ont été utilisées. Pour modéliser la vitesse et la direction du vent à un pas de temps infra-journalier sur plusieurs stations, Ailliot and Monbet (2012) permettent aux transitions entre les états cachés de dépendre du temps de façon périodique, de façon à reproduire le cycle diurne du vent. Une autre façon d'introduire une inhomogénéité dans la chaîne de Markov est de supposer que les probabilités de transitions entre les états s'expriment comme fonctions d'une variable exogène, comme dans Hughes et al. (1999) ou Bellone et al. (2000) pour les précipitations. Les HMM peuvent aussi être généralisés en abandonnant l'hypothèse d'indépendance conditionnelle. Par

exemple, on peut autoriser la variable cible à l'instant t à ne pas dépendre uniquement de l'état au même instant, mais aussi de la variable cible à l'instant précédent. Un tel modèle, appelé MS-AR (Markov Switching Auto Regressive), a été utilisé pour la modélisation du vent par Ailliot and Monbet (2012), ou encore par Kirshner (2005) pour la modélisation multisite des précipitations. La flexibilité des HMM provient aussi de la possibilité de choisir une modélisation non paramétrique pour les lois conditionnelles. Par exemple, Lambert et al. (2003) introduisent un modèle de Markov caché non paramétrique pour les précipitations.

Autres types de modèles

Les modèles à espace d'état ne sont pas la seule manière de générer des variables climatiques. Une autre grande catégorie de générateurs de temps est constituée par les méthodes de rééchantillonage. Elles consistent par exemple à générer une série synthétique en tirant aléatoirement des valeurs parmi un ensemble d'*analogues*, issus des observations. Dans Rajagopalan and Lall (1999), les auteurs utilisent un critère des plus proches voisins pour définir ces analogues. Une autre manière d'obtenir un rééchantillonage est d'utiliser le *bootstrap par blocs*. Celui-ci consiste en la construction d'une série en tirant aléatoirement des blocs d'observations consécutives, les longueurs des blocs étant éventuellement aléatoires (Politis and Romano, 1994). L'avantage de ces méthodes est leur capacité à produire des séries synthétiques réalistes, pourvu que les effets de la saisonnalité soient correctement gérés. Ils présentent en revanche l'inconvénient d'une moindre variabilité puisqu'ils sont incapables de générer des valeurs qui ne figurent pas déjà parmi les observations. Dans le chapitre 6, nous comparons les performances d'une méthode d'analogues avec celles d'un modèle de Markov caché, dans le cadre d'une application hydrologique. Récemment, des modèles hybrides ont été développés, mêlant processus stochastiques calibrés statistiquement et équations issues de la physique liant les différentes variables à modéliser. Ainsi, dans Peleg et al. (2017), les auteurs introduisent un modèle capable de générer température, rayonnement, humidité, pression atmosphérique, vent, précipitations et couverture nuageuse, à haute résolution spatiale et temporelle.

Traitement de la saisonnalité

Au pas de temps journalier, les séries de variables météorologiques ne sont pas stationnaires : elles présentent toujours des saisonnalités de période 1 an, et parfois des tendances liées au changement climatique. Ces saisonnalités peuvent se présenter sous différentes formes. Par exemple, la température présente une saisonnalité en moyenne, mais aussi en variance, laquelle est plus élevée en hiver qu'en été. Même lorsque l'on considère la série centrée et réduite, il subsiste un comportement saisonnier, notamment dans les auto-corrélations. On observe aussi, sur toute série de température suffisamment longue, une tendance croissante, illustration du réchauffement climatique. Les séries de précipitations présentent également une composante saisonnière, dont la forme et l'amplitude sont très variables selon la zone climatique. Cette saisonnalité intervient de deux manières : dans l'occurrence des précipitations (à certaines périodes de l'année il pleut plus fréquemment qu'à d'autres) et dans leur intensité. Ainsi en Europe on observe généralement des précipitations plus fréquentes mais en moyenne moins intenses en hiver qu'en été, avec toutefois des comportements différents entre l'Europe du Nord et l'Europe du Sud. On constate parfois, mais pas toujours, la présence de tendances lorsque l'on considère certaines statistiques liées aux précipitations : cumuls annuels, estivaux, hivernaux, occurrence, intensité, valeurs extrêmes (Van den Besselaar et al., 2013)... Ces tendances peuvent être dans un sens ou dans l'autre selon les régions (là aussi, on constate des différences entre l'Europe du Nord et l'Europe du Sud [1]). Le vent et le rayonnement présentent eux aussi de très nettes saisonnalités. On comprend donc

1. Voir par exemple https://www.eea.europa.eu/data-and-maps/indicators/european-precipitation-2/assessment

que ces phénomènes non stationnaires doivent être pris en compte lors de la construction d'un générateur de temps. Pour ce faire, nous avons distingué trois types d'approche :

1. La façon la plus simple de procéder, et aussi de loin la plus fréquente dans la littérature sur les générateurs de temps, est de regrouper les données par mois ou par "saison". Les saisons peuvent être les 4 saisons "classiques" de 3 mois (DJF, MAM, JJA, SON) comme dans Wilson et al. (1992), mais d'autres choix peuvent être faits (saisons de 2 mois dans Zucchini and Guttorp (1991), 6 mois dans Parlange and Katz (2000)...) selon les particularités de la série que l'on cherche à modéliser. Par exemple, si l'on partitionne les données en les 12 mois de l'année, les paramètres du modèle seront ajustés séparément pour chaque mois (voir Wilks (1998) par exemple). Cette approche a l'avantage de la simplicité puisqu'aucun effort de modélisation de la saisonnalité n'est à fournir. Cependant elle repose sur l'hypothèse forte de la stationnarité sur chaque sous-période. Le choix de la longueur des "blocs" de données peut alors être vu comme un compromis biais-variance. Des blocs courts rendent plus crédible l'hypothèse de stationnarité mais réduisent la quantité de données disponibles pour estimer les paramètres des sous-modèles, qui sont alors plus nombreux. A contrario, des blocs longs (6 mois par exemple) permettent d'avoir davantage de données pour l'estimation, mais au risque de mettre en péril la stationnarité. Le choix de regrouper les données en 12 mois ou en 4 saisons peut alors paraître arbitraire. D'autre part, la distribution des variables météorologiques varie en général de façon lisse au cours de l'année : celle du mois de janvier est proche de celle du mois de février par exemple. Traiter chaque mois/saison indépendemment revient à considérer que l'évolution de cette distribution est constante par morceaux, et induit donc une perte d'information. Le modèle global (sur l'année entière) risque donc de comporter plus de paramètres que nécessaire. Enfin, puisque l'objectif est la simulation de longues séries temporelles, il est préférable d'être capable de simuler une année entière avec un unique modèle, et ainsi d'éviter d'effectuer des "recollements" à la fin de chaque mois/saison.
2. Une autre approche courante est le prétraitement des données brutes pour obtenir un résidu stationnaire. On parle alors de *stationnarisation* ou de *désaisonnalisation*. L'exemple le plus simple d'un tel traitement est la décomposition moyenne-variance suivante :

$$X_t = m(t) + \sigma(t)Z_t,$$

où X_t est la variable à modéliser, $m(t)$ et $\sigma(t)$ sont des fonctions déterministes à estimer pouvant inclure des tendances et/ou des saisonnalités, et Z_t est un bruit centré de variance unitaire. Après avoir estimé $m(\cdot)$ et $\sigma(\cdot)$, on modélise Z_t, on le simule, puis on construit une simulation de la variable brute en utilisant les saisonnalités et tendances estimées. Une telle approche est utilisée par Dacunha-Castelle et al. (2015) pour la modélisation de la température. Dans leur modèle, $m(\cdot)$ et $\sigma(\cdot)$ comportent tous deux une tendance et une saisonnalité. Malgré une procédure sophistiquée de stationnarisation, le résidu présente toujours une saisonnalité dans les moments d'ordre supérieurs et dans les auto-corrélations. Cet exemple montre que la stationnarisation des variables climatiques peut être une entreprise difficile à cause de la complexité de leur structure. Dans Lambert et al. (2003), les auteurs désaisonnalisent les précipitations mois par mois, en retirant la moyenne mensuelle et en divisant par la variance mensuelle. Le résidu est alors modélisé par un HMM non paramétrique. Au lieu de la moyenne et de la variance, les médiane et écart moyen absolu, plus robustes aux valeurs extrêmes, sont parfois utilisés pour la stationnarisation. C'est par exemple le choix retenu par Flecher et al. (2010) pour désaisonnaliser température et rayonnement solaire.
3. Enfin, on peut modéliser la saisonnalité en introduisant directement des coefficients périodiques dans le modèle. Souvent, comme les variations saisonnières sont relativement lisses, il

suffit pour cela d'utiliser des polynômes trigonométriques de faible degré (1 ou 2) dont les coefficients constituent alors des paramètres à estimer. Dans le cadre des modèles de Markov cachés ou leurs généralisations, la saisonnalité peut être introduite à deux niveaux : dans les paramètres des lois d'émission et/ou dans les probabilités de transition (la chaîne de Markov cachée sous-jacente devient alors non-homogène). Dans Ailliot and Monbet (2012), les auteurs modélisent la vitesse du vent au pas de temps infra-journalier (4 observations par jour) en utilisant un modèle MS-AR (Markov Switching Auto-Regressive) dont les coefficients sont formés à partir de fonctions trigonométriques avec deux périodes différentes : annuelle et journalière. Ils envisagent également des probabilités de transition périodiques sous la forme

$$P(S_t = j \mid S_{t-1} = i) \propto q_{ij} \exp\left(\kappa_j \cos\left(\frac{2\pi}{T} t - \phi_j\right)\right).$$

Toujours dans le cadre des HMM, la saisonnalité peut être introduite de manière implicite dans les probabilités de transition par l'intermédiaire d'une variable exogène, qui elle est observée. Par exemple, dans Bellone et al. (2000), les précipitations sont modélisées à partir d'un modèle de Markov caché non homogène dont les transitions sont données par

$$P(S_t = j \mid S_{t-1} = i, \mathbf{X}_t) \propto \gamma_{ij} \exp\left[-\frac{1}{2}(\mathbf{X}_t - \boldsymbol{\mu}_{ij})\boldsymbol{\Sigma}^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_{ij})^\top\right],$$

où $\mathbf{X}_t$ est un vecteur de variables atmosphériques comprenant hauteurs de géopotentiel, températures et humidités relatives. [2]

Dans le cadre de cette thèse, et pour les raisons que nous avons données, nous avons préféré la troisième approche aux deux premières. Ainsi les HMM que nous utilisons comportent tous des paramètres périodiques pour traiter la saisonnalité, dans les probabilités de transition, les lois d'émission, et parfois les deux à la fois. L'introduction de saisonnalités et de tendances a été l'occasion de développements théoriques sur ces modèles de Markov cachés non homogènes (voir les chapitres 3 et 4).

Validation des générateurs de temps.
Une fois un modèle défini et ses paramètres estimés, la performance du générateur de temps doit être évaluée : fait-il ce qu'on attend de lui ? Ce qu'on attend d'un générateur de temps, c'est que les simulations qu'il produit soient réalistes. On est alors amené à se demander : que signifie *réaliste* ? Il est clair qu'aucun générateur de temps, aussi sophistiqué soit-il, ne parvient à reproduire exactement la distribution des variables qu'il prétend simuler. Celles-ci ne sont pas générées par un simple processus stochastique mais sont le fruit de phénomènes physiques et d'interactions très complexes. Il faut donc se fixer un niveau d'exigence quant aux caractéristiques des séries de variables climatiques que nous souhaitons voir reproduites. Par exemple, si l'on souhaite modéliser les précipitations, les critères suivants sont couramment considérés : distribution des précipitations journalières, cumuls mensuels ou annuels, fréquence des précipitations, durée des séquences sèches ou pluvieuses. Les critères de validation que l'on considère dépendent avant tout des applications. Ainsi dans certains cas on s'intéressera seulement au comportement moyen des variables tandis que dans d'autres on souhaitera obtenir une bonne modélisation des extrêmes (vagues de froid ou canicules, vent fort, précipitations extrêmes). Si le modèle est multivarié, il conviendra également de vérifier que la dépendance entre les variables est bien reproduite. D'autre part, on attend d'un générateur de temps qu'il offre suffisamment de variabilité : si toutes les trajectoires qu'il simule sont presque identiques à la trajectoire observée, il perd son intérêt. En

2. Dans l'article, les données sont restreintes à l'hiver et donc une telle modélisation a davantage pour objectif d'effectuer une descente d'échelle que d'introduire la saisonnalité.

pratique, après avoir estimé les paramètres, on simule un certain nombre N de trajectoires de même longueur que la trajectoire observée en utilisant le générateur, et on les compare avec la trajectoire observée, selon les différents critères sélectionnés. C'est le principe du *bootstrap paramétrique*. Supposons par exemple que l'on s'intéresse aux extrêmes de température et que l'on souhaite obtenir une bonne modélisation du quantile d'ordre 0.95 de la distribution des températures. En utilisant la trajectoire observée, on calcule le quantile empirique q_{95}^{obs}. On fait de même avec chacune des trajectoires simulées et on obtient une collection de quantiles $q_{95}^{\text{sim},1}, \dots, q_{95}^{\text{sim},N}$. La distribution empirique des $(q_{95}^{\text{sim},i})_{1 \leq i \leq N}$ est alors une approximation de la distribution de q_{95} sous le modèle, et permet donc d'obtenir, par exemple, un intervalle de confiance. Il suffit alors de comparer q_{95}^{obs} à cette distribution. La même procédure est appliquée à différentes statistiques, et on identifie ainsi les points forts et les points faibles du modèle.

1.3 Contenu de la thèse

Objectif : concevoir un générateur de temps de résolution temporelle journalière, en un point de l'espace, capable de simuler de façon cohérente la température moyenne, les précipitations, la vitesse du vent et le rayonnement solaire.

Comme nous venons de le voir, il existe de nombreuses manières d'aborder ce problème. Dans cette thèse, nous avons choisi l'approche par modèles de Markov cachés, en raison de sa flexibilité. Dit simplement, un modèle de Markov caché est un processus $(X_t, Y_t)_{t \geq 1}$ tel que $(X_t)_{t \geq 1}$ est une chaîne de Markov d'ordre 1, et conditionnellement aux X_t, les Y_t sont indépendants et la loi de Y_t ne dépend que de X_t, et éventuellement de t. Les X_t sont appelés les *états cachés* car ils ne sont pas accessibles à l'observation : ce sont des variables latentes. Nous ne considérerons dans cette thèse que des HMM à espace d'état fini. La chaîne de Markov $(X_t)_{t \geq 1}$ est donc à valeurs dans un ensemble fini que nous identifierons à $\{1, \dots, K\}$ pour un certain entier $K \geq 1$. La loi du processus $(X_t)_{t \geq 1}$ est donc entièrement déterminée par la loi de X_1, appelée *loi initiale*, et par les probabilités de transitions $\mathbb{P}(X_{t+1} = j \mid X_t = i)$, pour $t \geq 1$ et $i, j \in \{1, \dots, K\}$. Notons que nous ne supposons pas la chaîne de Markov homogène, donc les probabilités de transition peuvent éventuellement varier au cours du temps. Les lois des observations Y_t conditionnellement aux états cachés sont appelées *lois d'émission*. Elles peuvent éventuellement varier au cours du temps. Ainsi on notera $\nu_k(t)$ la loi de $Y_t \mid \{X_t = k\}$, c'est-à-dire la k-ème loi d'émission à l'instant t. Lorsque les lois d'émission sont à densité par rapport à une même mesure, nous parlons de *densités d'émission*.

Cette thèse peut être divisée en deux parties :

- une partie appliquée dans laquelle nous introduisons différents générateurs de temps construit à partir du modèle de Markov caché, nous appliquons ces modèles à des données réelles et nous testons leur capacité à répondre à la problématique posée.
- une partie théorique dans laquelle nous nous attachons à formuler des garanties théoriques pour ces modèles : identifiabilité et convergence de l'estimateur du maximum de vraisemblance.

Le chapitre 2 comporte des généralités sur les modèles de Markov cachés. Ce chapitre ne contient pas de résultats nouveaux mais pose quelques bases utiles pour toute la suite. Après avoir défini les modèles de Markov cachés, nous nous concentrons sur quelques aspects pratiques concernant l'estimation de leurs paramètres. En particulier, nous détaillons l'algorithme EM et certaines de ses variantes, d'abord dans un cadre général, puis appliqué à celui des HMM. Cet algorithme classique nous a permis d'estimer par maximum de vraisemblance les paramètres des différents modèles étudiés dans cette thèse. Nous y détaillons également les algorithmes *forward-backward*

et de *Viterbi*, qui servent à "estimer" la séquence d'états, bien qu'elle ne soit pas observée. Ces algorithmes sont illustrés sur un exemple simple. La question importante, et qui ne manque pas de se poser en pratique, de la sélection du nombre d'états est également abordée. Enfin, ce chapitre contient quelques résultats théoriques concernant l'identifiabilité des modèles de Markov cachés d'une part, et la convergence presque sûre de l'estimateur du maximum de vraisemblance d'autre part. Dans la suite de la thèse, certains de ces résultats seront généralisés. Ainsi ce chapitre introductif est utile tant pour le lecteur intéressé par les aspects théoriques que par celui intéressé par les applications climatiques.

Dans le chapitre 3, nous nous intéressons à une classe particulière de HMM non homogènes pour lesquels les probabilités de transition et les lois d'émission sont des fonctions périodiques du temps. Autrement dit, il existe un entier T tel que pour tout $t \geq 1$ et pour tous $i, j \in \{1, \dots, K\}$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = \mathbb{P}(X_{t+1+T} = j \mid X_{t+T} = i)$ et $\nu_i(t) = \nu_i(t+T)$. De tels modèles sont bien entendu adaptés aux variables climatiques, lesquelles présentent une saisonnalité annuelle. Deux résultats nouveaux y sont prouvés. D'une part, nous donnons des conditions suffisantes, et faibles, pour garantir l'identifiabilité de ces modèles. Nous généralisons ainsi le résultat d'identifiabilité de Gassiat et al. (2016a) qui concernait des HMM stationnaires. La preuve repose sur une méthode spectrale. D'autre part, nous montrons que dans ce cadre, l'estimateur du maximum de vraisemblance converge presque sûrement vers le vrai paramètre : il est dit fortement consistant. Le théorème de consistance repose sur des hypothèses similaires à celles de Douc et al. (2004) sur les HMM autorégressifs homogènes. Nous montrons en effet que l'on peut se ramener au cas d'un HMM stationnaire en augmentant la dimension de l'espace d'état et de l'espace des observations. Les démonstrations sont suivies d'exemples concrets de modèles pour lesquels les hypothèses des théorèmes d'identifiabilité et de consistance sont vérifiées. Les lois d'émission considérées dans ces exemples sont des mélanges de gaussiennes et des mélanges de loi exponentielles et seront utilisées dans nos applications. La consistance de l'estimateur du maximum de vraisemblance est illustrée sur des données simulées. Enfin, nous proposons une application sur données réelles : nous modélisons les précipitations en utilisant un modèle de Markov caché non homogène.

Les saisonnalités ne sont pas la seule source de non-stationnarité des séries chronologiques de variables climatiques : on observe aussi des tendances liées au changement climatique, en particulier en ce qui concerne la température. Il n'est pas difficile d'introduire une tendance dans un HMM. Par exemple, considérons un HMM homogène $(X_t, Z_t)_{t\geq 1}$ à valeurs dans $\{1, \dots, K\} \times \mathbb{R}$, de matrice de transition Q et de densités d'émission $\gamma_1, \dots, \gamma_K$. Soit $T : \mathbb{N}^* \longrightarrow \mathbb{R}$ une fonction et $Y_t = T(t) + Z_t$. Alors $(X_t, Y_t)_{t\geq 1}$ est un HMM non-homogène, de matrice de transition Q, dont les densités d'émission à l'instant t sont les $\gamma_k(\cdot - T(t))$ et dont la tendance est T. Un cas un peu plus général est celui où l'on autorise la tendance à dépendre de l'état : si $T_1, \dots, T_K$ est une famille de fonctions $\mathbb{N}^* \longrightarrow \mathbb{R}$, on considère le HMM $(X_t, Y_t)_{t\geq 1}$ avec $Y_t = T_{X_t}(t) + Z_t$. Pour $k \in \{1, \dots, K\}$ et $t \geq 1$, la densité d'émission dans l'état k à l'instant t est alors $\gamma_k(\cdot - T_k(t))$. Si l'algorithme EM permet d'approcher l'estimateur du maximum de vraisemblance pour un tel modèle, jusqu'à présent aucun résultat théorique n'existait sur cet estimateur. Le chapitre 4, issu d'un travail en collaboration avec Luc Lehéricy, traite du cas où les tendances sont polynomiales. Nous montrons que dans ce cadre, et sous certaines hypothèses sur l'espace des paramètres, l'estimateur du maximum de vraisemblance est fortement consistant : presque sûrement, il converge quand le nombre d'observations tend vers l'infini vers le paramètre constitué de la vraie matrice de transition, des vraies densités d'émission du HMM homogène $(X_t, Z_t)_{t\geq 1}$, et des vraies tendances. L'idée principale de la démonstration repose sur le fait que les tendances étant polynomiales, soit elles diffèrent d'une constante, soit elles divergent les unes des autres. On a donc des *blocs* de tendances (deux tendances sont dans le même bloc si elles diffèrent d'une constante). Par conséquent, quand le nombre d'observations est grand, l'EMV va fortement pénaliser les paramètres

liés à des tendances éloignées des vraies tendances. On montre donc que pour un nombre d'observations n assez grand, l'EMV $\hat{\theta}_n$ se situe dans une région de l'espace des paramètres pour laquelle les vraies tendances sont bien approchées, au sens de la norme de la convergence uniforme. Puis on montre qu'asymptotiquement, dans cette région, la vraisemblance se comporte comme celle du HMM où les blocs $(B_t)_{t\geq 1}$ seraient observés, c'est-à-dire le processus $(X_t, (Y_t, B_t))$. On a alors $Z_t \simeq Y_t - T^{\theta}_{B_t}(t)$, où θ est un paramètre qui approche bien les vraies tendances. On montre enfin que le processus des observations détendancées $Y_t - T^{\theta}_{B_t}$ est "presque" homogène, ce qui permet de retrouver tous les paramètres. Après la démonstration, des simulations illustrent la convergence de l'EMV. Ces simulations montrent en particulier que l'EMV peut retrouver les vrais paramètres même si les tendances n'ont pas encore divergé. Ce point est important pour la pratique, puisqu'on n'a jamais accès à un nombre illimité d'observations.

Dans le chapitre 5, on introduit un générateur de temps destiné à la simulation conjointe de la température et des précipitations au pas journalier. Il est basé sur un modèle de Markov caché non-homogène incluant tendances et saisonnalités. Les probabilités de transition entre les états cachés sont périodiques de période 365 jours et les lois d'émission dépendent du temps. Plus précisément, ce sont des mélanges de la forme

$$\nu_k(t) = \sum_{m=1}^{M} p_{km} \nu_{km}^{\text{precip}}(t) \otimes \nu_{km}^{\text{temp}}(t),$$

où les $(p_{km})_{1\leq m\leq M}$ sont les poids du mélange, les $\nu_{km}^{\text{precip}}(t)$ sont des lois de probabilité sur $\mathbb{R}_+$ et les $\nu_{km}^{\text{temp}}(t)$ sont des lois de probabilité sur $\mathbb{R}$. Cette forme particulière a été choisie de façon à pouvoir approcher, pourvu que M soit assez grand, des distributions bivariées quelconques conditionnellement à chaque état. On peut ainsi modéliser finement la dépendance entre la température et les précipitations. Néanmoins, si on ne veut pas devoir choisir M exagérément grand, il convient de choisir soigneusement les lois des marginales, c'est-à-dire les $\nu_{km}^{\text{precip}}(t)$ et les $\nu_{km}^{\text{temp}}(t)$. Les précipitations doivent pouvoir être nulles avec probabilité strictement positive, puisqu'il ne pleut pas tous les jours. Ainsi les premières composantes du mélange correspondent à des précipitations nulles, c'est-à-dire $\nu_{km}^{\text{precip}}(t) = \delta_0$ pour $1 \leq m \leq M_1$, où $M_1 \in \{1, \ldots, M-1\}$ et δ_0 désigne la mesure de Dirac en 0. Les autres composantes du mélange sont des lois exponentielles dont le paramètre est modulé par une fonction périodique : pour $M_1 < m \leq M$,

$$\nu_{km}^{\text{precip}}(t) = \mathcal{E}\left(\frac{\lambda_{km}}{1+\sigma_k(t)}\right).$$

Notons que la fonction périodique $\sigma_k(\cdot)$ dépend de l'état mais pas de la composante du mélange. On obtient ainsi la modélisation des variations saisonnières de l'intensité des précipitations. La saisonnalité dans la fréquence des précipitations est obtenue via celle de la matrice de transition, le poids total de la masse de Dirac étant différent selon les états. Du côté de la température, les composantes sont des gaussiennes de la forme

$$\nu_{km}^{\text{temp}}(t) = \mathcal{N}\left(T_k(t) + S_k(t) + \mu_{km}, \sigma_{km}^2\right).$$

Les S_k sont des fonctions périodiques du temps spécifiques à l'état et les T_k sont des tendances qui dépendent aussi de l'état et qui servent à modéliser le réchauffement climatique. L'impact de ce dernier n'étant pas partout identique, la forme paramétrique des tendances doit être choisie en fonction du lieu dont on souhaite modéliser la température, et en fonction de la période sur laquelle on souhaite réaliser des simulations. Des tendances linéaires peuvent convenir dans certains cas. Des tendances linéaires par morceaux sont une alternative possible. Dans ce cas, il

faut choisir la date de rupture de pente. Le modèle étant défini, nous avons réalisé l'estimation des paramètres à partir des données de précipitations et de température issues de plusieurs stations situées dans différentes régions d'Europe (Finlande, Norvège, France, Allemagne, Espagne, Italie), de façon à vérifier sa robustesse sous différents climats. Enfin, des tests de validation ont été menés pour évaluer la performance du modèle.

Le générateur bivarié décrit et testé dans le chapitre 5 peut en particulier être utilisé dans le cadre de la production hydroélectrique. Les précipitations et la température sont en effet les deux variables météorologiques qui déterminent le débit des cours d'eau, et donc l'énergie produite par les installations hydroélectriques. Dans le chapitre 6, nous proposons une application de notre générateur bivarié température/précipitations à ce contexte. Ce chapitre est issu d'une collaboration avec la Direction Technique Générale (DTG) d'EDF. La DTG utilise un modèle hydrologique qui permet, à partir de chroniques de précipitations et de températures pour un bassin versant donné, d'obtenir une chronique de débit pour le cours d'eau correspondant. Il est ainsi possible, à partir de séries de précipitations et de températures simulées par notre générateur HMM, de simuler des chroniques de débit. Si le générateur de temps et le modèle hydrologique sont tous deux performants, les séries de débit ainsi obtenues sont réalistes, ce qui permet de générer des scénarios de production hydroélectrique. Dans le chapitre 6, nous comparons, pour deux bassins versants dans les Alpes, les performances de notre générateur bivarié avec celles du générateur utilisé par la DTG, du point de vue des précipitations, des températures, de l'enneigement, et des débits produits en sortie du modèle hydrologique. Nous montrons en particulier que notre générateur constitue une alternative crédible. Dans un second temps, nous utilisons notre générateur pour simuler des séries de températures et débits hivernaux au pas de temps annuel, sur un bassin versant, et sur la période 1965-2010. Un hiver froid et avec de faibles débits est problématique dans le cadre de la production hydroélectrique "au fil de l'eau" à cause de la conjonction d'une consommation élevée et d'une moindre production. A partir des simulations, nous obtenons une estimation de la distribution bivariée des températures et débits hivernaux, et donc une estimation de la probabilité d'observer un tel hiver. On s'intéresse également à l'évolution de cette probabilité liée au changement climatique, ce qu'il est difficile de faire sans simulateur à cause du manque d'observations. Dans la dernière partie du chapitre, nous utilisons l'algorithme de Viterbi pour estimer la séquence d'états, interprétés comme types de temps, et nous donnons une interprétation à cette classification en considérant des champs de variables météorologiques à l'échelle continentale : géopotentiel, précipitations, température. Cette interprétation vient compléter celle obtenue à l'échelle locale.

Le chapitre 7 introduit la modélisation de la vitesse du vent, variable déterminante pour la production d'énergie éolienne. La première partie du chapitre est consacrée à la modélisation univariée de la vitesse du vent à l'aide d'un modèle de Markov caché. Comme la vitesse du vent présente une saisonnalité, ce HMM est nécessairement inhomogène. Classiquement, nous nous sommes tournés vers la loi de Weibull pour les lois d'émission. La non-homogénéité peut être introduite dans les probabilités de transition ou dans les paramètres des lois d'émission. Ce choix n'étant pas exclusif, 7 modèles différents sont possibles. Nous avons testé deux d'entre eux : celui ou seules les probabilités de transition sont périodiques, et celui ou seul le paramètre d'échelle de la loi de Weibull est périodique. Ces deux modèles ont été appliquées aux observations de 6 stations européennes. Nous avons montré qu'ils étaient tous deux performants pour produire des chroniques réalistes de vent. Dans la seconde partie du chapitre, nous nous sommes basés sur le modèle bivarié température/précipitations du chapitre 5 et sur les résultats de la première partie pour introduire un modèle trivarié température/précipitations/vitesse du vent, que nous avons testé sur les six mêmes stations.

Première partie

Modèles de Markov caché non-homogènes

Chapitre 2

Généralités sur les modèles de Markov cachés

2.1 Introduction

Les modèles de Markov cachés (HMM pour Hidden Markov Model) ont été introduits par Baum and Petrie (1966) sous le nom de *probabilistic functions of Markov chains*. Ils ont par la suite été abondamment étudiés d'un point de vue théorique et ont été utilisés dans de nombreuses applications, de la reconnaissance vocale à la modélisation climatique, en passant par la génomique et la finance. Ils doivent leur succès à leur simplicité, leur grande flexibilité, leur interprétabilité et à leur facilité d'implémentation. Rabiner and Juang (1986) constitue une introduction simple aux modèles de Markov cachés. Pour un exposé plus complet et plus récent, on pourra consulter l'ouvrage Cappé et al. (2009). Ce chapitre introductif n'est pas une revue exhaustive des connaissances actuelles sur les modèles de Markov cachés, mais contient les bases nécessaires à la lecture des chapitre suivants. Après avoir défini ces modèles, nous nous intéresserons aux questions soulevées par leur mise en œuvre pratique. En particulier, nous décrirons les principales méthodes d'inférence des HMM, dont la plus populaire, celle du maximum de vraisemblance. Celle-ci nécessite l'utilisation de l'algorithme EM. Cet algorithme, souvent utilisé pour estimer des modèles à variables latentes, prend une forme particulière dans le cadre des HMM, que nous décrirons de façon détaillée et que nous utiliserons tout au long de cette thèse. Dans un second temps, nous rappellerons quelques résultats théoriques concernant l'identifiabilité des HMM, et la convergence de l'estimateur du maximum de vraisemblance (EMV). Enfin, nous aborderons la question de la sélection du nombre d'états cachés, qui ne manque pas de se poser en pratique.

2.2 Définition d'un modèle de Markov caché

Commençons par rappeler quelques définitions élémentaires.

Définition 2.1 (Modèle statistique)**.** *Un* modèle statistique *sur un espace mesurable* $(\mathsf{Y}, \mathcal{Y})$ *est une famille de lois de probabilité* $(\mathbb{P}_\theta)_{\theta\in\Theta}$ *sur* $(\mathsf{Y}, \mathcal{Y})$.

L'espace Θ qui indexe les lois de probabilité du modèle dans la définition 2.1 est l'espace des *paramètres*. Lorsque cet espace est un sous-ensemble d'un espace de dimension finie, on dit que le modèle est *paramétrique*. Dans le cas contraire, il est dit *non paramétrique*.

En pratique, on dispose d'observations $Y_1, \dots, Y_n$ qui sont des réalisations d'un processus qui suit une certaine loi inconnue $\mathbb{P}^*$. S'il existe un paramètre $\theta^* \in \Theta$ tel que $\mathbb{P}_{\theta^*} = \mathbb{P}^*$, le modèle $(\mathbb{P}_\theta)_{\theta\in\Theta}$ est dit *bien spécifié*.

Définition 2.2 (Modèle de Markov caché)**.** *Pour un entier $K \geq 1$, on définit l'*espace d'états $\mathsf{X} = \{1, \dots, K\}$*, qu'on munit de la tribu $\mathcal{X}$ de l'ensemble de ses parties. Soit $(\mathsf{Y}, \mathcal{Y})$ un espace mesurable quelconque, et $(X_t, Y_t)_{t\geq 1}$ un processus stochastique à temps discret défini sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans l'espace $(\mathsf{X}\times\mathsf{Y}, \mathcal{X}\otimes\mathcal{Y})$. On suppose que le processus $(X_t)_{t\geq 1}$ est une chaîne de Markov d'ordre* 1*. La loi de ce processus est entièrement déterminée par sa loi initiale ξ (i.e. la loi de X_1), et par ses* probabilités de transition *définies, pour $i, j \in \mathsf{X}$ et $t \geq 1$, par $Q_{ij}(t) = \mathbb{P}(X_{t+1} = j \mid X_t = i)$. Les X_t sont appelés les* états cachés *car ils ne sont pas observés. On suppose de plus que pour tout $t \geq 1$, la loi de Y_t conditionnellement à $(X_s)_{s\geq 1}$ ne dépend que de X_t et de t et que les Y_t sont indépendants conditionnellement à $(X_s)_{s\geq 1}$. Pour $k \in \mathsf{X}$ et $t \geq 1$, on note $\nu_k(t)$ la loi de Y_t conditionnellement à $\{X_t = k\}$. Ces lois conditionnelles sont appelées* lois d'émission*. Sous ces conditions, le processus $(X_t, Y_t)_{t\geq 1}$ est appelé* modèle de Markov caché*, ce que nous abrègerons en HMM (pour* Hidden Markov Model*) dans la suite de cette thèse.*

Quelques remarques importantes découlent directement de cette définition :

- Un HMM est un processus de Markov : conditionnellement à $(X_s, Y_s)_{1\leq s\leq t-1}$, la loi de (X_t, Y_t) ne dépend que de X_{t-1}.
- Si $(X_t, Y_t)_{t\geq 1}$ est un HMM, la loi marginale de Y_t s'écrit sous la forme d'un mélange des lois d'émission :
$$Y_t \sim \sum_{k\in\mathsf{X}} \mathbb{P}(X_t = k)\nu_k(t).$$
Ainsi les HMM sont une généralisation des modèles de mélange, pour lesquels le processus des états cachés (aussi appelé variable latente) est une suite indépendante et identiquement distribuée (i.i.d.).
- Les Y_t sont indépendants conditionnellement aux états cachés X_t mais le processus $(Y_t)_{t\geq 1}$ est bien dépendant, par l'intermédiaire des états cachés.

Signalons aussi que cette définition des HMM, ainsi que de nombreux résultats les concernant, s'étendent à des espaces d'état infinis, dénombrables ou non (par exemple, $\mathsf{X} = \mathbb{R}$). Dans cette thèse, nous ne considérerons que des HMM à espace d'état fini.

Le cas usuel en pratique est celui où il existe une mesure dominante μ définie sur $(\mathsf{Y}, \mathcal{Y})$ telle que les lois d'émission soient absolument continues par rapport à μ. Il existe alors des *densités d'émission*, notées $f_{k,t}$, telles que $\frac{d\nu_k(t)}{d\mu} = f_{k,t}$. La Figure 2.1 résume schématiquement les relations de dépendance entre les variables dans un HMM.

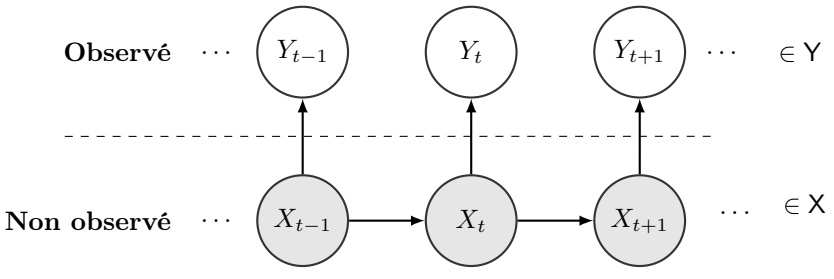

FIGURE 2.1 – Dynamique d'un modèle de Markov caché.

HMM homogène Le cas le plus usuel, et le plus étudié, est celui où la chaîne de Markov $(X_t)_{t\geq 1}$ est homogène et où les lois d'émission ne dépendent pas du temps : pour tous $s, t \geq 1$ et $i, j \in \mathsf{X}$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = \mathbb{P}(X_{s+1} = j \mid X_s = i)$ et $\nu_k(t) = \nu_k(s)$. Dans ce cas nous parlerons de *HMM homogène*. La loi d'un tel HMM est donc déterminée par sa loi initiale $\xi = \mathbb{P}(X_1 = \cdot)$, sa matrice de transition $Q \in \mathbb{R}^{K\times K}$ telle que $Q_{ij} = \mathbb{P}(X_2 = j \mid X_1 = i)$ et ses K lois d'émissions $(\nu_k = \mathbb{P}(Y_1 \in \cdot \mid X_1 = k))_{k\in\mathsf{X}}$.

HMM stationnaire

Définition 2.3. *Un processus stochastique $(Z_t)_{t\geq 1}$ est dit* stationnaire *lorsque pour tout entier $\tau \geq 0$, pour tout $k \geq 1$ et pour tous $t_1, \ldots, t_k \geq 1$, $(Z_{t_1}, \ldots, Z_{t_k})$ et $(Z_{t_1+\tau}, \ldots, Z_{t_k+\tau})$ ont même loi.*

Remarquons qu'un HMM homogène $(X_t, Y_t)_{t\geq 1}$ est stationnaire si et seulement si la chaîne de Markov sous-jacente $(X_t)_{t\geq 1}$ l'est. Dans ce cas, sa loi initiale ξ est une loi invariante pour sa matrice de transition Q. Pour un tel HMM, la loi marginale des observations est un mélange qui ne dépend pas du temps : pour tout $t \geq 1$,

$$Y_t \sim \sum_{k\in\mathsf{X}} \xi_k \nu_k.$$

HMM paramétrique ou non paramétrique Nous avons défini au début de ce chapitre la notion de modèle paramétrique ou non paramétrique. Nous dirons qu'un modèle de Markov caché est paramétrique lorsque l'on suppose que ses lois d'émission appartiennent à un modèle paramétrique sur $(\mathsf{Y}, \mathcal{Y})$ (par exemple : lois d'émission gaussiennes). Dans le cas contraire nous parlerons de HMM non paramétrique (exemple : densités d'émission de carré intégrable sur $[0, 1]$). Dans cette thèse, la plupart des HMM que nous étudions sont paramétriques, à l'exception du Chapitre 4 dans lequel nous considérons un ensemble de lois d'émission plus général. En revanche nous considérerons souvent des HMM inhomogènes (et donc non stationnaires).

2.3 Algorithmes liés aux HMM

Dans cette partie, nous nous attachons aux aspects pratiques de l'estimation des HMM. Plus précisément, nous allons présenter successivement les trois algorithmes suivants :

- L'algorithme forward-backward permet un calcul récursif des lois conditionnelles des états cachés, sachant une séquence d'observations.
- L'algorithme de Viterbi permet de calculer la séquence d'états cachés la plus probable conditionnellement à une séquence d'observations. Il est important dans les applications où le but est de retrouver la séquence d'états (par exemple, un signal Markovien est bruité et l'on cherche à retrouver le signal à partir de la seule observation du signal bruité). Nous l'utiliserons dans le Chapitre 6 pour donner une interprétation physique aux états cachés.
- L'algorithme EM permet le calcul pratique de l'estimateur du maximum de vraisemblance pour un HMM.

Notation Dans toute la suite, pour $1 \leq r \leq s$, $Y_{r:s}$ désigne le vecteur $(Y_r, Y_{r+1}, \ldots, Y_s)$.

2.3.1 Filtrage et lissage, algorithme forward-backward

Soit $(X_t, Y_t)_{t\geq 1}$ un modèle de Markov caché et supposons connus ses paramètres : on connaît sa loi initiale ξ, ses probabilités de transition $Q_{ij}(t)$ et ses lois d'émission $\nu_{k,t}$. Supposons en

avoir observé une trajectoire de longueur n : on dispose des observations $Y_1, \dots, Y_n$. Les états cachés correspondant $X_1, \dots, X_n$ ne sont pas observés, il est donc impossible de les connaître avec certitude. Cependant les observations $Y_1, \dots Y_n$ fournissent de l'information sur les X_t puisque chaque Y_t a été généré selon une loi ne dépendant que de X_t. Nous allons voir qu'il est possible de calculer efficacement, pour tout $t \in \{1, \dots n\}$ la loi de X_t conditionnellement au vecteur d'observations $Y_{1:n}$. Ces probabilités sont appelées probabilités de *lissage*. Par définition, pour tout $1 \leq t \leq n$ et $k \in \mathsf{X}$,

$$\pi_{t|n}(k) = \mathbb{P}(X_t = k \mid Y_{1:n}).$$

La distribution de lissage $\mathbb{P}(X_t \in \cdot \mid Y_{1:n})$ est une loi de probabilité sur X, identifiée à un vecteur du simplexe $\Delta_K := \{(p_1, \dots, p_K) \in [0,1]^K, \sum_{k=1}^K p_k = 1\}$. Nous noterons $\pi_{t|n}$ ce vecteur. Les probabilités de *filtrage* sont quant à elles définies par

$$\pi_t(k) = \mathbb{P}(X_t = k \mid Y_{1:t}).$$

De même, la loi de probabilité correspondante est identifiée à un vecteur noté π_t. Le calcul des probabilités de lissage s'effectue grâce à l'algorithme *forward-backward*. Nous allons présenter cet algorithme dans le cas particulier où l'espace Y est fini et où le HMM est homogène, comme dans Rabiner and Juang (1986), puis nous en donnerons la version plus générale que nous allons utiliser. Ce cadre simplifié permet de mettre en évidence les relations de récurrence sur lesquelles reposent l'algorithme tout en évitant un excès de formalisme. L'algorithme forward-backward s'effectue en deux phases. Dans la phase *forward*, on calcule les probabilités de filtrage. Celles-ci sont ensuite utilisées dans la phase *backward* pour obtenir les probabilités de lissage. Pour $1 \leq t \leq n$, $k \in \mathsf{X}$ et $y_{1:t} \in \mathsf{Y}^t$, on définit

$$\alpha_t(k) = \mathbb{P}(Y_1 = y_1, \dots, Y_t = y_t, X_t = k)$$

La formule de Bayes implique alors que

$$\pi_t(k) = \frac{\alpha_t(k)}{\sum_{j=1}^K \alpha_t(j)}.$$

On a donc l'égalité vectorielle

$$\pi_t = \frac{\alpha_t}{\mathbf{1}^\top \alpha_t},$$

où $\pi_t = (\pi_t(1), \dots, \pi_t(K))^\top$, $\alpha_t = (\alpha_t(1), \dots, \alpha_t(K))^\top$, et $\mathbf{1} = (1, \dots, 1)^\top$. Les α_t, appelées *variables forward*, peuvent être facilement calculées récursivement. En effet, pour $t \geq 2$ et $i \in \mathsf{X}$, en utilisant la formule de Bayes et la propriété d'indépendance conditionnelle des HMM,

$$\begin{aligned}
\alpha_t(i) &= \mathbb{P}(Y_1 = y_1, \dots Y_t = y_t, X_t = i) \\
&= \sum_{j=1}^{K} \mathbb{P}(Y_1 = y_1, \dots Y_t = y_t, X_t = i, X_{t-1} = j) \\
&= \sum_{j=1}^{K} \mathbb{P}(Y_t = y_t \mid Y_1 = y_1, \dots Y_{t-1} = y_{t-1}, X_t = i, X_{t-1} = j) \\
&\quad \times \mathbb{P}(Y_1 = y_1, \dots Y_{t-1} = y_{t-1}, X_t = i, X_{t-1} = j) \\
&= \sum_{j=1}^{K} \mathbb{P}(Y_t = y_t \mid X_t = i) \mathbb{P}(X_t = i \mid Y_1 = y_1, \dots Y_{t-1} = y_{t-1}, X_{t-1} = j) \\
&\quad \times \mathbb{P}(Y_1 = y_1, \dots Y_{t-1} = y_{t-1}, X_{t-1} = j) \\
&= \sum_{j=1}^{K} \mathbb{P}(Y_t = y_t \mid X_t = i) Q_{ji} \alpha_{t-1}(j)
\end{aligned}$$

En notant $\Upsilon(y) \in \mathbb{R}^{K \times K}$ la matrice diagonale définie par $\Upsilon(y)_{ii} = \mathbb{P}(Y_t = y \mid X_t = i)$, on obtient donc que

$$\begin{aligned}
\alpha_1 &= \Upsilon(y_1)\xi \\
\alpha_t &= \Upsilon(y_t) Q^\top \alpha_{t-1} = F_t \alpha_{t-1},\ t \geq 2,
\end{aligned}$$

où $F_t = \Upsilon(y_t)Q^\top$ est le *noyau forward*. On en déduit alors les probabilités de filtrage $\pi_t = \frac{\alpha_t}{\mathbf{1}^\top \alpha_t}$. Les variables forward α_t tendent vers 0 à vitesse exponentielle, de sorte qu'en pratique, on atteint rapidement la limite de la précision des calculs d'un ordinateur (*underflow*). Il convient donc, pour éviter ce problème, de renormaliser au fur et à mesure. On définit donc, pour $2 \leq t \leq n$,

$$\begin{aligned}
\tilde{\pi}_t &= F_t \pi_{t-1} \\
c_t &= \mathbf{1}^\top \tilde{\pi}_t = \frac{\mathbf{1}^\top \alpha_t}{\mathbf{1}^\top \alpha_{t-1}}
\end{aligned}$$

Alors,

$$c_t^{-1} \tilde{\pi}_t = F_t \frac{\alpha_{t-1}}{\mathbf{1}^\top \alpha_{t-1}} \frac{\mathbf{1}^\top \alpha_{t-1}}{\mathbf{1}^\top \alpha_t} = F_t \frac{\alpha_{t-1}}{\mathbf{1}^\top \alpha_t} = \frac{\alpha_t}{\mathbf{1}^\top \alpha_t} = \pi_t.$$

On en déduit l'algorithme *forward*, qui est la première partie de l'algorithme *forward-backward*. Il prend en entrée les paramètres du HMM et un vecteur d'observation $y_{1:n}$ et renvoie les distributions de filtrage $(\pi_t)_{1 \leq t \leq n}$.

Algorithm 1: Algorithme forward

$c_1 \leftarrow \mathbf{1}^\top \Upsilon(y_1)\xi$;
$\pi_1 \leftarrow \Upsilon(y_1)\xi / c_1$;
for $t = 2, \dots, n$ **do**
  $\tilde{\pi}_t \leftarrow \Upsilon(y_t) Q^\top \pi_{t-1}$;
  $c_t \leftarrow \mathbf{1}^\top \tilde{\pi}_t$;
  $\pi_t \leftarrow \tilde{\pi}_t / c_t$;
end

De plus, on a $c_1 = \mathbb{P}(Y_1 = y_1)$ et pour $t \geq 2$,

$$c_t = \frac{\mathbf{1}^\top \alpha_t}{\mathbf{1}^\top \alpha_{t-1}} = \frac{\mathbb{P}(Y_1 = y_1, \ldots, Y_t = y_t)}{\mathbb{P}(Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1})} = P(Y_t = y_t \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1})$$

On en déduit que

$$\prod_{t=1}^{n} c_t = \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n),$$

qui est la vraisemblance de la séquence d'observations $y_{1:n}$. Donc un sous-produit de l'algorithme *forward* est la log-vraisemblance

$$\log \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n) = \sum_{t=1}^{n} \log c_t.$$

Pour $k, l \in \mathsf{X}$ et $t \geq 1$, on définit les probabilités de lissage bivariées

$$\pi_{t,t+1|n}(k, l) = \mathbb{P}(X_t = k, X_{t+1} = l \mid Y_1 = y_1, \ldots, Y_n = y_n)$$

et on note $\pi_{t,t+1|n} \in \mathbb{R}^{K \times K}$ la matrice ainsi définie. Ces quantité seront utilisées dans l'étape E de l'algorithme EM. Remarquons que les distributions de lissage peuvent être obtenues par $\pi_{t|n} = \pi_{t,t+1|n}\mathbf{1}$.

Soit $1 \leq t < n$. On a :

$$\begin{aligned}
\pi_{t|n}(i) &= \mathbb{P}(X_t = i \mid Y_1 = y_1, \ldots, Y_n = y_n) \\
&= \frac{\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n, X_t = i)}{\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n)} \\
&= \frac{\alpha_t(i)\mathbb{P}(Y_{t+1} = y_{t+1}, \ldots Y_n = y_n \mid X_t = i)}{\mathbb{P}(Y_1 = y_1, \ldots Y_n = y_n)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{P(Y_1 = y_1, \ldots, Y_n = y_n)}
\end{aligned}$$

où l'on a posé :

$$\begin{aligned}
&\beta_n(i) = 1 \\
&\beta_t(i) = \mathbb{P}(Y_{t+1} = y_{t+1}, \ldots Y_n = y_n \mid X_t = i), \quad 1 \leq t < n
\end{aligned}$$

Les $(\beta_t)_{1 \leq t \leq n}$ sont les variables *backward*. Par un calcul similaire à celui fait sur les variables *forward*, on obtient facilement la relation de récurrence suivante :

$$\begin{aligned}
&\beta_n = \mathbf{1} \\
&\beta_t = Q\Upsilon(y_{t+1})\beta_{t+1},\ 1 \leq t < n
\end{aligned}$$

D'autre part, en utilisant la formule de Bayes et la définition d'un HMM, on obtient la relation

$$\pi_{t,t+1|n} = \frac{\mathrm{diag}(\alpha_t) Q \Upsilon(y_{t+1}) \mathrm{diag}(\beta_{t+1})}{\mathbf{1}^\top \alpha_n}$$

Théoriquement c'est suffisant, mais pour les mêmes raisons de précision numérique, on préfère renormaliser au fur et à mesure. On définit pour cela la suite $\tilde{\beta}$ par

$$\begin{aligned}\tilde{\beta}_n &= \frac{\mathbf{1}}{c_n} \\ \tilde{\beta}_{n-t} &= \frac{1}{c_{n-t}} Q\Upsilon(y_{n-t+1})\tilde{\beta}_{n-t+1}, \quad 1 \leq t < n.\end{aligned}$$

Alors on montre par récurrence que :

$$\tilde{\beta}_{n-t} = \frac{\beta_{n-t}}{c_n \cdots c_{n-t}}, \quad 1 \leq t < n,$$

avec :

$$c_n \cdots c_{n-t} = \frac{\mathbf{1}^\top \alpha_n}{\mathbf{1}^\top \alpha_{n-t-1}}$$

On en déduit :

$$\begin{aligned}\mathrm{diag}(\pi_{n-t})Q\Upsilon(y_{n-t+1})\mathrm{diag}(\tilde{\beta}_{n-t+1}) &= \frac{\mathrm{diag}(\alpha_{n-t})Q\Upsilon(y_{n-t+1})\mathrm{diag}(\beta_{n-t+1})}{\mathbf{1}^\top \alpha_{n-t}} \frac{1}{c_n \cdots c_{n-t+1}} \\ &= \frac{1}{\mathbf{1}^\top \alpha_{n-t}} \mathrm{diag}(\alpha_{n-t})Q\Upsilon(y_{n-t+1})\mathrm{diag}(\beta_{n-t+1}) \frac{\mathbf{1}^\top \alpha_{n-t}}{\mathbf{1}^\top \alpha_n} \\ &= \pi_{n-t,n-t+1|n}.\end{aligned}$$

D'où finalement l'algorithme forward-backward (avec renormalisation). Il prend en entrée les paramètres du HMM et une séquence d'observations $y_{1:n}$ et fournit :

- les distributions de filtrage $(\pi_t)_{1 \leq t \leq n}$,
- les distributions de lissage bivariées $(\pi_{t,t+1|n})_{1 \leq t \leq n-1}$,
- les distributions de lissage $(\pi_{t|n})_{1 \leq t \leq n}$,
- la log-vraisemblance $\log \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n)$.

Algorithm 2: Algorithme forward-backward

$c_1 \leftarrow \mathbf{1}^\top \Upsilon(y_1)\xi$;
$\pi_1 \leftarrow \Upsilon(y_1)\xi / c_1$;
for $t = 2, \ldots, n$ **do**
  $\tilde{\pi}_t \leftarrow \Upsilon(y_t)Q^\top \pi_{t-1}$;
  $c_t \leftarrow \mathbf{1}^\top \tilde{\pi}_t$;
  $\pi_t \leftarrow \tilde{\pi}_t / c_t$;
end
$\tilde{\beta}_n \leftarrow \mathbf{1}/c_n$;
for $t = 1, \ldots n-1$ **do**
  $\tilde{\beta}_{n-t} \leftarrow Q\Upsilon(y_{n-t+1})\tilde{\beta}_{n-t+1}$;
  $\pi_{n-t,n-t+1|n} \leftarrow \mathrm{diag}(\pi_{n-t})Q\Upsilon(y_{n-t+1})\mathrm{diag}(\tilde{\beta}_{n-t+1})$;
  $\pi_{n-t|n} \leftarrow \pi_{n-t,n-t+1|n}\mathbf{1}$;
end

Nous avons introduit l'algorithme forward-backward dans le cas particulier d'un HMM homogène dont l'espace des observations Y est fini. On le généralise facilement à un HMM non-homogène

avec un espace d'observations quelconque. Il suffit de remplacer, dans l'algorithme, Q par $Q(t)$, et la matrice $\Upsilon(y)$ par $\Upsilon_t(y)$, la matrice diagonale dont le k-ième élément diagonal est $f_{k,t}(y)$, la densité d'émission dans l'état k, à l'instant t, évaluée en y. C'est cette version plus générale que nous utiliserons dans toutes nos applications.

2.3.2 Algorithme de Viterbi

Un problème classique qui se pose dans les applications des HMM est celui de retrouver la séquence d'états, non observée, à partir des observations. Nous avons vu que l'algorithme *forward-backward* permettait de calculer, à partir des observations $Y_1, \dots, Y_n$ et des paramètres du HMM, les probabilités de lissage $\pi_{t|n}(k) = \mathbb{P}(X_t = k \mid Y_{1:n})$. Ainsi, à chaque instant t, il est possible d'obtenir l'état le plus probable en calculant le *maximum a posteriori* :

$$\hat{X}_t^{\text{MAP}} = \underset{1 \leq k \leq K}{\arg\max}\, \pi_{t|n}(k) \tag{2.1}$$

Une première idée pour retrouver la séquence d'états serait donc de considérer $(\hat{X}_t^{\text{MAP}})_{1 \leq t \leq n}$. De cette façon on maximise le nombre d'états correctement "décodés". Cependant il se peut que la suite d'états $(\hat{X}_t^{\text{MAP}})_{1 \leq t \leq n}$ ne soit pas une suite d'états possible. Par exemple, on pourrait avoir $Q_{ij} = 0$ avec $\hat{X}_1^{\text{MAP}} = i$ et $\hat{X}_2^{\text{MAP}} = j$. Il est donc nécessaire d'adopter une autre approche si l'on veut obtenir la séquence d'états la plus probable compte tenu des observations. Ainsi on définit la suite $(\hat{X}_t)_{1 \leq t \leq n}$ vérifiant

$$(\hat{X}_1, \dots, \hat{X}_n) \in \underset{(x_1,\dots,x_n) \in \mathsf{X}^n}{\arg\max}\, \mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n \mid Y_{1:n}\right), \tag{2.2}$$

c'est-à-dire la suite d'états la plus probable conditionnellement aux observations, connaissant les paramètres du HMM. Une recherche exhaustive est inenvisageable dès que n est raisonnablement grand, mais l'algorithme de Viterbi (Viterbi, 1967; Forney, 1973) permet de résoudre ce problème récursivement. Nous allons présenter cet algorithme dans le cas d'un HMM homogène dont l'espace des observations Y est fini. Comme pour l'algorithme *forward-backward*, la généralisation est immédiate. Supposons les paramètres du HMM connus et soit $y_{1:n}$ la séquence d'observations. Remarquons que maximiser $\mathbb{P}\left(X_{1:n} = x_{1:n} \mid Y_{1:n} = y_{1:n}\right)$ équivaut à maximiser $\mathbb{P}\left(X_{1:n} = x_{1,n}, Y_{1:n} = y_{1:n}\right)$. Pour $1 \leq t \leq n$ et $i \in \mathsf{X}$, on définit

$$\delta_t(i) = \max_{x_{1:t-1}} \mathbb{P}\left(X_1 = x_1, \dots X_{t-1} = x_{t-1}, X_t = i, Y_{1:t} = y_{1:t}\right).$$

Il vient alors, pour $j \in \mathsf{X}$,

$$\delta_{t+1}(j) = \left(\max_{i \in \mathsf{X}} \delta_t(i) Q_{ij}\right) \mathbb{P}(Y_{t+1} = y_{t+1} \mid X_{t+1} = j)$$

et on note $\psi_t(j)$ l'état qui réalise ce maximum. L'algorithme de Viterbi est le suivant :

— Initialisation : pour $i \in \mathsf{X}$,

$$\begin{aligned} \delta_1(i) &= \xi_i \mathbb{P}(Y_1 = y_1 \mid X_1 = i) \\ \psi_1(i) &= 0 \end{aligned}$$

— Récursion : pour $2 \leq t \leq n$ et $j \in \mathsf{X}$,

$$\delta_t(j) = \left(\max_{i \in \mathsf{X}} \delta_{t-1}(i) Q_{ij}\right) \mathbb{P}(Y_t = y_t \mid X_t = j)$$
$$\psi_t(j) = \arg\max_{i \in \mathsf{X}} \delta_{t-1}(i) Q_{ij}$$

— La probabilité maximale est alors donnée par :

$$P^* = \max_{i \in \mathsf{X}} \delta_n(i)$$

et le chemin le plus probable se termine par l'état :

$$\hat{X}_n = \arg\max_{i \in \mathsf{X}} \delta_n(i)$$

— Les états précédents sont donnés, pour $t = n-1, \ldots, 1$, par :

$$\hat{X}_t = \psi_{t+1}(\hat{X}_{t+1})$$

En pratique, on passe au logarithme des probabilités pour éviter les problèmes d'*underflow*. On a donc l'algorithme suivant :

Algorithm 3: Algorithme de Viterbi

for $i = 1, \ldots K$ **do**
  $\delta_1(i) \leftarrow \log \xi_i + \log \mathbb{P}(Y_1 = y_1 \mid X_1 = i)$;
end
for $t = 2, \ldots, n$ **do**
  for $j = 1, \ldots, K$ **do**
    $\psi_t(j) \leftarrow \arg\max_{1 \le i \le K} (\delta_{t-1}(i) + \log Q_{ij})$;
    $\delta_t(j) \leftarrow \delta_{t-1}(\psi_t(j)) + \log Q_{\psi_t(j)j} + \log \mathbb{P}(Y_t = y_t \mid X_t = j)$;
  end
end
$\hat{X}_n \leftarrow \arg\max_{1 \le i \le K} \delta_n(i)$;
for $t = n-1, \ldots, 1$ **do**
  $\hat{X}_t \leftarrow \psi_{t+1}(\hat{X}_{t+1})$;
end

Remarquons qu'un modèle de Markov caché peut être utilisé, via l'algorithme de Viterbi, pour traiter un problème de classification non supervisée (clustering) d'une série temporelle puisque l'on obtient une suite de segments d'observations homogènes dans le temps. Un exemple d'une telle application sera donné dans le Chapitre 6.

2.3.3 Algorithme EM

Principe général Sauf cas particulier, en pratique les paramètres du HMM sont inconnus, il faut donc les estimer. Considérons un HMM à espace d'état fini $(X_t, Y_t)_{t \ge 1}$ dans un cadre paramétrique : les matrices de transition et les lois d'émission sont fonctions d'un paramètre $\theta \in \Theta$, où Θ est un sous-ensemble d'un espace vectoriel de dimension finie. Un estimateur classique de θ est l'estimateur du maximum de vraisemblance (EMV). L'algorithme Espérance-Maximisation (abrégé en *EM*), introduit par Dempster et al. (1977), permet de calculer récursivement une approximation de l'estimateur du maximum de vraisemblance dans des modèles à variables latentes

pour lesquels une maximisation directe de la vraisemblance serait difficile. Son cadre d'application dépasse donc largement celui des HMM. Nous allons d'abord présenter son principe général, puis nous traiterons le cas particulier des HMM, avant de discuter de ses limitations et variantes.

Considérons un modèle indexé par un paramètre $\theta \in \Theta$. Soit $Y = (Y_1, \dots, Y_n)$ le vecteur des observations et $X = (X_1, \dots, X_n)$ les variables cachées correspondantes. Notons $L_n(\theta; Y)$ la vraisemblance du vecteur d'observations Y lorsque le paramètre est θ. L'EMV est alors défini par

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \log L_n(\theta; Y).$$

On note $L_n(\theta; (X, Y))$ la *vraisemblance complétée*, c'est-à-dire la vraisemblance, sous le paramètre θ, si on avait observé X. Comme X n'est pas observé, la vraisemblance complétée n'est pas calculable, mais on peut calculer son espérance conditionnellement à Y. A partir d'un point initial $\theta^{(0)} \in \Theta$, l'algorithme EM construit par récurrence une suite de paramètres $(\theta^{(q)})_{q \geq 0}$ en alternant deux étapes :

— L'étape **E** (Espérance) est le calcul de la *quantité intermédiaire*. Elle est définie, pour $\theta, \theta^{(q)} \in \Theta$, par

$$R(\theta, \theta^{(q)}) = \mathbb{E}^{\theta^{(q)}} \left[\log L_n(\theta; (X, Y)) \mid Y \right],$$

c'est-à-dire l'espérance de la log-vraisemblance complétée sous la loi *a posteriori* et sous le paramètre $\theta^{(q)}$. Lorsque $\theta^{(q)}$ est proche du "vrai" paramètre, on peut voir cette quantité comme une approximation de la log-vraisemblance complétée $\log L_n(\theta; (X, Y))$, laquelle est inaccessible.

— L'étape **M** (Maximisation) est la mise à jour des paramètres par la maximisation en θ de ce substitut de la log-vraisemblance complétée :

$$\theta^{(q+1)} \in \arg\max_{\theta \in \Theta} R(\theta, \theta^{(q)}).$$

Ces deux étapes sont répétées jusqu'à atteindre un critère d'arrêt. Typiquement, on choisit $\varepsilon > 0$ et on arrête l'algorithme dès que

$$\frac{L\left(\theta^{(q)}; Y\right) - L\left(\theta^{(q-1)}; Y\right)}{L\left(\theta^{(q-1)}; Y\right)} < \varepsilon.$$

Notre estimateur est alors $\theta^{(q)}$. Il découle de l'inégalité de Jensen que le suite des vraisemblances $\left(L_n(\theta^{(q)}; Y)\right)_{q \geq 0}$ est croissante. D'autre part, il a été établi par Wu (1983) que sous certaines conditions de régularité, cette suite converge vers un maximum local de la fonction de vraisemblance. Insistons sur le fait que s'il y a convergence, c'est vers un maximum local, mais en général rien ne garantit qu'il s'agit d'un maximum global. En particulier, dans des cas particulièrement défavorables et lorsque la fonction de vraisemblance est compliquée (par exemple si la dimension de l'espace des paramètres est élevée), il se peut que l'algorithme EM converge vers un maximum local dont la vraisemblance est très inférieure à la vraisemblance maximale. C'est l'une des limitations majeures de cet algorithme. Comme l'EM est déterministe, tout va dépendre du point initial $\theta^{(0)}$. Une solution classique, mais coûteuse en temps de calcul, pour contourner ce problème est de lancer plusieurs fois l'algorithme EM en utilisant des initialisations différentes, puis de sélectionner parmi tous les estimateurs obtenus (qui sont souvent tous distincts) le paramètre qui correspond à la meilleure vraisemblance. On peut par exemple tirer aléatoirement ces différentes initialisations. Souvent, en pratique, on a une idée de la région de Θ dans laquelle se situe le "vrai" paramètre car on peut lui donner un sens physique, ce qui permet d'initialiser

l'EM intelligemment. Par exemple si on estime la matrice de transition d'un HMM et qu'on s'attend à des états plutôt stables, on initialisera l'EM avec une matrice de transition à diagonale dominante. Si l'un des paramètres représente une température atmosphérique moyenne en degrés Celcius, on l'initialisera à une valeur réaliste. Dans Biernacki et al. (2003), les auteurs comparent plusieurs procédures d'initialisation de l'algorithme EM, en utilisant des variantes telles que SEM (Broniatowski et al., 1983).

Un autre inconvénient de l'algorithme EM est que sa convergence, si elle existe, peut être lente. Si de plus à chaque étape **M** il est nécessaire d'employer une routine d'optimisation numérique, le temps de calcul peut devenir important, surtout si l'EM est lancé plusieurs fois pour explorer différents points initiaux. Pour remédier à ces inconvénients, plusieurs variantes de l'algorithme EM ont été introduites.

Variantes de l'algorithme EM L'algorithme GEM (Generalized EM), introduit en même temps que l'algorithme EM par Dempster et al. (1977), est une modification de l'étape **M** : au lieu de chercher à maximiser la quantité intermédiaire $R\left(\theta, \theta^{(q)}\right)$, on cherche un $\theta^{(q+1)}$ tel que $R\left(\theta^{(q+1)}, \theta^{(q)}\right) > R\left(\theta^{(q)}, \theta^{(q)}\right)$, ce qui implique que la vraisemblance croît. Cela peut être utile lorsqu'il est difficile de maximiser $\theta \mapsto R\left(\theta, \theta^{(q)}\right)$, ou bien dans les premières itérations de l'EM, lorsque l'on veut s'approcher rapidement d'un maximum local sans chercher une grande précision.

L'algorithme ECM (Meng and Rubin, 1993) est un cas particulier de l'algorithme GEM qui consiste à remplacer l'étape **M** par une série de maximisations conditionnelles. Plus précisément, si θ s'écrit sous la forme $\theta = (\theta_1, \ldots, \theta_d)$, on maximise la quantité intermédiaire successivement en chacune de ses variables, en fixant toutes les autres. Ainsi la maximisation, éventuellement complexe, de l'étape **M** est subdivisée en plusieurs maximisations plus simples.

Plusieurs variantes randomisées de l'algorithme EM permettent de limiter le risque d'aboutir à un maximum local non global de la vraisemblance (Celeux et al., 1995; Delyon et al., 1999). La plus simple d'entre elles est l'algorithme SEM (Broniatowski et al., 1983; Celeux and Diebolt, 1986). Dans cette variante, à l'étape q, au lieu de calculer l'espérance de la log-vraisemblance complétée sous la loi *a posteriori* (c'est-à-dire la loi de X sachant Y) et sous $\theta^{(q)}$ comme dans l'algorithme EM, on *simule* $\mathbf{x}$ selon cette même loi, sous le paramètre courant. L'étape **M** est alors la maximisation de $L_n[\theta; (\mathbf{x}, Y)]$, la log-vraisemblance complétée par les variables latentes simulées. La randomisation permet de garder la possibilité de s'échapper du bassin d'attraction d'un maximum local, ce qui est impossible avec l'EM, déterministe.

L'algorithme MCEM (Wei and Tanner, 1990) est une généralisation de SEM dans laquelle la quantité intermédiaire est remplacée par une approximation Monte-Carlo. Cela est utile dans les cas où l'espérance dans l'étape **E** n'admet pas de forme explicite. Au lieu de maximiser $R\left(\theta, \theta^{(q)}\right)$, on maximise

$$\tilde{R}\left(\theta, \theta^{(q)}\right) \frac{1}{m} \sum_{j=1}^{m} \log L_n\left[\theta; (\mathbf{x}^{(j)}, Y)\right],$$

où chaque $\mathbf{x}^{(j)}$ a été généré comme dans SEM, de façon indépendante. Lorsque $m = 1$ on retrouve l'algorithme SEM, et lorsque m est grand, par la loi des grands nombres, on retrouve l'algorithme EM. Comme signalé dans Celeux et al. (1995), il est pertinent de faire croître m au cours des itérations du MCEM, c'est-à-dire $m = m(q)$. Ainsi lors des premières itérations on s'autorise à explorer l'espace des paramètres car la variance de $\tilde{R}$ est grande, puis lorsque m devient grand, la variance de $\tilde{R}$ diminue et on se rapproche du comportement d'un EM classique car on est raisonnablement sûr d'être au voisinage d'un "bon" maximum local : il s'agit d'une

forme particulière de recuit simulé (voir par exemple Kirkpatrick et al. (1983)) dans laquelle le paramètre de température est $\frac{1}{m}$. Toute la difficulté est donc de déterminer une bonne vitesse de croissance de $m(q)$.

L'algorithme EM dans le cadre des HMM Dans le cas particulier d'un HMM, l'algorithme EM a été introduit par Baum et al. (1970). Il est aussi connu sous le nom d'algorithme de *Baum-Welch*. Dans Baum et al. (1970), les auteurs montront que l'algorithme EM converge vers un maximum local de la vraisemblance. Considérons un HMM homogène d'espace d'état $\mathsf{X} = \{1, \dots, K\}$, de loi initiale ξ, de matrice de transition Q et dont les lois d'émission sont paramétrées par $\theta_Y \in \Theta_Y$ et à densité par rapport à une mesure dominante μ. Pour $k \in \mathsf{X}$, on note $f_k^{\theta_Y}$ la densité d'émission dans l'état k lorsque le paramètre est θ_Y. L'espace des paramètres est alors $\Delta_K \times \Sigma_K \times \Theta_Y$, où Σ_K est l'ensemble des matrices stochastiques de taille K, et on note $\theta = (\xi, Q, \theta_Y)$. La vraisemblance sous le paramètre θ s'écrit

$$p^\theta(Y_{1:n}) = \sum_{x_1,\dots,x_n} \xi_{x_1} f_{x_1}^{\theta_Y}(Y_1) \prod_{t=2}^{n} Q_{x_{t-1},x_t} f_{x_t}^{\theta_Y}(Y_t).$$

On comprend donc que la maximisation directe de cette vraisemblance est infaisable numériquement dès que n est raisonnablement grand puisque la somme comporte K^n termes. En revanche la log-vraisemblance complétée s'exprime de façon beaucoup plus agréable, ce qui permet une implémentation facile de l'algorithme EM. En effet,

$$\log p^\theta(X_{1:n}, Y_{1:n}) = \log \xi_{X_1} + \sum_{t=1}^{n-1} \log Q_{X_t X_{t+1}} + \sum_{t=1}^{n} \log f_{X_t}^{\theta_Y}(Y_t).$$

A partir d'un paramètre initial $\theta^{(0)}$, on alterne les étapes **E** et **M** de la façon suivante. A l'itération q, l'étape **E** est le calcul de

$$R\left(\theta, \theta^{(q)}\right) = \mathbb{E}^{\theta^{(q)}}\left[\log p^\theta(X_{1:n}, Y_{1:n}) \mid Y_{1:n}\right],$$

soit

$$R\left(\theta, \theta^{(q)}\right) = \sum_{k=1}^{K} \pi_{1|n}^{\theta^{(q)}}(k) \log \xi_k + \sum_{k,l=1}^{K} \sum_{t=1}^{n-1} \pi_{t,t+1|n}^{\theta^{(q)}}(k,l) \log Q_{kl} + \sum_{k=1}^{K} \sum_{t=1}^{n} \pi_{t|n}^{\theta^{(q)}}(k) \log f_k^{\theta_Y}(Y_t),$$

où les $\pi_{t|n}^{\theta^{(q)}}$ et $\pi_{t,t+1|n}^{\theta^{(q)}}(k,l)$ sont les probabilités de lissage calculées sous le paramètre courant $\theta^{(q)}$. L'étape **E** se résume donc au calcul de ces probabilités de lissage, et donc à l'algorithme forward-backward (voir paragraphe 2.3.1). L'étape **M** est la maximisation en $\theta = (\xi, Q, \theta_Y)$ de $\theta \mapsto R(\theta, \theta^{(q)})$, sous les contraintes $\sum_{k=1}^{K} \xi_k = 1$ et pour tout $k \in \mathsf{X}$, $\sum_{l=1}^{K} Q_{kl} = 1$. Remarquons que chacun des trois termes composant $R(\theta, \theta^{(q)})$ peut être maximisé séparement. Les deux premiers donnent :

$$\begin{aligned} \xi^{(q+1)} &= \pi_{1|n}^{(q)} \\ Q_{kl}^{(q+1)} &= \frac{\sum_{t=1}^{n-1} \pi_{t,t+1|n}^{(q)}(k,l)}{\sum_{t=1}^{n-1} \pi_{t|n}^{(q)}(k)}, \quad 1 \le k, l \le K \end{aligned}$$

Ces deux formules s'interprètent facilement. La nouvelle valeur de Q_{kl} est le rapport entre le nombre moyen, sous $\theta^{(q)}$, de transitions de l'état k vers l'état l, et le nombre moyen de transitions

à partir de l'état k. Notons que bien que l'EM effectue à chaque itération une mise à jour de la loi initiale ξ à travers la première probabilité de lissage, ce paramètre du HMM ne peut pas être estimé si l'on n'observe, comme c'est généralement le cas en pratique, qu'une seule trajectoire du HMM. En d'autres termes, dans un tel cas, l'EM fournit davantage une estimation de X_1 que de sa loi.

Dans certains cas, il est possible d'obtenir une solution explicite au problème d'optimisation

$$\arg\max_{\theta_Y \in \Theta_Y} \sum_{k=1}^{K} \sum_{t=1}^{n} \pi_{t|n}^{\theta^{(q)}}(k) \log f_k^{\theta_Y}(Y_t). \tag{2.3}$$

Dans le cas contraire, on a recours à un algorithme d'optimisation numérique pour trouver une solution approchée à ce problème : on peut par exemple utiliser une méthode *quasi-Newton* telle que BFGS lorsque le gradient de la fonction objectif se calcule facilement. Un cas favorable et courant en pratique est celui où θ_Y s'écrit sous la forme $(\theta_Y^{(1)}, \dots, \theta_Y^{(K)})$ et où $f_k^{\theta_Y}$ ne dépend de θ_Y qu'à travers $\theta_Y^{(k)} \in \Theta_Y^{(k)}$. Usuellement, on a même $\Theta_Y^{(1)} = \cdots = \Theta_Y^{(K)}$, ce qui signifie que toutes les lois d'émissions appartiennent à une même famille paramétrique. Dans un tel cas, le problème (2.3) peut se scinder en K problèmes indépendants et de dimension plus faible :

$$\arg\max_{\theta_Y^{(k)} \in \Theta_Y^{(k)}} \sum_{t=1}^{n} \pi_{t|n}^{\theta^{(q)}}(k) \log f_k^{\theta_Y^{(k)}}(Y_t), \quad 1 \leq k \leq K. \tag{2.4}$$

Cela est particulièrement intéressant d'un point de vue numérique puisque ces K optimisations étant indépendantes, elles peuvent être menées en parallèle, ce qui représente un important gain en temps de calcul.

Cas particulier : lois d'émission mélanges Considérons le cas où les lois d'émission sont des mélanges finis à M composantes, par exemple des mélanges de gaussiennes. Dans ce cas les densités d'émissions s'écrivent sous la forme

$$f_k = \sum_{m=1}^{M} p_{km} f_{km},$$

où $p_k = (p_{k1}, \dots, p_{kM}) \in \Delta_M$ est un vecteur de probabilité représentant les poids du mélange. En termes de probabilités, l'interprétation est la suivante : si U est une variable aléatoire à valeurs dans $\{1, \dots, M\}$ telle que $\mathbb{P}(U = m) = p_{km}$ et si V est une variable aléatoire à valeurs dans Y telle que $V \mid \{U = m\} \sim f_{km}$, alors $V \sim f_k$. Dans ce cas, nous allons voir que l'algorithme EM prend une forme particulière. Pour chaque instant t, on introduit une seconde variable latente Z_t correspondant au choix d'une population dans le mélange dans l'état k : elle vérifie $\mathbb{P}(Z_t = m \mid X_t = k) = p_{km}$ et les Z_t sont indépendantes conditionnellement aux états X_t.

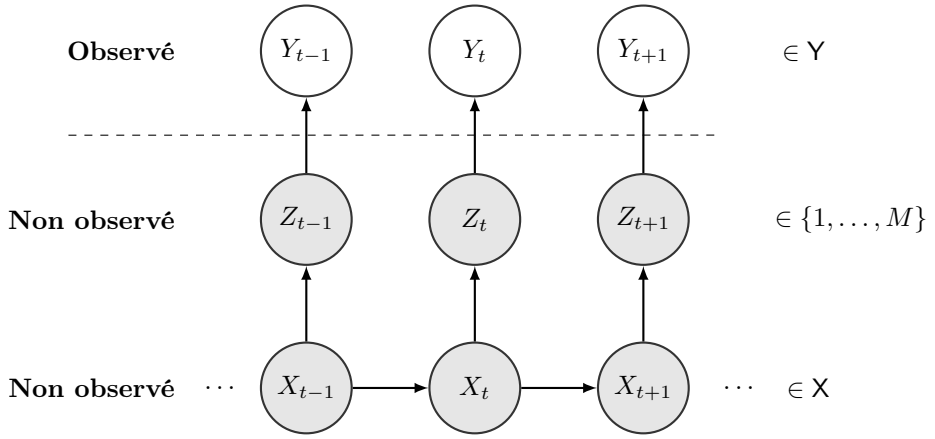

FIGURE 2.2 – Dynamique d'un modèle de Markov caché dont les lois d'émission sont des mélanges.

Pour $t \geq 1$, on introduit la variable aléatoire $\tilde{X}_t = (X_t, Z_t)$, à valeurs dans $\mathsf{X} \times \{1, \dots, M\}$. Alors le HMM dont les lois d'émission sont des mélanges peut aussi être vu comme un HMM dont les états cachés sont les $\tilde{X}_t$, et dont les densités d'émission sont les f_{km}. En effet les deux représentations induisent la même loi pour le processus $(Y_t)_{t\geq 1}$. On introduit alors les probabilités de lissage correspondant à la seconde représentation : pour $t \geq 1$, $k \in \mathsf{X}$ et $m \in \{1, \dots, M\}$,

$$\gamma_t(k, m) = \mathbb{P}\left(\tilde{X}_t = (k, m) \mid Y_{1:n}\right).$$

Ces quantités se calculent facilement de la façon suivante :

$$\gamma_t(k, m) = \pi_{t|n}(k) \frac{p_{km} f_{km}(Y_t)}{\sum_{m'=1}^{M} p_{km'} f_{km'}(Y_t)},$$

où l'on rappelle que $\pi_{t|n}(k) = \mathbb{P}(X_t = k \mid Y_{1:n})$ est calculée grâce à l'algorithme forward-backward. Voyons maintenant comment estimer les paramètres d'un tel HMM grâce à l'algorithme EM. On note $\mathbf{p} = (p_1, \dots, p_K) \in (\Delta_M)^K$. Le paramètre est

$$\theta = (\xi, Q, \mathbf{p}, \theta_Y) \in \Delta_K \times \Sigma_K \times (\Delta_M)^K \times \Theta_Y.$$

La log-vraisemblance complétée s'écrit :

$$\log p^\theta(\tilde{X}_{1:n}, Y_{1:n}) = \log \xi_{X_1} + \sum_{t=1}^{n-1} \log Q_{X_t X_{t+1}} + \sum_{t=1}^{n} \log p_{X_t Z_t} + \sum_{t=1}^{n} \log f^{\theta_Y}_{X_t Z_t}(Y_t).$$

D'où l'expression de la quantité intermédiaire de l'EM :

$$
\begin{aligned}
R\left(\theta, \theta^{(q)}\right) &= \mathbb{E}^{\theta^{(q)}}\left[\log p^{\theta}(\tilde{X}_{1:n}, Y_{1:n}) \mid Y_{1:n}\right] \\
&= \sum_{k=1}^{K} \pi_{1|n}^{\theta^{(q)}}(k) \log \xi_k \\
&+ \sum_{t=1}^{n-1} \sum_{k,l=1}^{K} \pi_{t,t+1|n}^{\theta^{(q)}}(k,l) \log Q_{kl} \\
&+ \sum_{t=1}^{n} \sum_{k=1}^{K} \sum_{m=1}^{M} \gamma_t^{\theta^{(q)}}(k,m) \log p_{km} \\
&+ \sum_{t=1}^{n} \sum_{k=1}^{K} \sum_{m=1}^{M} \gamma_t^{\theta^{(q)}}(k,m) \log f_{km}^{\theta_Y}(Y_t)
\end{aligned}
$$

Ainsi, et c'est tout l'intérêt d'avoir introduit une nouvelle variable latente Z_t, dans l'étape **M** on peut optimiser séparément en $\mathbf{p}$ et en θ_Y, au lieu d'optimiser simultanément tous les paramètres des lois d'émission. Mieux : on peut optimiser séparément sur chaque état k. Ainsi on résout, pour tout $k \in \mathsf{X}$,

$$
\arg\max_{p_k \in \Delta_M} \sum_{t=1}^{n} \sum_{m=1}^{M} \gamma_t^{\theta^{(q)}}(k,m) \log p_{km}.
$$

Cela fournit, pour tout $m \in \{1, \dots, M\}$ et $k \in \mathsf{X}$,

$$
p_{km}^{(q+1)} = \frac{\sum_{t=1}^{n} \gamma_t^{\theta^{(q)}}(k,m)}{\sum_{m'=1}^{M} \sum_{t=1}^{n} \gamma_t^{\theta^{(q)}}(k,m')} = \frac{\sum_{t=1}^{n} \gamma_t^{\theta^{(q)}}(k,m)}{\sum_{t=1}^{n} \pi_{t|n}^{\theta^{(q)}}(k)}.
$$

Signalons à nouveau que cette formule a une interprétation claire en termes de nombre moyen de passages dans l'état (k,m). Les autres paramètres sont mis à jour de la même façon que dans le cas général.

Remarque Tous les algorithmes que nous venons de présenter pour des HMM homogènes s'adaptent très facilement à des HMM non-homogènes.

2.3.4 Exemple sur simulations

Illustrons sur un exemple simple de HMM les différents algorithmes que nous venons de présenter. Soit $(X_t, Y_t)_{t \geq 1}$ un HMM à deux états, i.e. $K = 2$ et $\mathsf{X} = \{1, 2\}$, dont l'espace des observations est $\mathsf{Y} = \mathbb{R}$. On suppose que les lois d'émission sont gaussiennes. On a donc $\theta_Y = \left(\theta_Y^{(1)}, \theta_Y^{(2)}\right)$ avec, pour $k \in \mathsf{X}$, $\theta_Y^{(k)} = (m_k, \sigma_k^2) \in \mathbb{R} \times \mathbb{R}_+^*$, $\Theta_Y = \left(\mathbb{R} \times \mathbb{R}_+^*\right)^K$, et, pour $k \in \mathsf{X}$ et $y \in \mathbb{R}$,

$$
f_k^{\theta_Y}(y) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(y - m_k)^2}{\sigma_k^2}\right].
$$

On simule une réalisation de longueur $n = 1000$ de ce HMM en utilisant les *vrais* paramètres suivants :

$$
Q^* = \begin{pmatrix} 0.85 & 0.15 \\ 0.1 & 0.9 \end{pmatrix}, \quad \xi^* = (0.3; 0.7), \quad (m_1, m_2)^* = (-1; 1), \quad (\sigma_1^2, \sigma_2^2)^* = (1; 0.5).
$$

Dans un premier temps, supposons connus les vrais paramètres, et comparons les vrais états (ici on y a accès puisqu'on les a simulés), les états obtenus par la méthode MAP (Equation (2.1)) via l'algorithme forward-backward, et les états obtenus par l'algorithme de Viterbi (Equation (2.2)). Les résultats sont présentés sur la Figure 2.3 pour les 200 premières observations.

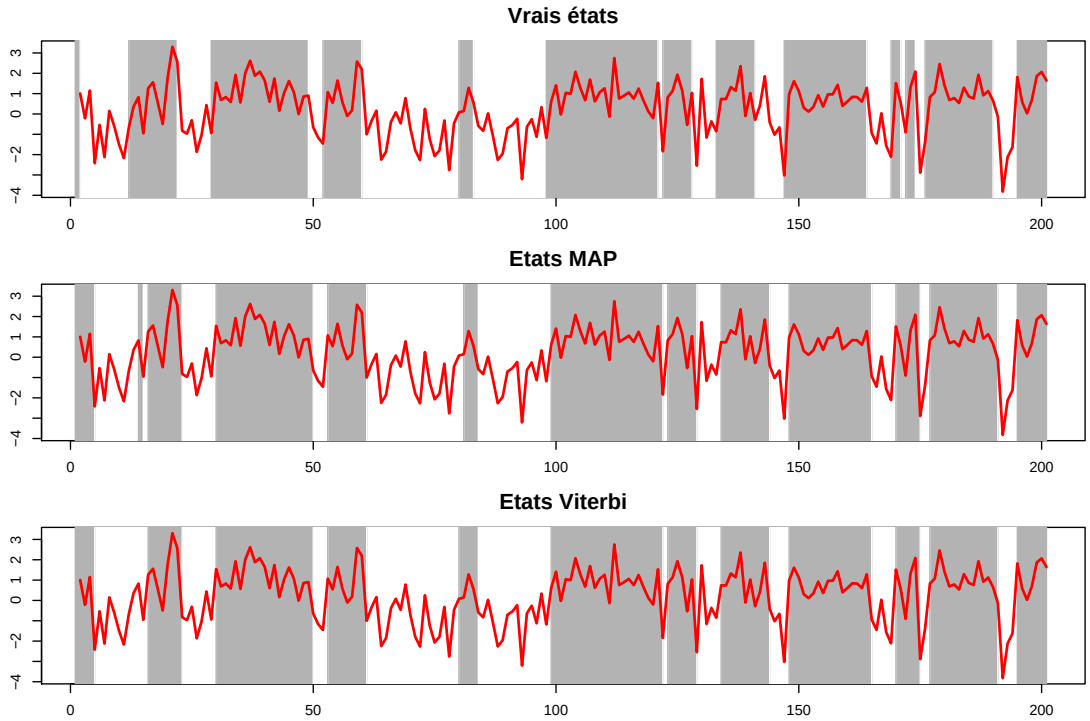

FIGURE 2.3 – Comparaison des états fournis par le MAP et l'algorithme de Viterbi. Les observations sont en rouge, l'état 1 est figuré par un fond blanc et l'état 2 par un fond gris.

On constate que globalement les états sont bien retrouvés par les algorithmes mais qu'ils commettent parfois quelques erreurs : sur 1000 états inconnus, MAP a commis 47 erreurs d'affectation et Viterbi 48. Ce taux de réussite élevé peut être imputé au fait que dans cet exemple les lois d'émission sont bien séparées, il est donc plus facile d'attribuer un état à chaque observation. Supposons à présent les paramètres inconnus et estimons-les en utilisant l'algorithme EM. On l'initialise avec les paramètres suivant :

$$Q^{(0)} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \quad \xi^{(0)} = (0.5; 0.5), \quad (\sigma_1^2, \sigma_2^2)^{(0)} = (2; 2).$$

et (m_1, m_2) est initialisé aléatoirement en tirant selon une loi normale centrée réduite.
Ici l'étape **M** ne nécessite pas d'optimisation numérique en θ_Y puisque le problème de maximisation (2.4) possède une solution explicite. Plus précisément, à l'itération q de l'algorithme EM, la mise à jour des paramètres (m_1, m_2) et (σ_1^2, σ_2^2) se fait de la façon suivante. Pour $k \in \{1, 2\}$,

$$m_k^{(q+1)} = \frac{\sum_{t=1}^n \pi_{t|n}^{(q)}(k) Y_t}{\sum_{t=1}^n \pi_{t|n}^{(q)}(k)}, \quad (\sigma_k^2)^{(q+1)} = \frac{\sum_{t=1}^n \pi_{t|n}^{(q)}(k) \left(Y_t - m_k^{(q+1)}\right)^2}{\sum_{t=1}^n \pi_{t|n}^{(q)}(k)}.$$

Notons là-aussi que les formules obtenues sont intuitives, puisqu'on obtient respectivement les moyennes et variances empiriques des observations, pondérées par les probabilités de lissage. Les résultats des 20 premières itérations de l'algorithme EM sont présentées sur la figure 2.4.

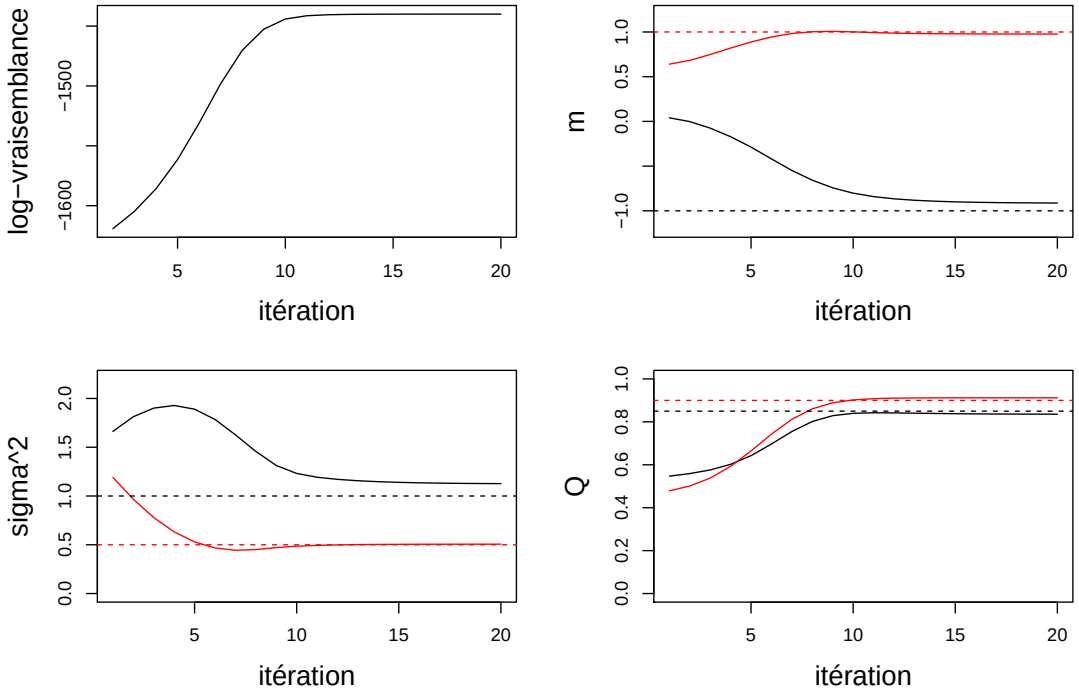

FIGURE 2.4 – En haut à gauche : croissance de la log-vraisemblance au cours des itérations de l'EM. Les autres graphes représentent les paramètres estimé à chaque itération. En haut à droite, m_1 (noir) et m_2 (rouge). En bas à gauche, σ_1^2 (noir) et σ_2^2 (rouge). En bas à droite, Q_{11} (noir) et Q_{22} (rouge). Les pointillés représentent les vraies valeurs.

Conformément à la théorie, la log-vraisemblance augmente à chaque itération. De plus, l'algorithme a convergé vers des paramètres proches des vrais paramètres en une quinzaine d'itérations (temps de calcul de l'ordre de la seconde). La précision pourrait être améliorée en augmentant la taille n du vecteur d'observations.

2.4 Résultats d'identifiabilité et de convergence de l'EMV

Dans cette section nous rappelons quelques résultats théoriques concernant les modèles de Markov cachés.

2.4.1 Identifiabilité

Commençons par rappeler la définition la plus générale de l'identifiabilité, avant de traiter le cas des HMM.

Définition 2.4. *Un modèle statistique* $(\mathbb{P}_\theta)_{\theta\in\Theta}$ *est dit* identifiable *lorsque l'application* $\theta \mapsto \mathbb{P}_\theta$ *est injective.*

Ainsi lorsqu'un modèle est identifiable, on peut retrouver le paramètre à partir de la loi. Si l'on suppose qu'il existe un *vrai* paramètre, c'est-à-dire un $\theta^* \in \Theta$ tel que les observations ont été générées par la loi $\mathbb{P}_{\theta^*}$ (modèle bien spécifié), on peut alors considérer une notion plus faible d'identifiabilité : pour tout $\theta \in \Theta$, $\mathbb{P}_\theta = \mathbb{P}_{\theta^*}$ si et seulement si $\theta = \theta^*$. Si cette condition n'est pas vérifiée, il est inutile d'espérer inférer les paramètres à partir des observations. En effet, sans identifiabilité, même la connaissance parfaite de la loi des observations ne fournit pas le paramètre.

Si $(X_t, Y_t)_{t\geq 1}$ est un HMM homogène à K états, on note $\mathbb{P}_Y^{(\xi,Q,\nu)}$ la loi du processus des observations $(Y_t)_{t\geq 1}$ lorsque la loi de X_1 est ξ, la matrice de transition de la chaîne $(X_t)_{t\geq 1}$ est Q et le vecteur des densités d'émission est $\nu = (\nu_1, \dots, \nu_K)$. Supposons le modèle bien spécifié : il existe un *vrai* paramètre (ξ^*, Q^*, ν^*). Peut-on retrouver les vrais paramètres à partir de la loi des observations ? Autrement dit, a-t-on, pour tout paramètre (ξ, Q, ν),

$$\left(\mathbb{P}_Y^{(\xi,Q,\nu)} = \mathbb{P}_Y^{(\xi^*,Q^*,\nu^*)}\right) \implies ((\xi, Q, \nu) = (\xi^*, Q^*, \nu^*)) \ ?$$

Dans le cas général, la réponse est négative car la loi de $(Y_t)_{t\geq 1}$ est invariante par permutation des états. Formellement, soit σ une permutation de $\mathsf{X} = \{1, \dots, K\}$ et (ξ, Q, ν) un paramètre. On définit, pour tous $k, l \in \mathsf{X}$,

$$\tilde{\xi}_k = \xi_{\sigma(k)}, \quad \tilde{Q}_{kl} = Q_{\sigma(k)\sigma(l)}, \quad \tilde{\nu}_k = \nu_{\sigma(k)}.$$

En d'autres termes, $(\tilde{\xi}, \tilde{Q}, \tilde{\nu})$ est obtenu à partir de (ξ, Q, ν) en changeant la numérotation des états cachés. On a alors

$$\mathbb{P}_Y^{(\tilde{\xi},\tilde{Q},\tilde{\nu})} = \mathbb{P}_Y^{(\xi,Q,\nu)}.$$

Par conséquent, dans le cas général, on ne peut pas obtenir mieux qu'une identifiabilité à permutation près des états cachés.

Des résultats d'identifiabilité pour les HMM stationnaires non paramétriques ont été montrés dans Gassiat et al. (2016b) et Gassiat et al. (2016a). Le premier concerne les HMM dont les lois d'émission sont obtenues par translation d'une loi non paramétrique inconnue. Dans ce cadre, il est possible d'identifier tous les paramètres (et même le nombre d'états) à partir de la loi de (Y_1, Y_2). Dans le second les auteurs établissent l'identifiabilité à permutation des états cachés près sous les conditions d'inversibilité de la matrice de transition et d'indépendance linéaire des lois d'émission, à partir de la loi de trois observations consécutives, c'est-à-dire, pour un HMM stationnaire, la loi de (Y_1, Y_2, Y_3). Nous noterons $\mathbb{P}_{Y_{1:3}}^{(\xi,Q,\nu)}$ cette loi sous les paramètres (ξ, Q, ν).

Théorème 2.1 (Gassiat et al. (2016a))**.** *Soit* $(X_t, Y_t)_{t\geq 1}$ *un HMM stationnaire à* K *états (*K *connu), de loi initiale* ξ^**, de matrice de transition* Q^* *et de lois d'émissions* $(\nu_1^*, \dots, \nu_K^*)$*. On suppose que* Q^* *est inversible, que les lois d'émissions* $(\nu_k)_{1\leq k\leq K}$ *sont indépendantes et que pour tout* $k \in \mathsf{X}$*,* $\xi_k > 0$*. Alors les paramètres sont identifiables à partir de la loi de* (Y_1, Y_2, Y_3)*, à permutation des états cachés près. Plus précisément, si* (ξ, Q, ν) *sont des paramètres tels que* $\mathbb{P}_{Y_{1:3}}^{(\xi,Q,\nu)} = \mathbb{P}_{Y_{1:3}}^{(\xi^*,Q^*,\nu^*)}$*, alors il existe* σ *une permutation de* $\{1, \dots, K\}$ *telle que pour tous* $k, l \in \{1, \dots, K\}$*,*

$$\xi_k = \xi^*_{\sigma(k)}, \quad Q_{kl} = Q^*_{\sigma(k)\sigma(l)}, \quad \nu_k = \nu^*_{\sigma(k)}.$$

Remarque . Les hypothèses ne s'appliquent qu'aux vrais paramètres.

Un théorème similaire a été démontré par Alexandrovich et al. (2016). Dans leur version, l'hypothèse d'indépendance linéaire des lois d'émission est relâchée : ils font seulement l'hypothèse qu'elles sont distinctes. En contrepartie, la matrice de transition n'est plus seulement supposée inversible mais aussi ergodique, et l'identification se fait à partir de la loi $(Y_1, \ldots, Y_{2K+1})$ au lieu de (Y_1, Y_2, Y_3). Les démonstrations de ces deux théorèmes d'identifiabilité s'appuient sur des arguments développés dans Allman et al. (2009), dont les auteurs montrent des résultats d'identifiabilité pour différents modèles à variables latentes, dont des HMM discrets, en se basant sur des résultats algébriques de Kruskal (1977).

Dans le chapitre 3, nous généraliserons le Théorème 2.1 à une classe particulière de HMM non-homogènes.

2.4.2 Propriétés asymptotiques de l'estimateur du maximum de vraisemblance

Les propriétés asymptotiques de l'estimateur du maximum de vraisemblance pour les HMM ont été largement étudiées. La consistance et la normalité asymptotique de l'EMV ont été obtenues par Baum and Petrie (1966) dans le cadre de HMM stationnaires dont l'espace d'état et l'espace des observations sont tous deux finis. Le résultat de consistance forte est étendu à un espace d'observations continu par Leroux (1992) et la normalité asymptotique est obtenue par Le Gland and Mevel (1997) et Bickel et al. (1998). Ces résultats ont ensuite été généralisés au cas où l'espace d'états n'est plus nécessairement fini mais est seulement supposé être un espace topologique compact par Douc et al. (2001) (voir aussi Douc et al. (2011) pour la consistance forte). Les résultats de Douc et al. (2001) sont étendus à des HMM auto-régressifs dans Douc et al. (2004). Les propriétés asymptotiques de l'EMV ont aussi été étudiées dans le cas (important en pratique) où le modèle est mal spécifié, ce qui signifie que la loi des observations n'appartient pas au modèle : voir par exemple Mevel and Finesso (2004) pour la consistance et la normalité asymptotique dans le cas d'un espace d'état fini et Douc et al. (2012) pour la consistance dans le cas d'un espace d'état plus général.

Tous les travaux que nous venons de citer concernent des HMM homogènes, voire stationnaires. Le cas inhomogène n'a été abordé que récemment. Dans Ailliot and Pene (2015), les auteurs obtiennent un résultat de consistance forte de l'EMV pour une généralisation des HMM auto-régressifs dans laquelle le noyau de transition n'est pas constant mais dépend des valeurs précédentes des observations et/ou de variables exogènes. Pouzo et al. (2016) étudient un modèle similaire en abandonnant l'hypothèse de bonne spécification mais en se restreignant à un espace d'états fini, et obtiennent la normalité asymptotique de l'estimateur du maximum de vraisemblance. Dans Diehn et al. (2018), les auteurs étudient un autre type de non-homogénéité. Ils considèrent des processus $(X_t, Y_t, Z_t)_t$ où seul Z_t est observé, $(X_t, Y_t)_t$ est un HMM homogène et $(X_t, Z_t)_t$ est un HMM inhomogène tel que la distance entre Y_t et Z_t tend vers 0 lorsque t tend vers l'infini. Autrement dit, le HMM inhomogène est asymptotiquement "proche" d'un HMM homogène. La consistance forte de l'EMV est démontrée dans un tel cadre. Dans le chapitre 3, nous étudions un autre cas particulier de HMM inhomogène, celui où les matrices de transition et les lois d'émission dépendent périodiquement du temps. Dans le chapitre 4, nous considérons le cas où les lois d'émission possèdent une tendance polynomiale.

2.5 Autres méthodes d'estimation des HMM

Bien que la méthode du maximum vraisemblance soit très populaire pour l'estimation des paramètres des HMM, notamment en raison de la facilité d'implémentation de l'algorithme EM,

elle n'est pas la seule. D'autres méthodes ont été développées et étudiées au cours des dernières années. Dans cette section nous présentons brièvement deux d'entre elles : la méthode spectrale et la méthode des moindre carrés.

2.5.1 Estimateur spectral

La méthode *spectrale* d'estimation des HMM (Hsu et al., 2012; Anandkumar et al., 2012) est étroitement liée aux résultats d'identifiabilité présentés dans la Section 2.4.1. Le théorème 2.1 établit que sous certaines hypothèses sur les paramètres d'un HMM stationnaire, ceux-ci peuvent être obtenus à partir de la loi de trois observations consécutives. La méthode spectrale montre comment le faire explicitement en fournissant un algorithme d'estimation. Plus précisément, c'est une méthode des moments : à partir des moments de la loi de trois observation consécutives, on montre que les paramètres du HMM sous-jacent peuvent être obtenus par des opérations algébriques simples. Effectuer ces mêmes opérations en utilisant les moments empiriques (les moments théoriques sont inacessibles puisqu'ils dépendent de la loi inconnue) fournit un estimateur, dit *estimateur spectral*. Contrairement à l'estimation par maximum de vraisemblance via l'algorithme EM, l'algorithme spectral ne souffre pas du problème de convergence vers un maximum local. D'autre part, il est adapté au cadre non paramétrique : il permet d'estimer des densités d'émission dans un espace de dimension infinie. L'algorithme spectral pour des HMM non paramétriques est détaillé dans De Castro et al. (2017) et De Castro et al. (2016). Celui-ci permet de retrouver les projections des densités d'émission sur une suite croissante de sous-espaces de dimension finie.

La méthode spectrale sera utilisée dans le chapitre 3 pour montrer un résultat d'identifiabilité pour des HMM non homogènes.

2.5.2 Estimateur des moindre carrés pénalisés

L'estimateur des moindre carrés pénalisées pour des HMM stationnaires non paramétriques dont les densités d'émission sont dans l'espace $L^2(\mathsf{Y})$ des fonctions de carré intégrable sur Y est notamment présenté dans De Castro et al. (2016) et Lehéricy (2019). Comme les paramètres d'un HMM (loi initiale ξ, matrice de transition Q, densités d'émission f_k) peuvent être identifiés à partir de la loi de trois observations consécutives, l'idée est d'estimer la densité correspondante, notée g^* en minimisant en t le contraste empirique

$$\gamma_n(t) := \|t\|_2^2 - \frac{2}{n}\sum_{s=1}^{n} t(Y_s, Y_{s+2}, Y_{s+3}),$$

version empirique de la perte quadratique $\|t - g^*\|_2^2 - \|g^*\|_2^2$, minimale en $t = g^*$. Comme pour l'estimation spectrale, on considère une suite croissante $(\mathfrak{P}_M)_{M\in\mathcal{M}\subset\mathbb{N}}$ de sous-espaces de $L^2(\mathsf{Y})$ dont l'union est dense dans $L^2(\mathsf{Y})$, et pour $M \in \mathcal{M}$,

$$\hat{g}_M \in \operatorname*{arg\,min}_{t\in\mathcal{S}_M} \gamma_n(t),$$

où $\mathcal{S}_M$ est l'ensemble des densités de trois observations consécutives, lorsque les lois d'émission sont restreintes à $\mathfrak{P}_M$. Une optimisation numérique est nécessaire pour obtenir $\hat{g}_M$, et son initialisation est importante si l'on veut éviter un maximum local sous-optimal. L'algorithme spectral peut être utilisé pour obtenir une bonne initialisation. La taille du modèle est choisie en sélectionnant $\hat{M}$ minimisant $\gamma_n(\hat{g}_M) + \mathrm{pen}(n, M)$, où $\mathrm{pen}(n, M)$ est une pénalité à préciser, et l'estimateur final est alors $\hat{g}_{\hat{M}}$.

2.6 Sélection du nombre d'états

Le praticien désireux d'utiliser un HMM à espace d'état fini pour modéliser un phénomène se trouve confronté au problème du choix du nombre K d'états. Ce problème est un cas particulier d'un problème très général en statistique, celui de la *sélection de modèle*. Dans certain cas, ce choix est simple car il est directement guidé par la nature du phénomène à modéliser. Dans d'autres cas, comme dans les applications qui nous intéressent dans cette thèse, aucun nombre d'états n'est à privilégier a priori. Le nombre K d'états influe directement sur la complexité du modèle. La matrice de transition comporte $K(K-1)$ paramètres libres et on a K lois d'émissions. Le nombre de paramètres d'un HMM est donc quadratique en K. Une bonne modélisation implique donc un choix correct de K. En effet, un nombre trop élevé d'états pose plusieurs problèmes. D'un point de vue pratique, si le nombre de paramètres est trop élevé, leur estimation devient difficile, d'une part parce que le nombre de données nécessaires pour obtenir une bonne précision devient trop important, et d'autre part parce que la fonction de vraisemblance est alors très complexe et comporte de nombreux maxima locaux, pièges dans lequels l'algorithme EM ne manquera pas de tomber. L'estimation devient aussi plus difficile en pratique parce que cette complexité induit un temps de calcul plus élevé. Enfin, dans de nombreuses applications, on accorde une importance particulière à l'interprétabilité des états, en rapport avec le phénomène que l'on cherche à modéliser. Par exemple, en écologie, des HMM sont utilisés pour modéliser les déplacements d'un animal. On s'attend alors à ce que les états correspondent à des comportements bien identifiables de l'animal (chasser, dormir, migrer...). Si le nombre d'états est choisi trop élevé, certains états au moins n'auront pas d'interprétation claire. A contrario, un nombre d'états trop faible, et donc un modèle trop simple, implique un mauvais ajustement aux données et une perte d'information concernant leur structure. Certains "comportements" des données échapperont au modèle. Il est donc crucial de choisir K dans un intervalle "acceptable".

D'un point de vue théorique, peu de résultats existent concernant la sélection de modèle dans le cadre des HMM, et ces résultats ne sont pas toujours applicables en pratique ou se limitent à des cas particuliers. Pour une revue récente de ces travaux, le lecteur pourra se reporter à l'introduction de Lehéricy (2019). Dans cet article, l'auteur établit deux procédures de sélection de modèle pour des HMM non paramétriques : l'une par moindre carrés pénalisés, l'autre par seuillage des valeurs propres d'une matrice dans l'estimation spectrale. Il montre que ces méthodes sont consistantes au sens où, si les observations sont réellement issues d'un HMM et donc s'il existe un "vrai" nombre d'états K^*, le nombre d'états sélectionné par ces procédures converge presque sûrement vers K^* lorsque le nombre d'observations tend vers l'infini. Cependant ces procédures ne concernent pas l'estimation par maximum de vraisemblance, méthode la plus populaire pour l'estimation des HMM, et celle que nous utilisons dans cette thèse. En outre, ces résultats concernent des modèles *bien spécifiés* : il est supposé que les observations ont été générées par un HMM à K^* états. Or en pratique, lorsque l'on traite des données réelles, ce n'est pas le cas. Les phénomènes réels peuvent être modélisés, avec plus ou moins de succès, par des HMM, mais ils ne sont pas des réalisations d'un HMM. Les phénomènes météorologiques qui motivent cette thèse n'y font pas exception. Des procédures automatiques de sélection de modèle, même justifiées théoriquement, ne peuvent donc être le seul critère de décision.

D'un point de vue pratique, une méthode populaire pour la sélection du nombre d'états d'un HMM (et plus généralement pour la sélection de modèle) est la pénalisation de la vraisemblance. L'idée sous-jacente est la suivante. Lorsque l'on choisit un HMM avec davantage d'états, la vraisemblance du modèle (c'est-à-dire la fonction de vraisemblance évaluée en l'EMV) augmente, ce qui traduit un meilleur ajustement aux données. Ainsi, en considérant le seul critère de la maximisation de la vraisemblance, on serait amené à choisir des modèles toujours plus complexes.

D'où l'idée de "pénaliser" les modèles trop complexes et de maximiser non pas la vraisemblance, mais un critère de vraisemblance pénalisée, de façon à réaliser un compromis entre l'ajustement aux données et la parcimonie du modèle. Plus précisément, si l'on doit choisir un modèle m parmi un ensemble $\mathcal{M}$ de modèles (typiquement, l'ensemble des valeurs possibles pour le nombre K d'états), on considère le critère de vraisemblance pénalisée suivant :

$$\text{crit(m)} = \ell_n^{(m)}\left(\hat{\theta}_n^{(m)}\right) - \text{pen}(m),$$

où $\ell_n^{(m)}$ désigne la log-vraisemblance dans le modèle m, $\hat{\theta}_n^{(m)}$ l'EMV dans le modèle m et pen$(\cdot)$ est une pénalité à préciser, qui est une mesure de la complexité du modèle. On choisit alors un modèle

$$\hat{m} \in \underset{m\in\mathcal{M}}{\arg\max}\, \text{crit}(m)$$

et l'estimateur retenu est alors $\hat{\theta}_n^{\hat{m}}$. Les critères de vraisemblance pénalisée les plus connus sont AIC (Akaike Information Criterion), où la pénalité est égale au nombre p de paramètres libres du modèle, BIC (Bayesian Information Criterion, Schwarz et al. (1978)) où la pénalité est de la forme $\frac{p}{2}\log(n)$. Un critère proche de BIC, qui utilise une approximation de la log-vraisemblance complétée, et utilisé dans le cadre des modèles de mélange, est ICL (Integrated Completed Likelihood, Biernacki et al. (2000)). Notons qu'aucun de ces critères n'est appuyé par une justification théorique dans le cadre des HMM. Dans Pohle et al. (2017), une étude sur simulations a été réalisée pour évaluer les performances de ces trois critères dans le cadre des HMM dans différents cas de mauvaise spécification (par exemple, mauvais choix de famille paramétrique pour les lois d'émission, chaîne de Markov d'ordre 2 pour la séquence d'états, non-homogénéité temporelle...). Les auteurs en concluent que très souvent, quand le modèle est mal spécifié, les critères de sélection automatique surestiment le nombre d'états. Intuituivement, le modèle cherche à "expliquer" un comportement qui ne devrait pas être (la cause de la mauvaise spécification) en ajoutant des états supplémentaires. De ce point de vue, le critère ICL semble plus robuste que BIC et (surtout) AIC. Néanmoins, malgré ses défauts, le critère BIC semble être largement utilisé par les praticiens des HMM, y compris pour modéliser des variables météorologiques (voir par exemple Bellone et al. (2000), Hughes et al. (1999), Bessac et al. (2016), Robertson et al. (2003)). Plusieurs éléments peuvent expliquer la popularité du critère BIC pour la sélection du nombre d'états dans le cadre des HMM. Premièrement, c'est un critère de pénalisation de la vraisemblance. Or l'estimation par maximum de vraisemblance est la méthode la plus utilisée pour l'inférence des HMM. Deuxièmement, c'est un critère facile à calculer : il ne nécessite que de calculer la log-vraisemblance maximale pour différentes valeurs de K, et le nombre de paramètres. Or la log-vraisemblance est fournie pas l'étape E de l'algorithme EM (algorithme *forward*). C'est un avantage par rapport à la validation croisée par exemple (Celeux and Durand, 2008). Enfin, l'expérience montre qu'en pratique, le critère BIC donne souvent des résultats satisfaisants en sélectionnant des modèles parcimonieux et qui s'ajustent bien aux données.

Dans Pohle et al. (2017), les auteurs proposent une procédure en 6 étapes pour choisir le nombre d'états d'un HMM. Leur approche est orientée vers la pratique et permet de prendre en compte différentes considérations : intepprétation des états par l'analyse des paramètres et estimation de la séquence d'états (algorithme de Viterbi), réalisme du modèle, critère statistique de sélection de modèle, objectif de la modélisation... Cette procédure en 6 étapes, bien qu'introduisant une dose de subjectivité, nous semble plus raisonnable que l'utilisation aveugle d'un critère de sélection automatique du nombre d'états non justifié par la théorie.

Par conséquent, **dans cette thèse**, à chaque fois qu'un HMM a été ajusté à des données réelles, les paramètres correspondant à différents nombres d'états (par exemple de 2 à 6 pour un modèle univarié) ont été estimés, et les critères suivant ont guidé notre choix :

— Le critère BIC.
— La possibilité d'interpréter physiquement les états, évaluée par analyse des paramètres estimés (transitions et lois d'émission). Par exemple, l'existence d'états très rares et/ou instables, de lois d'émission proches les unes des autres, voire identiques, d'états qui ne correspondent à aucune situation météorologique identifiable sont des indications d'une surestimation du nombre d'états.
— La capacité du modèle à atteindre l'objectif fixé : produire des séries réalistes de variables météorologiques.
— Le temps nécessaire à l'exécution de l'algorithme EM a tendance à nous faire préférer des modèles plus simples.

Chapitre 3

Convergence de l'Estimateur de Maximum de Vraisemblance pour un modèle de Markov caché périodique

Convergence of the Maximum Likelihood Estimator for a Seasonal Hidden Markov model

This chapter has been published as an article in *Statistics and Computing* (Touron, 2018).

3.1 Introduction

Hidden Markov models (HMM) have been applied to various fields during the last decades : finance (Mamon and Elliott, 2007), ecology (Patterson et al., 2017), climate modelling (Wilks, 1998), speech recognition (Gales and Young, 2008), genomics (Yoon, 2009) and many more. Let X be a finite set and $(\mathsf{Y}, \mathcal{Y})$ a measurable space. A *hidden Markov model* (HMM) with state space X is a $\mathsf{X} \times \mathsf{Y}$-valued stochastic process $(X_t, Y_t)_{t\geq 1}$ where $(X_t)_{t\geq 1}$ is a Markov chain and $(Y_t)_{t\geq 1}$ are Y-valued random variables that are independent conditionnally on $(X_t)_{t\geq 1}$ and such that for all $j \geq 1$, the conditionnal distribution of Y_j given $(X_t)_{t\geq 1}$ only depends on X_j. The law of the Markov chain $(X_t)_{t\geq 1}$ is determined by its initial distribution π and its transition matrix $\mathbf{Q}$. For all $k \in \mathsf{X}$, the distribution of Y_1 given $X_1 = k$ is called the *emission distribution* in state k. The Markov chain $(X_t)_{t\geq 1}$ is called the *hidden* Markov chain because it is not accessible to observation. The process $(Y_t)_{t\geq 1}$ only is observed. See Rabiner and Juang (1986) for an introduction to HMM and Cappé et al. (2009) for a more general formulation. One very common approach to fit such models is to give a parametric form to the emission distributions and to infer the parameters by maximizing the likelihood function. The asymptotic properties of such an estimator have been widely studied. In Baum and Petrie (1966), the authors proved the consistency and the asymptotic normality of the maximum likelihood estimator (MLE) when the emission distributions have a finite support. Since then, these results have been extended to

more general hidden Markov models : see e.g. Douc et al. (2004) and references therein.

Motivation In many applications, especially in climate modeling, simple stationary HMM are not adapted because the data exhibit non-stationarities such as trends and seasonal behaviours. Obviously, temperature has a seasonal component, but this is also the case of precipitations or wind speed for example. It is sometimes possible to preprocess the data in order to obtain a stationary residual. However, finding the right form for the non-stationarity can be very tricky, as well as testing the stationarity of the residual. In the case where the non-stationarity is caused by the presence of a seasonality, a popular solution to avoid this pitfall is to split the period into several sub-periods, and to assume stationarity over each sub-period. For example, in the case of an annual cycle, one may consider each month separately and fit twelve different models. However, this is not entirely satisfactory, for several reasons :

- A choice has to be made for the length of the time blocks.
- The stationarity assumption over each sub-period may not be satisfied.
- We have to fit independently several sub-models, which requires a lot of data.
- The time series used to fit each of the sub-models is obtained by concatenation of data that do not belong to the same year. For exemple, if the time periods are months, the 31st of January of year n will be followed by the first of January of year $n+1$. This is a problem if we use a Markovian model, which exhibits time dependence.
- If the purpose is simulation, it is preferable to be able to simulate a full year using only one model.

Therefore, we prefer to use an extension of HMM that allows seasonality, both in the transition probabilities and in the emission distributions. This will be refered to as a Seasonal Hidden Markov Model (SHMM).

Our contribution In this paper, we will first detail the mathematical framework of a general SHMM. Although this generalization of hidden Markov models is very useful in practice, as far as we know there exists no theoritical result concerning this class of models. The first question that arises is the identifiability of such models. The identifiability of stationary hidden Markov models is not obvious and has been solved only recently. In Gassiat et al. (2016a), it is proved that the transition matrix and the emission distributions of a hidden Markov model are identifiable from the joint distribution of three consecutive observations, provided that the transition matrix is non-singular and that the emission distributions are linearly independent. Alexandrovich et al. (2016) proved that the identifiability can be obtained with the weaker assumption that the emission distributions are distinct. In this paper, we extend the result of Gassiat et al. (2016a) by proving that the SHMM are identifiable up to state labelling, under similar assumptions. To achieve this, we use a spectral method as described in (Hsu et al., 2012). Once we have proved that SHMM are identifiable, it is natural to seek a consistent estimator for their parameters. Regarding HMM, it has been proved by Douc et al. (2011) that under weak assumptions, the maximum likelihood estimator in the framework of HMM is strongly consistent. In this paper, we generalize this result to SHMM. This is done by applying the result of Douc et al. to a well chosen hidden Markov model. We then give several examples of specific models for which our consistency result can be applied. The practical computation of the maximum likelihood estimator for HMM, and a fortiori for SHMM, is not straightforward. We address this problem by adapting the classical EM algorithm to our framework. Then we run this algorithm on simulated data to illustrate the convergence of the MLE to the true parameters. Finally, we successfully fit a SHMM to precipitation data and show that our model is able to reproduce the statistical behaviour of precipitation.

Outline In Section 3.2, we give the general formulation of a SHMM in a parametric framework. Using a spectral method (Hsu et al., 2012) and a general identifiability result for HMM (Gassiat et al., 2016a), we show that under weak assumptions, SHMM are identifiable up to state labelling. Then, we generalize the existing results on the convergence of the MLE in HMM to prove our main result, that is the strong consistency of the MLE in SHMM. We also give two examples of models (using mixtures of exponential distributions and mixtures of Gaussian distributions) for which this convergence theorem is applicable. In Section 3.3, we describe the EM algorithm used for the numerical computation of the maximum likelihood estimator and we illustrate the convergence of the MLE for a simple SHMM using simulated data. Finally, in Section 3.4, we fit a SHMM to precipitation data and we show that such a model can be used to simulate realistic times series of weather variables.

3.2 Consistency result

3.2.1 Model description

For positive integers p and q, we shall denote by $\mathbb{R}^{p\times q}$ the set of matrices with real entries, p rows and q columns. Let K be a positive integer and $(X_t)_{t\geq 1}$ a non-homogeneous Markov chain whose state space is $\mathsf{X} = \{1,\dots,K\}$ and initial distribution π, that is $\pi_k = \mathbb{P}(X_1 = k)$. For $t \geq 1$ and $1 \leq i,j \leq K$, let

$$Q_{ij}(t) := \mathbb{P}(X_{t+1} = j \mid X_t = i).$$

$Q(t) \in \mathbb{R}^{K\times K}$ is the transition matrix from X_t to X_{t+1}. Let us assume that $Q(\cdot)$ is a T-periodic function, so that there exists some integer $T \geq 1$ such that for all $t \geq 1$, $Q(t+T) = Q(t)$. Let Y be a Polish space, $\mathcal{Y}$ its Borel σ-algebra, and $(Y_t)_{t\geq 1}$ a Y-valued stochastic process. Assume that conditionally to $(X_t)_{t\geq 1}$, the $(Y_t)_{t\geq 1}$ are independent and that the distribution of Y_s conditionally to the $(X_t)_{t\geq 1}$ only depends on X_s and s. We shall denote by $\nu_{k,t}$ the distribution of Y_t given $X_t = k$. We also assume that for all $t \geq 1$ and for all $k \in \mathsf{X}$, $\nu_{k,t+T} = \nu_{k,t}$. Then the process $(X_t, Y_t)_{t\geq 1}$ is called a *seasonal hidden Markov model* (SHMM) and the $\nu_{k,t}$ are its *emission distributions*.

The law of the process $(X_t, Y_t)_{t\geq 1}$ is determined by the distribution π of X_1, the transition matrices $Q(1), \dots, Q(T)$ and the emission distributions $\nu_{k,1}, \dots, \nu_{k,T}$ for $1 \leq k \leq K$.

Choosing $T = 1$, we retrieve the classical hidden Markov model (HMM). A general SHMM includes periodicity in both the transition probabilitites and the emission distributions. However, for some applications, it may be enough to consider periodic transitions and constant emission distributions, or vice versa (see e.g. Section 3.4).

The following remark will be the key to the proof of our main result. Given $(X_t, Y_t)_{t\geq 1}$ a seasonal hidden Markov model, two (classical) hidden Markov models naturally arise.

— For any $t \in \{1,\dots,T\}$ the process

$$(X_{jT+t}, Y_{jT+t})_{j\geq 0} \in (\mathsf{X}\times\mathsf{Y})^{\mathbb{N}}$$

is a hidden Markov model with transition matrix $Q(t)Q(t+1)\dots Q(T)Q(1)\dots Q(t-1)$ and emission distributions $(\nu_{k,t})_{k\in\mathsf{X}}$.

— For $j \geq 0$, let $U_j := (X_{jT+1}, \dots, X_{jT+T})$ and $W_j := (Y_{jT+1}, \dots, Y_{jT+T})$. Then

$$(U_j, W_j)_{j\geq 0} \in \left(\mathsf{X}^T \times \mathsf{Y}^T\right)^{\mathbb{N}}$$

is a hidden Markov model. The conditional independence property implies that for $u = (u_1, \ldots, u_T) \in \mathsf{X}^T$, the conditional distribution of W_0 given $U_0 = u$ is $\bigotimes_{t=1}^{T} \nu_{u_t,t}$. The transition matrix of the homogeneous Markov chain $(U_j)_{j\geq 0}$ is given by

$$\tilde{Q}_{uv} := \mathbb{P}(U_1 = v \mid U_0 = u) = Q_{u_T v_1}(T) Q_{v_1 v_2}(1) \ldots Q_{v_{T-1} v_T}(T-1),$$

with

$$u = (u_1, \ldots, u_T),\ v = (v_1, \ldots, v_T) \in \mathsf{X}^T.$$

Parametric framework Assume that there exists a measure μ defined on $\mathcal{Y}$ such that all the emission distributions are absolutely continuous with respect to μ and let $f_{k,t} := \frac{d\nu_{k,t}}{d\mu}$ be the *emission densities*. We consider that all the emission distributions and the transition matrices depend on a parameter $\theta \in \Theta$, where Θ is a compact subset of some finite-dimensional vector space, e.g. $\mathbb{R}^q$. Let us precise the structure of Θ.

— The function $t \mapsto Q(t)$ belongs to a (known) parametric family of T-periodic functions indexed by a parameter β that we wish to estimate.

Example

$$Q_{ij}(t) \propto \exp\left(P_{ij}(t)\right) \tag{3.1}$$

where

$$P_{ij}(t) = \sum_{l=0}^{d} \left(a_{ijl} \cos\left(\frac{2\pi l t}{T}\right) + b_{ijl} \sin\left(\frac{2\pi l t}{T}\right) \right)$$

is a trigonometric polynomial. In this example, $\beta = (a_{ijl}, b_{ijl})_{1\leq i,j\leq K, 0\leq l\leq d}$ and all the transition matrices are entirely determined by β.

— For any $t \geq 1$, the emission densities $f_{k,t}$ belong to a (known) parametric family (which does not depend on t) indexed by a parameter $\theta^Y(t)$. In addition, we assume that the T-periodic function $t \mapsto \theta^Y(t)$ itself belongs to a parametric family indexed by a parameter δ.

Example Denoting by $\mathcal{E}(\alpha)$ the exponential distribution with parameter α,

$$\nu_{k,t} = \mathcal{E}\left[\delta_k \left(1 + \cos\left(\frac{2\pi t}{T}\right)\right)\right]$$

In such a case, $\theta^Y(t) = \left(\delta_k \left(1 + \cos\left(\frac{2\pi t}{T}\right)\right)\right)_{k\in\mathsf{X}}$ and $\delta = (\delta_1, \ldots, \delta_K) \in \mathbb{R}^K$.

— Hence we can define $\theta = (\beta, \delta)$.

We shall denote by $\mathbb{P}^{\pi,\theta}$ the law of the process $(Y_t)_{t\geq 1}$ when the parameter is θ and the distribution of X_1 is π, and by $\mathbb{E}^{\pi,\theta}(\cdot)$ the corresponding expected value. If the initial distribution π is the stationary distribution associated with the transition matrix $Q(1) \cdots Q(T)$, then the two HMM described above are stationary. In such a case, we will simply write $\mathbb{P}^\theta$ and $\mathbb{E}^\theta(\cdot)$ for the law of $(Y_t)_{t\geq 1}$ and the corresponding expected value. Our purpose is to infer θ from a vector of observations $(Y_1, \ldots, Y_n) \in \mathsf{Y}^n$. We assume that the model is well specified, which means that there exists a *true* initial distribution π^* and a *true* parameter $\theta^* = (\beta^*, \delta^*)$ in the interior of Θ such that the observed vector $(Y_1, \ldots, Y_n)$ is generated by the SHMM defined by π^* and θ^*. We denote by $Q^*(t)$ and $\nu^*_{k,t}$ the corresponding transition matrices and emission distributions. Note that we consider the number K of hidden states to be known.

3.2.2 Identifiability

In Gassiat et al. (2016a), the authors prove that the transition matrix and the emission distributions of a stationary HMM are identifiable (up to state labelling) from the law of three consecutive observations, provided that the transition matrix has full rank and that the emission distributions are linearly independent. Still in the context of HMM, the authors of Alexandrovich et al. (2016) use the weaker assumption that the emission distributions are all distinct to obtain identifiability up to state labelling. However, it requires to consider the law of more than three consecutive observations. In this paragraph, we show that under similar assumptions, the SHMM described above is identifiable : we can retrieve the transition matrices and the emission distributions from the law of the process $(Y_t)_{t\geq 1}$ up to permutations of the states. We will use the following assumptions :

(A1). *For $1 \leq t \leq T$, the transition matrix $Q^*(t)$ is invertible.*

(A2). *The matrix $Q^*(1)\cdots Q^*(T)$ is irreducible and its unique stationary distribution π^* is the distribution of X_1.*

(A3). *For $1 \leq t \leq T$, the K emission distributions $(\nu^*_{k,t})_{k\in\mathsf{X}}$ are linearly independent.*

Remark These assumptions only involve the data generating process and its corresponding parameter θ^*, they are not about the whole parameter space Θ.

Theorem 3.1. *Assume that the set of true parameters $(\pi^*, Q^*(t), \nu^*_{k,t})_{1\leq t\leq T, k\in\mathsf{X}}$ satisfies Assumptions (A1)-(A3). Let $(\tilde{\pi}, \tilde{Q}(t), \tilde{\nu}_{k,t})_{1\leq t\leq T, k\in\mathsf{X}}$ be another set of parameters. Let us assume that the joint distribution of $(Y_1, \ldots, Y_{T+2})$ is the same under both sets of parameters. Then there exist $\sigma_1, \ldots, \sigma_T$ permutations of X such that for all $k \in \mathsf{X}$, $\tilde{\pi}_k = \pi_{\sigma_1(k)}$ and for all $t \in \{1, \ldots, T\}$, $k, l \in \mathsf{X}$, $\tilde{\nu}_{k,t} = \nu_{\sigma_t(k),t}$ and $\tilde{Q}(t)_{kl} = Q(t)_{\sigma_t(k),\sigma_{t+1}(l)}$, with $\sigma_{T+1} = \sigma_1$.*

Proof. We shall follow the spectral method as presented in Hsu et al. (2012) (see also De Castro et al. (2016) and De Castro et al. (2017)). The proof is based on a method of moments : using simple linear algebra techniques such as singular value decomposition, we show that we can retrieve the transitions matrices as well as the emission distributions from matrices that are computable if the distribution of $(Y_1, \ldots, Y_{T+1})$ is known. Before going through the spectral algorithm, let us first introduce some notations. Let $(\phi_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions such that for any probability measures ν_1, ν_2 on $(\mathsf{Y}, \mathcal{Y})$,

$$\left(\int_{\mathsf{Y}} \phi_n d\nu_1\right)_{n\in\mathbb{N}} = \left(\int_{\mathsf{Y}} \phi_n d\nu_2\right)_{n\in\mathbb{N}} \implies \nu_1 = \nu_2.$$

Such a sequence exists because Y is a Polish space. Let $t \geq 2$ be a time index and $N \geq 1$. In the following definitions, $k \in \{1, \ldots, K\}$ is a state index and $a, b, c \in \{1, \ldots, N\}$. We shall consider the following matrices :

Let $O_t = O_t^{(N)} \in \mathbb{R}^{N\times K}$ be the matrix defined by

$$(O_t)_{ak} = \mathbb{E}\left[\phi_a(Y_t) \mid X_t = k\right] = \int_Y \phi_a d\nu^*_{k,t}.$$

Let $L(t) \in \mathbb{R}^N$ be the vector such that, for

$$1 \leq a \leq N,\ L_a(t) = \mathbb{E}\left[\phi_a(Y_t)\right].$$

Let $N(t) \in \mathbb{R}^{N\times N}$ be the matrix defined by

$$N_{ab}(t) = \mathbb{E}\left[\phi_a(Y_t)\phi_b(Y_{t+1})\right].$$

Let $P(t) \in \mathbb{R}^{N\times N}$ be the matrix defined by

$$P_{ac}(t) = \mathbb{E}\left[\phi_a(Y_{t-1})\phi_c(Y_{t+1})\right].$$

For $b \in \{1,\dots,N\}$, let $M_t(\cdot,b,\cdot) \in \mathbb{R}^{N\times N}$ be the matrix defined by

$$M_t(a,b,c) = \mathbb{E}[\phi_a(Y_{t-1})\phi_b(Y_t)\phi_c(Y_{t+1})].$$

Notice that all these quantities can be computed from the law of $(Y_t)_{t\geq 1}$, except O_t which requires the emission distributions. Although they all depend on N, we do not indicate it, for the sake of readability. The aim of the proof is to show that the matrices O_t and $Q(t)$ can be obtained from the matrices $L(t)$, $M_t(\cdot,b,\cdot)$, $N(t)$ and $P(t)$, using linear algebra arguments. Using Assumption (A3), we see that there exists an integer $N_0 > K$ such that for all $N \geq N_0$, the matrices $O_t^{(N)}$ have full rank. From now on, we will consider that $N \geq N_0$. We denote by $\pi^*(t)$ the (unconditional) distribution of X_t.

The following equations, obtained through elementary calculations, are relations between the known quantities $L(t)$, $N(t)$, $P(t)$ and $(M_t(\cdot,b,\cdot))_{1\leq b\leq N}$ and the quantities we want to identify : $Q^*(t)$ and O_t.

$$L(t) = O_t\pi^*(t) \tag{3.2}$$
$$N(t) = O_t\mathrm{diag}(\pi^*(t))Q^*(t)O_{t+1}^\top \tag{3.3}$$
$$P(t) = O_{t-1}\mathrm{diag}(\pi^*(t-1))Q^*(t-1)Q^*(t)O_{t+1}^\top \tag{3.4}$$

For all $1 \leq b \leq N$,

$$M_t(\cdot,b,\cdot) = O_{t-1}\mathrm{diag}(\pi^*(t-1))Q^*(t-1)\mathrm{diag}[O_t(b,\cdot)]Q^*(t)O_{t+1}^\top. \tag{3.5}$$

Here $\mathrm{diag}(v)$ is the diagonal matrix whose diagonal entries are those of the vector v.

Now let us define matrices, denoted by $B(b)$ hereafter, whose eigenvalues will be proved to be the entries of O_t. Note that thanks to Assumption (A2), all the entries of $\pi^*(t)$ are positive, so that the matrix $\mathrm{diag}(\pi^*(t))$ is invertible. In addition, Assumption (A1) and equations (3.3) and (3.4) show that the matrices $P(t)$ and $N(t)$ also have rank K. Let $P(t) = U\Sigma V^\top$ be a singular value decomposition (SVD) of $P(t)$: U and V are matrices of size $N \times K$ whose columns are orthonormal families being the left (resp. right) singular vectors of $P(t)$ associated with its K non-zero singular values, and $\Sigma = U^\top P(t)V$ is an invertible diagonal matrix of size K containing these singular values. Note that such a decomposition is not unique, as we may choose arbitrarily the order of the diagonal entries of Σ, which is equivalent to swapping the columns of U and V, using the same permutation of $\{1,\dots,K\}$. Let us define, for $1 \leq b \leq N$, the $K \times K$ matrix

$$B(b) := (U^\top P(t)V)^{-1}U^\top M_t(\cdot,b,\cdot)V.$$

For $1 \leq b \leq N$, the matrix $B(b)$ is computable from the distribution of $(Y_1,\dots,Y_{T+1})$. Moreover, using (3.4) and (3.5), we see that :

$$B(b) = \left(Q^*(t)O_{t+1}^\top V\right)^{-1}\mathrm{diag}[O_t(b,\cdot)]\left(Q^*(t)O_{t+1}^\top V\right),$$

so that there exists an invertible $K \times K$ matrix

$$R := \left(Q^*(t)O_{t+1}^\top V\right)^{-1}$$

such that for all $b \in \{1, \dots, N\}$,

$$\text{diag}[O_t(b, \cdot)] = R^{-1}B(b)R.$$

Besides, as O_t has rank K, there exist real numbers $\alpha_1, \dots, \alpha_N$ such that the eigenvalues of

$$B := \sum_{b=1}^{N} \alpha_b B(b)$$

are distinct. Hence the eigenvalue decomposition of B is unique up to permutation and scaling. As

$$R^{-1}BR = \sum_{b=1}^{B} \alpha_b \text{diag}[O_t(b, \cdot)],$$

the $K \times K$ matrix R is obtained (up to permutation and scaling) by computing the eigenvectors of B. Then we can deduce $O_t(b, \cdot) = R^{-1}B(b)R$ for all $b \in \{1, \dots, N\}$, up to a common permutation that we denote by σ_t. It follows that for all $t \geq 2$, the matrix O_t is computable from M_t, $N(t)$ and $P(t)$, up to permutation of its columns, corresponding to the states.

Then, since O_t has full rank, we obtain $\pi^*(t)$ from O_t and $L(t)$ thanks to equation (3.2). Again, $\pi^*(t)$ is only determined up to permutation of its entries : if R is replaced by RP_{σ_t} where P_{σ_t} is the matrix of the permutation σ_t, we get $P_{\sigma_t}^\top \pi^*(t)$ instead of $\pi^*(t)$.

We finally obtain the transition matrix :

$$\left(\tilde{U}^\top O_t \text{diag}(\pi^*(t))\right)^{-1} \tilde{U}^\top N(t) V \left(O_{t+1}^\top V\right)^{-1} = Q^*(t),$$

where $\tilde{U}$ is the matrix whose columns are the left singular vectors of $N(t)$. Replacing O_t by $O_t P_{\sigma_t}$, $\pi^*(t)$ by $P_{\sigma_t}^\top \pi^*(t)$ and O_{t+1} by $O_{t+1}P_{\sigma_{t+1}}$ in the last equation, we obtain $P_{\sigma_t}^\top Q^*(t) P_{\sigma_{t+1}}$ instead of $Q^*(t)$, which means that $Q^*(t)$ is only determined up to permutations of its lines and columns, those permutations being possibly different. Therefore, we have proved that if the law of $(Y_t)_{t\geq 1}$ is the same under both sets of parameters

$$(\pi^*, Q^*(t), \nu^*_{k,t})_{1\leq t\leq T, k\in \mathsf{X}} \text{ and } (\tilde{\pi}, \tilde{Q}(t), \tilde{\nu}_{k,t})_{1\leq t\leq T, k\in \mathsf{X}}$$

with the first one satisfying Assumptions (A1)-(A3), then there exists $\sigma_1^{(N)}, \dots, \sigma_T^{(N)}, \sigma_{T+1}^{(N)}$, permutations of $\{1, \dots, K\}$ such that for all $k \in \mathsf{X}$, $\tilde{\pi}_k = \pi^*_{\sigma_1^{(N)}(k)}$ and for all $t \in \{1, \dots, T\}$,

$$\tilde{Q}(t)_{kl} = Q^*(t)_{\sigma_t^{(N)}(k), \sigma_{t+1}^{(N)}(l)}$$

$$\tilde{O}_t^{(N)}(\cdot, k) = O_t^{(N)}\left(\cdot, \sigma_t^{(N)}(k)\right), \quad k, l \in \mathsf{X}$$

where $\tilde{O}_t$ is the analog of O_t with respect to the distributions $\tilde{\nu}_{k,t}$. The relationship between $\tilde{Q}(t)$ and $Q^*(t)$ shows that for all $N \geq N_0$ and for all $t \in \{1, \dots, T\}$, $\sigma_t^{(N)} = \sigma_t^{(N_0)}$. Hence we denote by σ_t this permutation. Thus, for all $N \geq N_0$, $\tilde{O}_t^{(N)} = O_t^{(N)} P_{\sigma_t}$. By definition of the sequence $(\phi_n)_{n\in\mathbb{N}}$, this implies that for all $t \in \{1, \dots, T\}$ and for all $k \in \mathsf{X}$, $\tilde{\nu}_{k,t} = \nu^*_{\sigma_t(k),t}$ and the theorem is proved. □ □

Remarks

— In Theorem 3.1, we need not assume that the set of parameters $(\tilde{\pi}, \tilde{Q}(t), \tilde{\nu}_{k,t})_{1\leq t\leq T, k\in \mathsf{X}}$ satisfies Assumptions (A1)-(A3) because it has to be the case. Indeed, as the two sets of parameters induce the same distribution of $(Y_1, \dots, Y_{T+1})$, we have, using the notations of the above proof,

$$\begin{aligned} N(t) &= O_t \mathrm{diag}(\pi(t)) Q(t) O_{t+1}^\top \\ &= \tilde{O}_t \mathrm{diag}(\tilde{\pi}(t)) \tilde{Q}(t) \tilde{O}_{t+1}^\top \end{aligned}$$

Therefore, if the second set of parameters does not satisfy Assumptions (A1)-(A3), there exists some t such that $\tilde{O}_t \mathrm{diag}(\tilde{\pi}(t)) \tilde{Q}(t) \tilde{O}_{t+1}^\top$ has not full rank. This is a contradiction since $O_t \mathrm{diag}(\pi(t)) Q(t) O_{t+1}^\top$ has full rank.

— We proved that we could identify the emission distributions and the transition matrices up to permutations, these permutations depending on the time step. However it is not possible to prove that there exist a single permutation of the states that is common to all the times steps. Indeed, if we choose another labelling at time t, the matrix O_t is replaced by $O_t P_\sigma$ and the matrix $Q(t)$ is replaced by $Q(t)P_\sigma$, with P_σ a permutation matrix. Then we see, using (3.5), that the matrices $M(\cdot, b, \cdot)$ remain unchanged, which means that the permutation σ cannot be identified from the distribution of (Y_{t-1}, Y_t, Y_{t+1}).

— The spectral method presented above provides a non-parametric moment estimator for the parameters of a HMM. The properties of such an estimator are studied in De Castro et al. (2017).

— The identifiability of the parameters β and δ depends on the parametric form chosen for $t \mapsto Q(t)$ and $t \mapsto \theta^Y(t)$. Hence this should be studied separately for each particular version of the SHMM.

3.2.3 Consistency

In this paragraph, we prove our main result, that is the strong consistency of the maximum likelihood estimator for SHMM. Assume that we have observed $(Y_1, \dots, Y_n)$ (recall that $X_1, \dots, X_n$ are not observed). For a probability distribution π on X and $\theta \in \Theta$, let $L_{n,\pi}[\theta; (Y_1, \dots, Y_n)]$ be the likelihood function when the parameter is θ and the distribution of X_1 is π. We define the maximum likelihood estimator by

$$\hat{\theta}_{n,\pi} := \arg\max_{\theta\in\Theta} L_{n,\pi}[\theta; (Y_1, \dots, Y_n)].$$

We will need the following assumptions :

(A4). *The parameter β can be identified from the transition matrices $Q(1), \dots, Q(T)$ and the parameter δ can be identified from the emission distributions $\nu_{k,t}$.*

(A5).

$$\alpha := \inf_{\theta\in\Theta} \inf_{1\leq t\leq T} \inf_{i,j\in\mathsf{X}} Q_{ij}(t) > 0$$

(A6). *The transition probabilities (resp. the emission densities) are continuous functions of β (resp. δ).*

(A7). *For all $y \in \mathsf{Y}$, $k \in \mathsf{X}$ and $t \in \{1, \dots, T\}$,*

$$\inf_{\theta\in\Theta} f_{k,t}^\theta(y) > 0, \quad \sup_{\theta\in\Theta} f_{k,t}^\theta(y) < \infty$$

(A8). *For $t \in \{1, \dots, T\}$ and $y \in \mathsf{Y}$, we define*

$$c_t(y) := \inf_{\theta \in \Theta} \sum_{k \in \mathsf{X}} f^{\theta}_{k,t}(y), \quad d_t(y) := \sup_{\theta \in \Theta} \sum_{k \in \mathsf{X}} f^{\theta}_{k,t}(y),$$

and we assume that

$$\mathbb{E}_{\nu^{\theta^*}_{k,t}} [-\log c_t(Y)] < \infty, \quad \mathbb{E}_{\nu^{\theta^*}_{k,t}} [\log d_t(Y)] < \infty$$

Let $\mathfrak{S}_K$ be the set of permutations of $\{1, \dots, K\}$. For $\sigma = (\sigma_1, \dots, \sigma_T) \in (\mathfrak{S}_K)^T$ and $\theta \in \Theta$, let us denote by $\sigma(\theta)$ the parameter obtained from θ by swapping the states according to the permutations $\sigma_1, \dots, \sigma_T$. More precisely, we compute the transition matrices and the emission distributions corresponding to the parameter θ, we swap them using the permutations $\sigma_1, \dots, \sigma_T$, and using (A4), we identify the parameter corresponding to the swapped matrices and emission distributions. This parameter is denoted by $\sigma(\theta)$. It follows from Theorem 3.1 that under Assumptions (A1) to (A4),

$$\Theta^* := \{\theta \in \Theta : \mathbb{P}^{\theta} = \mathbb{P}^{\theta^*}\} \subset \{\sigma(\theta^*) : \sigma \in (\mathfrak{S}_K)^T\}.$$

Note that due to the parametric form of the transition matrices and the emission distributions, the set $\{\sigma \in (\mathfrak{S}_K)^T : \mathbb{P}^{\theta^*} = \mathbb{P}^{\sigma(\theta^*)}\}$ may actually be much smaller than $(\mathfrak{S}_K)^T$. However, it contains at least $\{\sigma \in (\mathfrak{S}_K)^T : \sigma_1 = \cdots = \sigma_T\}$.

Theorem 3.2. *Under Assumptions (A1) to (A8), for any initial distribution π and $\mathbb{P}^{\theta^*}$-a.s., there exists $\left(\sigma^{(n)}\right)_{n \in \mathbb{N}}$ a $(\mathfrak{S}_K)^T$-valued sequence such that*

$$\lim_{n \to \infty} \sigma^{(n)} \left(\hat{\theta}_{n,\pi}\right) = \theta^*.$$

Proof. Let us consider $(U_j, W_j)_{j \geq 0}$ the HMM defined in paragraph 3.2.1. Recall that its transition matrix is given by

$$\tilde{Q}_{uv} = Q_{u_T v_1}(T) Q_{v_1 v_2}(1) \dots Q_{v_{T-1} v_T}(T-1),$$

for $u = (u_1, \dots, u_T) \in \mathsf{X}^T$ and $v = (v_1, \dots, v_T) \in \mathsf{X}^T$. When the parameter is θ, its emission densities are

$$g^{\theta}(w \mid u) = \prod_{t=1}^{T} f^{\theta^Y}_{u_t,t}(w_t),$$

where $w = (w_1, \dots, w_T) \in \mathsf{Y}^T$. Thus the law of the process $(U_j, W_j)_{j \geq 0}$ is entirely determined by θ and π and it is stationary under Assumption (A2). Denoting by $\mathbb{Q}^{\pi,\theta}$ the law of the process $(W_j)_{j \geq 0}$ when the parameter is θ and the distribution of X_1 is π, we notice that for any $\theta_1, \theta_2 \in \Theta$,

$$\mathbb{Q}^{\pi,\theta_1} = \mathbb{Q}^{\pi,\theta_2} \implies \mathbb{P}^{\pi,\theta_1} = \mathbb{P}^{\pi,\theta_2}.$$

Therefore, using Theorem 3.1 and Assumption (A4), we have

$$\mathbb{Q}^{\theta} = \mathbb{Q}^{\theta^*} \implies \exists\, \sigma \in (\mathfrak{S}_K)^T,\ \theta = \sigma(\theta^*). \tag{3.6}$$

We notice that for all $\theta \in \Theta$, $J \geq 0$ and initial distribution π, we have :

$$\tilde{L}_{J,\pi}[\theta; (W_0, \dots, W_J)] = L_{(J+1)T,\pi}[\theta; (Y_1, \dots, Y_{(J+1)T})],$$

where $\tilde{L}_{J,\pi}$ is the likelihood function corresponding to the model $(U_j, W_j)_{j\geq 0}$ when the distribution of X_1 is π. Let $\tilde{\theta}_{J,\pi}$ be a maximizer of $\tilde{L}_{J,\pi}$. If we are able to prove the strong consistency of $\tilde{\theta}_{J,\pi}$, by the same arguments, for all $s \in \{0, \dots, T-1\}$, we can prove that the estimator

$$\tilde{\theta}^s_{J,\pi} := \arg\max_{\theta\in\Theta} L_{(J+1)T+s,\pi}[\theta; Y_1, \dots, Y_{(T+1)J+s}]$$

is strongly consistent. From there we easily deduce that $\hat{\theta}_{n,\pi}$ is strongly consistent. Therefore it is sufficient to prove the strong consistency of the maximum likelihood estimator for the HMM $(U_j, W_j)_{j\geq 0}$. To this end, we shall use Theorem 3.3, stated and proved in Section 3.A, which is an adaptation of Theorem 13.14 in Douc et al. (2014). The following properties must hold in order to apply this theorem to the HMM $(U_j, W_j)_{j\geq 0}$:

1. $$\tilde{\alpha} := \inf_{\theta\in\Theta} \inf_{u,v\in \mathsf{X}^T} \tilde{Q}^\theta_{uv} > 0$$
2. For any $u, v \in \mathsf{X}^T$ and $w \in \mathsf{Y}^T$, the functions $\theta \mapsto \tilde{Q}^\theta_{uv}$ and $\theta \mapsto g^\theta(w \mid u)$ are continuous.
3. For all $w \in \mathsf{Y}^T$,
$$b_-(w) := \inf_{\theta\in\Theta} \sum_{u\in\mathsf{X}^T} g^\theta(w \mid u) > 0$$
$$b_+(w) := \sup_{\theta\in\Theta} \sum_{u\in\mathsf{X}^T} g^\theta(w \mid u) < \infty.$$
4. $$\mathbb{E}^{\theta^*}\left[|\log b_+(W_0)|\right] < \infty, \quad \mathbb{E}^{\theta^*}\left[|\log b_-(W_0)|\right] < \infty$$

The first three properties are straightforward consequences of Assumptions (A5) to (A7). Let us prove that $\mathbb{E}^{\theta^*}\left[|\log b_+(W_0)|\right] < \infty$. The proof that

$$\mathbb{E}^{\theta^*}\left[|\log b_-(W_0)|\right] < \infty$$

follows the same lines. We have :

$$\mathbb{E}^{\theta^*}[|\log b_+(W_0)|] = \sum_{u\in\mathsf{X}^T} \tilde{\pi}^{\theta^*}(u) \int g^{\theta^*}(w \mid u)|\log b_+(w)|\mu^{\otimes T}(dw),$$

where $\tilde{\pi}^{\theta^*}$ is the stationary distribution associated with $\tilde{Q}^{\theta^*}$. Hence it is enough to prove that for all $u \in \mathsf{X}^T$,

$$\int g^{\theta^*}(w \mid u)|\log b_+(w)|\mu^{\otimes T}(dw) < \infty$$

We have :

$$\begin{aligned}
&\int g^{\theta^*}(w \mid u)|\log b_+(w)|\mu^{\otimes T}(dw) \\
&= \int_{b_+>1} g^{\theta^*}(w \mid u)\log b_+(w)\mu^{\otimes T}(dw) \\
&+ \int_{b_+<1} g^{\theta^*}(w \mid u)(-\log b_+(w))\mu^{\otimes T}(dw)
\end{aligned}$$

Using Assumption (A7), we get $\inf_{w\in \mathsf{Y}^T} b_+(w) > 0$. Therefore, as $g^{\theta^*}(\cdot \mid u)$ is a probability density function, the second term is finite. In order to show that the first one is finite, it is enough to find a function C such that for all $w \in \{b_+ > 1\}$, $C(w) \geq b_+(w)$ and

$$\int_{b_+>1} g^{\theta^*}(w \mid u) \log C(w) \mu^{\otimes T}(dw) < \infty.$$

Let $C(w) = \prod_{t=1}^{T} d_t(w_t)$. For all $w \in \mathcal{Y}^T$, we get :

$$\begin{aligned}
C(w) &= \prod_{t=1}^{T} \sup_{\theta\in\Theta} \sum_{u_t=1}^{K} f_{u_t,t}^{\theta^Y}(w_t) \\
&\geq \sup_{\theta\in\Theta} \prod_{t=1}^{T} \sum_{u_t=1}^{K} f_{u_t,t}^{\theta^Y}(w_t) \\
&= \sup_{\theta\in\Theta} \sum_{u\in \mathsf{X}^T} \prod_{t=1}^{T} f_{u_t,t}^{\theta^Y}(w_t) \\
&= \sup_{\theta\in\Theta} \sum_{u\in \mathsf{X}^T} g^{\theta}(w \mid u) \\
&= b_+(w).
\end{aligned}$$

In addition :

$$\begin{aligned}
&\int_{b_+>1} g^{\theta^*}(w \mid u) \log C(w) \mu^{\otimes T}(dw) \\
&\leq \int_{\mathsf{Y}^T} g^{\theta^*}(w \mid u) \left(\sum_{t=1}^{T} \log d_t(w_t) \right) \mu^{\otimes T}(dw) \\
&= \sum_{t=1}^{T} \int_{\mathsf{Y}^T} \left(\prod_{s=1}^{T} f_{u_s,s}^{\theta_Y^*}(w_s) \right) \log d_t(w_t) \mu^{\otimes T}(dw) \\
&= \sum_{t=1}^{T} \int_{\mathsf{Y}} f_{u_t,t}^{\theta_Y^*}(w_t) \log d_t(w_t) \mu(dw_t) \quad \text{(Fubini)} \\
&= \sum_{t=1}^{T} \mathbb{E}_{\nu_{u_t,t}^{\theta^*}} [\log d_t(Y)] \overset{(A8)}{<} +\infty.
\end{aligned}$$

Hence $\mathbb{E}^{\theta^*}[|\log b_+(W_0)|] < \infty$. Thus we can apply Theorem 3.3 to get that for any initial distribution π, $\mathbb{P}^{\theta^*}$-a.s.,

$$\lim_{n\to\infty} d\left(\hat{\theta}_{n,\pi}, \left\{\theta \in \Theta : \mathbb{Q}^{\theta} = \mathbb{Q}^{\theta^*}\right\}\right) = 0,$$

where d is a distance on Θ. Combining this theorem with (3.6), we obtain the strong consistency of $\tilde{\theta}_{J,\pi}$ and then Theorem 3.2 is proved. □ □

Remark The strong consistency of the MLE does not depend on the choice of the initial distribution in the computation of the likelihood function. This result relies on the geometric *forgetting* rate of the initial distribution by the HMM. This is where (A5) is crucial. This forgetting property is proved in details in Douc et al. (2014).

3.2.4 Applications

In this section, we introduce two examples of SHMM and we show that under weak assumptions, Theorem 3.1 and Theorem 3.2 can be applied. To this end, we shall show that the assumptions of these two theorems are satisfied. The proofs will only be detailed for the first example, as they are very similar for the second one. Theorem 3.1 shows that as far as identifiability is concerned, the only necessary condition on emission distributions is their linear independence, so that a large variety of choices is possible. The first example deals with emission distributions which are mixtures of exponential distributions, whereas in the second one, emission distributions are mixtures of gaussian distributions. Mixtures are interesting examples because their flexibility allows them to approximate complex distributions (see e.g. Kruijer et al. (2010)). Also, these particular mixtures are useful in our application field. In Section 3.4, we use mixtures of exponential distributions as emission distributions to model daily precipitations. Mixtures of gaussian distributions can be used to model temperature.

Mixtures of exponential distributions

Let $(X_t, Y_t)_{t\geq 1}$ be a SHMM whose transitions are given by equation (3.1), and whose emission distributions are mixtures of exponential distributions. Denoting by $\mathcal{E}$ the exponential distribution, the emission distribution in state k and time t is

$$\nu_{k,t} = \sum_{m=1}^{M} p_{km} \mathcal{E}\left(\frac{\lambda_{km}}{1+\sigma_k(t)}\right),$$

where $(p_{k1}, \dots, p_{kM})$ is a vector of probability, the λ_{km} are positive and $\sigma_k(t)$ is a trigonometric polynomial whose constant term is zero and whose degree d is known. This polynomial can be interpreted as a periodic scaling factor. Let δ_k be the vector of coefficients of σ_k. Note that the emission densities are given by

$$f_{k,t}(y) = \sum_{m=1}^{M} p_{km} \frac{\lambda_{km}}{1+\sigma_k(t)} \exp\left(-\frac{\lambda_{km}}{1+\sigma_k(t)} y\right) \mathbb{1}_{y>0}. \tag{3.7}$$

The vector of parameters of this model is $\theta = (\beta, \mathbf{p}, \Lambda, \delta)$ where

$$\beta = (\beta_{ijl})_{i,j,l} \in \mathbb{R}^{K\times(K-1)\times(2d+1)}$$

is the parameter of the transition probabilities : $Q_{ij}(t)$ is proportional to

$$\exp\left[\beta_{ij1} + \sum_{l=1}^{d} \beta_{ij,2l} \cos\left(\frac{2\pi}{T} lt\right) + \beta_{ij,2l+1} \sin\left(\frac{2\pi}{T} lt\right)\right]$$

and

$$\mathbf{p} = (p_{km})_{k,m} \in [0,1]^{K\times(M-1)}$$

is the set of weights of the mixtures,

$$\Lambda = (\lambda_{km})_{k,m} \in (0,+\infty)^{K\times M}$$

is the set of parameters of the exponential distributions, and

$$\delta = (\delta_{kl})_{k,l} \in \mathbb{R}^{2d+1}$$

is the vector of coefficients of the trigonometric polynomials $(\sigma_k)_{k\in\mathsf{X}}$. We shall make the following assumptions about the *true* parameter θ^*. They ensure that the mixtures have exactly M components.

(A9). *For all* $k \in \mathsf{X}$, $\lambda^*_{k1} < \cdots < \lambda^*_{kM}$.

(A10). *For all* $k \in \mathsf{X}$ *and* $m \in \{1, \ldots, M\}$, $p^*_{km} > 0$.

Let us first show that the transition matrices and the emission distributions can be identified. Following the result of paragraph 3.2.2, it suffices to show that Assumptions (A1)-(A3) are satisfied. Clearly, for any time t, the transition matrix $Q^*(t)$ is irreducible as all its entries are positive. Let us show that for almost every $\beta \in \mathbb{R}^{K\times(K-1)\times(2d+1)}$, $Q^*(t)$ is invertible for all $t \in \{1, \ldots, T\}$. To this aim, we will need the following lemma regarding analytic functions. Recall that $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is said to be analytic if for all $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, there exists V a neighbourhood of a such that for all $x = (x_1, \ldots, x_n) \in V$,

$$f(x) = \sum_{\alpha\in\mathbb{N}^n} \frac{\partial^{\alpha_1+\cdots+\alpha_n} f}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}}(a) \frac{(x_1 - a_1)^{\alpha_1} \ldots (x_n - a_n)^{\alpha_n}}{\alpha_1! \ldots \alpha_n!}.$$

Lemma 3.1. *For* $n \geq 1$, *let us denote by* $\boldsymbol{\lambda}_n$ *the Lebesgue measure on* $\mathbb{R}^n$. *Let* $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ *be an analytic function and* $Z(f) := \{x \in \mathbb{R}^n : f(x) = 0\}$ *its zero-set. If* $\boldsymbol{\lambda}_n(Z(f)) > 0$ *then* $f \equiv 0$.

Proof. We proceed by induction. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a real analytic function such that $\boldsymbol{\lambda}_1(Z(f)) > 0$. Then $Z(f)$ is uncountable, hence it has an accumulation point. As f is analytic, this implies that $f \equiv 0$. Now let $n \geq 2$ and assume that the result holds for analytic functions on $\mathbb{R}^{n-1}$. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be an analytic function such that $\boldsymbol{\lambda}_n(Z(f)) > 0$. By Fubini's theorem, we have :

$$\begin{aligned} 0 &< \boldsymbol{\lambda}_n(Z(f)) \\ &= \int_{\mathbb{R}^n} \mathbb{1}_{Z(f)}(x)dx \\ &= \int_{\mathbb{R}} dx_n \int_{\mathbb{R}^{n-1}} \mathbb{1}_{Z(f)}(x_1, \ldots, x_n)dx_1 \ldots dx_{n-1}. \end{aligned}$$

This implies that there exists $A \subset \mathbb{R}$ with $\boldsymbol{\lambda}_1(A) > 0$ such that for all $x_n \in A$,

$$\int_{\mathbb{R}^{n-1}} \mathbb{1}_{Z(f)}(x_1, \ldots, x_{n-1}, x_n)dx_1 \ldots dx_{n-1} > 0,$$

that is

$$\forall x_n \in A,\ \boldsymbol{\lambda}_{n-1}\left[Z(f) \cap (\mathbb{R}^{n-1} \times \{x_n\})\right] > 0.$$

Thus, for all $x_n \in A$, the function

$$f(\cdot, x_n) : (x_1, \ldots, x_{n-1}) \mapsto f(x_1, \ldots, x_{n-1}, x_n)$$

is analytic on $\mathbb{R}^{n-1}$ and vanishes on a set with positive Lebesgue measure. Therefore, for all $x_n \in A$, $f(\cdot, x_n) \equiv 0$. Let $y \in \mathbb{R}^{n-1}$. The function

$$f(y, \cdot) : x_n \mapsto f(y, x_n)$$

is real analytic and vanishes on A with $\boldsymbol{\lambda}_1(A) > 0$. Hence $f(y, \cdot) \equiv 0$. As this holds for any y, it shows that $f \equiv 0$ and the lemma is proved. □ □

Now recall Leibniz's formula for the determinant :

$$\det Q(t) = \sum_{\sigma\in\mathfrak{S}_K} \epsilon(\sigma) \prod_{i=1}^{K} Q_{\sigma(i),i}(t),$$

where $\mathfrak{S}_K$ is the set of permutations of $\{1,\dots,K\}$ and $\epsilon(\sigma)$ is the signature of the permutation σ. Looking at the definition of $Q(t)$, we see that $\det Q(t) = 0$ if and only if

$$\sum_{\sigma\in\mathfrak{S}_K} \varepsilon(\sigma) \exp\left(\sum_{i=1}^{K-1} Z(t)\cdot\beta_{\sigma(i),i}\right) = 0,$$

where $\cdot$ is the standard inner product in $\mathbb{R}^{2d+1}$,

$$Z(t) = \left(1, \cos\left(\frac{2l\pi}{T}t\right), \sin\left(\frac{2l\pi}{T}t\right)\right)_{1\le l\le d} \in \mathbb{R}^{2d+1},$$

and

$$\beta_{\sigma(i),i} = \left(a_{\sigma(i),i,0}; a_{\sigma(i),i,1}; b_{\sigma(i),i,1}; \cdots; a_{\sigma(i),i,d}; b_{\sigma(i),i,d}\right) \in \mathbb{R}^{2d+1}.$$

For all $t \in \{1,\dots,T\}$, the function

$$\phi_t : \beta \mapsto \sum_{\sigma\in\mathfrak{S}_K} \varepsilon(\sigma) \exp\left(\sum_{i=1}^{K-1} Z(t)\cdot\beta_{\sigma(i),i}\right)$$

is analytic, as it is a sum of exponentials of linear combinations of its parameters. Therefore, by Lemma 3.1, either $\phi_t(\beta) = 0$ for all β, either the Lebesgue measure of its zero-set is zero. Let us define $\bar\beta$ by

$$\bar\beta_{ijl} = \mathbb{1}_{i=j}\mathbb{1}_{l=1}\log K.$$

The corresponding transition matrix for any time t is

$$\frac{1}{2K-1}\begin{pmatrix} K & 1 & \cdots & 1 \\ 1 & K & \cdots & 1 \\ \vdots & & \ddots & \vdots \\ 1 & \cdots & 1 & K \end{pmatrix},$$

which is invertible as it is a diagonally dominant matrix. Hence, for all $t \in \{1,\dots,T\}$, $\phi_t(\bar\beta) \neq 0$. Thus, for any time t, the zero-set of ϕ_t is negligible. As there is a finite number of such functions, we get the desired result and Assumption (A1) is satisfied.

As for all $t \in \{1,\dots,T\}$, the entries of $Q^*(t)$ are positive, the matrix $Q^*(1)\dots Q^*(T)$ is ergodic. Thus Assumption (A2) is satisfied provided that the distribution of X_1 is the stationary distribution of that matrix.

Checking Assumption (A3) regarding the linear independence of the emission distributions requires the following lemma (the second part of the lemma will be useful in the second application with gaussian mixtures) :

Lemma 3.2. *1. Let $\lambda_1,\dots,\lambda_n$ be pairwise distinct positive numbers and let us denote by $\mathcal{E}(\lambda)$ the exponential distribution with parameter λ. Then the distributions $\mathcal{E}(\lambda_1),\dots,\mathcal{E}(\lambda_n)$ are linearly independent.*

2. Let $m_1,\dots,m_n$ be real numbers and $\sigma_1^2,\dots,\sigma_n^2$ be pairwise distinct positive numbers. For any $k \in \{1,\dots,n\}$, let us denote by μ_k the Gaussian distribution with mean m_k and variance σ_k^2. Then the distributions $\mu_1,\dots,\mu_n$ are linearly independent.

Proof. Without loss of generality, we can assume that $\lambda_1 < \lambda_2 < \cdots < \lambda_n$. Let $(a_1, \ldots, a_n) \in \mathbb{R}^n$ such that

$$a_1\mathcal{E}(\lambda_1) + \cdots + a_n\mathcal{E}(\lambda_n) = 0.$$

For $x > 0$, we apply this equality to the Borel set $(-\infty, x]$ and we take the derivative with respect to x. We get that for any $x > 0$,

$$a_1\lambda_1 e^{-\lambda_1 x} + \cdots + a_n\lambda_n e^{-\lambda_n x} = 0. \tag{3.8}$$

This implies that for any $x > 0$,

$$a_1\lambda_1 + a_2\lambda_2 e^{-(\lambda_2-\lambda_1)x} + \cdots + a_n\lambda_n e^{-(\lambda_n-\lambda_1)x} = 0.$$

Hence, letting x go to infinity, we obtain that $a_1 = 0$, and equation (3.8) reduces to

$$a_2\lambda_2 e^{-\lambda_2 x} + \cdots + a_n\lambda_n e^{-\lambda_n x} = 0.$$

Thus, step by step, we show that $a_1 = \cdots = a_n = 0$, which ends the proof of the first statement. The second one can be proved using the same arguments. □ □

Now we are able to show that Assumption (A3) is easily satisfied. This is the purpose of Lemma 3.3.

Lemma 3.3. *Assume that for all $k \in \mathsf{X}$,*

$$\lambda_{k1}^* < \lambda_{k2}^* < \cdots < \lambda_{kM}^*.$$

For $t \in \{1, \ldots, T\}$, let us define the set

$$E_t = \left\{ \frac{\lambda_{km}^*}{1 + \sigma_k(t)} : 1 \leq k \leq K,\ 1 \leq m \leq M \right\}.$$

If E_t has cardinality at least K, Assumption (A3) is generically satisfied.

Before proving the lemma, let us clarify what we mean by *generically*. Here, *genericity* is to be understood in the algebraic sense and relates to the weights of the mixtures p_{km}. This notion of genericity is explained in the paragraph *Algebraic terminology* in Allman et al. (2009). Simply put, a property is true generically if it is true everywhere except on a proper algebraic subvariety, that is a set where a finite number of polynomials simultaneously vanish. In our case, this means that under the assumptions of Lemma 3.3, the set of weights $(p_{km})_{k\in\mathsf{X}, m\in\{1,\ldots,M\}}$ for which Assumption (A3) does not hold lies in a finite union of roots of polynomials. Note that this implies that it is negligible in the measure-theoretic sense (for the Lebesgue measure).

Proof. For $t \in \{1, \ldots, T\}$, let us denote by $p(t)$ the cardinality of E_t. Thus $K \leq p(t) \leq KM$ and we can write $E_t = \{\tilde{\lambda}_1^*, \ldots, \tilde{\lambda}_{p(t)}^*\}$. For all $k \in \{1, \ldots K\}$, $\nu_{k,t}$ is a linear combination of exponential distributions whose parameters belong to E_t. Hence,

$$\nu_{k,t} = \sum_{j=1}^{p(t)} B_{kj}(t)\mathcal{E}(\tilde{\lambda}_j^*),$$

where the $B_{kj}(t)$ are among the p_{km}^* and the matrix $B(t)$ has K rows and $p(t)$ columns. The $\tilde{\lambda}_j^*$ being pairwise distinct, the family $\left(\mathcal{E}(\tilde{\lambda}_j^*)\right)_{1\leq j\leq p(t)}$ is linearly independent, using Lemma 3.2.

Hence, the family $(\nu_{k,t})_{k\in\mathsf{X}}$ is linearly independent if and only if the rank of the matrix $B(t)$ is K (this requires that $p(t) \geq K$). Assume this is not true. Then all the minors of order K of $B(t)$ are zero (there are $p(t) - K + 1$ such minors). As these minors are polynomials in the p^*_{km}, this means that $B(t)$ has full rank except if the p^*_{km} belong to the set of common roots of a finite number of polynomials (which is an algebraic subvariety). Moreover, as the entries of $B(t)$ are among the p^*_{km}, the range of the map $t \mapsto B(t)$ is finite. Thus we obtain linear independence of the $(\nu_{k,t})_{k\in\mathsf{X}}$ for all t, except if the p^*_{km} belong to a finite union of algebraic subvarieties, which is itself an algebraic subvariety. Hence the generic linear independence. □ □

Then, we can identify the parameter themselves from the emission distributions and the transitions matrices. Let us denote by $\mathbb{V}$ the variance operator. For $t \in \{1, \dots, T\}$ and $k \in \mathsf{X}$, let

$$\tilde{s}(t) := \frac{1 + \sigma_k(t)}{1 + \sigma_k(1)} = \sqrt{\frac{\mathbb{V}(Y_t \mid \{X_t = k\})}{\mathbb{V}(Y_1 \mid \{X_1 = k\})}}.$$

If the emission distributions are known, $\tilde{s}(t)$ can be computed for any time step t. Let c be the constant coefficient of the trigonometric polynomial $\tilde{s}$. It follows from the above that $c = \frac{1}{1+\sigma_k(1)}$. Hence we get $\sigma_k(t) = \frac{\tilde{s}(t)}{c} - 1$. From this we can obtain δ^*_k. Using Assumptions (A9)-(A10) and equation (3.7), we have

$$\lim_{y\to\infty} \frac{\log f^*_{k,1}(y)}{y} = -\frac{\lambda^*_{k1}}{1 + \sigma_k(1)},$$

from which we find λ^*_{k1}. Then we can determine p^*_{k1} by computing

$$\exp\left[\lim_{y\to\infty}\left(\log f^*_{k,1}(y) + \frac{\lambda^*_{k1} y}{1 + \sigma_k(1)} - \log\frac{\lambda^*_{k1}}{1 + \sigma_k(1)}\right)\right].$$

Step by step, following the same method, we identify the rest of the parameters of the emission distributions.

It remains to show that we can retrieve β from the transition matrices $(Q(t))_{1\leq t\leq T}$. To this end, notice that for any $t \in \{1, \dots, T\}$ and for any $i \in \mathsf{X}$,

$$\sum_{j=1}^{K-1} Q_{ij}(t) = \frac{\sum_{j=1}^{K-1} \exp\left(P_{ij}(t)\right)}{1 + \sum_{j=1}^{K-1} \exp\left(P_{ij}(t)\right)},$$

with

$$P_{ij}(t) = \sum_{l=0}^{d}\left(a_{ijl}\cos\left(\frac{2\pi l t}{T}\right) + b_{ijl}\sin\left(\frac{2\pi l t}{T}\right)\right),$$

so that

$$\sum_{j=1}^{K-1}\exp\left[\sum_{l=0}^{d}\left(a_{ijl}\cos\left(\frac{2\pi l t}{T}\right) + b_{ijl}\sin\left(\frac{2\pi l t}{T}\right)\right)\right] = \frac{\sum_{j=1}^{K-1} Q_{ij}(t)}{1 - \sum_{j=1}^{K-1} Q_{ij}(t)}.$$

Then, for any $j \in \{1, \dots, K-1\}$,

$$\exp\left[\sum_{l=0}^{d}\left(a_{ijl}\cos\left(\frac{2\pi l t}{T}\right) + b_{ijl}\sin\left(\frac{2\pi l t}{T}\right)\right)\right] = Q_{ij}(t)\frac{\sum_{j'=1}^{K-1} Q_{ij'}(t)}{1 - \sum_{j'=1}^{K-1} Q_{ij'}(t)}.$$

As a trigonometric polynomial of degree d has at most $2d$ zeros over a period, this implies that we can retrieve β, as long as $T > 2d$. Hence Assumption (A4) is satisfied.

In order to prove the strong consistency of the maximum likelihood estimator, it remains to check that Assumptions (A5) to (A8) are satisfied. Clearly, Assumption (A6) is satisfied. Assuming that there exists $\beta_{\min}$ and $\beta_{\max}$ such that for all $\theta \in \Theta$ and for all i, j, l, $\beta_{ijl} \in [\beta_{\min}, \beta_{\max}]$, Assumption (A5) is satisfied. Assumption (A7) is satisfied provided that :

- $\inf_{\theta\in\Theta} \inf_{k\in\mathsf{X},m\in\{1,\dots,M\}} p_{km} > 0$,
- there exists positive numbers $\lambda_{\min}$ and $\lambda_{\max}$ such that for all $\theta \in \Theta$, for all k, m, $\lambda_{km} \in [\lambda_{\min}, \lambda_{\max}]$,
- there exists positive numbers $\sigma_{\min}$ and $\sigma_{\max}$ such that for all $\theta \in \Theta$, for all $k \in \mathsf{X}$ and for all $t \in \{1, \dots, T\}$, $\sigma_k(t) \in [\sigma_{\min}, \sigma_{\max}]$. Implicitly, this is a boundedness condition on the parameter δ.

Under the same boundedness assumptions, we obtain Assumption (A8). Thus, under weak conditions on the parameters, we can apply our identifiability and convergence results to this particular model.

Mixtures of Gaussian distributions

We choose the same transition matrices as in the previous example, and the emission distribution in state k at time t writes :

$$\nu_{k,t} = \sum_{m=1}^{M} p_{km} \mathcal{N}\left(m_k(t), \sigma_{km}^2\right)$$

where m_k is a trigonometric polynomial with (known) degree d, $(p_{k1}, \dots, p_{kM})$ is a vector of probability and $\mathcal{N}(m, \sigma^2)$ refers to the Gaussian distribution with mean m and variance σ^2. The following lemma ensures that Assumption (A3) is easily satisfied.

Lemma 3.4. *Assume that for all $k \in \mathsf{X}$, $\sigma_{k1}^2, \dots, \sigma_{kM}^2$ are pairwise distinct. Let*

$$E = \{\sigma_{km}^2, 1 \le m \le M,\ 1 \le k \le K\}.$$

If E has at least K elements, then, for all $t \in \{1, \dots, T\}$, the distributions $\nu_{1,t}, \dots, \nu_{K,t}$ are generically linearly independent.

Here, genericity has the same meaning as in Lemma 3.3.

Proof. Using the second statement of Lemma 3.2, the proof is the same as in Lemma 3.3. □ □

Hence we can identify the transition matrices and the emission distributions, up to state labelling. Thus, using the fact that finite Gaussian mixtures are identifiable, we can identify, for each state k and each time step t, the vector $(p_{k1}, \dots, p_{kM})$, the mean $m_k(t)$ and the variances σ_{km}^2, up to permutation of the components of the mixture. Finally, for each k, we can identify the coefficients of the trigonometric polynomial $m_k(\cdot)$ from its values $(m_k(1), \dots, m_k(T))$, so that Assumption (A4) is satisfied. Then, under boundedness conditions on the parameters that are very similar to those of the previous example, we see that Assumptions (A5) to (A8) are satisfied. Hence the strong consistency of the maximum likelihood estimator.

3.3 Simulation study

3.3.1 Computation of the maximum likelihood estimator

We have already shown the strong consistency of the maximum likelihood estimator. This paragraph deals with its practical computation. Assume that we have observed a trajectory of the

process $(Y_t)_{t\geq 1}$ with length n. Let $X := (X_1, \ldots, X_n)$ and $Y := (Y_1, ..., Y_n)$ and recall that X is not observed. The likelihood function with initial distribution π is then

$$L_{n,\pi}[\theta; Y] = \sum_{\mathbf{x}\in \mathsf{X}^T} \pi_{x_1} f^{\theta_Y}_{x_1,1}(Y_1) \prod_{t=2}^{n} Q_{x_{t-1}x_t}(t-1) f^{\theta_Y}_{x_t,t}(Y_t),$$

where $\mathbf{x} = (x_1, \ldots, x_n)$. As X is not observed, we use the Expectation Maximization (EM) algorithm to find a local maximum of the log-likelihood function. The EM algorithm is a classical algorithm to perform maximum likelihood inference with incomplete data. See Dempster et al. (1977) for a general formulation of the EM algorithm and (Baum et al., 1970) for its application to HMM. For any initial distribution π, we define the *complete* log-likelihood by :

$$\log L_{n,\pi}\left[\theta; (X, Y)\right] := \log \pi_{X_1} + \sum_{t=1}^{n-1} \log Q_{X_t X_{t+1}}(t) + \sum_{t=1}^{n} \log f^{\theta_Y}_{X_t,t}\left(Y_t\right).$$

This would be the log-likelihood function if X were observed. The algorithm starts from an initial vector of parameters $(\theta^{(0)}, \pi^{(0)})$ and alternates between two steps to construct a sequence of parameters $\left(\theta^{(q)}, \pi^{(q)}\right)_{q\geq 0}$.

The **E** step is the computation of the *intermediate quantity* defined by :

$$\mathbf{Q}\left[(\theta, \pi), \left(\theta^{(q)}, \pi^{(q)}\right)\right] := \mathbb{E}^{\pi^{(q)},\theta^{(q)}}\left[\log L_{n,\pi}\left(\theta; (X, Y)\right) \mid Y\right].$$

This requires to compute the *smoothing probabilties*, that are the *a posteriori* distributions of X given Y. More precisely, the following quantities are to be computed :

$$\pi^{(q)}_{t|n}(k) := \mathbb{P}^{\pi^{(q)},\theta^{(q)}}\left(X_t = k \mid Y\right)$$

for all $k \in \mathsf{X}$ and $1 \leq t \leq n$, and

$$\pi^{(q)}_{t,t+1|n}(k, l) := \mathbb{P}^{\pi^{(q)},\theta^{(q)}}\left(X_t = k, X_{t+1} = l \mid Y\right)$$

for $k, l \in \mathsf{X}$ and $1 \leq t \leq n-1$. The computation of the smoothing probabilities can be done efficiently using the *forward-backward* algorithm. See Rabiner and Juang (1986) or Cappé et al. (2009) for a description of this algorithm in the framework of HMM. The adaptation of the forward-backward algorithm for SHMM is straightforward. The intermediate quantity writes :

$$\begin{aligned}
&\mathbf{Q}\left[(\theta, \pi), \left(\theta^{(q)}, \pi^{(q)}\right)\right] \\
&= \mathbb{E}^{\pi^{(q)},\theta^{(q)}}\left[\log L_{n,\pi}\left(\theta; (X, Y)\right) \mid Y\right] \\
&= \sum_{k=1}^{K} \pi^{(q)}_{1|n}(k) \log \pi_k \\
&+ \sum_{t=1}^{n-1} \sum_{k=1}^{K} \sum_{l=1}^{K} \pi^{(q)}_{t,t+1|n}(k, l) \log Q_{kl}(t) \\
&+ \sum_{t=1}^{n} \sum_{k=1}^{K} \pi^{(q)}_{t|n}(k) \log f^{\theta_Y}_{k,t}\left(Y_t\right).
\end{aligned}$$

The **M** step consists in finding $\left(\theta^{(q+1)}, \pi^{(q+1)}\right)$ maximizing the function

$$(\theta, \pi) \mapsto \mathbf{Q}\left[(\theta, \pi), \left(\theta^{(q)}, \pi^{(q)}\right)\right],$$

or at least increasing it. Depending on the specific models chosen for $t \mapsto Q(t)$ and $t \mapsto \theta^Y(t)$ it is sometimes possible to find an analytic formula for the solution of this maximization problem. However, in most cases, a numerical optimization algorithm is required.

It can be shown that $\left(L_n\left[\theta^{(q)}; Y\right]\right)_{q\geq 0}$ is an increasing sequence and that under regularity conditions, it converges to a local maximum of the likelihood function (Wu, 1983). We alternate the two steps of the EM algorithm until we reach a stopping criterion. For example, we can stop the algorithm when the relative difference $\frac{L_{n,\pi^{(q)}}(\theta^{(q+1)};Y) - L_{n,\pi^{(q)}}(\theta^{(q)};Y)}{L_{n,\pi^{(q)}}(\theta^{(q)};Y)}$ drops below some threshold ε. The last computed term of the sequence $\left(\theta^{(q)}\right)_{q\geq 0}$ is then an approximation of the maximum likelihood estimator. However, if the EM algorithm does converge, it only guarantees that the limit is a *local* maximum of the likelihood function, which may not be global. Therefore it is a common practice to run the algorithm a large number of times, starting from different (e.g. randomly chosen) initial points and select the parameter with the largest likelihood. In Biernacki et al. (2003), the authors compare several procedures to initialize the EM algorithm, using variants such as SEM (Broniatowski et al., 1983). Introducing randomness in the EM algorithm provides a way to escape from local maxima.

3.3.2 Example

In order to illustrate our consistency result, we perform the maximum likelihood inference with simulated data. We used a simple SHMM with two states and gaussian emission distributions.

Model Let us consider the SHMM $(X_t, Y_t)_{t\geq 1}$ with two states ($K = 2$) and $T = 365$. The transition matrix at time t, denoted by $Q(t)$, is determined by its first column, given by :

$$Q_{i1}(t) = \frac{\exp\left(\beta_{i1} + \beta_{i2}\cos\left(\frac{2\pi t}{T}\right) + \beta_{i3}\sin\left(\frac{2\pi t}{T}\right)\right)}{1 + \exp\left(\beta_{i1} + \beta_{i2}\cos\left(\frac{2\pi t}{T}\right) + \beta_{i3}\sin\left(\frac{2\pi t}{T}\right)\right)}$$

Thus the law of the Markov chain $(X_t)_{t\geq 1}$ is determined by π (the distribution of X_1) and the β_{il} for $1 \leq i \leq K$ and $1 \leq l \leq 3$. The emission distributions are Gaussian :

$$Y_t \mid \{X_t = k\} \sim \mathcal{N}\left(m_k(t), \sigma_k^2\right),$$

where the mean $m_k(t)$ is given by

$$m_k(t) = \mu_k + \delta_{k1}\cos\left(\frac{2\pi t}{T}\right) + \delta_{k2}\sin\left(\frac{2\pi t}{T}\right).$$

Thus the parameters of the emission distributions are the μ_k, δ_{k1}, δ_{k2} and σ_k^2, for $1 \leq k \leq K$. Note that this is a special case of the model introduced in paragraph 3.2.4, with $M = d = 1$. The parameter θ is the vector containing both the parameters of the transitions and the parameters of the emission distributions. Then for any choice of θ and π, it is easy to simulate a realization of $(X_t, Y_t)_{1\leq t\leq n_{\max}}$. First we simulate the Markov chain (X_t), then we simulate (Y_t) conditionally to (X_t). We chose $n_{\max} = 500000$.

True parameters The transition probabilities are given by

$$\begin{aligned}
\pi^* &= (0.5, 0.5),\\
\beta_1^* &:= (\beta_{1l}^*)_{1\leq l\leq 3} = (1, 0.7, 0.5),\\
\beta_2^* &:= (\beta_{2l}^*)_{1\leq l\leq 3} = (-1, -0.6, 0.7).
\end{aligned}$$

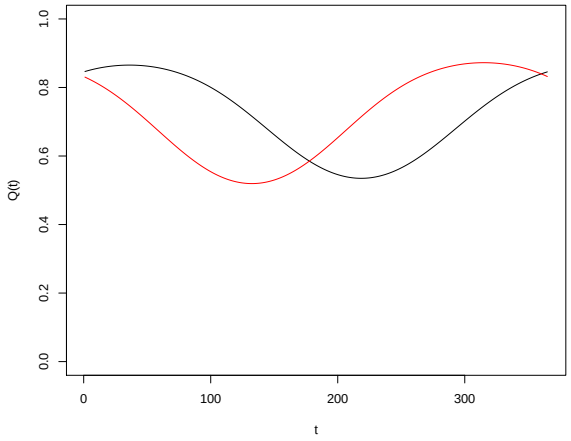

FIGURE 3.1 – Transition probabilities $Q^*_{11}(t)$ (black) and $Q^*_{22}(t)$ (red)

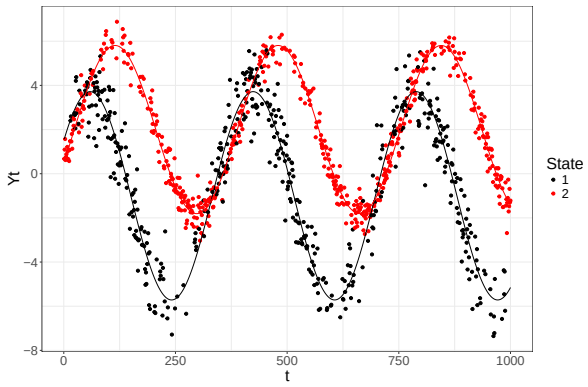

FIGURE 3.2 – Realization of $Y_1, \ldots, Y_{1000}$

The graphs of the functions $t \mapsto Q^*_{11}(t)$ (black) and $t \mapsto Q^*_{22}(t)$ (red) for $1 \leq t \leq 365$ are depicted in Figure 3.1. The parameters of the emission distributions are given by

$$\begin{aligned} \mu^* &= (-1, 2) \\ (\sigma^2)^* &= (1, 0.25) \\ \delta^*_1 &= (2.5, 4) \\ \delta^*_2 &= (-1.5, 3.5) \end{aligned}$$

Figure 3.2 depicts a simulation of $(Y_t)_{1 \leq t \leq 1000}$ using the parameter θ^*. The lines correspond to the conditionnal means $m_k(t)$.

Estimation In order to compute the MLE in this model, we use the EM algorithm described in paragraph 3.3.1. The estimation procedure is described below :

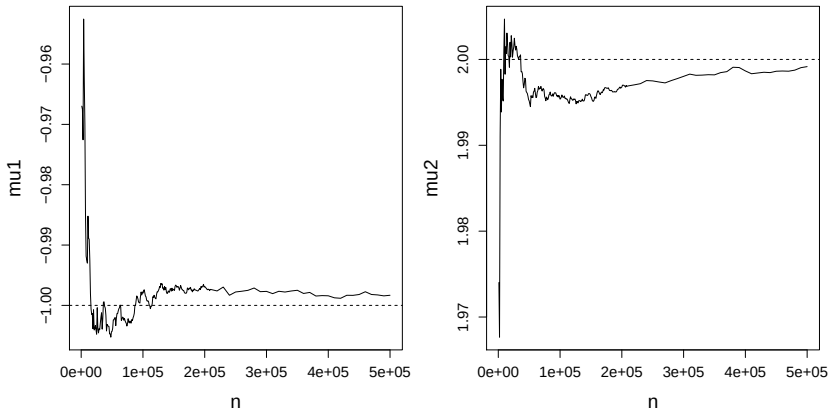

FIGURE 3.3 – Estimators $\hat{\mu}_{1,n}$ (left) and $\hat{\mu}_{2,n}$ (right)

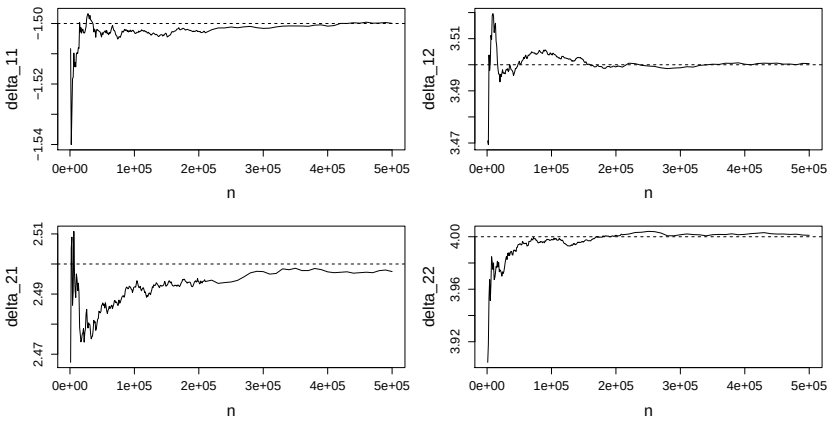

FIGURE 3.4 – Estimators $\hat{\delta}_{k,n}$

1. Select a random initial parameter and run the EM algorithm using $Y_1, \ldots, Y_{500}$, with 50 iterations starting from this initial point.
2. Repeat the first step 30 times.
3. Among the 30 initial points candidates, select the one that led to the largest log-likelihood after the EM.
4. Run a long EM using the selected initial point and $Y_1, \ldots, Y_n$. Stop when the relative difference in log-likelihoods drops below 10^{-7}. Then the result of the last **M**-step is $\hat{\theta}_n$.

Hence we can compute a sequence of MLE $\left(\hat{\theta}_{n_p}\right)_{p\geq 1}$. We chose $n_p = 1000p$, for $1 \leq p \leq 200$ and then we increased n_p by steps of 10000 until we reach $n_p = n_{\max} = 500000$.

Results Figures 3.3 to 3.5 show the estimated parameters for the emission distributions, i.e. the μ_k, δ_k and σ_k^2. The dashed lines represent the true parameters.
We can also check that the periodic means $m_k(t)$ have been well estimated. In Figure 3.6 are

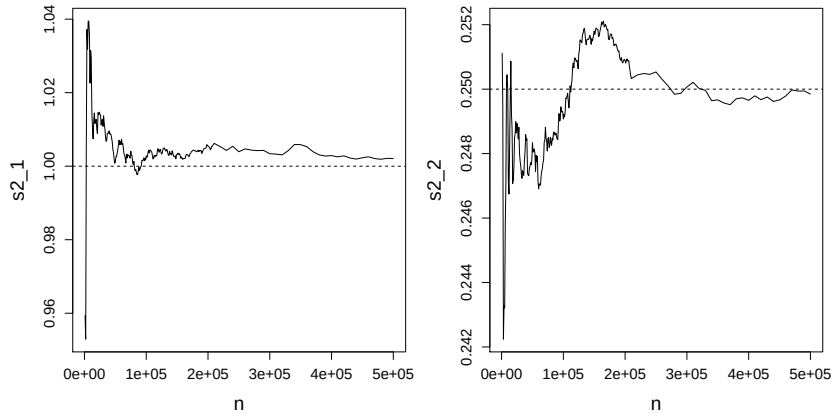

FIGURE 3.5 – Estimators $\hat{\sigma}^2_{1,n}$ (left) and $\hat{\sigma}^2_{2,n}$ (right)

drawn the graphs of the true means (solid line) and their estimated counterparts (dashed line), for each state.

The coefficients β_{kl} for the transition matrices, as well as the transition matrices themselves, are well estimated, as shown in Figures 3.7 and 3.8.

Remarks

- We had to swap the two states before producing these graphs. This illustrates the fact that the emission distributions and transitions matrices are identifiable only up to label swapping.
- The convergence of some of the parameters seems to be slower (see e.g. μ_1, δ_{21} or β_{23}) as more observations are required to achieve good precision.
- These are not the *true* MLE (as no closed-form expression is available) but only good approximations given by the EM algorithm. Hence, reducing the error requires not only to increase the number of observations but also the number of iterations of the EM algorithm, which is costly in terms of computing time.

3.4 Application to precipitation data

This work was motivated by the design of a *stochastic weather generator*. A *stochastic weather generator* (Katz, 1996) is a statistical model used whenever we need to quickly produce synthetic time series of weather variables. These series can then be used as input for physical models (e.g. electricity consumption models, hydrological models...), to study climate change, to investigate on extreme values (Yaoming et al., 2004)... A good weather generator produces times series that can be considered *realistic*. By realistic we mean that they mimic the behaviour of the variables they are supposed to simulate, according to various criteria. For example, a temperature generator may need to reproduce daily mean temperatures, the seasonality of the variability of the temperature, its global distribution, the distribution of the extreme values, its temporal dependence structure... and so on. The criteria that we wish to consider largely depend on applications. In this section, we introduce a univariate stochastic weather generator that focuses on rainfall.

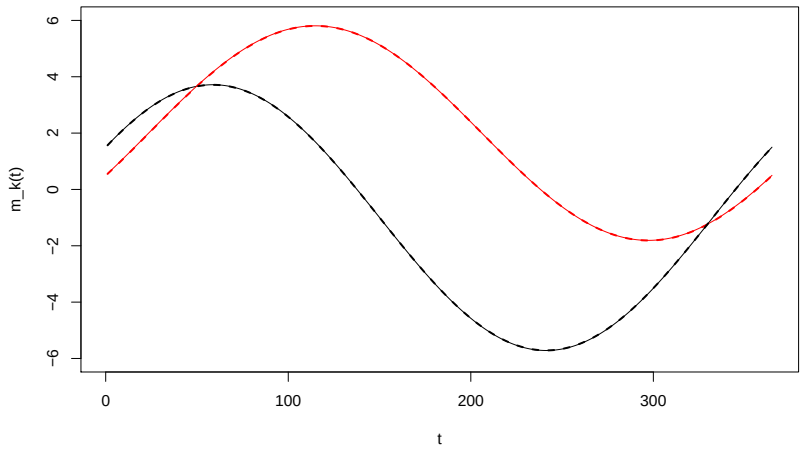

FIGURE 3.6 – True means $m_k(t)$ and their estimators $\hat{m}_k(t)$ (here computed with $n = n_{\max}$)

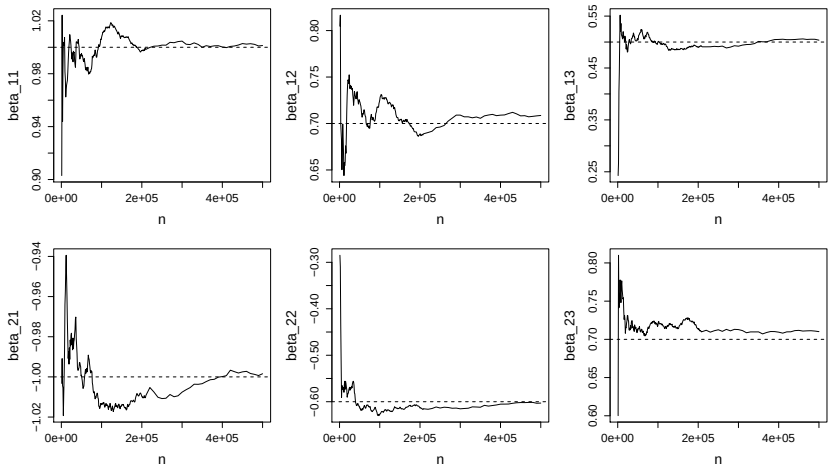

FIGURE 3.7 – Estimators $\hat{\beta}_{kl}$

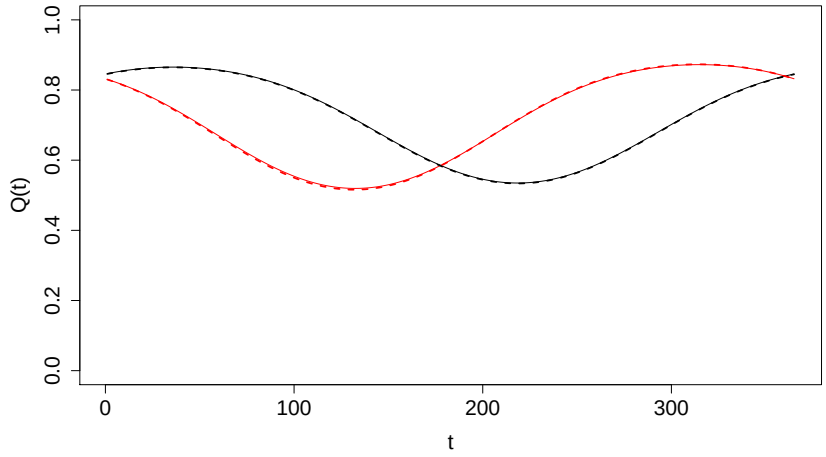

FIGURE 3.8 – True transitions $Q_{11}(t)$ and $Q_{22}(t)$ (solid line) and their estimated counterparts (dashed line, computed with $n = n_{\max}$)

Data We use data from the *European Climate Assessment and Dataset* (ECA&D project : http://www.ecad.eu). It consists in daily rainfalls measurements (in millimeters) at the weather station of Bremen, Germany, from 1/1/1950 to 12/31/2015. We remove the 16 February 29 so that every year of the period of observation has 365 days. Thus there are 24090 data points left. Missing data are replaced by drawing at random a value among those corresponding to the same day of the year. The distribution of daily precipitation amounts naturally appears as a mixture of a mass at 0 corresponding to dry days, and a continuous distribution with support on $\mathbb{R}_+$ corresponding to the intensity of precipitations on rainy days. Also, the data exhibits a seasonal behaviour with an annual cycle. Rainfalls tend to be less frequent and heavier in summer than in winter. Hence, a simple HMM cannot be used to model this data, but using a SHMM seems appropriate.

Model We use a model that is very similar to the one introduced in paragraph 3.2.4. To account for dry days, for each state, we replace the first component of the mixture of exponential distributions by a Dirac mass at 0, so that the emission distribution in state k is

$$\nu_k = p_{k1}\delta_0 + \sum_{m=2}^{M} p_{km}\mathcal{E}\left(\lambda_{km}\right).$$

Notice that the emission densities do not depend on t, as introducing periodic emission distributions is not necessary to generate realalistic time series of precipitations. It is enough to consider periodic transitions. Here the dominating measure is $\mu = \delta_0 + \boldsymbol{\lambda}$ where $\boldsymbol{\lambda}$ is the Lebesgue measure over $(0, +\infty)$. Hence the emission densities are given by

$$f_k(y) = p_{k1}\mathbb{1}_{y=0} + \sum_{m=2}^{M} p_{km}\lambda_{km}e^{-\lambda_{km}y}\mathbb{1}_{y>0}.$$

Recall that $(p_{k1}, \ldots, p_{kM})$ is a vector of probability and the transition probabilities between the hidden states are given by equation (3.1).

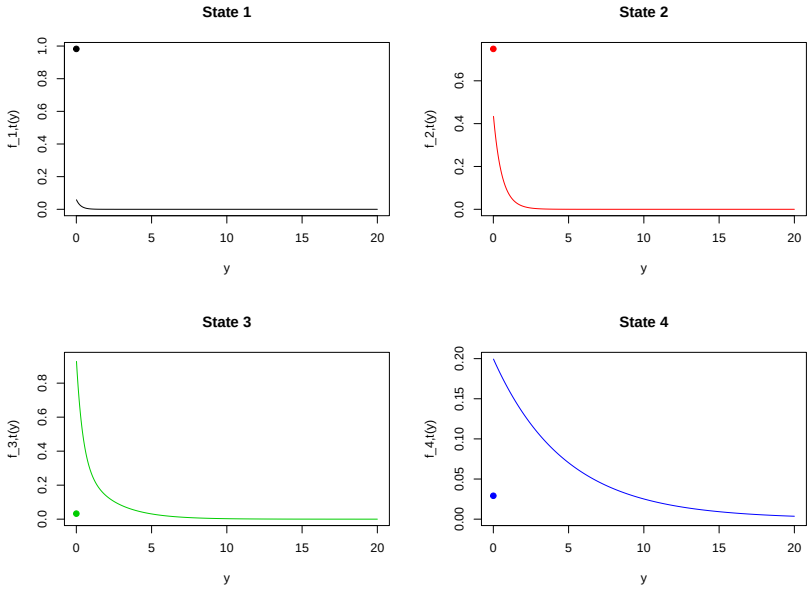

FIGURE 3.9 – Estimated emission densities

Results We estimate the parameters of the model by maximum likelihood inference, using the EM algorithm described in paragraph 3.3.1. We choose $K = 4$ (four states), $M = 3$ (mixtures of two exponential distributions and a Dirac mass at 0) and $d = 2$ (complexity of the seasonal components) as these parameters give good validation results. Figure 3.9 displays the estimated emission densities. They correspond to the following estimators :

$$\hat{\Lambda} = \begin{pmatrix} 3.455 & 3.455 \\ 1.162 & 1.825 \\ 2.330 & 0.481 \\ 0.086 & 0.214 \end{pmatrix}, \quad \hat{\mathbf{p}} = \begin{pmatrix} 0.983 & 0.003 & 0.014 \\ 0.749 & 0.025 & 0.226 \\ 0.032 & 0.258 & 0.709 \\ 0.029 & 0.059 & 0.912 \end{pmatrix}$$

The physical interpretation of the four states is straightforward. State 1 is mostly a dry state : in this state, it rarely rains and when it does rain, the rainfalls amounts are small. On the opposite, state 4 is a rainy state, with heavy rainfalls. Between these two extremes, state 2 and state 3 are intermediate states with moderate precipitation amounts. However, they differ by their precipitations frequency, as state 2 is dry most of the time whereas state 3 is almost always rainy. Figure 3.10 shows the transition probabilites between the four states as functions of time. It is also interesting to look at the relative frequencies of the four states (Figure 3.11). These vary quite a lot throughout the year. In particular, we see that in summer, dry states 1 and 2 are less frequent whereas state 4 is the most visited state. It means that in summer, we observe either dry days or heavy rain. It is the opposite in winter, where rainfalls are more frequent but also lighter compared to the summer. These interpretations of the states are consistent with climatology.

Validation As we wish to generate realistic simulations of daily rainfall amounts, i.e. simulations whose statistical properties mimic those of the real data, we evaluate the model by comparing the simulations produced by the model using the estimated parameters to the observed

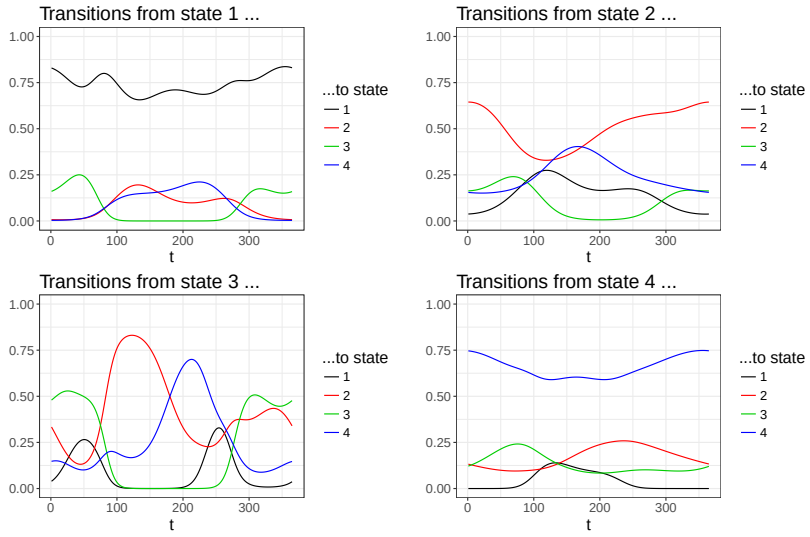

FIGURE 3.10 – Estimated transition probabilities

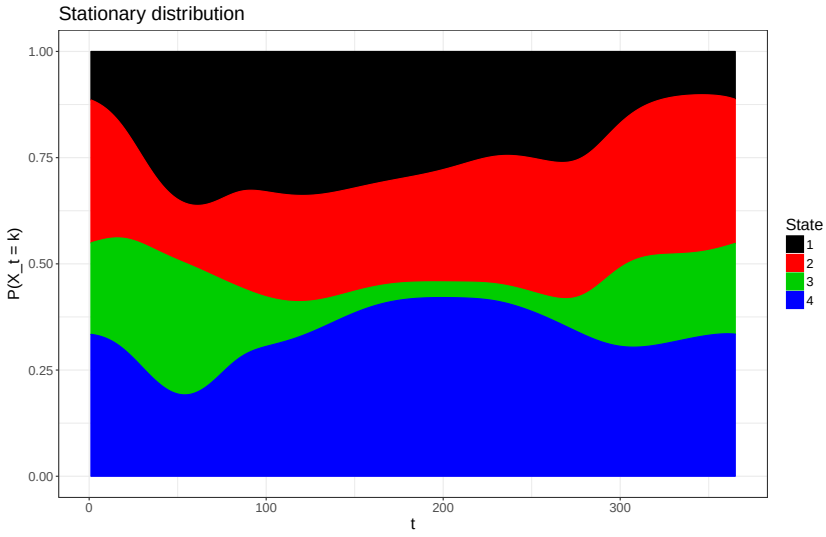

FIGURE 3.11 – Relative frequencies of states

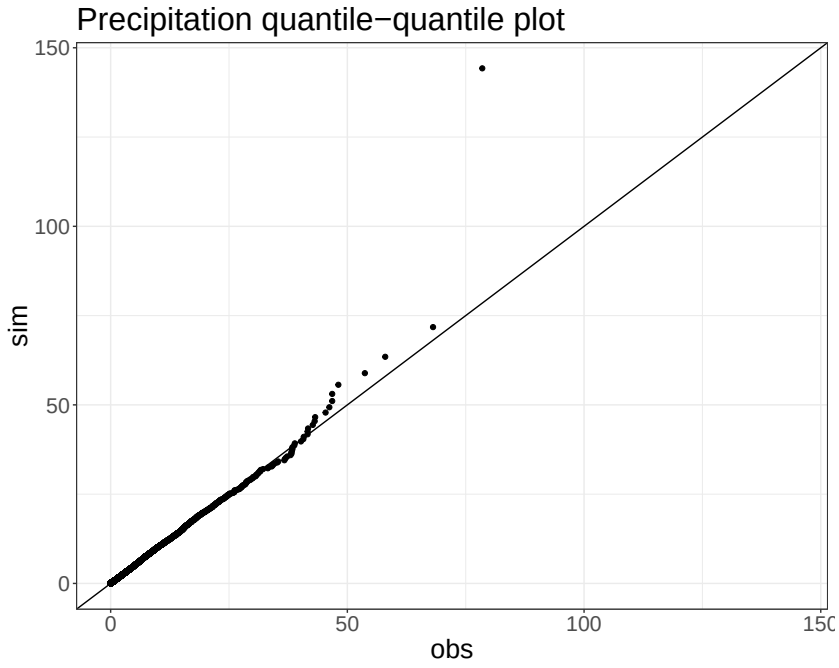

FIGURE 3.12 – Quantile-quantile plot

time series. To be specific, 1000 independent simulations are produced, each of them having the same length as the observed series. To perform a simulation, we first simulate a Markov chain $\left(X_t^{\text{sim}}\right)_{t\geq 1}$ with transition matrices $\hat{Q}(t)$. Then we simulate the observation process $\left(Y_t^{\text{sim}}\right)_{t\geq 1}$ using the estimated densities $f^{\hat{\theta}_Y}_{X_t^{\text{sim}},t}$. Several criteria can be considered to carry out the comparison : daily statistics (moments, quantiles, maxima, rainfall occurrence), overall distribution of precipitations, distribution of annual maximum, interannual variability, distribution of the length of dry and wet spells... The choice of the criteria mostly depends on the specific application of the model. Each of these statistics is computed from the simulations, which provides an approximation of the distribution of the quantity of interest under the law of the generator (in other words, we use *parametric bootstrap*), hence a 95% prediction interval. Then this distribution is compared to the value of the same statistic computed using the data. Let us first compare the overall distributions of the real precipitation amounts and the simulated ones by looking at the quantile-quantile plot (see Figure 3.12).

The match is correct, except in the upper tail of the distribution. The last point corresponds to the maximum of the simulated values, which is much larger than the maximum observed value. This should not be considered as a problem : a good weather generator should be able to (sometimes) generate values that are larger than those observed. We then focus on daily distributions. Figure 3.13 shows the results obtained for the first four daily moments and for the daily frequency of rainfall. It shows that these statistics are well reproduced by the model. Even though we did not introduce seasonal coefficients in emission densities, seasonalities appear both in the frequency of rainfall and the amounts. This is only due to the seasonality of the transition probabilities between the states.

The distribution of the duration of *dry and wet spells* is another quantity of interest when one studies precipitation. A wet (resp. dry) spell is a set of consecutive rainy (resp. dry) days.

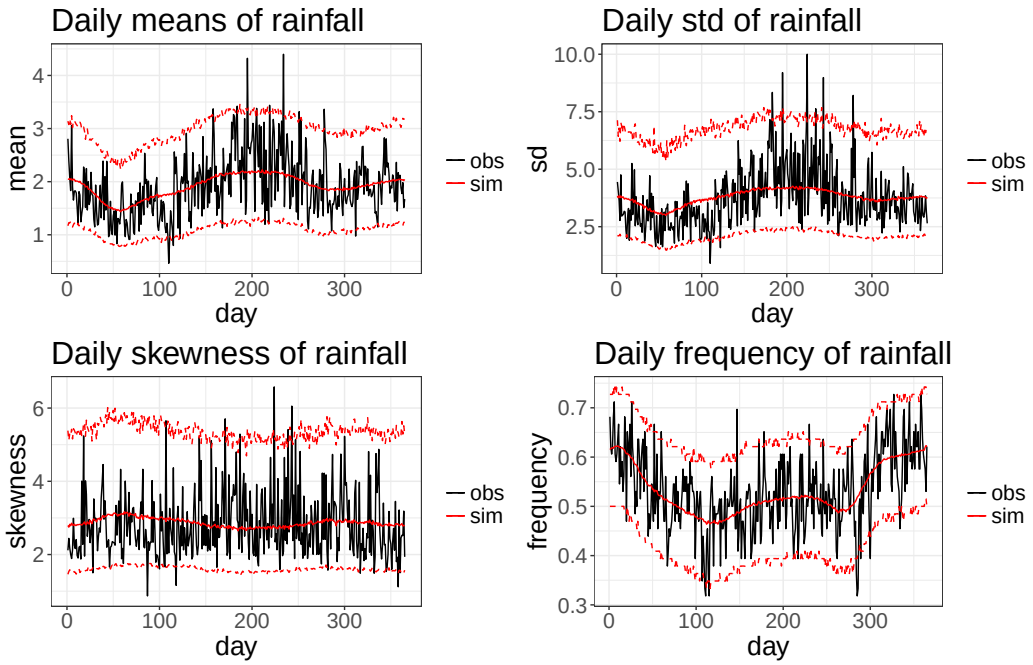

FIGURE 3.13 – Daily moments and frequency of precipitations. The black line relates to observations, the red solid line is the mean over all simulations, and the dashed lines depict an estimated 95% prediction interval under the model.

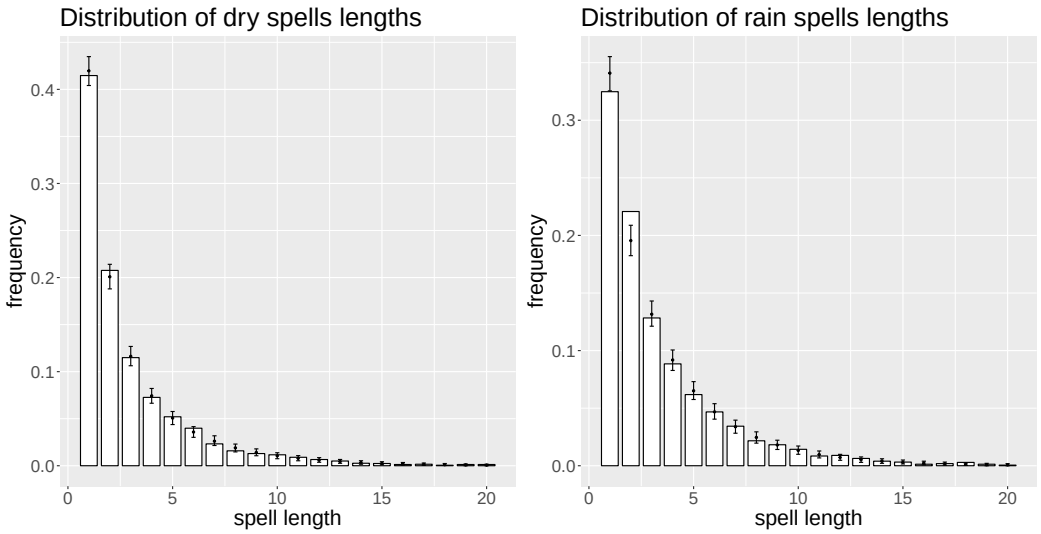

FIGURE 3.14 – Distribution of the lengths of dry (left plot) and wet (right plot) spells : observed (bars) versus simulated (error bars). The dots represent the means of the simulations

This statistic provides a way to measure the time dependence of the occurrence process. The results are presented in Figure 3.14. The dry spells are well modelled, whereas there is a slight underestimation of the frequency of 2-day wet spells while the single day events frequency is slightly overestimated.

3.5 Conclusion

We introduced a variant of hidden Markov models called SHMM, adapted to data with a seasonal behaviour. In these models, the transition probabilities between the states are periodic, as well as the emission distributions. We gave sufficient conditions of identifiabiity for SHMM and we proved that under reasonable assumptions on the parameter space, the maximum likelihood estimator is strongly consistent, thus generalizing previous results on HMM. Two specific models for which those conditions are satisfied were given as examples. In order to compute the maximum likelihood estimator, we described the EM algorithm adapted to the framework of SHMM and we used it with simulated data to illustrate our consistency result. In the last section, we applied a SHMM with zero-inflated mixtures of exponential distributions as emission laws to precipitation data. We showed that such a model provides a good example of a stochastic weather generator, as the statistical properties of time series generated by the model are close to those of the observations.

In this paper, we considered the number of states of the hidden process as a known parameter. However, in most real world applications, it is unknown. This model selection problem has yet to be adressed. In many applications, the data exhibit trends (e.g. climate change) in addition to seasonalities. However, the techniques presented in this paper cannot be directly adapted to deal with trends, so that this case requires further investigation.

Appendices

3.A The consistency theorem for a finite state space HMM

The proof of Theorem 3.2 relies on the application of a consistency result for (classical) hidden Markov models. This result is very close from Theorem 13.14 in Douc et al. (2014). Let $(X_t, Y_t)_{t\geq 0}$ be a stationary $\mathsf{X} \times \mathsf{Y}$-valued hidden Makov model, and (Θ, d) a compact metric space of parameters. Recall that the state space $\mathsf{X} = \{1, \dots, K\}$ is finite. Let us denote by Q^θ the transition matrix when the parameter is θ and $g^\theta(\cdot \mid x)$ the emission density in state $x \in \mathsf{X}$ when the parameter is θ. As the process $(X_t, Y_t)_{t\geq 0}$ is stationary, we can extend it to negative time indices and consider X_{-m} and Y_{-m} for $m > 0$. Let us make the following assumptions :

(A11). $\sigma_- := \inf_{\theta\in\Theta} \inf_{x,x'\in\mathsf{X}} Q^\theta_{xx'} > 0$

(A12). *For any $x, x' \in \mathsf{X}$ and $y \in \mathsf{Y}$, the functions $\theta \mapsto Q^\theta_{xx'}$ and $\theta \mapsto g^\theta(y \mid x)$ are continuous.*

(A13). *For all $y \in \mathsf{Y}$,*

$$b_-(y) := \inf_{\theta\in\Theta} \sum_{x\in\mathsf{X}} g^\theta(y \mid x) > 0 \tag{3.9}$$

$$b_+(y) := \sup_{\theta\in\Theta} \sum_{x\in\mathsf{X}} g^\theta(y \mid x) < \infty \tag{3.10}$$

(A14).

$$\mathbb{E}^{\theta^*}\left[|\log b_-(Y_0)|\right] < \infty \tag{3.11}$$

$$\mathbb{E}^{\theta^*}\left[|\log b_+(Y_0)|\right] < \infty \tag{3.12}$$

Let $\hat{\theta}_{n,\pi}$ be the maximum likelihood estimator in this model when the initial distribution is π. Let $\mathbb{P}^\theta$ be the distribution of the process $(Y_t)_{t\geq 0}$ when the initial distribution is the invariant distribution of Q^θ (whose existence and uniqueness is guaranteed by Assumption (A11)). We also define $\Theta^* = \{\theta \in \Theta : \mathbb{P}^\theta = \mathbb{P}^{\theta^*}\}$, where θ^* is the true parameter.

Theorem 3.3. *Under Assumptions (A11) to (A14), for any initial distribution π, $\mathbb{P}^{\theta^*}$-a.s.,*

$$\lim_{n\to\infty} d(\hat{\theta}_{n,\pi}, \Theta^*) = 0$$

Proof. The difference between Theorem 3.3 and Theorem 13.14 in Douc et al. (2014) lies in Assumption (A14). They assume that

$$\tilde{b}_+ := \sup_{\theta\in\Theta} \sup_{(x,y)\in\mathsf{X}\times\mathsf{Y}} g^\theta(y \mid x) < \infty, \tag{3.13}$$

instead of our integrability assumption (3.12). Also, they work in a general state space, whereas we restrict ourselves to a finite state space. Now let us show that the result still holds with our slightly different assumption. As the proof is almost identical, we shall concentrate on the parts where it differs.
Following the notations of Douc et al. (2014), for any probability measure ξ on X, $r, s \in \mathbb{Z}$, observations $y_r, \dots, y_s \in \mathsf{Y}$ and $\theta \in \Theta$ we denote by

$$p_\xi^\theta(y_{r:s}) := \sum_{x_r,\dots,x_s} \xi(x_r) g^\theta(y_r \mid x_r) \prod_{p=r+1}^{s} Q^\theta_{x_{p-1}x_p} g^\theta(y_p \mid x_p)$$

the likelihood of the observations $y_r, \dots, y_s$ when the parameter is θ and the distribution of X_r is ξ. For $r < t \leq s$, the conditional likelihood is defined by

$$p_\xi^\theta(y_{t:s} \mid y_{r:t-1}) = \frac{p_\xi^\theta(y_{r:s})}{p_\xi^\theta(y_{r:t-1})}.$$

Lemma 13.4 in Douc et al. (2014) states that

$$\sup_{\theta\in\Theta} \sup_{m\geq 0} p_\xi^\theta(y_t \mid y_{-m:t-1}) \leq \tilde{b}_+. \tag{3.14}$$

With our alternative assumption, we just get

$$\sup_{\theta\in\Theta} \sup_{m\geq 0} p_\xi^\theta(y_t \mid y_{-m:t-1}) \leq b_+(y_t). \tag{3.15}$$

Lemma 13.4 in Douc et al. (2014) implies that $\mathbb{P}^{\theta^*}$-a.s., $\lim_{m\to\infty} \log p_\xi^\theta(Y_t \mid Y_{-m:t-1})$ exists and does not depend on ξ. This limit is denoted by $\log p^\theta(Y_t \mid Y_{-\infty:t-1})$.
In their proof, Douc et al. (2014) show that for all $\theta \in \Theta$ and $Y_{-m}, \dots, Y_t$,

$$\sigma_- b_-(Y_t) \leq p_\xi^\theta(Y_t \mid Y_{-m:t-1}) \leq \tilde{b}_+ \tag{3.16}$$

These inequalities then justify the integrability of $\log p^\theta(Y_t \mid Y_{-\infty:t-1})$, thus the ergodic theorem can be applied to the stationary ergodic process

$$\left(\log p^\theta(Y_t \mid Y_{-\infty:t-1})\right)_t.$$

Using our assumptions, we have

$$\sigma_- b_-(Y_t) \leq p_\xi^\theta(Y_t \mid Y_{-m:t-1}) \leq b_+(Y_t) \tag{3.17}$$

and we still obtain the integrability of $\log p^\theta(Y_t \mid Y_{-\infty:t-1})$ using Assumption (A14).
Then, the uniform bound (3.13) is used to show the continuity of $\theta \mapsto p_\xi^\theta(Y_{r:s})$, by using the dominated convergence theorem and Assumption (A12). In the case of a finite state space, we do not need (3.13) because the likelihood function is expressed as a finite sum, not an integral, so we don't need the dominated convergence theorem to obtain the continuity, it is sufficient to use (A12).
Finally, in the proof of Theorem 13.7, they need to show that

$$\mathbb{E}\left[\left(\sup_{\theta\in\Theta} \log p^\theta(Y_t \mid Y_{-\infty:t-1})\right)_+\right] < \infty.$$

To this end, they use (3.14) but we get the same result using (3.15) and the integrability assumption (A14). The rest of the proof is identical. □

Acknowledgements We thank the anonymous reviewer for his/her comments. The author would like to thank Yohann De Castro, Élisabeth Gassiat, Sylvain Le Corff and Luc Lehéricy from Université Paris-Sud for fruitful discussions and valuable suggestions. This work is supported by EDF. We are grateful to Thi-Thu-Huong Hoang and Sylvie Parey from EDF R&D for providing this subject and for their useful advice.

Chapitre 4

Convergence de l'Estimateur du Maximum de Vraisemblance pour un HMM avec tendances

Convergence of the Maximum Likelihood Estimator for HMM with trends

This chapter is a joint work with Luc Lehéricy (Laboratoire de Mathématiques d'Orsay, Université Paris-Sud).

4.1 Introduction

Most existing results on hidden Markov model rely heavily on the homogeneity of the process. In practice, some processes cannot be assumed to be stationary. In hidden Markov models and most of their generalizations, the joint process $(X_t, Y_t)_{t\geq 1}$ is a Markov chain. We say that the process is *inhomogeneous* when this chain is inhomogeneous, that is when the distribution of (X_t, Y_t) conditionally to (X_{t-1}, Y_{t-1}) depends on t. In inhomogeneous HMM, the transition matrix and emission densities may vary over time.

This chapter is motivated by the study of meteorological data recordings spanning over several decades, in particular temperature. In this setting, it is necessary to include trends to account for the global warming, yet there is no theoretical guarantee that the corresponding maximum likelihood estimator is consistent.

Some inhomogeneous generalizations of HMMs have been studied recently. All of them deal with parametric models. Diehn et al. (2018) focus on the case where a rapidly fading phenomenon affects the distribution of the observations. Their model is a trivariate process $(X_t, Y_t, Z_t)_{t\geq 1}$ where only $(Z_t)_{t\geq 1}$ is observed, such that $(X_t, Y_t)_{t\geq 1}$ is an homogeneous HMM and $(X_t, Z_t)_{t\geq 1}$ is an inhomogeneous HMM. Their key assumption is that the distance between Z_t and Y_t tends to zero fast enough when t tends to infinity. In this sense, the process $(Z_t)_{t\geq 1}$ can be seen as

a perturbation of the process $(Y_t)_{t\geq 1}$ by a rapidly fading inhomogeneous noise. The authors introduce two estimators. The first one is the usual maximum likelihood estimator. The second one is a so-called quasi-maximum likelihood estimator. Let us write $p^\theta_{Y_1^n}$ (resp. $p^\theta_{Z_1^n}$) the density of the vector of random variables Y_1^n (resp. Z_1^n) under the parameter θ. The maximum likelihood estimator is

$$\arg\max_{\theta\in\Theta} \frac{1}{n} \log p^\theta_{Z_1^n}(Z_1^n)$$

and the quasi-maximum likelihood estimator is

$$\arg\max_{\theta\in\Theta} \frac{1}{n} \log p^\theta_{Y_1^n}(Z_1^n).$$

Thus, the second estimator is obtained by doing as if the perturbation wasn't here. The main theoretical result of their article is that both estimators are consistent. The core idea is that the asymptotic properties of $(Y_t)_{t\geq 1}$ can be transferred to the process $(Z_t)_{t\geq 1}$, in particular the ergodicity and the convergence of the log-likelihood. This makes it possible to adapt existing proofs for homogeneous HMMs. In practice, the quasi-maximum likelihood estimator does not need to know the structure of the inhomogeneous noise, which makes it easier to use. Their result can also be seen as a proof that the maximum likelihood estimator is robust to a temporary perturbation of the data, in which case the natural estimator is the quasi-maximum likehood one.

Consistency results for non-homogeneous Markov-switching models have also been obtained by Pouzo et al. (2016) and Ailliot and Pene (2015) for instance. These models are a generalization of HMMs where the hidden state X_t depends both of the previous hidden state X_{t-1} and on previous observations, let's say Y_{t-1} for an order one model, and where the observation Y_t depends both on the corresponding hidden state X_t and on previous observations. Such models are called "non-homogeneous" because the transition kernel of the hidden Markov chain depends on time through previous observations. However, this model is actually homogeneous in the previous sense, since the joint process $(X_t, Y_t)_{t\geq 1}$ is an homogeneous Markov chain.

In this chapter, we introduce a new inhomogeneous generalization : HMMs with trends. These models allow to deal with non periodic and non vanishing inhomogeneities. A HMM with trends is a trivariate process $(X_t, Y_t, Z_t)_{t\geq 1}$ taking values in $\mathcal{X}\times\mathbb{R}^d\times\mathbb{R}^q$ for some integer q (which we assume to be 1 for the sake of simplicity) where only the process $(Y_t)_{t\geq 1}$ is observed. In addition, we assume that $(X_t, Z_t)_{t\geq 1}$ is an homogeneous HMM and that there exists a vector of functions $(T_x)_{x\in\mathcal{X}}$ from $\mathbb{N}^*$ to $\mathbb{R}^q$, the *trends*, such that

$$Y_t = T_{X_t}(t) + Z_t.$$

We consider polynomial trends. As a consequence, they may diverge. We show that the maximum likelihood estimator recovers the parameter of the homogeneous HMM $(X_t, Z_t)_{t\geq 1}$ as well as the trends with respect to the supremum norm.

First, we introduce the notion of "blocks" : a trend belongs to the block of a true trend if it remains at a bounded distance of this true trend during the first n time steps. The blocks depend on the number n of observations since the distance is measured on the first n times steps. However, the bound on the distance must not depend on n.

The blocks $(B_t)_{t\geq 1}$ are unobserved, yet when the trends are close enough to the true trends, it is possible to assume that they are observed : the first step of the proof shows that the log-likelihood of the process $(Y_t)_{t\geq 1}$ is asymptotically the same as the one of the process $(Y_t, B_t)_{t\geq 1}$, see Theorem 4.2.

The second step of the proof shows that the estimated trends are eventually close enough to the true trend for the above result to apply, see Theorem 4.3.

Once the blocks $(B_t)_{t\geq 1}$ are observed, it is possible to de-trend the observations by substracting a representative of the block for the parameter θ :

$$Y_t - T^{\theta}_{B_t}(t) = Z_t + \left[T^*_{X_t}(t) - T^{\theta}_{B_t}(t)\right].$$

The observations defined in this way are not perfectly de-trended, since the residual trends $\left[T^*_{X_t}(t) - T^{\theta}_{B_t}(t)\right]$ are not necessarily constant. Fortunately, as the number of observations grow, these residuals become closer to constant functions, at least locally. Thus, the de-trended process behaves locally like a homogeneous HMM, which allows to explicitly compute the limit of the log-likelihood, see Theorem 4.4. The identifiability of the model and the consistency of the maximum likelihood estimator follow from the properties of this limit.

We illustrate our main result using simulated data. Two experiments are conducted. In the first one, we introduce a HMM with quadratic trends and we illustrate the convergence of the MLE. In the second experiment, we observe the convergence of the MLE in the case where trends have not diverged yet. This proves that even though we explicitly use the divergence of the trends to prove our consistency result, this divergence is actually not needed in practice to obtain a good estimator.

Outline The model and the assumptions used are detailed in Section 4.2. In Section 4.3, we state our consistency result as well as the main steps leading to it, without proving them, for clarity. Section 4.4 contains our study on simulations. Most of the proofs are in the appendices. Appendix 4.A contains the proof of Theorem 4.2, while Theorem 4.3 is proved in Appendix 4.B. The proof of Theorem 4.4, leading to Theorem 4.1, is detailed in Appendix 4.C. Finally, Appendix 4.D contains the proofs of two technical lemmas.

Notation. For each positive integer K, $[K]$ denotes the set $\{1, \dots, K\}$. For $a \leq b$ integers, we write Y_a^b instead of $(Y_a, \dots, Y_b)$.

4.2 Model and assumptions

Let K^* be a positive integer and $\mathcal{X}^* = [K^*]$. Let $\gamma^* = (\gamma^*_{x^*})_{x^*\in\mathcal{X}^*}$ be a vector of probability densities on $\mathbb{R}$ with respect to the Lebesgue measure. Let $(X_t)_{t\geq 1}$ be a Markov chain on $\mathcal{X}^*$ with transition matrix Q^* and initial distribution π^*. For all $x^* \in \mathcal{X}^*$, let $(Z_t^{x^*})_{t\geq 1}$ be a sequence of i.i.d. random variables in $\mathbb{R}$ such that these sequences are mutually independent and independent on $(X_t)_{t\geq 1}$ and such that for all $x^* \in \mathcal{X}^*$, $Z_1^{x^*}$ has density $\gamma^*_{x^*}$ with respect to the Lebesgue measure. Let $Z_t^{\max} = \max_{x^*\in\mathcal{X}^*} Z_t^{x^*}$ and $Z_t = Z_t^{X_t}$. Finally, let $T^* = (T^*_{x^*})_{x^*\in\mathcal{X}^*}$ be a family of functions $\mathbb{N}^* \longrightarrow \mathbb{R}$ and let $Y_t = Z_t + T^*_{X_t}(t)$ for all $t \geq 1$. The $(T^*_{x^*})_{x^*\in\mathcal{X}^*}$ are called *trends*.
The process $(X_t, Z_t)_{t\geq 1}$ is a homogeneous hidden Markov model with parameter $(\mathcal{X}^*, \pi^*, Q^*, \gamma^*)$ and $(X_t, Y_t)_{t\geq 1}$ is a hidden Markov model with trends with parameter $(\mathcal{X}^*, \pi^*, Q^*, \gamma^*, T^*)$.

Remark. *The random variables $Z_t^{\max}$ are i.i.d. and independent of $(X_t)_{t\geq 1}$. They allow to bound Z_t uniformly for all possible values of X_t.*

Consider a sample $(Y_1, \dots, Y_n)$ generated by a hidden Markov model with trends $(X_t, Y_t)_{t\geq 1}$ with parameter $\theta^* := (\mathcal{X}^*, \pi^*, Q^*, \gamma^*, T^*)$, which we call the *true parameter*. The goal is to recover

this parameter. In the following, we write $\mathbb{P}^*$ the distribution of the process $(X_t, Y_t)_{t\geq 1}$ and $\mathbb{E}^*$ the corresponding expectation.

Let $\sigma_- \in (0,1)$ be a positive constant.

(Aerg) A stochastic matrix Q satisfies **(Aerg)** when all its coefficients are lower bounded by σ_- :

$$\forall x, x', \quad Q(x,x') \geq \sigma_-. \tag{4.1}$$

For each $K \in \mathbb{N}^*$, let $\Sigma_K^{\sigma_-}$ be the set of stochastic matrices of size K which satisfy **(Aerg)** and Δ_K be the set of probability vectors of size K. Let Γ be a set of density functions on $\mathbb{R}$ and for each $d \in \mathbb{N}$, write $\mathbb{R}_d[X]$ the set of polynomials whose degree is at most d.
Let K be a positive integer. The model considered in this chapter is

$$\Theta = \bigcup_{K'=1}^{K} \{[K']\} \times \Delta_{K'} \times \Sigma_{K'}^{\sigma_-} \times \Gamma^{K'} \times (\mathbb{R}_d[X])^{K'}$$

Each $\theta = (\mathcal{X}^\theta, \pi^\theta, Q^\theta, \gamma^\theta, T^\theta) \in \Theta$ is a parameter of a HMM with trends. For all $K' \in \mathbb{N}^*$, let $\Theta_{K'} := \{\theta \in \Theta \text{ s.t. } |\mathcal{X}^\theta| = K'\}$. Write $\mathbb{P}^\theta$ the distribution under the parameter θ. Assume that the true parameter belong to the model : $\theta^* \in \Theta$.
Assume that Γ is compact with respect to the pointwise convergence topology (i.e. $g_n \underset{n\to\infty}{\longrightarrow} g$ if and only if for all $z \in \mathbb{R}$, $g_n(z) \underset{n\to\infty}{\longrightarrow} g(z)$) and equip Θ with the product topology, so that the mappings $\theta \mapsto Q^\theta$ and $\theta \mapsto \gamma^\theta$ are continuous.
We will use the following assumptions.

(Amax) *Envelope function.* There exists a nonincreasing function $g : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that $g \underset{+\infty}{\longrightarrow} 0$ and

$$\forall \theta \in \Theta \quad \forall x \in \mathcal{X}^\theta \quad \forall z \in \mathbb{R} \quad \gamma_x^\theta(z) \leq g(|z|).$$

(Amin) *Lower bound function.* There exists a nonincreasing function $m : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that

$$\forall \theta \in \Theta \quad \forall x \in \mathcal{X}^\theta \quad \forall z \in \mathbb{R} \quad \gamma_x^\theta(z) \geq m(|z|) > 0.$$

(Aint) *Integrability of the lower bound function.* The function m defined in **(Amin)** satisfies

$$\forall M > 0, \quad \mathbb{E}^* |\log m(M + |Z_t^{\max}|)| < \infty.$$

(Acentering) *Centering of the emission densities.* 0 is a median of the emission densities, that is :

$$\forall \theta \in \Theta, \quad \forall x \in \mathcal{X}^\theta, \quad \int_{z\leq 0} \gamma_x^\theta(z)dz = \frac{1}{2}.$$

(Aid) *Identifiability.* Q^* is invertible and the couples $(\gamma_{x^*}^*(\cdot - \Delta(x^*)), \mathbf{b}^*(x^*))_{x^* \in \mathcal{X}^*}$ are pairwise distinct, where the functions $\mathbf{b}^*$ and Δ are defined in Section 4.A.

(Areg) *Regularity of the emission densities.* There exists a modulus of continuity ω (that is a nondecreasing function $\mathbb{R}_+ \longrightarrow \mathbb{R}_+ \cup \{+\infty\}$ that is continuous at 0 and such that $\omega(0) = 0$) and a nondecreasing function $L : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that

$$\forall z, \eta \in \mathbb{R}, \quad \forall \theta \in \Theta, \quad \forall x \in \mathcal{X}^\theta, \quad \left|\log \frac{\gamma_x^\theta(z+\eta)}{\gamma_x^\theta(z)}\right| \leq L(|z|)\omega(|\eta|)$$

and such that

$$\forall M > 0, \quad \mathbb{E}^*[L(M + |Z_1^{\max}|)] < \infty.$$

4.3 Main result

4.3.1 Consistency of the MLE

Let $n \in \mathbb{N}^*$. The maximum likelihood estimator $\hat{\theta}_n = (\mathcal{X}^{\hat{\theta}_n}, \pi^{\hat{\theta}_n}, Q^{\hat{\theta}_n}, \gamma^{\hat{\theta}_n}, T^{\hat{\theta}_n})$ is an element of

$$\arg\max_{\theta \in \Theta} \frac{1}{n} \log p^{\theta}_{Y_1^n}(Y_1^n).$$

where $p^{\theta}_{Y_1^n}$ is the density of the distribution of Y_1^n with respect to the Lebesgue measure under the parameter θ.

Theorem 4.1. *Assume that $|\mathcal{X}^*|$ is known and that **(Amax)**, **(Amin)**, **(Aint)**, **(Acentering)**, **(Aid)** and **(Areg)** hold.*
Then almost surely, up to permutation of the hidden states :

$$\begin{cases} Q^{\hat{\theta}_n} \underset{n\to\infty}{\longrightarrow} Q^*, \\ \forall z \in \mathbb{R}, \quad \forall x \in [K], \quad \gamma_x^{\hat{\theta}_n}(z) \underset{n\to\infty}{\longrightarrow} \gamma_x^*(z), \\ \forall x \in [K], \quad \|T_x^{\hat{\theta}_n} - T_x^*\|_{\infty,[0,n]} \underset{n\to\infty}{\longrightarrow} 0. \end{cases}$$

Up to permutation of the hidden states means that there exists some relabelling of the states, i.e. a permutation of $[K]$ such that the claims of the theorem hold.

Remark. *Note that trends converge in supremum norm. This is considerably stronger than assuming that the coefficients of the polynomial converge. It is also intrinsic in the sense that it shows the convergence of the trends when seen as continuous functions and not just as polynomials. Therefore, it could be extended to other types of continuous trends.*

4.3.2 Overview of the proof

For all $M > 0$ and $n \in \mathbb{N}$, let $\Theta_n^{\mathrm{OK}}(M)$ be the subset of Θ defined by

$$\forall \theta \in \Theta_n^{\mathrm{OK}}(M), \qquad \forall x^* \in \mathcal{X}^*, \quad \exists x \in \mathcal{X}^{\theta}, \quad \|T^*_{x^*} - T^{\theta}_x\|_{\infty,[0,n]} \leq M \tag{4.2}$$

$$\text{and} \quad \forall x \in \mathcal{X}^{\theta}, \quad \exists x^* \in \mathcal{X}^*, \quad \|T^*_{x^*} - T^{\theta}_x\|_{\infty,[0,n]} \leq M. \tag{4.3}$$

We shall denote by $x(x^*, \theta, n)$ (resp. $x^*(x, \theta, n)$) the smallest suitable x (resp. x^*) in Equation (4.2) (resp. (4.3)).

$\Theta_n^{\mathrm{OK}}(M)$ is the set of parameters whose related trends are close to the true trends on the segment $[0, n]$.

Définition 4.1 (Blocks of trends)**.** *Let $\mathcal{R}^*$ be the equivalence relation defined on $\mathcal{X}^*$ by $x\mathcal{R}^*x'$ if and only if $T^*_x - T^*_{x'}$ is constant. Let*

$$\mathcal{B}^* := \mathcal{X}^*/\mathcal{R}^*$$

be the set of "blocks" of true trends. For $b \in \mathcal{B}^*$, we shall use the notation abuse T^*_b to indicate $T^*_{x^*}$ where x^* is the element of b associated with the smallest trend in the class.
Let us denote by $\mathbf{b}^* : \mathcal{X}^* \longrightarrow \mathcal{B}^*$ the quotient mapping, let

$$B_t := \mathbf{b}^*(X_t)$$

be the block from which the observation Y_t is generated and

$$\Delta : x^* \in \mathcal{X}^* \longmapsto T^*_{x^*}(1) - T^*_{\mathbf{b}^*(x^*)}(1)$$

be the function which maps the index of a trend to the difference between the corresponding trend and the reference trend of its block.
For all $\theta \in \Theta$, write

$$\ell_n(\theta) := \log p^{\theta}_{Y_1^n}(Y_1^n) \tag{4.4}$$

$$\ell_n^{(Y,B)}(\theta) := \log p^{\theta}_{(Y,B)_1^n}((Y,B)_1^n) \tag{4.5}$$

the log-likelihood of θ with respect to the processes $(Y_t)_{t\geq 1}$ and $(Y_t, B_t)_{t\geq 1}$.

The following theorem states that under a parameter that closely approximates the true trends and provided that the number of observations is large enough, the log-likelihood associated to the observed process $(Y_t)_{t\geq 1}$ can be approximated by the log-likelihood of the process $(Y_t, B_t)_{t\geq 1}$ where the blocks of trends are observed.

Theorem 4.2 (Adding block information)**.** *Assume **(Amax)**, **(Amin)** and **(Aint)**. Then for all $M > 0$, almost surely,*

$$\sup_{\theta \in \Theta_n^{OK}(M)} \left| \frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n^{(Y,B)}(\theta) \right| \underset{n\to+\infty}{\longrightarrow} 0.$$

Theorem 4.2 in proved in Appendix 4.A, together with the following corollary.

Corollary 4.1. *Assume **(Amax)**, **(Amin)** and **(Aint)**. Then there exists a finite $\ell(\theta^*)$ such that*

$$\frac{1}{n}\ell_n(\theta^*) \underset{n\to\infty}{\longrightarrow} \ell(\theta^*).$$

Corollary 4.1 is a key argument in the proof of the following result, stating that asymptotically, the maximum likelihood estimator approximates the true trends.

Theorem 4.3 (Localization of the maximum likelihood estimator)**.** *Assume **(Amax)**, **(Amin)** and **(Aint)**. Then there exists $M > 0$ and $n_{loc} \in \mathbb{N}$ such that almost surely, for all $n \geq n_{loc}$,*

$$\hat{\theta}_n \in \Theta_n^{OK}(M).$$

Let us give an intuition on the last steps of the proof. Informally, when n is large enough and θ lies in $\Theta_n^{\mathrm{OK}}(M)$, we have the following sequence of approximations (the new notations are explained in the following paragraph) :

$$\frac{1}{n}\ell_n(\theta) \simeq \frac{1}{n}\ell_n^{(Y,B)}(\theta) \tag{4.6}$$

$$\simeq \frac{1}{n}\ell_n^{(Y,B)}[N](\theta) \tag{4.7}$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} \ell^{\hom}\left(\theta, \frac{i}{N}\right) \tag{4.8}$$

$$\simeq \int_{[0,1]} \ell^{\hom}(\theta, u)du. \tag{4.9}$$

Approximation (4.6) results from Theorem 4.2. $\frac{1}{n}\ell_n^{(Y,B)}[N](\theta)$ is an approximation of $\frac{1}{n}\ell_n^{(Y,B)}(\theta)$ where each residual trend (that is the difference between an estimated trend and the true trend of the same block) has been made constant on the N segments of the form $[i\frac{n}{N},(i+1)\frac{n}{N}]$ for $i \in \{0,\dots,N-1\}$. Hence it is the normalized log-likelihood of a locally (with respect to time) homogeneous HMM. Approximation (4.8) comes from the convergence of the normalized log-likelihood of a homogeneous HMM to some continuous function ℓ^{hom}. Finally, the last one is the convergence of a Riemann sum towards its integral. Now let us state these intermediate results more rigorously.

For all $K' \in \mathbb{N}^*$, for all K'-uple $\gamma = (\gamma_x)_{x\in[K']}$ of measurable functions and for all $\mathbf{D} = (D_x)_{x\in[K']} \in \mathbb{R}^{K'}$, let

$$\tau(\gamma,\mathbf{D}) := (z' \longmapsto \gamma_x(z'-D_x))_{x\in[K']}$$

be the vector of functions γ translated by the vector $\mathbf{D}$.

Définition 4.2 (Log-likelihood of homogeneous HMM)**.** *Let $K' \in \mathbb{N}^*$ be a positive integer, π be a probability measure on $[K']$, Q be a transition matrix of size K', γ be a vector of K' emission densities on $\mathbb{R}$ and $\mathbf{b}$ be a function $[K'] \longrightarrow \mathcal{B}^*$. Let $(X_t,(\tilde{Z}_t,\tilde{B}_t))_{t\geq 1}$ be a homogeneous HMM taking values in $[K']\times(\mathbb{R}\times\mathcal{B}^*)$ with parameter $(\pi,Q,(\gamma_x\otimes\mathbf{1}_{\mathbf{b}(x)})_{x\in[K']})$.*
Denote by $\frac{1}{n}\ell_n^{hom}(\pi,Q,\gamma,\mathbf{b})\{(\tilde{z},\tilde{b})_1^n\}$ (resp. $\ell^{hom}(Q,\gamma,\mathbf{b})$) its normalized log-likelihood associated with the observations $(\tilde{z},\tilde{b})_1^n$ (resp. the limit of the log-likelihood, if it exists), that is

$$\frac{1}{n}\ell_n^{hom}(\pi,Q,\gamma,\mathbf{b})\{(\tilde{z},\tilde{b})_1^n\} = \frac{1}{n}\log\sum_{x_1^n\in[K']^n}\pi(x_1)Q(x_1,x_2)\dots Q(x_{n-1},x_n)\prod_{t=1}^n\gamma_{x_t}(\tilde{z}_t)\mathbf{1}_{\mathbf{b}(x_t)=\tilde{b}_t}$$

and

$$\ell^{hom}(Q,\gamma,\mathbf{b}) = \lim_{n\to\infty}\frac{1}{n}\ell_n^{hom}(\pi,Q,\gamma,\mathbf{b})\{(\tilde{Z},\tilde{B})_1^n\}. \tag{4.10}$$

The following Lemma ensures the existence of the limit of the normalized log-likelihood in Definition 4.2. It relies on results from Douc et al. (2004) on homogeneous HMM.

Lemma 4.1. *Assume **(Amax)**, **(Amin)** and **(Aint)**. Let $K' \in \mathbb{N}^*$. Then almost surely, for all $Q \in \Sigma_{K'}^{\sigma-}$, $\gamma \in \Gamma^{K'}$, $\mathbf{D} \in \mathbb{R}^{K'}$ and $\mathbf{b} : [K'] \longrightarrow \mathcal{B}^*$, the quantity*

$$\ell^{hom}(Q,\tau(\gamma,\mathbf{D}),\mathbf{b})$$

from Equation (4.10) exists, does not depend on the choice of the initial measure π and is finite when $(\tilde{Z}_t,\tilde{B}_t)_t = (Y_t - T^*_{B_t}(t),B_t)_t$.

Définition 4.3 (Integrated log-likelihood)**.** *We call* integrated log-likelihood *the following mapping :*

$$\ell^{int} : (Q,\gamma,\mathfrak{D},\mathbf{b}) \in \bigcup_{K'=1}^K \Sigma_{K'}^{\sigma-}\times\Gamma^{K'}\times\mathbf{L}^\infty([0,1])^{K'}\times(\mathcal{B}^*)^{K'} \longmapsto \int_0^1 \ell^{hom}(Q,\tau(\gamma,\mathfrak{D}(u)),\mathbf{b})du$$

For all $n\in\mathbb{N}$, $\theta\in\Theta$ and $x\in\mathcal{X}^\theta$, write

$$D_x^{\theta,n} : u\in[0,1] \longmapsto T_x^\theta(nu) - T^*_{\mathbf{b}^\theta(x)}(nu) \tag{4.11}$$

and for all $M>0$

$$\mathcal{D}(M) := \bigcup_{n\geq 4K(d+1)}\left\{D_x^{\theta,n} \,|\, \theta\in\Theta_n^{\text{OK}}(M), x\in\mathcal{X}^\theta\right\}. \tag{4.12}$$

By definition of $\Theta_n^{\text{OK}}(M)$, $\mathcal{D}(M)$ is a subset of $\mathbf{L}^\infty([0,1])$ uniformly bounded by $M+\|\Delta\|_\infty$. Let $\text{Cl}(\mathcal{D}(M))$ be its closure in $\mathbf{L}^\infty([0,1])$.

Proposition 4.1. *Let $M > 0$. Assume **(Amax)** and **(Amin)**. Then $Cl(\mathcal{D}(M))$ is a compact subset of $\mathcal{C}^0([0,1])$.*

Theorem 4.4. *Assume **(Amax)**, **(Amin)**, **(Aint)** and **(Areg)**. For all $M > 0$, the function ℓ^{int} is continuous on $\bigcup_{K'=1}^{K} \Sigma_{K'}^{\sigma_-} \times \Gamma^{K'} \times Cl(\mathcal{D}(M))^{K'} \times (\mathcal{B}^*)^{K'}$ and almost surely,*

$$\sup_{\theta \in \Theta_n^{OK}(M)} \left| \frac{1}{n} \ell_n(\theta) - \ell^{int}(Q^\theta, \gamma^\theta, (D_x^{\theta,n})_x, \mathbf{b}^\theta) \right| \underset{n\to\infty}{\longrightarrow} 0. \tag{4.13}$$

Proposition 4.1 and Theorem 4.4 are proved in the first part of Appendix 4.C.

Now, assume that $|\mathcal{X}^*|$ is known and take $K = |\mathcal{X}^*|$. The following proposition ensures that the only maximizer of the integrated log-likelihood is the true parameter.

Proposition 4.2. *Assume that **(Amax)**, **(Amin)**, **(Aint)**, **(Acentering)**, **(Aid)** and **(Areg)** hold. Let $(Q, \gamma, \mathfrak{D}, \mathbf{b}) \in \Sigma_K^{\sigma_-} \times \Gamma^K \times Cl(\mathcal{D})^K \times (\mathcal{B}^*)^K$ be a maximizer of ℓ^{int}, then $\mathfrak{D}$ is constant and $(Q, \gamma, \mathfrak{D}, \mathbf{b}) = (Q^*, \gamma^*, \Delta, \mathbf{b}^*)$ up to permutation of the hidden states.*

The proof of Proposition 4.2 is in the second part of Appendix 4.C.

We may now prove the consistency of the maximum likelihood estimator. By Theorem 4.3, there exists $M > 0$ such that almost surely, there exists a (random) integer n_{loc} such that for all $n \geq n_{\text{loc}}$, $\hat{\theta}_n \in \Theta_n^{\text{OK}}(M)$. For $n \geq n_{\text{loc}}$, let

$$\begin{cases} (Q_n, \gamma_n) := (Q^{\hat{\theta}_n}, \gamma^{\hat{\theta}_n}), \\ \mathfrak{D}_n := (D_x^{\hat{\theta}_n, n})_{x \in [K]}, \\ \mathbf{b}_n := \mathbf{b}^{\hat{\theta}_n}. \end{cases}$$

For all $n \geq n_{\text{loc}}$, $(Q_n, \gamma_n, \mathfrak{D}_n, \mathbf{b}_n) \in \Sigma_K^{\sigma_-} \times \Gamma^K \times \text{Cl}(\mathcal{D})^K \times (\mathcal{B}^*)^K$, and this set is compact by compactness of Γ and Proposition 4.1. Let $(Q, \gamma, \mathfrak{D}, \mathbf{b})$ be the limit of a convergent subsequence $(Q_{\varphi(n)}, \gamma_{\varphi(n)}, \mathfrak{D}_{\varphi(n)}, \mathbf{b}_{\varphi(n)})_{n\geq 1}$, then by continuity of ℓ^{int} and by the uniform convergence of equation (4.13), one has

$$\frac{1}{\varphi(n)} \ell_{\varphi(n)}(\hat{\theta}_{\varphi(n)}) \underset{n\to\infty}{\longrightarrow} \ell^{\text{int}}(Q, \gamma, \mathfrak{D}, \mathbf{b}) \leq \ell^{\text{int}}(Q^*, \gamma^*, \Delta, \mathbf{b}^*) = \ell(\theta^*),$$

by Proposition 4.2, and by definition of the maximum likelihood estimator

$$\frac{1}{\varphi(n)} \ell_{\varphi(n)}(\theta^*) \leq \frac{1}{\varphi(n)} \ell_{\varphi(n)}(\hat{\theta}_{\varphi(n)}).$$

Hence $\ell^{\text{int}}(Q, \gamma, \mathfrak{D}, \mathbf{b}) = \ell(\theta^*) = \ell^{\text{int}}(Q^*, \gamma^*, \Delta, \mathbf{b}^*)$. By Proposition 4.2, this means that up to permutation of the hidden states, $(Q, \gamma, \mathfrak{D}, \mathbf{b}) = (Q^*, \gamma^*, \Delta, \mathbf{b}^*)$. Thus, the maximum likelihood estimators sequence $(Q_n, \gamma_n, \mathfrak{D}_n, \mathbf{b}_n)_{n\geq 1}$ has only one possible limit : the true parameter. Theorem 4.1 follows.

4.4 Simulations

4.4.1 First experiment : the trends have diverged

In this first example, we consider the following HMM with trends $(X_t, Y_t)_{t\geq 1}$ with $K^* = 3$ states. The emission distributions are centered Gaussian distributions with respective variances $(\sigma_1^*)^2 = 5$, $(\sigma_2^*)^2 = 10$ and $(\sigma_3^*)^2 = 15$. The trends are given by

$$T_1^*(t) = \alpha(t + 10^4)^2, \quad T_2^*(t) = T_1^*(t) - 5, \quad T_3^*(t) = 3T_1^*(t),$$

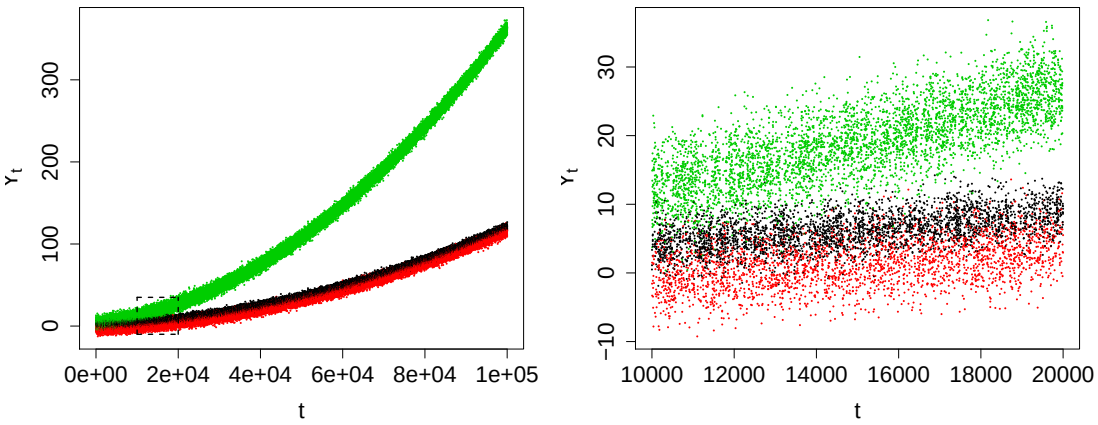

FIGURE 4.1 – Left panel : a simulated trajectory of length 100000 of the observations of the HMM with trends $(X_t, Y_t)_{t\geq 1}$. Each color corresponds to a different state (state 1 : black, state 2 : red, state 3 : green). Right panel : a focus on $10000 \leq t \leq 20000$.

with $\alpha = 10^{-8}$. Thus T_1^* and T_2^* belong to the same block while T_3^* diverges from the two other trends. Finally, the transition matrix is

$$Q^* = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}.$$

The model is as follows. The number of states $K^* = 3$ is assumed known. The set Γ of possible emission distributions is taken as the set of centered Gaussian distributions, without constraint on the variance. The lower bound on the transition probabilities is chosen as $\sigma_- = 0$. Even if Γ is not compact and σ_- is not positive, contrary to the theoretical result, the maximum likelihood estimator is still able to recover the parameters. Finally, the maximum degree of the trends is $d = 4$. Note that the model is over-parametrized, as it contains all the trends in $\mathbb{R}_4[X]$ while the degree of the true trends is only 2. This reflects the fact that in practice, we may not know the degree of the true trends.

We simulated 100 realizations $(X_t, Y_t)_{1\leq t\leq n_{\max}}$ with $n_{\max} = 5.10^5$. Figure 4.1 shows the first 10^5 data points of one of these realizations. Figure 4.2 illustrates the fact that for each $k \in [K^*]$, $\|T_k^* - T_k^{\hat{\theta}_n}\|_{\infty,[0,n]} \underset{n\to\infty}{\longrightarrow} 0$, where $T_k^{\hat{\theta}_n}$ is the maximum likelihood estimator of T_k^* computed from the first n observations.

For each simulated trajectory and for several $n \in \{1, \ldots, n_{\max}\}$, we compute :

- The errors on the trends $\|T_k^* - T_k^{\hat{\theta}_n}\|_{\infty,[0,n]}$, $1 \leq k \leq K^*$,
- The error on the transition matrix $\|Q^* - Q^{\hat{\theta}_n}\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm,
- The error on the variances $\max_k |(\sigma_k^*)^2 - \hat{\sigma}_k^2|$.

We plotted the logarithm of these errors against $\log n$ (see Figure 4.3) for large values of n. These graphs suggest a linear decrease of the logarithm of the errors with respect to $\log n$. Therefore, we can conjecture that as n tends to infinity, $n^\alpha \|Q^* - Q^{\hat{\theta}_n}\|_F$ is bounded in probability for some

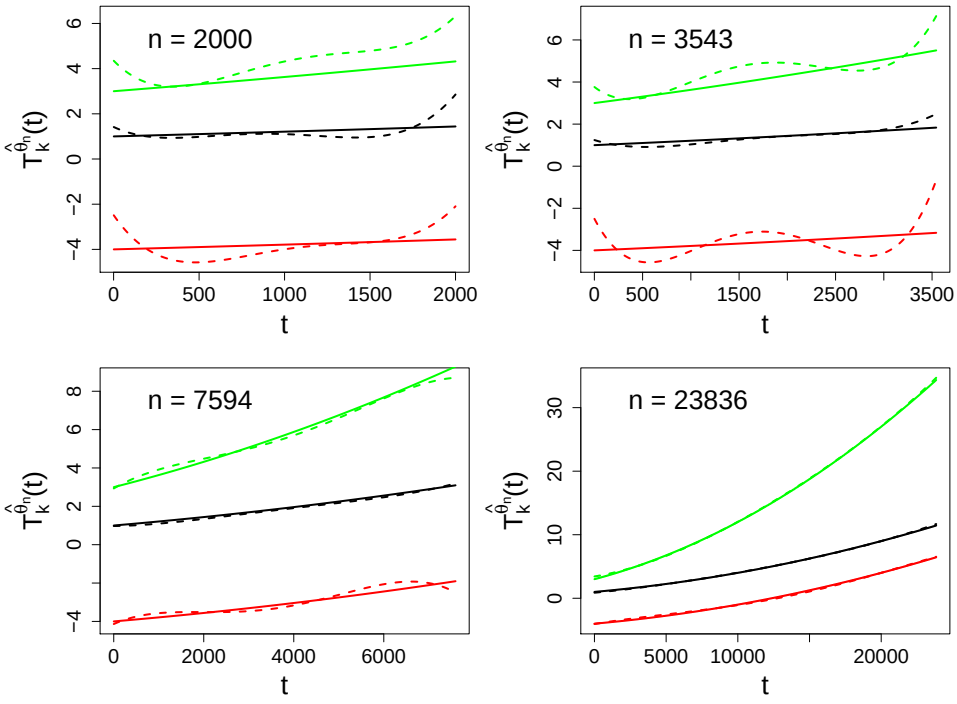

FIGURE 4.2 – Convergence of the estimated trends $T_k^{\hat{\theta}_n}$ (dashed lines) to the true trends T_k^* (solid lines).

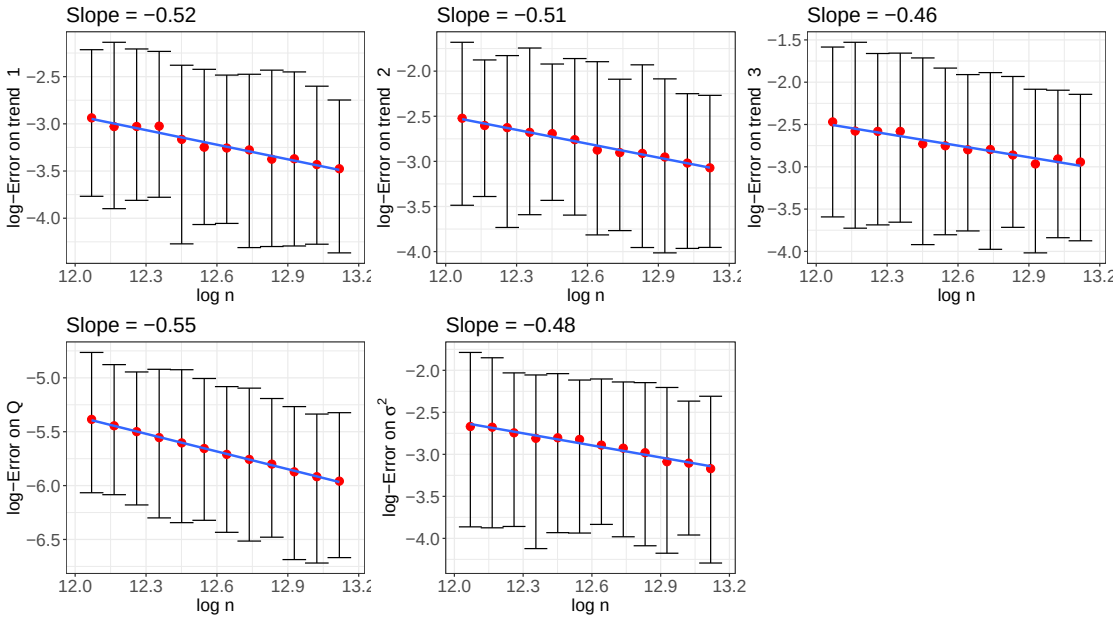

FIGURE 4.3 – Rate of convergence of the maximum likelihood estimator. The red dots are the means of the log-errors across the 100 simulations, the blue line is the linear fit to these means and the black bars contain, for each value of n, 95% of the simulated errors.

$\alpha > 0$, and that the same property holds for the other parameters (possibly with a different α). Based on this experiment, it seems reasonable to conjecture that the maximum likelihood estimator achieves the parametric rate of convergence $\alpha = 0.5$. However, proving this claim would require further investigation that go beyond the scope of this chapter. It is also worth noting that in this example, we chose simple parametric emission distributions, whereas the model described in Section 4.2 for which our main result holds only requires the set of emission densities to be compact, not necessarily finite-dimensional.

4.4.2 Second experiment : the trends have not diverged yet

In this section, we consider a HMM with trends whose trends have not diverged during the experiment, which is an assumption on which the proofs rely heavily. We show that the MLE is still able to recover the trends and the homogeneous parameter accurately. This is especially relevant for practical applications where one may not have enough time to see the trends diverge. Fix the maximum number of observations to $n = 10000$ and consider the following HMM with trends $(X_t, Y_t)_{t\geq 1}$ with $K^* = 2$ states. The trends are defined by $T_1^*(t) = 0$ and $T_2^*(t) = 3\left(\frac{t-\frac{n}{2}}{\frac{n}{2}}\right)^2 - 1$. The emission distributions are centered gaussian distributions with respective variances $(\sigma_1^*)^2 = 1$ and $(\sigma_2^*)^2 = 2$ and the transition matrix is

$$Q^* = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}.$$

Figure 4.4 shows the simulated observations $(Y_t)_{1\leq t\leq n}$ as well as the true trends T_1^* and T_2^*. The two states are not clearly separated : the trends will eventually diverge, but we don't have

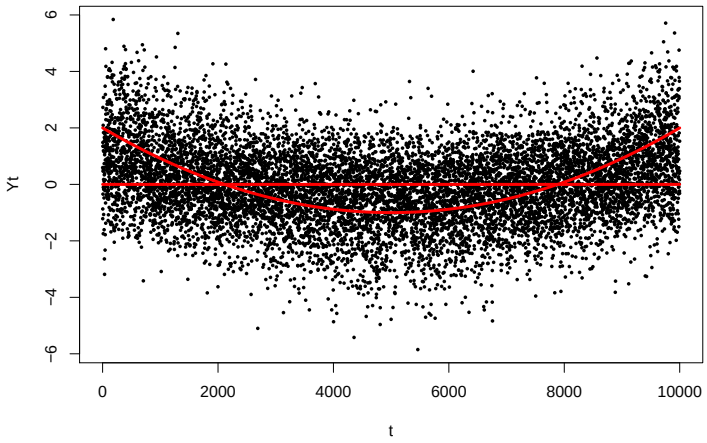

FIGURE 4.4 – Simulated data points. The red lines are the true trends.

enough observations to make use of this. However, the maximum likelihood estimator is able to recover the trends even in this situation.
The model is as follows. The number of states $K^* = 2$ is assumed known. As in the previous section, we take $\sigma_- = 0$, Γ as the set of centered Gaussian distributions and $d = 4$.
Figure 4.5 shows the true and estimated trends, obtained using the EM algorithm (see Section 2.3.3). The estimated transition matrix and variances are

$$\hat{Q} = \begin{pmatrix} 0.74 & 0.26 \\ 0.22 & 0.78 \end{pmatrix}, \quad (\hat{\sigma}_1^2, \hat{\sigma}_2^2) = (1.13, 2.11).$$

The precision of these estimations can be further improved by increasing the number of data points. The trends and the homogeneous parameter are already well estimated. An intuition to explain this convergence even when the trends have not diverged is that the trends vary slowly enough for the homogeneous approximation of Theorem 4.4 to hold.

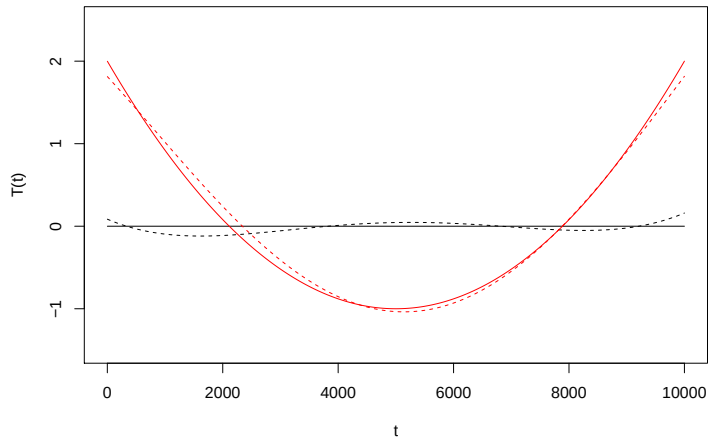

FIGURE 4.5 – True (full lines) and estimated (dashed lines) trends.

Appendices

4.A Block approximation

In this appendix, we shall prove Theorem 4.2 and Corollary 4.1. Let us begin with a few definitions. Denote by

$$E : t \in \mathbb{N}^* \longmapsto \inf_{b,b' \in \mathcal{B}^* \text{ s.t. } b \neq b'} |T_b^*(t) - T_{b'}^*(t)|$$

the minimum difference between two distinct blocks of true trends at time t. Note that $E(t)$ diverges to $+\infty$ since the true trends are polynomials.
Let $M > 0$ and

$$n_1(M) := \inf\{n \in \mathbb{N}^* \,|\, \forall t \geq n, E(t) > 4M\}.$$

By Equation (4.3), for all $n \geq n_1(M)$, $\theta \in \Theta_n^{\text{OK}}(M - \|\Delta\|_\infty)$ and $x, x' \in \mathcal{X}^\theta$,

$$\|S_x^{\theta,n} - S_{x'}^{\theta,n}\|_\infty \leq 2M \iff \mathbf{b}^*(x^*(\theta, x, n)) = \mathbf{b}^*(x^*(\theta, x', n)), \tag{4.14}$$

where for all $u \in [0,1]$, $\theta \in \Theta$, $n \in \mathbb{N}^*$ and $x \in \mathcal{X}^\theta$, $S_x^{\theta,n}(u) := T_x^\theta(nu)$.

Let $n \geq n_1(M)$ and $\theta \in \Theta_n^{\text{OK}}(M - \|\Delta\|_\infty)$. Consider the quotient space

$$\mathcal{B}^{\theta,n} = \mathcal{X}^\theta / \mathcal{R}_M$$

where the equivalence relation $\mathcal{R}_M$ is defined by $x \mathcal{R}_M x'$ if and only if $\|S_x^{\theta,n} - S_{x'}^{\theta,n}\|_\infty \leq 2M$ (Equation 4.14 shows that this is indeed an equivalence relation). $\mathcal{B}^{\theta,n}$ is the set of trend blocks associated with θ.
Equation (4.14) proves that there is an injection

$$b \in \mathcal{B}^{\theta,n} \longmapsto \mathbf{b}^*(x^*(\theta, x_b, n)) \in \mathcal{B}^*$$

where x_b is a representative of b (it does not matter which one). This mapping is also surjective : Equations (4.2) and (4.3) imply that for all $x^* \in \mathcal{X}^*$, there exists $x \in \mathcal{X}^\theta$ such that

$$\|S_{x^*}^{*,n} - S_{x^*(\theta,x,n)}^{*,n}\|_\infty \leq 2(M - \|\Delta\|_\infty),$$

so that $\mathbf{b}^*(x^*) = \mathbf{b}^*(x^*(\theta, x, n))$. Thus, this mapping is a bijection. In other words, $\mathcal{B}^{\theta,n}$ can be identified to $\mathcal{B}^*$, for all $n \geq n_1(M)$ and $\theta \in \Theta_n^{\text{OK}}(M - \|\Delta\|_\infty)$.

Définition 4.4. *For all $M > 0$, $n \geq n_1(M)$ and $\theta \in \Theta_n^{OK}(M - \|\Delta\|_\infty)$, denote by $\mathbf{b}^\theta : \mathcal{X}^\theta \longrightarrow \mathcal{B}^*$ the function that maps a state to its equivalence class.*

Note that by Equation (4.3), for all $M > 0$, $n \geq n_1(M)$, $\theta \in \Theta_n^{\mathrm{OK}}(M - \|\Delta\|_\infty)$ and $x \in \mathcal{X}^\theta$,

$$\sup_{t \in \{1,\dots,n\}} |T^*_{\mathbf{b}^\theta(x)}(t) - T^\theta_x(t)| \leq M. \tag{4.15}$$

The idea of the proof is to make the following approximations rigorous.

$$\begin{aligned}\log p^\theta_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) &\approx \log p^\theta_{Y_t,B_t|Y_1^{t-1}}(Y_t, B_t|Y_1^{t-1}) \\ &\approx \log p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(Y_t, B_t|Y_1^{t-1}, B_1^{t-1}).\end{aligned}$$

Hence the two steps that follow.

4.A.1 Step 1 : introduction of the trend block B_t in the log-likelihood

Assume **(Amax)**, **(Amin)** and **(Aint)**. Let us show that the following quantity tends to 0 uniformly in θ.

$$\begin{aligned}&\log p^\theta_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^\theta_{Y_t,B_t|Y_1^{t-1}}(Y_t, B_t|Y_1^{t-1}) \\ &\qquad = \log\left(\frac{\displaystyle\sum_{x_t \in \mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Y_t - T^\theta_{x_t}(t))}{\displaystyle\sum_{x_t \in \mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Y_t - T^\theta_{x_t}(t))\mathbf{1}_{\mathbf{b}^\theta(x_t)=B_t}}\right).\end{aligned}$$

This can be rewritten as

$$\begin{aligned}&\log p^\theta_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^\theta_{Y_t,B_t|Y_1^{t-1}}(Y_t, B_t|Y_1^{t-1}) \\ &\qquad = \log(p^\theta_{B_t|Y_1^t}(B_t|Y_1^t)^{-1}) \\ &\qquad = \log\left(\left\{1 - p^\theta_{B_t|Y_1^t}\left(\{b \in \mathcal{B}^* \text{ s.t. } b \neq B_t\}|Y_1^t\right)\right\}^{-1}\right).\end{aligned} \tag{4.16}$$

Intuitively, when t is large, since the trends get further from one another, the probability to get the wrong block converges to zero.

Lemma 4.2. *Assume **(Amax)** and **(Amin)**. Then for all $t \in \mathbb{N}^*$,*

$$\begin{aligned}&\sup_{\theta \in \Theta_n^{OK}(M - \|\Delta\|_\infty)} \left|\log p^\theta_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^\theta_{Y_t,B_t|Y_1^{t-1}}(Y_t, B_t|Y_1^{t-1})\right| \\ &\leq \log\left(\left\{1 - \frac{g(\{E(t) - M - |Z_t^{\max}| - \|\Delta\|_\infty\}_+)}{\sigma_- m(M + |Z_t^{\max}| + \|\Delta\|_\infty)}\right\}_+^{-1} \wedge \frac{g(0)}{\sigma_- m(M + |Z_t^{\max}| + \|\Delta\|_\infty)}\right) \\ &=: h(E(t), Z_t^{\max})\end{aligned}$$

with the convention $\{z\}_+^{-1} = +\infty$ if $z \leq 0$.

The first part of the infimum can be understood thanks to equation (4.16). The second one ensures that the upper bound is integrable.

Proof. Proof in Section 4.A.4. □

After summing, one gets

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}(M-\|\Delta\|_\infty)}\left|\frac{1}{n}\ell_n(\theta)-\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right|\leq\frac{1}{n}\sum_{t=1}^{n}h(E(t),Z_t^{\max}).$$

The function $e\in\mathbb{R}_+\longmapsto h(e,z)$ is non-negative, non-increasing for all $z\in\mathbb{R}$ and tends to 0 as e tends to $+\infty$. Moreover, under Assumption **(Aint)**, $h(0,Z_1^{\max})$ is integrable by definition of h. Thus, under Assumption **(Aint)**, the law of large numbers implies that for all $E>0$

$$\limsup_{n\to\infty}\sup_{\theta\in\Theta_n^{\mathrm{OK}}(M-\|\Delta\|_\infty)}\left|\frac{1}{n}\ell_n(\theta)-\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right|\leq\mathbb{E}^*[h(E,Z_1^{\max})].$$

Hence the dominated convergence theorem ensures that $\mathbb{E}^*[h(E,Z_1^{\max})]\longrightarrow 0$ as $E\longrightarrow+\infty$. Thus, we obtain the following uniform approximation of the normalized log-likelihood.

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}(M-\|\Delta\|_\infty)}\left|\frac{1}{n}\ell_n(\theta)-\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right|\underset{n\to\infty}{\longrightarrow}0. \tag{4.17}$$

4.A.2 Step 2 : conditioning on the blocks B_1^{t-1}

Assume **(Amax)** and **(Amin)**. The following lemma is a consequence of the lower bound on the transition matrices, see for instance Lemma 1 and Corollary 1 of Douc et al. (2004).

Lemma 4.3 (Exponential forgetting)**.** *There exists $C>0$ such that for all $n\in\mathbb{N}^*$, $y_1^n\in\mathbb{R}^n$, $\theta\in\Theta$ and for all probability measures π,π' on $\mathcal{X}^\theta$:*

$$\sum_{x\in\mathcal{X}^\theta}|p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi)-p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi')|\leq C\rho^n$$

with $\rho=1-\frac{\sigma_-}{1-\sigma_-}\in(0,1)$.

Besides, under **(Aerg)**, for all $\theta\in\Theta$, $x\in\mathcal{X}^\theta$, $y_1^{n-1}\in\mathbb{R}^n$ and for all probability measure π on $\mathcal{X}^\theta$:

$$p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi)\geq\sigma_-.$$

Hence, using the inequality $|\log x-\log y|\leq\frac{|x-y|}{x\wedge y}$ for all $x,y>0$: for all $n\in\mathbb{N}^*$, $\theta\in\Theta$, $y_1^n\in\mathbb{R}^n$, $b\in\mathcal{B}^*$ and for all probability measures π,π' on $\mathcal{X}^\theta$:

$$\begin{aligned}
&|\log p^{\theta}_{Y_n,B_n|Y_1^{n-1}}(y_n,b|y_1^{n-1},X_0\sim\pi)-\log p^{\theta}_{Y_n,B_n|Y_1^{n-1}}(y_n,b|y_1^{n-1},X_0\sim\pi')|\\
&\leq\frac{\displaystyle\sum_{x\in\mathcal{X}^\theta}|p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi)-p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi')|p^{\theta}_{Y_n,B_n|X_n}(y_n,b|x)}{\sigma_-\displaystyle\sum_{x\in\mathcal{X}^\theta}p^{\theta}_{Y_n,B_n|X_n}(y_n,b|x)}\\
&\leq\frac{C}{\sigma_-}\rho^n.
\end{aligned}$$

Changing the constant C if necessary, one has for all $a \in \mathbb{N}^*$:

$$\sup_{\theta \in \Theta_n^{\mathrm{OK}}(M-\|\Delta\|_\infty)} \left| \frac{1}{n} \sum_{t=1}^{n} \log p^{\theta}_{Y_t, B_t | Y_1^{t-1}}(Y_t, B_t | Y_1^{t-1}) \right. \\ \left. - \frac{1}{n} \sum_{t=1}^{n} \log p^{\theta}_{Y_t, B_t | Y_1^{t-1}, B_1^{t-a}}(Y_t, B_t | Y_1^{t-1}, B_1^{t-a}) \right| \leq C\rho^a. \quad (4.18)$$

It remains to condition on B_{t-a+1}^{t-1}.

Lemma 4.4. *Assume **(Amax)** and **(Amin)**. Then for all $a \in \mathbb{N}^*$,*

$$\sup_{\theta \in \Theta_n^{OK}(M-\|\Delta\|_\infty)} \left| \frac{1}{n} \sum_{t=1}^{n} \log p^{\theta}_{Y_t, B_t | Y_1^{t-1}, B_1^{t-a}}(Y_t, B_t | Y_1^{t-1}, B_1^{t-a}) \right. \\ \left. - \frac{1}{n} \sum_{t=1}^{n} \log p^{\theta}_{Y_t, B_t | Y_1^{t-1}, B_1^{t-1}}(Y_t, B_t | Y_1^{t-1}, B_1^{t-1}) \right| \\ \leq \frac{2aK^2}{\sigma_-^3} \frac{1}{n} \sum_{i=1}^{n-1} \left(1 \wedge \frac{g(\{E(i) - M - |Z_i^{\max}| - \|\Delta\|_\infty\}_+)}{m(M + |Z_i^{\max}| + \|\Delta\|_\infty)} \right) \\ =: \frac{2aK^2}{\sigma_-^3} \frac{1}{n} \sum_{i=1}^{n-1} h'(E(i), Z_i^{\max}).$$

Proof. We show that when **(Aerg)** holds, one has for all $\theta \in \Theta_n^{\mathrm{OK}}(M - \|\Delta\|_\infty)$, $t \leq n$, $a \in \mathbb{N}^*$, $y_1^t \in \mathcal{Y}^t$ and $b_1^t \in (\mathcal{B}^*)^t$,

$$\left| \log p^{\theta}_{Y_t, B_t | Y_1^{t-1}, B_1^{t-a}}(y_t, b_t | y_1^{t-1}, b_1^{t-a}) - \log p^{\theta}_{Y_t, B_t | Y_1^{t-1}, B_1^{t-1}}(y_t, b_t | y_1^{t-1}, b_1^{t-1}) \right| \\ \leq \frac{2}{\sigma_-} p^{\theta}_{B^{t-1}_{(t-a+1)\vee 1} | Y_1^{t-1}, B_1^{t-a}} \left((\mathcal{B}^*)^{(a \wedge t)-1} \setminus \{ b^{t-1}_{(t-a+1)\vee 1} \} | y_1^{t-1}, b_1^{t-a} \right) \quad (4.19)$$

$$\leq \frac{2K^2}{\sigma_-^3} \sum_{i=(t-a+1)\vee 1}^{t-1} \frac{\sup_{x_i \in \mathcal{X}^\theta \text{ s.t. } \mathbf{b}^\theta(x_i) \neq b_i} \gamma^\theta_{x_i}(y_i - T^\theta_{x_i}(i))}{\sum_{x \in \mathcal{X}^\theta} \gamma^\theta_x (y_i - T^\theta_x(i))}. \quad (4.20)$$

Then, we show that under **(Amax)** and **(Amin)**, one has for all $\theta \in \Theta_n^{\mathrm{OK}}(M - \|\Delta\|_\infty)$ and $i \leq n$

$$\frac{\sup_{x_i \in \mathcal{X}^\theta \text{ s.t. } \mathbf{b}^\theta(x_i) \neq B_i} \gamma^\theta_{x_i}(Y_i - T^\theta_{x_i}(i))}{\sum_{x \in \mathcal{X}^\theta} \gamma^\theta_x (Y_i - T^\theta_x(i))} \leq 1 \wedge \frac{g(\{E(i) - M - |Z_i^{\max}| - \|\Delta\|_\infty\}_+)}{m(M + |Z_i^{\max}| + \|\Delta\|_\infty)} \quad (4.21)$$

$$=: h'(E(i), Z_i^{\max}),$$

and the lemma follows by summing over t and i. The details of the proof can be found in Section 4.A.4. □

Therefore, one has almost surely

$$\begin{aligned}\limsup_{n\to+\infty}\sup_{\theta\in\Theta_n^{\mathrm{OK}}(M-\|\Delta\|_\infty)}\Bigg|\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})&\\ -\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(Y_t,B_t|Y_1^{t-1},B_1^{t-1})\Bigg|&\\ \leq\limsup_{n\to+\infty}\left(C\rho^a+\frac{2aK^2}{\sigma_-^3}\frac{1}{n}\sum_{i=1}^{n-1}h'(E(i),Z_i^{\max})\right)&\\ \leq C\rho^a+\frac{2aK^2}{\sigma_-^3}\mathbb{E}^*[h'(E,Z_1^{\max})]&\\ \leq C'\left(-\mathbb{E}^*[h'(E,Z_1^{\max})]\log\mathbb{E}^*[h'(E,Z_1^{\max})]\right)&\end{aligned}$$

for some explicit constant C' and for all E sufficiently large to have $\mathbb{E}^*[h'(E,Z_1^{\max})]<1/2$ using Assumption **(Adiv)**, the law of large numbers, the fact that the mapping $e\longmapsto h'(e,z)$ is non-negative, bounded by 0 and 1 and non-increasing for all z, and by taking $a=\lceil\frac{\log\mathbb{E}^*[h'(E,Z_1^{\max})]}{\log\rho}\rceil$. Since the function $e\longmapsto h'(e,z)$ tends to 0 as e tends to $+\infty$ for all z, the dominated convergence theorem ensures that almost surely,

$$\begin{aligned}\sup_{\theta\in\Theta_n^{\mathrm{OK}}(M-\|\Delta\|_\infty)}\Bigg|\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})&\\ -\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(Y_t,B_t|Y_1^{t-1},B_1^{t-1})\Bigg|\underset{n\to\infty}{\longrightarrow}0.&\end{aligned}$$

Let $\frac{1}{n}\ell_n^{(Y,B)}(\theta)=\frac{1}{n}\log p^{\theta}_{(Y,B)_1^n}((Y,B)_1^n)$. Combining the above equation with equation (4.17) yields Theorem 4.2.

4.A.3 Application : existence and finiteness of the relative entropy rate

Theorem 4.2 implies that Corollary 4.1 holds : since $\theta^*\in\Theta_n^{\mathrm{OK}}(M)$ for all $n\in\mathbb{N}^*$ and $M>0$,

$$\left|\frac{1}{n}\ell_n(\theta^*)-\frac{1}{n}\ell_n^{(Y,B)}(\theta^*)\right|\underset{n\to+\infty}{\longrightarrow}0.$$

Let $Z'_t=Y_t-T^*_{B_t}(t)$. Then $\frac{1}{n}\ell_n^{(Y,B)}(\theta^*)=\frac{1}{n}\ell_n^{(Z',B)}(\theta^*)$. Moreover the process $(X_t,(Z'_t,B_t))_{t\geq1}$ is a homogeneous and ergodic HMM under θ^*, with emission densities $((z',b)\longmapsto\gamma^*_{x^*}(z'-\Delta(x^*))\otimes\mathbf{1}_{\mathbf{b}^*(x^*)}(b))_{x^*\in\mathcal{X}^*}$ with respect to the measure $\mathrm{Leb}\otimes\mu_{\mathcal{B}^*}$, where $\mu_{\mathcal{B}^*}$ is the counting measure on $\mathcal{B}^*$.
Since it is homogeneous and ergodic, Barron (1985) shows that there exists $\ell(\theta^*)>-\infty$ such that

$$\frac{1}{n}\ell_n^{(Z',B)}(\theta^*)\longrightarrow\ell(\theta^*).$$

Then, all emission densities are upper bounded by $g(0)$ under **(Amax)**, so that the positive part of their logarithm is integrable. Therefore, Leroux (1992) implies that $\ell(\theta^*)<+\infty$ and Corollary 4.1 is proved.

4.A.4 Proofs

Proof of Lemma 4.2 (current block)

First note that this quantity is non-negative : the denominator contains less terms, and all of them are non-negative. Hence it is enough to find an upper bound. To this aim we will use Assumptions **(Amax)**, **(Amin)** and **(Aerg)** :

$$\begin{aligned}
&\left|\log p^\theta_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^\theta_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right| \\
&\leq \log\left(\frac{g(0)}{\sigma_- \sup\limits_{x_t\in\mathcal{X}^\theta \text{ s.t. } \mathbf{b}^\theta(x_t)=B_t} m(|Y_t - T^\theta_{x_t}(t)|)}\right) \\
&\leq \log\frac{g(0)}{\sigma_-} + \sum_{x_t^*\in\mathcal{X}^*} \mathbf{1}_{X_t=x_t^*}\left(-\log m\left(\inf_{x_t\in\mathcal{X}^\theta \text{ s.t. } \mathbf{b}^\theta(x_t)=\mathbf{b}^*(x_t^*)} |Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t)|\right)\right).
\end{aligned}$$

When $X_t = x_t^*$,

$$\begin{aligned}
&\inf_{x_t\in\mathcal{X}^\theta \text{ s.t. } \mathbf{b}^\theta(x_t)=\mathbf{b}^*(x_t^*)} |Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t)| \\
&\qquad\leq |Z_t| + \inf_{x_t\in\mathcal{X}^\theta \text{ s.t. } \mathbf{b}^\theta(x_t)=\mathbf{b}^*(x_t^*)} |T^*_{\mathbf{b}^*(x_t^*)}(t) - T^\theta_{x_t}(t)| + \Delta(x_t^*) \\
&\qquad\leq |Z_t^{\max}| + M + \|\Delta\|_\infty
\end{aligned}$$

using equation (4.15), hence

$$-\log m\left(\inf_{x_t\in\mathcal{X}^\theta \text{ s.t. } \mathbf{b}^\theta(x_t)=\mathbf{b}^*(x_t^*)} |Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t)|\right) \leq -\log m(M + |Z_t^{\max}| + \|\Delta\|_\infty).$$

This yields

$$\left|\log p^\theta_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^\theta_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right| \leq \log\frac{g(0)}{\sigma_- m(M + |Z_t^{\max}| + \|\Delta\|_\infty)}.$$

Let us show the second bound. We can rewrite it as

$$\begin{aligned}
&\log p^\theta_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^\theta_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1}) \\
&= -\log\left(1 - \frac{\sum\limits_{x_t\in\mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Y_t - T^\theta_{x_t}(t))\mathbf{1}_{\mathbf{b}^\theta(x_t)\neq B_t}}{\sum\limits_{x_t\in\mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Y_t - T^\theta_{x_t}(t))}\right) \\
&= -\log\left(1 - \sum_{x_t^*\in\mathcal{X}^*} \mathbf{1}_{X_t=x_t^*} \underbrace{\frac{\sum\limits_{x_t\in\mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t))\mathbf{1}_{\mathbf{b}^\theta(x_t)\neq B_t}}{\sum\limits_{x_t\in\mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t))}}_{(*)}\right).
\end{aligned}$$

Using **(Amax)**, **(Amin)** and **(Aerg)**,

$$\begin{aligned}(*) &\leq \frac{\sup_{x_t\in\mathcal{X}^\theta \text{ s.t. } \mathbf{b}^\theta(x_t)\neq B_t} g(|Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t)|)}{\sigma_- \sum_{x_t\in\mathcal{X}^\theta} m(|Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t)|)} \\ &\leq \frac{\sup_{x_t\in\mathcal{X}^\theta \text{ s.t. } \mathbf{b}^\theta(x_t)\neq B_t} g(\{|T^*_{x_t^*}(t) - T^\theta_{x_t}(t)| - |Z_t^{\max}|\}_+)}{\sigma_- \sup_{x_t\in\mathcal{X}^\theta} m(|T^*_{x_t^*}(t) - T^\theta_{x_t}(t)| + |Z_t^{\max}|)}.\end{aligned}$$

Let $x \in \mathcal{X}^\theta$ such that $\mathbf{b}^\theta(x) \neq B_t$. When $X_t = x_t^*$,

$$\begin{aligned}|Y_t - T^\theta_x(t)| &= |Z_t + T^*_{x_t^*}(t) - T^\theta_x(t)| \\ &\geq |T^*_{B_t}(t) - T^\theta_x(t)| - |Z_t^{\max}| - \Delta(x_t^*) \\ &\geq |T^*_{B_t}(t) - T^*_{\mathbf{b}^\theta(x)}(t)| - M - |Z_t^{\max}| - \|\Delta\|_\infty \\ &\geq E(t) - M - |Z_t^{\max}| - \|\Delta\|_\infty\end{aligned}$$

and

$$\begin{aligned}\sup_{x_t\in\mathcal{X}^\theta} m(|T^*_{x_t^*}(t) - T^\theta_{x_t}(t)| + |Z_t^{\max}|) &\leq m(\inf_{x_t\in\mathcal{X}^\theta} |T^*_{x_t^*}(t) - T^\theta_{x_t}(t)| + |Z_t^{\max}|) \\ &\leq m(M + |Z_t^{\max}| + \|\Delta\|_\infty)\end{aligned}$$

using equation (4.15). We finally obtain

$$(*) \leq \frac{g(\{E(t) - M - |Z_t^{\max}| - \|\Delta\|_\infty\}_+)}{\sigma_- m(M + |Z_t^{\max}| + \|\Delta\|_\infty)}. \tag{4.22}$$

Proof of Lemma 4.4 (recent blocks)

Proof of equation (4.19). Without loss of generality, one may assume $a \leq t$ (otherwise, the proof holds by replacing a by $a \wedge t$).

$$\begin{aligned}&p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(y_t, b_t|y_1^{t-1}, b_1^{t-a}) \\ &\quad= p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t, b_t|y_1^{t-1}, b_1^{t-1}) \\ &\qquad\times p^\theta_{B_{t-a+1}^{t-1}|Y_1^{t-1},B_1^{t-a}}(b_{t-a+1}^{t-1}|y_1^{t-1}, b_1^{t-a}) \\ &\qquad+ p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-a},B_{t-a+1}^{t-1}}(y_t, b_t|y_1^{t-1}, b_1^{t-a}, (\mathcal{B}^*)^{a-1} \setminus \{b_{t-a+1}^{t-1}\}) \\ &\qquad\times p^\theta_{B_{t-a+1}^{t-1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1} \setminus \{b_{t-a+1}^{t-1}\}|y_1^{t-1}, b_1^{t-a}),\end{aligned}$$

hence

$$\begin{aligned}&|p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(y_t, b_t|y_1^{t-1}, b_1^{t-a}) - p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t, b_t|y_1^{t-1}, b_1^{t-1})| \\ &\quad\leq p^\theta_{B_{t-a+1}^{t-1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1} \setminus \{b_{t-a+1}^{t-1}\}|y_1^{t-1}, b_1^{t-a}) \Big[p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t, b_t|y_1^{t-1}, b_1^{t-1}) \\ &\qquad\qquad+ p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t, b_t|y_1^{t-1}, b_1^{t-a}, (\mathcal{B}^*)^{a-1} \setminus \{b_{t-a+1}^{t-1}\})\Big] \\ &\quad\leq 2p^\theta_{B_{t-a+1}^{t-1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1} \setminus \{b_{t-a+1}^{t-1}\}|y_1^{t-1}, b_1^{t-a}) \sum_{x\in\mathcal{X}^\theta} p^\theta_{Y_t,B_t|X_t}(y_t, b_t|x).\end{aligned}$$

Finally, since under **(Aerg)**

$$p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(y_t,b_t|y_1^{t-1},b_1^{t-a}) = \sum_{x\in\mathcal{X}} p^\theta_{Y_t,B_t|X_t}(y_t,b_t|x)p^\theta_{X_t|Y_1^{t-1},B_1^{t-a}}(x|y_1^{t-1},b_1^{t-a})$$
$$\geq \sigma_- \sum_{x\in\mathcal{X}} p^\theta_{Y_t,B_t|X_t}(y_t,b_t|x)$$

and the same inequality holds for $p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t,b_t|y_1^{t-1},b_1^{t-1})$, we obtain equation (4.19) using $|\log x - \log y| \leq \frac{|x-y|}{x\wedge y}$ for all $x,y>0$.

Proof of equation (4.20). Since

$$(\mathcal{B}^*)^{a-1}\setminus\{b_{t-a+1}^{t-1}\} = \bigcup_{i=t-a+1}^{t-1} (\mathcal{B}^*)^{i-(t-a+1)}\times(\mathcal{B}^*\setminus\{b_i\})\times(\mathcal{B}^*)^{t-1-i},$$

by union bound,

$$\begin{aligned}
&p^\theta_{B_{t-a+1}^{t-1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1}\setminus\{b_{t-a+1}^{t-1}\}|y_1^{t-1},b_1^{t-a})\\
&\leq \sum_{i=t-a+1}^{t-1} p^\theta_{B_i|Y_1^{t-1},B_1^{t-a}}(\mathcal{B}^*\setminus\{b_i\}|y_1^{t-1},b_1^{t-a})\\
&= \sum_{i=t-a+1}^{t-1}\sum_{x_i\in\mathcal{X}^\theta} p^\theta_{B_i|X_i}(\mathcal{B}^*\setminus\{b_i\}|x_i)p^\theta_{X_i|Y_1^{t-1},B_1^{t-a}}(x_i|y_1^{t-1},b_1^{t-a})\\
&= \sum_{i=t-a+1}^{t-1}\sum_{x_i\in\mathcal{X}^\theta} \mathbf{1}_{\mathbf{b}^\theta(x_i)\neq b_i}\sum_{x_{i-1},x_{i+1}\in\mathcal{X}^\theta} p^\theta_{X_i|Y_i,X_{i-1},X_{i+1}}(x_i|y_i,x_{i-1},x_{i+1})\\
&\qquad\times p^\theta_{X_{i-1},X_{i+1}|Y_1^{t-1},B_1^{t-a}}(x_{i-1},x_{i+1}|y_1^{t-1},b_1^{t-a}).
\end{aligned}$$

Then, use that for all $x_{i-1},x_{i+1}\in\mathcal{X}^\theta$,

$$p^\theta_{X_{i-1},X_{i+1}|Y_1^{t-1},B_1^{t-a}}(x_{i-1},x_{i+1}|y_1^{t-1},b_1^{t-a})\leq 1$$

and that by the Markov property and **(Aerg)** for all $x_{i-1},x,x_{i+1}\in\mathcal{X}^\theta$

$$\begin{aligned}
p^\theta_{X_i|X_{i-1},X_{i+1}}(x|x_{i-1},x_{i+1}) &= \frac{p^\theta_{X_{i+1}|X_i}(x_{i+1}|x)p^\theta_{X_i|X_{i-1}}(x|x_{i-1})}{p^\theta_{X_{i+1}|X_{i-1}}(x_{i+1}|x_{i-1})}\\
&\geq Q^\theta(x,x_{i+1})Q^\theta(x_{i-1},x)\\
&\geq \sigma_-^2,
\end{aligned}$$

so that

$$\begin{aligned}
&p^\theta_{X_i|Y_i,X_{i-1},X_{i+1}}(x_i|y_i,x_{i-1},x_{i+1})\\
&\qquad= \frac{p^\theta_{X_i|X_{i-1},X_{i+1}}(x_i|x_{i-1},x_{i+1})\gamma^\theta_{x_i}(y_i-T^\theta_{x_i}(i))}{\sum_{x\in\mathcal{X}^\theta} p^\theta_{X_i|X_{i-1},X_{i+1}}(x|x_{i-1},x_{i+1})\gamma^\theta_x(y_i-T^\theta_x(i))}\\
&\qquad\leq \frac{p^\theta_{X_i|X_{i-1},X_{i+1}}(x_i|x_{i-1},x_{i+1})\gamma^\theta_{x_i}(y_i-T^\theta_{x_i}(i))}{\sigma_-^2\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_x(y_i-T^\theta_x(i))}.
\end{aligned}$$

Thus,

$$
\begin{aligned}
&p^{\theta}_{B_{t-a+1}^{t-1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1}\setminus\{b_{t-a+1}^{t-1}\}|y_1^{t-1},b_1^{t-a})\\
&\leq \sum_{i=t-a+1}^{t-1}\sum_{x_{i-1},x_{i+1}\in\mathcal{X}^\theta}\frac{\sum_{x_i\in\mathcal{X}^\theta}\mathbf{1}_{\mathbf{b}^\theta(x_i)\neq b_i}p^{\theta}_{X_i|X_{i-1},X_{i+1}}(x_i|x_{i-1},x_{i+1})\gamma^\theta_{x_i}(y_i-T^\theta_{x_i}(i))}{\sigma_-^2\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_x(y_i-T^\theta_x(i))}\\
&\leq \frac{K^2}{\sigma_-^2}\sum_{i=t-a+1}^{t-1}\frac{\sup_{x_i\in\mathcal{X}^\theta\text{ s.t. }\mathbf{b}^\theta(x_i)\neq b_i}\gamma^\theta_{x_i}(y_i-T^\theta_{x_i}(i))}{\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_x(y_i-T^\theta_x(i))}.
\end{aligned}
$$

Proof of equation (4.21). Using **(Amax)** and **(Amin)**,

$$
\begin{aligned}
&\frac{\sup_{x_i\in\mathcal{X}^\theta\text{ s.t. }\mathbf{b}^\theta(x_i)\neq B_i}\gamma^\theta_{x_i}(Y_i-T^\theta_{x_i}(i))}{\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_x(Y_i-T^\theta_x(i))}\leq\left(1\wedge\frac{\sup_{x_i\in\mathcal{X}^\theta\text{ s.t. }\mathbf{b}^\theta(x_i)\neq B_i}g(|Y_i-T^\theta_{x_i}(i)|)}{\sup_{x\in\mathcal{X}^\theta}m(|Y_i-T^\theta_x(i)|)}\right)\\
&\leq\sum_{x_i^*\in\mathcal{X}^*}\mathbf{1}_{X_t=x_i^*}\left(1\wedge\frac{\sup_{x_i\in\mathcal{X}^\theta\text{ s.t. }\mathbf{b}^\theta(x_i)\neq\mathbf{b}^*(x_i^*)}g(|Z_i+T^*_{x_i^*}(i)-T^\theta_{x_i}(i)|)}{\sup_{x\in\mathcal{X}^\theta}m(|Z_i+T^*_{x_i^*}(i)-T^\theta_x(i)|)}\right)\\
&\leq 1\wedge\frac{g(\{E(t)-M-|Z_i^{\max}|-\|\Delta\|_\infty\}_+)}{m(M+|Z_i^{\max}|+\|\Delta\|_\infty)}
\end{aligned}
$$

by the same arguments as in the control of **(*)** in equation (4.22).

4.B Localization of the MLE

In this section we shall prove Theorem 4.3. Assume **(Aerg)**, **(Amin)**, **(Amax)** and **(Aint)**.

4.B.1 Preliminary : compactness results

Recall that for all $M>0$ and $n\in\mathbb{N}$, $\Theta_n^{\mathrm{OK}}(M)$ is the subset of Θ defined by

$$
\forall\theta\in\Theta_n^{\mathrm{OK}}(M),\quad\forall x^*\in\mathcal{X}^*,\quad\exists x\in\mathcal{X}^\theta,\quad\|T^*_{x^*}-T^\theta_x\|_{\infty,[0,n]}\leq M \tag{4.23}
$$

$$
\text{and}\quad\forall x\in\mathcal{X}^\theta,\quad\exists x^*\in\mathcal{X}^*,\quad\|T^*_{x^*}-T^\theta_x\|_{\infty,[0,n]}\leq M. \tag{4.24}
$$

To prove that the maximum likelihood estimator belongs to such a set, we consider relaxed versions of (4.23) and (4.24). Let

$$
\begin{aligned}
I_{n,D}(x,x^*,\theta)&:=\left\{t\in\{1,\dots,n\}:|T^*_{x^*}(t)-T^\theta_x(t)|\leq D\right\},\\
\mathcal{T}(\alpha,n,D)&:=\left\{\theta\in\Theta:\left|\bigcap_{x^*\in\mathcal{X}^*}\bigcup_{x\in\mathcal{X}^\theta}I_{n,D}(x,x^*,\theta)\right|\geq n\alpha\right\},\\
\mathcal{U}(\alpha,n,D)&:=\left\{\theta\in\Theta:\forall x\in\mathcal{X}^\theta,\left|\bigcup_{x^*\in\mathcal{X}^*}I_{n,D}(x,x^*,\theta)\right|\geq n\alpha\right\}.
\end{aligned}
$$

$\mathcal{T}$ corresponds to a relaxation of (4.23) and contains the parameters θ such that all true trends are close to at least one parameter trend during most of the first time steps. Likewise, $\mathcal{U}$ corresponds to a relaxation of (4.24) and contains the parameters θ whose trends are close to at least one true trend during most of the first time steps.
By construction of these sets, for each $\theta \in \mathcal{T}(\alpha, n, D)$ (resp. $\theta \in \mathcal{U}(\alpha, n, D)$) and for each $x^* \in \mathcal{X}^*$ (resp. $x \in \mathcal{X}^\theta$), there exists at least one $x \in \mathcal{X}^\theta$ (resp. $x^* \in \mathcal{X}^*$) such that $I_{n,D}(x, x^*, \theta) \geq n\alpha/K$.

Proposition 4.3. *For all $\alpha \in (0, 1]$ and $D > 0$, there exists $M(\alpha, D) > 0$ and $n_0 := 4K(d+1)/\alpha$ such that*

$$\forall n \geq n_0,\ \forall \theta \in \Theta,\ \forall x \in \mathcal{X}^\theta,\ \forall x^* \in \mathcal{X}^*,$$
$$|I_{n,D}(x, x^*, \theta)| \geq n\alpha/K \ \Rightarrow\ \|T^*_{x^*} - T^\theta_x\|_{\infty,[0,n]} \leq M(\alpha, D).$$

It follows that $\theta \in \mathcal{T}(\alpha, n, D)$ (resp. $\theta \in \mathcal{U}(\alpha, n, D)$) is equivalent to (4.23) (resp. (4.24)) up to changing M :

Définition 4.5. *Let $\alpha \in (0, 1]$, $n \geq 1$ and $D > 0$.*
- *For all $\theta \in \mathcal{T}(\alpha, n, D)$ and $x^* \in \mathcal{X}^*$, let $x(\theta, x^*, n, \alpha, D)$ be the smallest $x \in \mathcal{X}^\theta$ such that $I_{n,D}(x, x^*, \theta) \geq n\alpha/K$.*
- *For all $\theta \in \mathcal{U}(\alpha, n, D)$ and $x \in \mathcal{X}^\theta$, let $x^*(\theta, x, n, \alpha, D)$ be the smallest $x^* \in \mathcal{X}^*$ such that $I_{n,D}(x, x^*, \theta) \geq n\alpha/K$.*

In the following, we omit the dependency in α and D and write $x^*(\theta, x, n)$ and $x(\theta, x^*, n)$.

Corollary 4.2. *For all $\alpha \in (0, 1]$ and $D > 0$, there exists $M(\alpha, D) > 0$ and $n_0 := 4K(d+1)/\alpha$ such that*

$$\forall n \geq n_0,\ \begin{cases} \theta \in \mathcal{T}(\alpha, n, D) \Rightarrow \forall x^* \in \mathcal{X}^*,\ \|T^*_{x^*} - T^\theta_{x(\theta,x^*,n)}\|_{\infty,[0,n]} \leq M(\alpha, D), & (4.25) \\ \theta \in \mathcal{U}(\alpha, n, D) \Rightarrow \forall x \in \mathcal{X}^\theta,\ \|T^*_{x^*(\theta,x,n)} - T^\theta_x\|_{\infty,[0,n]} \leq M(\alpha, D). & (4.26) \end{cases}$$

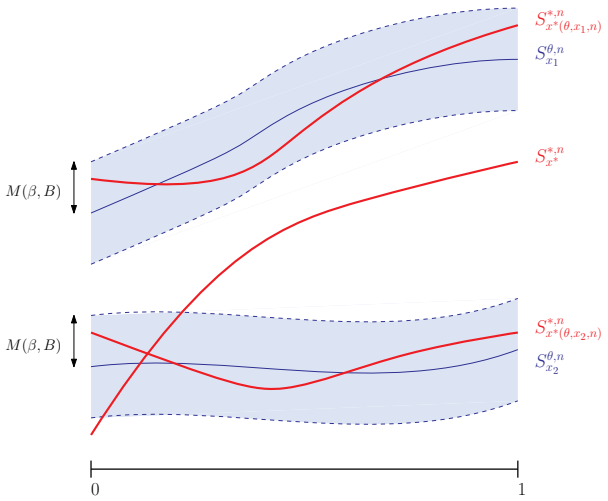

FIGURE 4.6 – Rescaled trends of a parameter in $\mathcal{U}(\beta, n, B)$. Every parameter trend is at bounded distance of at least one true trend. However, some true trends may be far from all parameter trends.

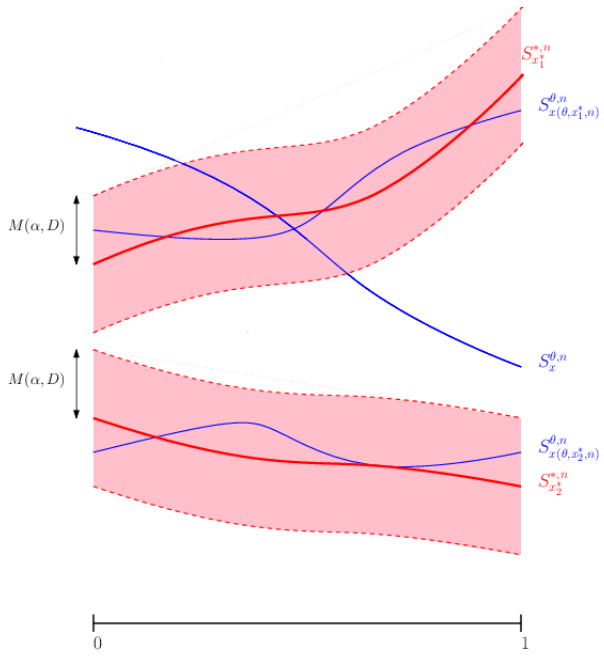

FIGURE 4.7 – Rescaled trends of a parameter in $\mathcal{T}(\alpha, n, D)$. Every true trend is at bounded distance of at least one parameter trend. However, some parameter trends may be far from all true trends.

Hence, for all $\alpha, \beta \in (0, 1]$, for all $M > 0$ and $D, B \geq M$ and for all n,

$$\Theta_n^{\mathrm{OK}}(M) \subset \mathcal{T}(\alpha, n, D) \cap \mathcal{U}(\beta, n, B) \subset \Theta_n^{\mathrm{OK}}(M(\alpha, D) \vee M(\beta, B)). \tag{4.27}$$

Proposition 4.3 is a direct consequence of Arzelà–Ascoli's theorem and of the following result.

Définition 4.6. *For all $\theta \in \Theta$, $n \in \mathbb{N}^*$ and $x \in \mathcal{X}^\theta$, let*

$$S_x^{\theta,n} : u \in [0, 1] \longmapsto T_x^\theta(nu)$$

be the trend T_x^θ rescaled from $[0, n]$ to $[0, 1]$. Likewise, define $S_{x^*}^{*,n}$ the true rescaled trend corresponding to the state x^*.

Theorem 4.5. *Let $\varepsilon \in (0, 1]$ and $D > 0$. Then the set*

$$\mathcal{S} := \bigcup_{n \geq n_0} \bigcup_{\theta \in \Theta} \bigcup_{(x,x^*) \in \mathcal{X}^\theta \times \mathcal{X}^* \text{ s.t. } |I_{n,D}(x,x^*,\theta)| \geq n\varepsilon} \{S_x^{\theta,n} - S_{x^*}^{*,n}\}$$

is relatively compact in the set of continuous functions $\left(\mathcal{C}^0([0, 1]), \|\cdot\|_\infty\right)$.

Remark. *This result, together with Arzelà–Ascoli's theorem, entails that $\mathcal{S}$ is uniformly equicontinuous in addition to being a bounded subset of $\mathbf{L}^\infty([0, 1])$. This will be used in Section 4.C.*

Proof. Let $\varepsilon \in (0, 1]$ and $D > 0$. Let us first give another representation of the elements of $\mathcal{S}$.

Lemma 4.5. *For any $S \in \mathcal{S}$, there exists $(u_1^S, \ldots, u_{d+1}^S) \in [0, 1]^{d+1}$ such that*

$$\begin{cases} \forall i \neq j, \quad |u_i^S - u_j^S| \geq \frac{\varepsilon}{4K(d+1)}, \\ \forall i, \quad |S(u_i^S)| \leq D. \end{cases}$$

Proof. For $S \in \mathcal{S}$, let $n(S) \geq n_0$, θ, $x \in \mathcal{X}^\theta$ and $x^* \in \mathcal{X}^*$ be such that $S = S_x^{\theta,n} - S_{x^*}^{*,n}$ and $|I_{n,D}(x, x^*, \theta)| \geq n\varepsilon$. Let us define $(u_1^S, \dots, u_{d+1}^S)$ iteratively. Let $\mathcal{A}_0 = I_{n,D}(x, x^*, D)$ and, for all $i \geq 1$,

— $t_i^S \in \mathcal{A}_{i-1}$;

— $\mathcal{A}_i = \mathcal{A}_{i-1} \setminus \bar{B}\left(t_i^S, \frac{\varepsilon}{4K(d+1)}n\right)$.

The closed ball $\bar{B}\left(t_i^S, \frac{\varepsilon}{4K(d+1)}n\right)$ contains at most $1 + \left\lfloor 2\frac{\varepsilon}{4K(d+1)}n \right\rfloor \leq 3\frac{\varepsilon}{4K(d+1)}n$ elements since $\frac{\varepsilon}{4K(d+1)}n \geq 1$. Thus, for all $i \geq 0$,

$$|\mathcal{A}_i| \geq n\frac{\varepsilon}{K}\left(1 - \frac{3}{4(d+1)}i\right).$$

In particular, $\mathcal{A}_i \neq \emptyset$ for all $i \in \{0, \dots, d+1\}$, which makes it possible to define $(t_i^S)_{1 \leq i \leq d+1}$. Taking $u_i^S = \frac{t_i^S}{n(S)}$ for all $i \in \{1, \dots, d+1\}$ concludes the proof. □

The next lemma is a straightforward consequence of the Lagrange form of the interpolation polynomial.

Lemma 4.6. *The mapping*

$$\begin{array}{ccc} \left\{(u_i)_i \in [0,1]^{d+1} \text{ s.t. } \inf_{i \neq j} |u_i - u_j| > 0\right\} \times \mathbb{R}^{d+1} & \longmapsto & \left(\mathcal{C}^0([0,1]), \|.\|_\infty\right) \\ (u_1, \dots, u_{d+1}, s_1, \dots, s_{d+1}) & \longmapsto & P_{u,s} \end{array} \tag{4.28}$$

is continuous, where $P_{u,s}$ is the only polynomial with degree at most d such that $P(u_i) = s_i$ for all $i \in \{1, \dots, d+1\}$.

To conclude, note that $\mathcal{S}$ is a subset of the image of the compact set

$$\{(u_i)_i \in [0,1]^{d+1} \text{ s.t. } \inf_{i \neq j} |u_i - u_j| \geq \varepsilon/(4K(d+1))\} \times [-D, D]^{d+1}$$

by the mapping (4.28). □

4.B.2 The MLE is in $\mathcal{T}(\alpha, n, D)$

The key idea of this section is that if one of the true trends is far from all parameter trends, then the observations coming from this true trend will significantly reduce the likelihood.

Let $\alpha \in (0,1)$, $n \geq 1$, $D > 0$ and $\theta \notin \mathcal{T}(\alpha, n, D)$, then

$$\left| \bigcup_{x^* \in \mathcal{X}^*} \bigcap_{x \in \mathcal{X}^\theta} I_{n,D}(x, x^*, \theta)^\complement \right| > n(1-\alpha)$$

with the notations of Section 4.B.1. In particular, there exists $x_{\mathcal{T}}^*(\theta) \in \mathcal{X}^*$ such that

$$\left| \bigcap_{x \in \mathcal{X}^\theta} I_{n,D}(x, x_{\mathcal{T}}^*(\theta), \theta)^\complement \right| \geq n\frac{1-\alpha}{K^*}. \tag{4.29}$$

Write $I_n^{\text{far}}(\theta) := \bigcap_{x\in\mathcal{X}^\theta} I_{n,D}(x, x^*_{\mathcal{T}}(\theta),\theta)^{\complement}$, then

$$\begin{aligned}\frac{1}{n}\ell_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n}\log p^\theta(Y_t|Y_1^{t-1})\\ &\le \log g(0) + \frac{1}{n}\sum_{t\in I_n^{\text{far}}(\theta)} \mathbf{1}_{X_t=x^*_{\mathcal{T}}(\theta)}\log\frac{g(\{D-|Z_t^{\max}|\}_+)}{g(0)}.\end{aligned} \tag{4.30}$$

We used the fact that under Assumption **(Amax)**,

$$\begin{aligned}p^\theta(Y_t|Y_1^{t-1}) &= \sum_{x\in\mathcal{X}^\theta} p^\theta(X_t=x|Y_1^{t-1})\gamma_x^\theta(Y_t - T_x^\theta(t))\\ &\le \sum_{x^*\in\mathcal{X}^*}\mathbf{1}_{X_t=x^*}\sup_{x\in\mathcal{X}^\theta}\gamma_x^\theta(Z_t + T^*_{x^*}(t) - T_x^\theta(t))\\ &\le \sum_{x^*\in\mathcal{X}^*}\mathbf{1}_{X_t=x^*}\sup_{x\in\mathcal{X}^\theta} g(\{|T^*_{x^*}(t) - T_x^\theta(t)| - |Z_t^{\max}|\}_+)\\ &\le \sum_{x^*\in\mathcal{X}^*}\mathbf{1}_{X_t=x^*} g(\inf_{x\in\mathcal{X}^\theta}\{|T^*_{x^*}(t) - T_x^\theta(t)| - |Z_t^{\max}|\}_+).\end{aligned}$$

Lemma 4.7. *For all $\theta\in\Theta$, $D>0$ and $x^*\in\mathcal{X}^*$, the set*

$$\bigcap_{x\in\mathcal{X}^\theta}\{u\in[0,n] : |T^*_{x^*}(t) - T_x^\theta(t)| > D\} \tag{4.31}$$

has at most $A := (d+1)^K$ connected components.

Note that $I_n^{\text{far}}(\theta) = (4.31)\cap\{1,\dots,n\}$.

Proof. The functions $(t\longmapsto T^*_{x^*}(t) - T_x^\theta(t))_{x\in\mathcal{X}^\theta}$ are polynomials whose degree is at most d. Their derivatives vanish at most $d-1$ times, and the set of times t where they are larger than D in absolute value is a union of segments containing either, a zero of their derivative, $+\infty$ or $-\infty$. Hence there are at most $d+1$ such segments. Thus, $\tilde{I}_{n,D,x^*}(\theta)$ is an intersection of at most K sets, each of them having at most $d+1$ connected components. Therefore, one may take $A=(d+1)^K$. □

Définition 4.7. *For all $n\in\mathbb{N}^*$, $D>0$, $x^*\in\mathcal{X}^*$ and $\theta\in\Theta$, write $J(n,D,x^*,\theta)$ the largest connected component of (4.31). In case of tie, choose the first one for the usual order in $\mathbb{R}$.*

Thus, by the pigeonhole principle and equation (4.29),

$$|J(n,D,x^*_{\mathcal{T}}(\theta),\theta)\cap\{1,\dots,n\}| \ge n\frac{1-\alpha}{AK^*}.$$

Write $J_n^{\text{far}}(\theta) := J(n,D,x^*_{\mathcal{T}}(\theta),\theta)\cap\{1,\dots,n\}$, then using equation (4.30) :

$$\frac{1}{n}\ell_n(\theta) \le \log g(0) + \frac{1}{n}\sum_{t\in J_n^{\text{far}}(\theta)}\mathbf{1}_{X_t=x^*_{\mathcal{T}}(\theta)}\log\frac{g(\{D-|Z_t^{\max}|\}_+)}{g(0)}. \tag{4.32}$$

Lemma 4.8. *Let $\delta>0$ and assume **(Aerg)**. Then, almost surely,*

$$\liminf_{n\to\infty}\inf_{\substack{S\subset\{1,\dots,n\}\\ S\ \text{segment}\\ |S|\ge\delta n}}\inf_{x^*\in\mathcal{X}^*}\frac{1}{n}\sum_{t\in S}\mathbf{1}_{X_t=x^*} \ge \frac{\delta\sigma_-}{4}.$$

By "segment", we mean a set of the form $[a,b]\cap\mathbb{Z}$ for some $a,b\in\mathbb{R}$.

Proof. The idea is to split $\{1,\dots,n\}$ into segments of size $\frac{\delta}{2}n$ and to control the infimum of the empirical mean over each segment. Each segment of size larger than δn contains at least one of those segments. The proof is detailled in Section 4.D.1. □

Applying Lemma 4.8 to $S = J_n^{\text{far}}(\theta)$, one gets that almost surely,

$$\liminf_{n\to\infty} \inf_{\theta\notin\mathcal{T}(\alpha,n,D)} \frac{1}{n} \sum_{t\in J_n^{\text{far}}(\theta)} \mathbf{1}_{X_t = x^*_{\mathcal{T}}(\theta)} \geq \frac{(1-\alpha)\sigma_-}{4AK^*}. \tag{4.33}$$

Lemma 4.9. *Let $\delta \in (0,1)$, $(U_t)_{t\geq 1}$ a sequence of i.i.d. non-positive integrable random variables and $(\delta_n)_{n\geq 1}$ a non-decreasing sequence of $[0,1]$-valued random variables such that almost surely, $\liminf_{n\to\infty}\delta_n \geq \delta$. For all $\beta\in[0,1]$, let us denote by $q_U(\beta)$ the β-quantile of U_1, i.e.*

$$q_U(\beta) = \inf\{u \text{ s.t. } \mathbb{P}(U_1 \leq u) \geq \beta\}.$$

Then, almost surely,

$$\limsup_{n\to\infty} \sup_{\substack{S\subset\{1,\dots,n\}\\ |S|\geq\delta_n n}} \frac{1}{n}\sum_{t\in S} U_t \leq \mathbb{E}[U_1 \mathbf{1}_{U_1 > q_U(1-\delta)}].$$

Equivalently, if $(V_t)_{t\geq 1}$ is a sequence of non-negative i.i.d. integrable random variables and $(\delta_n)_{n\geq 1}$ a non-increasing sequence of $[0,1]$-valued random variables such that $\limsup_{n\to\infty}\delta_n \leq \delta$ a.s., one has almost surely

$$\limsup_{n\to\infty} \sup_{\substack{S\subset\{1,\dots,n\}\\ |S|\leq\delta_n n}} \frac{1}{n}\sum_{t\in S} V_t \leq \mathbb{E}[V_1 \mathbf{1}_{V_1 \geq q_V(1-\delta)}].$$

Remark. *The supremum is taken over all subsets S, not only segments.*

Proof. Proof in Section 4.D.2. □

For all $t\geq 1$ and $D>0$, let $U_t^D = \log\frac{g(\{D-|Z_t^{\max}|\}_+)}{g(0)}$. U_t^D is non-positive by definition. Then, taking

$$\begin{cases} \delta = \frac{(1-\alpha)\sigma_-}{4AK^*}, \\ \delta_n = \inf\limits_{m\geq n} \inf\limits_{\theta\notin\mathcal{T}(\alpha,m,D)} \frac{1}{m} \sum\limits_{t\in J_m^{\text{far}}(\theta)} \mathbf{1}_{X_t = x^*_{\mathcal{T}}(\theta)}, \end{cases}$$

one has $\liminf_{n\to\infty}\delta_n \geq \delta$ by equation (4.33). Therefore, Lemma 4.9 combined with equation (4.32) implies that almost surely,

$$\limsup_{n\to\infty} \sup_{\theta\notin\mathcal{T}(\alpha,n,D)} \frac{1}{n}\ell_n(\theta) \leq \log g(0) + \mathbb{E}[U_1^D \mathbf{1}_{U_1^D \geq q_{U^D}(1-\frac{(1-\alpha)\sigma_-}{4AK^*})}].$$

Note that $U_t^D = f^D(Z_t^{\max})$ where $f^D : z\in\mathbb{R}_+ \longmapsto \log\frac{g(\{D-z\}_+)}{g(0)}$. f^D is non-decreasing, so for all x and q, $f^D(x) > f^D(q)$ implies $x > q$. Hence

$$\mathbb{P}\left(f^D(|Z_1^{\max}|) > f^D(q_{|Z^{\max}|}(1-\delta))\right) \leq \mathbb{P}\left(|Z_1^{\max}| > q_{|Z^{\max}|}(1-\delta)\right) \leq \delta.$$

In other words, $\mathbb{P}(U_1^D \leq f^D(q_{|Z^{\max}|}(1-\delta))) \geq 1-\delta$, hence $q_{U^D}(1-\delta) \leq f^D(q_{|Z^{\max}|}(1-\delta))$ by definition of quantiles. Thus, for all $z \geq 0$,

$$\mathbf{1}_{z \geq q_{|Z^{\max}|}(1-\delta)} \leq \mathbf{1}_{f^D(z) \geq f^D(q_{|Z^{\max}|}(1-\delta))} \leq \mathbf{1}_{f^D(z) \geq q_{U^D}(1-\delta)}$$

since f^D is non-decreasing. Therefore,

$$\mathbb{E}[U_1^D \mathbf{1}_{U_1^D \geq q_{U^D}(1-\frac{(1-\alpha)\sigma_-}{4AK^*})}] \leq \mathbb{E}[U_1^D \mathbf{1}_{|Z_1^{\max}| \geq q_{|Z^{\max}|}(1-\frac{(1-\alpha)\sigma_-}{4AK^*})}].$$

Then, for all $\delta > 0$, the monotone convergence theorem applied to the right-hand side entails

$$\mathbb{E}[U_1^D \mathbf{1}_{U_1^D \geq q_{U^D}(1-\delta)}] \underset{D \to +\infty}{\longrightarrow} -\infty.$$

Thus, under the assumptions of Corollary 4.1, there exists $D(\alpha) < \infty$ such that

$$\limsup_{n \to \infty} \sup_{\theta \notin \mathcal{T}(\alpha, n, D(\alpha))} \frac{1}{n} \ell_n(\theta) \leq \ell(\theta^*) - 1,$$

so that almost surely, for n large enough,

$$\hat{\theta}_n \in \mathcal{T}(\alpha, n, D(\alpha)). \tag{4.34}$$

4.B.3 The MLE is in $\mathcal{U}(\beta, n, B)$

Let $\alpha, \beta \in (0,1)$, $n \geq \frac{4K(d+1)}{1-\alpha}$, $D > 0$, $B > M(\alpha, D)$ and $\theta \in \mathcal{T}(\alpha, n, D) \setminus \mathcal{U}(\beta, n, B)$. Since $\theta \notin \mathcal{U}(\beta, n, B)$, one of its trends is far from all true trends. The key of the proof is to show that removing this trend increases the likelihood of the observations. An interpretation is that the likelihood "expects" some observations to come from this trend because of **(Aerg)**, but that it never observes it.
By definition of $\mathcal{U}(\beta, n, B)$, there exists $x_{\mathcal{U}}(\theta) \in \mathcal{X}^\theta$ such that

$$\left| \bigcap_{x^* \in \mathcal{X}^*} I_{n,B}(x_{\mathcal{U}}(\theta), x^*, \theta)^{\complement} \right| > n(1-\beta). \tag{4.35}$$

Write $I_n^U(\theta) := \bigcap_{x^* \in \mathcal{X}^*} I_{n,B}(x_{\mathcal{U}}(\theta), x^*, \theta)^{\complement}$ and let θ^U be the parameter defined by

$$\begin{cases} \mathcal{X}^{\theta^U} = \mathcal{X}^\theta \setminus \{x_{\mathcal{U}}(\theta)\}, \\ \forall x \in \mathcal{X}^{\theta^U}, \quad \pi^{\theta^U}(x) = \frac{\pi^\theta(x)}{1-\pi^\theta(x_{\mathcal{U}}(\theta))}, \\ \forall x, x' \in \mathcal{X}^{\theta^U}, \quad Q^{\theta^U}(x,x') = \frac{Q^\theta(x,x')}{1-Q^\theta(x,x_{\mathcal{U}}(\theta))}, \\ \forall x \in \mathcal{X}^{\theta^U}, \quad \gamma_x^{\theta^U} = \gamma_x^\theta, \\ \forall x \in \mathcal{X}^{\theta^U}, \quad T_x^{\theta^U} = T_x^\theta. \end{cases}$$

By construction of Θ, $\theta^U \in \Theta$. Note that for all $x, x' \in \mathcal{X}^{\theta^U}$,

$$\begin{aligned} \pi^{\theta^U}(x) &= \mathbb{P}^\theta(X_1 = x \,|\, X_1 \neq x_{\mathcal{U}}(\theta)) \\ &= \mathbb{P}^\theta(X_1 = x \,|\, \forall t \geq 1, X_t \neq x_{\mathcal{U}}(\theta)) \end{aligned}$$

and

$$\begin{aligned} \forall s \geq 1, \quad Q^{\theta^U}(x,x') &= \mathbb{P}^\theta(X_{s+1} = x' \,|\, X_s = x, X_{s+1} \neq x_{\mathcal{U}}(\theta)) \\ &= \mathbb{P}^\theta(X_{s+1} = x' \,|\, X_s = x, \forall t \geq 1, X_t \neq x_{\mathcal{U}}(\theta)), \end{aligned}$$

so that

$$\forall \in \sigma(Y_t \,|\, t \geq 1), \quad \mathbb{P}^{\theta^U}(A) = \mathbb{P}^{\theta}(A \,|\, \forall t \geq 1, X_t \neq x_{\mathcal{U}}(\theta)).$$

Then

$$\frac{1}{n}\ell_n(\theta^U) = \frac{1}{n}\sum_{t=1}^{n} \log p^{\theta}(Y_t \,|\, Y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^t),$$

with the abuse of notation $x \in X_1^t \Leftrightarrow (\exists s \in \{1, \dots, t\}\ X_s = x)$. By Assumption **(Aerg)**,

$$\begin{aligned} p^{\theta}(Y_t \,|\, Y_1^{t-1}) &= p^{\theta}(Y_t \,|\, X_t = x_{\mathcal{U}}(\theta)) p^{\theta}(X_t = x_{\mathcal{U}}(\theta) \,|\, Y_1^{t-1}) \\ &\quad + p^{\theta}(Y_t \,|\, X_t \neq x_{\mathcal{U}}(\theta), Y_1^{t-1}) p^{\theta}(X_t \neq x_{\mathcal{U}}(\theta) \,|\, Y_1^{t-1}) \\ &\leq (1-\sigma_-) p^{\theta}(Y_t \,|\, X_t = x_{\mathcal{U}}(\theta)) + (1-\sigma_-) p^{\theta}(Y_t \,|\, X_t \neq x_{\mathcal{U}}(\theta), Y_1^{t-1}), \end{aligned}$$

hence

$$\frac{1}{n}\ell_n(\theta) - \underbrace{\frac{1}{n}\sum_{t=1}^{n} \log p^{\theta}(Y_t \,|\, X_t \neq x_{\mathcal{U}}(\theta), Y_1^{t-1})}_{(i)} \leq \log(1-\sigma_-) + \underbrace{\frac{1}{n}\sum_{t=1}^{n} \log\left(1 + \frac{p^{\theta}(Y_t \,|\, X_t = x_{\mathcal{U}}(\theta))}{p^{\theta}(Y_t \,|\, X_t \neq x_{\mathcal{U}}(\theta), Y_1^{t-1})}\right)}_{(ii)}.$$

The next steps are :

— Prove that (i) is close to $\frac{1}{n}\ell_n(\theta^U)$ for large enough n.

— Prove that (ii) goes to zero uniformly in $\theta \in \mathcal{T}(\alpha, n, D) \setminus \mathcal{U}(\beta, n, B)$.

First step : controlling (i) We shall prove that for an adequate choice of β and B, one has almost surely

$$\limsup_{n\to\infty} \sup_{\theta \in \mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)} \left|(i) - \frac{1}{n}\ell_n(\theta^U)\right| \leq \frac{-\log(1-\sigma_-)}{3}.$$

The following forgetting property allows to control what happens in the distant past.

Lemma 4.10. *For all $t \geq 1$, $\theta \in \Theta$, for any probability measures μ and ν on $\mathcal{X}^{\theta}$, for all $x \in \mathcal{X}^{\theta}$ and Y_0^t,*

$$|\log p^{\theta}(Y_t \,|\, Y_0^{t-1}, X_t \neq x, X_0 \sim \mu) - \log p^{\theta}(Y_t \,|\, Y_0^{t-1}, X_t \neq x, X_0 \sim \nu)| \leq C\rho^t$$

where $\rho = 1 - \frac{\sigma_-}{1-\sigma_-}$ and $C = \frac{2}{\rho(1-\rho)^3}$.

Proof. Proof in Section 4.B.4. □

Let $a \in \mathbb{N}^*$. It follows from Lemma 4.10 that for all t, almost surely,

$$\left|\log p^{\theta}(Y_t \,|\, Y_1^{t-1}, X_t \neq x_{\mathcal{U}}(\theta)) - \log p^{\theta}(Y_t \,|\, Y_1^{t-1}, X_t \neq x_{\mathcal{U}}(\theta), x_{\mathcal{U}}(\theta) \notin X_1^{t-a})\right| \leq C\rho^a. \tag{4.36}$$

It remains to add X_{t-a+1}^{t-1} to the conditioning. This is the goal of the following lemma.

Lemma 4.11. *Assume* ***(Amax)*** *and* ***(Amin)****. Then for all* $a \in \mathbb{N}^*$,

$$\begin{aligned}\sup_{\theta\in\mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)}\Bigg|\frac{1}{n}\sum_{t=1}^{n}\log p^\theta(Y_t|Y_1^{t-1},X_t\neq x_{\mathcal{U}}(\theta),x_{\mathcal{U}}(\theta)\notin X_1^{t-a})\\ -\frac{1}{n}\sum_{t=1}^{n}\log p^\theta(Y_t|Y_1^{t-1},x_{\mathcal{U}}(\theta)\notin X_1^t)\Bigg|\\ \leq\frac{2aK^2}{\sigma_-^3}\left(\beta+\frac{1}{n}\sum_{i=1}^{n-1}\left(1\wedge\frac{g(\{B-|Z_i^{\max}|\}_+)}{m(|Z_i^{\max}|+M(\alpha,D))}\right)\right)\\ =:\frac{2aK^2}{\sigma_-^3}\left(\beta+\frac{1}{n}\sum_{i=1}^{n-1}h_U(B,Z_i^{\max})\right).\end{aligned}$$

Proof. [Overview of the proof] First, show that for all $\theta\in\mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)$, $t\leq n$, $a\in\mathbb{N}^*$ and $y_1^t\in\mathcal{Y}^t$,

$$\begin{aligned}&\left|\log p^\theta(y_t|y_1^{t-1},X_t\neq x_{\mathcal{U}}(\theta),x_{\mathcal{U}}(\theta)\notin X_1^{t-a})-\log p^\theta(y_t|y_1^{t-1},x_{\mathcal{U}}(\theta)\notin X_1^t)\right|\\ &\quad\leq\frac{2}{\sigma_-}p^\theta\left(x_{\mathcal{U}}(\theta)\in X_{(t-a+1)\vee1}^{t-1}|y_1^{t-1},X_t\neq x_{\mathcal{U}}(\theta),x_{\mathcal{U}}(\theta)\notin X_1^{t-a}\right)\qquad(4.37)\\ &\quad\leq\frac{2K^2}{\sigma_-^3}\sum_{i=(t-a+1)\vee1}^{t-1}\frac{\gamma^\theta_{x_{\mathcal{U}}(\theta)}(y_i-T^\theta_{x_{\mathcal{U}}(\theta)}(i))}{\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_x(y_i-T^\theta_x(i))}.\qquad(4.38)\end{aligned}$$

Then, under **(Amax)** and **(Amin)**, for all $\theta\in\mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)$ and $i\leq n$

$$\frac{\gamma^\theta_{x_{\mathcal{U}}(\theta)}(Y_i-T^\theta_{x_{\mathcal{U}}(\theta)}(i))}{\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_x(Y_i-T^\theta_x(i))}\leq\begin{cases}1\wedge\dfrac{g(\{B-|Z_i^{\max}|\}_+)}{m(|Z_i^{\max}|+M(\alpha,D))}=:h_U(B,Z_i^{\max}) & \text{if } i\in I_n^U(\theta),\\ 1 & \text{if } i\notin I_n^U(\theta).\end{cases}\qquad(4.39)$$

so that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{\gamma^\theta_{x_{\mathcal{U}}(\theta)}(Y_i-T^\theta_{x_{\mathcal{U}}(\theta)}(i))}{\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_x(Y_i-T^\theta_x(i))}&\leq\frac{1}{n}\sum_{i\notin I_n^U(\theta)}1+\frac{1}{n}\sum_{i\in I_n^U(\theta)}h_U(B,Z_i^{\max})\\ &\leq\beta+\frac{1}{n}\sum_{i=1}^{n}h_U(B,Z_i^{\max})\end{aligned}$$

since $|I_n^U(\theta)|>n(1-\beta)$ by Equation (4.35) and $h_U(b,z)\geq0$ for all $b,z\geq0$. The lemma follows by summing equation (4.38) over t. The details of the proof can be found in Section 4.B.4. □

Thus, equation (4.36) and Lemma 4.11 imply

$$\limsup_{n\to\infty}\sup_{\theta\in\mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)}\left|(i)-\frac{1}{n}\ell_n(\theta^U)\right|\leq C\rho^a+\frac{2aK^2}{\sigma_-^3}\left(\beta+\mathbb{E}^*\left[h_U(B,Z_i^{\max})\right]\right).$$

Now choose a large enough so that

$$C\rho^a\leq\frac{-\log(1-\sigma_-)}{9},$$

then β such that

$$\frac{2aK^2}{\sigma_-^3}\beta \leq \frac{-\log(1-\sigma_-)}{9}.$$

Finally, note that $0 \leq h_U(b,z) \leq 1$ for all $b, z \geq 0$ and that $h_U(b,z) \longrightarrow 0$ when $b \longrightarrow \infty$ for all z, so that by the dominated convergence theorem, there exists B such that

$$\frac{2aK^2}{\sigma_-^3}\mathbb{E}^*\left[h_U(B, Z_i^{\max})\right] \leq \frac{-\log(1-\sigma_-)}{9},$$

which ensures that for all $\alpha \in (0,1)$ and $D > 0$, there exists $\beta \in (0,1)$ and $B > 0$ such that

$$\limsup_{n\to\infty} \sup_{\theta \in \mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)} \left|(i) - \frac{1}{n}\ell_n(\theta^U)\right| \leq \frac{-\log(1-\sigma_-)}{3}.$$

This concludes the proof of the first step.

Second step : controlling (ii)

Lemma 4.12. *Assume* ***(Amax)*** *and* ***(Amin)****. Then*

$$\begin{aligned}
&\sup_{\theta \in \mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)} (ii) \\
&\quad \leq \frac{1}{n}\sum_{t=1}^{n} \log\left(1 + \frac{g(\{B - |Z_t^{\max}|\}_+)}{\sigma_- m(|Z_t^{\max}| + M(\alpha,D))}\right) \qquad (4.40)\\
&\qquad + \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{t \notin I_n^U(\theta)} \log\left(\frac{g(0) + \sigma_- m(|Z_t^{\max}| + M(\alpha,D))}{g(\{B - |Z_t^{\max}|\}_+) + \sigma_- m(|Z_t^{\max}| + M(\alpha,D))}\right) \qquad (4.41)\\
&\quad =: \frac{1}{n}\sum_{t=1}^{n} h'_U(B, Z_t^{\max}) + \frac{1}{n}\sum_{t \notin I_n^U(\theta)} V_t^B.
\end{aligned}$$

Proof. We show that

$$\frac{p^\theta(Y_t \,|\, X_t = x_{\mathcal{U}}(\theta))}{p^\theta(Y_t \,|\, X_t \neq x_{\mathcal{U}}(\theta), Y_1^{t-1})} \leq \begin{cases} \dfrac{g(\{B - |Z_t^{\max}|\}_+)}{\sigma_- m(|Z_t^{\max}| + M(\alpha,D))} & \text{if } t \in I_n^U(\theta), \\ \dfrac{g(0)}{\sigma_- m(|Z_t^{\max}| + M(\alpha,D))} & \text{if } t \notin I_n^U(\theta). \end{cases}$$

The lemma follows by summing over t. The details of the proof can be found in Section 4.B.4. □

Note that under Assumption **(Aint)**,

$$\mathbb{E}^*[-\log m(|Z_t^{\max}| + M(\alpha,D))] < \infty.$$

Hence,

$$\begin{aligned}
\mathbb{E}^*|h'_U(0, Z_t^{\max})| &\leq \mathbb{E}^*\left[\{\log\left(\sigma_- m\left(|Z_t^{\max}| + M(\alpha,D)\right) + g(0)\right)\}_+\right] \\
&\quad + \mathbb{E}^*[-\log(\sigma_- m(|Z_t^{\max}| + M(\alpha,D)))] \\
&\leq \log 2 + |\log g(0)| + \mathbb{E}^*[-\log m(|Z_t^{\max}| + M(\alpha,D))] - \log\sigma_- \\
&< \infty.
\end{aligned}$$

Thus, since $b \longmapsto h'_U(b,z)$ is nonincreasing and converges to zero when $b \longrightarrow \infty$ for all z, the dominated convergence theorem together with the law of large numbers imply that there exists B such that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} h'_U(B, Z_t^{\max}) \leq \frac{-\log(1-\sigma_-)}{6}.$$

Then, apply Lemma 4.9 to the i.i.d. non-negative random variables $(V_t^B)_{t\geq 1}$ using the fact that $|I_n^U(\theta)| > n(1-\beta)$, which yields

$$\limsup_{n\to\infty} \sup_{\theta\in\mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)} \frac{1}{n}\sum_{t\notin I_n^U(\theta)} V_t^B \leq \mathbb{E}^*[V_1^B \mathbf{1}_{V_1^B \geq q_{V^B}(1-\beta)}].$$

Note that

$$\mathbb{E}^* V_1^B \leq \log((1+\sigma_-)g(0)) - \log\sigma_- + \mathbb{E}^*[-\log m(|Z_t^{\max}| + M(\alpha, D))],$$

which is finite thanks to **(Aint)**. Thus,

$$\mathbb{E}^*[V_1^B \mathbf{1}_{V_1^B \geq q_{V^B}(1-\beta)}] \underset{\beta\to 0}{\longrightarrow} 0,$$

so that there exists β such that

$$\limsup_{n\to\infty} \sup_{\theta\in\mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)} \frac{1}{n}\sum_{t\notin I_n^U(\theta)} V_t^B \leq \frac{-\log(1-\sigma_-)}{6}.$$

Hence, we proved that there exists $\beta(\alpha, D) \in (0,1)$ and $B(\alpha, D) > 0$ such that

$$\limsup_{n\to\infty} \sup_{\theta\in\mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)} (ii) \leq \frac{-\log(1-\sigma_-)}{3},$$

which ends the second step.
Putting together the results of the two steps, one gets that for all $\alpha \in (0,1)$ and $D > 0$, there exists $\beta \in (0,1)$ and $B > 0$ such that almost surely,

$$\limsup_{n\to\infty} \left(\sup_{\theta\in\mathcal{T}(\alpha,n,D)\setminus\mathcal{U}(\beta,n,B)} \frac{1}{n}\ell_n(\theta) - \sup_{\theta\in\Theta} \frac{1}{n}\ell_n(\theta) \right) \leq \frac{\log(1-\sigma_-)}{3} < 0,$$

so that for n large enough, $\hat{\theta}_n \notin (\mathcal{T}(\alpha,n,D) \setminus \mathcal{U}(\beta,n,B))$.
Together with Section 4.B.2, this implies that for all $\alpha \in (0,1)$, there exists $\beta \in (0,1)$ and $D, B > 0$ such that $\hat{\theta}_n \in \mathcal{T}(\alpha,n,D) \cap \mathcal{U}(\beta,n,B)$ for n large enough, which entails Theorem 4.3 by equation (4.27).

4.B.4 Proofs

Proof of Lemma 4.10

We shall use the inequality

$$|\log C_\mu - \log C_\nu| \leq \frac{|C_\mu - C_\nu|}{C_\mu \wedge C_\nu}$$

with $C_\mu = p^\theta(Y_t \mid Y_0^{t-1}, X_t \neq x, X_0 \sim \mu)$ and $C_\nu = p^\theta(Y_t \mid Y_0^{t-1}, X_t \neq x, X_0 \sim \nu)$.

$$\begin{aligned}|C_\mu - C_\nu| &= \left| \frac{p^\theta(Y_t, X_t \neq x \mid Y_0^{t-1}, X_0 \sim \mu)}{p^\theta(X_t \neq x \mid Y_0^{t-1}, X_0 \sim \mu)} - \frac{p^\theta(Y_t, X_t \neq x \mid Y_0^{t-1}, X_0 \sim \nu)}{p^\theta(X_t \neq x \mid Y_0^{t-1}, X_0 \sim \nu)} \right| \\ &=: \left| \frac{B_\mu}{A_\mu} - \frac{B_\nu}{A_\nu} \right|.\end{aligned}$$

One has

$$\begin{aligned}B_\mu &= \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x') p^\theta(X_t = x' \mid Y_0^{t-1}, X_0 \sim \mu) \\ &= \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x') \sum_{x'' \in \mathcal{X}^\theta} Q^\theta_{x''x'} p^\theta(X_{t-1} = x'' \mid Y_0^{t-1}, X_0 \sim \mu),\end{aligned}$$

which yields

$$\sigma_- \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x') \leq B_\mu \leq (1 - \sigma_-) \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x')$$

and the same result holds for B_ν. Besides,

$$\begin{aligned}A_\mu &= \sum_{x \neq x'} p^\theta(X_t = x' \mid Y_0^{t-1}, X_0 \sim \mu) \\ &= \sum_{x' \neq x} \sum_{x'' \in \mathcal{X}^\theta} Q^\theta_{x''x'} p^\theta(X_{t-1} = x'' \mid Y_0^{t-1}, X_0 \sim \mu).\end{aligned}$$

Hence,

$$\sigma_- \leq A_\mu \leq 1 - \sigma_-$$

and the same result holds for A_ν. Then, letting $\phi_\mu(x') = p^\theta(X_{t-1} = x' \mid Y_0^{t-1}, X_0 \sim \mu)$, we get, using the above expressions :

$$\begin{aligned}|A_\mu - A_\nu| &\leq (1 - \sigma_-) \|\phi_\mu - \phi_\nu\|_1 \\ |B_\mu - B_\nu| &\leq (1 - \sigma_-) \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x') \|\phi_\mu - \phi_\nu\|_1.\end{aligned}$$

Thus,

$$\begin{aligned}|C_\mu - C_\nu| &= \left| \frac{B_\mu}{A_\mu} - \frac{B_\nu}{A_\nu} \right| \\ &\leq \frac{1}{A_\mu A_\nu} \left(B_\mu |A_\mu - A_\nu| + A_\mu |B_\mu - B_\nu| \right) \\ &\leq \frac{2(1 - \sigma_-)^2}{\sigma_-^2} \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x') \|\phi_\mu - \phi_\nu\|_1.\end{aligned}$$

Furthermore,

$$\frac{1}{C_\mu \wedge C_\nu} \leq \frac{(1-\sigma_-)}{\sigma_- \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x')}$$

Finally,

$$|\log C_\mu - \log C_\nu| \leq \frac{2}{(1-\rho)^3} \|\phi_\mu - \phi_\nu\|_1.$$

It remains to prove that $\|\phi_\mu - \phi_\nu\|_1 \leq \rho^{t-1}$, which follows from the geometric ergodicity of the HMM. See for instance Corollary 1 of Douc et al. (2004) or Proposition 2.1 of De Castro et al. (2017).

Proof of Lemma 4.11

Proof of equation (4.37). For all $t \geq 1$ and $y_1^t \in \mathbb{R}^t$,

$$\begin{aligned}
&p^\theta(y_t \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) \\
&\quad = p^\theta(y_t \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t}) p^\theta(x_{\mathcal{U}}(\theta) \notin X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) \\
&\qquad + p^\theta(y_t \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta), x_{\mathcal{U}}(\theta) \in X_{(t-a+1)\vee 1}^{t-1}) \\
&\qquad\quad \times p^\theta(x_{\mathcal{U}}(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)),
\end{aligned}$$

so that

$$\begin{aligned}
&|p^\theta(y_t \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) - p^\theta(y_t \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t})| \\
&\leq p^\theta(x_{\mathcal{U}}(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) \\
&\quad \times \Big(p^\theta(y_t \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t}) + p^\theta(y_t \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta), x_{\mathcal{U}}(\theta) \in X_{t-a+1}^{t-1}) \Big) \\
&\leq 2p^\theta(x_{\mathcal{U}}(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) \sum_{x \in \mathcal{X}^\theta \setminus \{x_{\mathcal{U}}(\theta)\}} \gamma_x^\theta(y_t - T_x^\theta(t)).
\end{aligned}$$

In addition, under **(Aerg)**,

$$\begin{aligned}
&p^\theta(y_t \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) \\
&\qquad = \sum_{x \in \mathcal{X}^\theta \setminus \{x_{\mathcal{U}}(\theta)\}} p^\theta(y_t | X_t = x) p^\theta(X_t = x \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}) \\
&\qquad \geq \sigma_- \sum_{x \in \mathcal{X}^\theta \setminus \{x_{\mathcal{U}}(\theta)\}} \gamma_x^\theta(y_t - T_x^\theta(t))
\end{aligned}$$

and the same holds for $p^\theta(y_t \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t})$, so that using that $|\log x - \log y| \leq \frac{|x-y|}{x \wedge y}$ for all $x, y > 0$, we obtain that for all $t \geq 1$ and $y_1^t \in \mathbb{R}^t$

$$\begin{aligned}
&|\log p^\theta(Y_t \,|\, X_1^{t-a} \neq x_{\mathcal{U}}(\theta), X_t \neq x_{\mathcal{U}}(\theta), Y_1^{t-1}) - \log p^\theta(Y_t \,|\, X_1^{t} \neq x_{\mathcal{U}}(\theta), Y_1^{t-1})| \\
&\qquad \leq \frac{2}{\sigma_-} p^\theta(x_{\mathcal{U}}(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)). \quad (4.42)
\end{aligned}$$

Proof of equation (4.38)**.** By union bound,

$$
\begin{aligned}
&p^\theta(x_{\mathcal{U}}(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) \\
&\leq \sum_{i=(t-a+1)\vee 1}^{t-1} p^\theta(X_i = x_{\mathcal{U}}(\theta) \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) \\
&= \sum_{i=(t-a+1)\vee 1}^{t-1} \sum_{x_{i-1}, x_{i+1} \in \mathcal{X}^\theta} p^\theta(X_i = x_{\mathcal{U}}(\theta) \,|\, y_i, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \\
&\qquad \times p^\theta(X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1} \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) \\
&\leq \sum_{i=(t-a+1)\vee 1}^{t-1} \sum_{x_{i-1}, x_{i+1} \in \mathcal{X}^\theta} p^\theta(X_i = x_{\mathcal{U}}(\theta) \,|\, y_i, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \\
&= \sum_{i=(t-a+1)\vee 1}^{t-1} \sum_{x_{i-1}, x_{i+1} \in \mathcal{X}^\theta} \frac{p^\theta(X_i = x_{\mathcal{U}}(\theta) \,|\, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \gamma^\theta_{x_{\mathcal{U}}(\theta)}(y_i - T^\theta_{x_{\mathcal{U}}(\theta)}(i))}{\sum_{x \in \mathcal{X}^\theta} p^\theta(X_i = x \,|\, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \gamma^\theta_x(y_i - T^\theta_x(i))}.
\end{aligned}
$$

Using the Markov property and **(Aerg)**, for all $x_{i-1}, x_{i+1} \in \mathcal{X}^\theta$,

$$
p^\theta(X_i = x \,|\, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \in [\sigma_-^2, 1].
$$

Hence,

$$
\begin{aligned}
&p^\theta(x_{\mathcal{U}}(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_{\mathcal{U}}(\theta) \notin X_1^{t-a}, X_t \neq x_{\mathcal{U}}(\theta)) \\
&\qquad \leq \sum_{i=(t-a+1)\vee 1}^{t-1} \sum_{x_{i-1}, x_{i+1} \in \mathcal{X}^\theta} \frac{\gamma^\theta_{x_{\mathcal{U}}(\theta)}(y_i - T^\theta_{x_{\mathcal{U}}(\theta)}(i))}{\sigma_-^2 \sum_{x \in \mathcal{X}^\theta} \gamma^\theta_x(y_i - T^\theta_x(i))} \\
&\qquad \leq \frac{K^2}{\sigma_-^2} \sum_{i=(t-a+1)\vee 1}^{t-1} \frac{\gamma^\theta_{x_{\mathcal{U}}(\theta)}(y_i - T^\theta_{x_{\mathcal{U}}(\theta)}(i))}{\sum_{x \in \mathcal{X}^\theta} \gamma^\theta_x(y_i - T^\theta_x(i))}
\end{aligned}
$$

which concludes the proof.

Proof of equation (4.39)**.** This quantity is always bounded by 1 since all terms are nonnegative. In addition, under Assumptions **(Amax)** and **(Amin)**, for all $i \in I_n^U(\theta)$,

$$
\begin{aligned}
\frac{\gamma^\theta_{x_{\mathcal{U}}(\theta)}(Y_i - T^\theta_{x_{\mathcal{U}}(\theta)}(i))}{\sum_{x \in \mathcal{X}^\theta} \gamma^\theta_x(Y_i - T^\theta_x(i))} &\leq \sum_{x \in \mathcal{X}^*} \mathbf{1}_{X_i = x^*} \left(1 \wedge \frac{\gamma^\theta_{x_{\mathcal{U}}(\theta)}(Z_i + T^*_{x^*}(i) - T^\theta_{x_{\mathcal{U}}(\theta)}(i))}{\sup_{x \in \mathcal{X}^\theta} \gamma^\theta_x(Z_i + T^*_{x^*}(i) - T^\theta_x(i))} \right) \\
&\leq 1 \wedge \frac{g\left(\left\{ \inf_{x^* \in \mathcal{X}^*} |T^*_{x^*}(i) - T^\theta_{x_{\mathcal{U}}(\theta)}(i)| - |Z_i^{\max}| \right\}_+ \right)}{m\left(|Z_i^{\max}| + \sup_{x^* \in \mathcal{X}^*} \inf_{x \in \mathcal{X}^\theta} |T^*_{x^*}(i) - T^\theta_x(i)| \right)}.
\end{aligned}
$$

Since $\theta \in \mathcal{T}(\alpha, n, D) \setminus \mathcal{U}(\beta, n, B)$, Corollary 4.2 ensures that $\inf_{x \in \mathcal{X}^\theta} |T^*_{x^*}(i) - T^\theta_x(i)| \leq M(\alpha, D)$ for all $x^* \in \mathcal{X}^*$ and $i \in \{1, \dots, n\}$. Moreover, by definition of $I_n^U(\theta)$, one has $\inf_{x^* \in \mathcal{X}^*} |T^*_{x^*}(i) -$

$T^\theta_{x_\mathcal{U}(\theta)}(i)| \geq B$ for all $i \in I_n^U(\theta)$, so that

$$\frac{\gamma^\theta_{x_\mathcal{U}(\theta)}(Y_i - T^\theta_{x_\mathcal{U}(\theta)}(i))}{\sum_{x\in\mathcal{X}^\theta} \gamma^\theta_x(Y_i - T^\theta_x(i))} \leq 1 \wedge \frac{g(\{B - |Z_i^{\max}|\}_+)}{m(|Z_i^{\max}| + M(\alpha, D))}$$

for all $i \in I_n^U(\theta)$, which concludes the proof.

Proof of Lemma 4.12

Under Assumption **(Amax)**,

$$\begin{aligned}
p^\theta(Y_t \,|\, X_t = x_\mathcal{U}(\theta)) &= \gamma^\theta_{x_\mathcal{U}(\theta)}(Y_t - T^\theta_{x_\mathcal{U}(\theta)}(t)) \\
&= \sum_{x^*\in\mathcal{X}^*} \mathbf{1}_{X_t = x^*} \gamma^\theta_{x_\mathcal{U}(\theta)} \left(Z_t + T^*_{x^*}(t) - T^\theta_{x_\mathcal{U}(\theta)}(t)\right) \\
&\leq \sup_{x^*\in\mathcal{X}^*} g\left(\left\{|T^*_{x^*}(t) - T^\theta_{x_\mathcal{U}(\theta)}(t)| - |Z_t^{\max}|\right\}_+\right) \\
&\leq g\left(\left\{\inf_{x^*\in\mathcal{X}^*} |T^*_{x^*}(t) - T^\theta_{x_\mathcal{U}(\theta)}(t)| - |Z_t^{\max}|\right\}_+\right),
\end{aligned}$$

hence, for all $\theta \in \mathcal{T}(\alpha, n, D) \setminus \mathcal{U}(\beta, n, B)$,

$$p^\theta(Y_t \,|\, X_t = x_\mathcal{U}(\theta)) \leq \begin{cases} g(\{B - |Z_t^{\max}|\}_+) & \text{if } t \in I_n^U(\theta), \\ g(0) & \text{otherwise.} \end{cases} \tag{4.43}$$

On the other hand, under Assumptions **(Amin)** and **(Aerg)**,

$$\begin{aligned}
p^\theta(Y_t \,|\, X_t \neq x_\mathcal{U}(\theta), Y_1^{t-1}) &= \frac{p^\theta(Y_t, X_t \neq x_\mathcal{U}(\theta) \,|\, Y_1^{t-1})}{p^\theta(X_t \neq x_\mathcal{U}(\theta) \,|\, Y_1^{t-1})} \\
&\geq \sum_{x\in\mathcal{X}^\theta, x\neq x_\mathcal{U}(\theta)} p^\theta(Y_t \,|\, X_t = x, Y_1^{t-1}) p^\theta(X_t = x \,|\, Y_1^{t-1}) \\
&\geq \sigma_- \sum_{x\in\mathcal{X}^\theta, x\neq x_\mathcal{U}(\theta)} p^\theta(Y_t \,|\, X_t = x) \\
&= \sigma_- \sum_{x\in\mathcal{X}^\theta, x\neq x_\mathcal{U}(\theta)} \gamma^\theta_x(Y_t - T^\theta_x(t)) \\
&= \sigma_- \sum_{x^*\in\mathcal{X}^*} \mathbf{1}_{X_t = x^*} \sum_{x\in\mathcal{X}^\theta, x\neq x_\mathcal{U}(\theta)} \gamma^\theta_x(Z_t + T^*_{x^*}(t) - T^\theta_x(t)) \\
&\geq \sigma_- \inf_{x^*\in\mathcal{X}^*} \sup_{x\in\mathcal{X}^\theta, x\neq x_\mathcal{U}(\theta)} m(|Z_t^{\max}| + |T^*_{x^*}(t) - T^\theta_x(t)|) \\
&\geq \sigma_- m\left(|Z_t^{\max}| + \sup_{x^*\in\mathcal{X}^*} \inf_{x\in\mathcal{X}^\theta, x\neq x_\mathcal{U}(\theta)} |T^*_{x^*}(t) - T^\theta_x(t)|\right).
\end{aligned}$$

Using $\theta \notin \mathcal{T}(\alpha, n, D)$ and Corollary 4.2, for all $x^* \in \mathcal{X}^*$,

$$\inf_{x\in\mathcal{X}^\theta, x\neq x_\mathcal{U}(\theta)} |T^*_{x^*}(t) - T^\theta_x(t)| \leq M(\alpha, D).$$

For example, we can choose $x = x(\theta, x^*, n)$ (defined in Definition 4.5). We know that $x_{\mathcal{U}}(\theta) \neq x(\theta, x^*, n)$ because we chose $B > M(\alpha, D)$, so that $|T^*_{x^*}(i) - T^\theta_{x_{\mathcal{U}}(\theta)}(i)| > M(\alpha, D)$ for at least one $i \in \{1, \dots, n\}$, and because $|T^*_{x^*}(i) - T^\theta_{x(\theta,x^*,n)}(i)| \leq M(\alpha, D)$ for all $i \in \{1, \dots, n\}$. Therefore, for all $\theta \in \mathcal{T}(\alpha, n, D) \setminus \mathcal{U}(\beta, n, B)$,

$$p^\theta(Y_t \,|\, X_t \neq x_{\mathcal{U}}(\theta), Y_1^{t-1}) \geq \sigma_- m(|Z_t^{\max}| + M(\alpha, D)),$$

which concludes the proof together with equation (4.43).

4.C Integrated log-likelihood

In this section, we use the fact that the observed process $(Y_t)_{t\geq 1}$ may be replaced by the process $(Y_t - T^\theta_{B_t}(t), B_t)_{t\geq 1}$. While this process is not homogeneous, its distribution varies slowly over time. We take advantage of this property to show the uniform convergence of the log-likelihood by approximating $(Y_t - T^\theta_{B_t}(t), B_t)_{t\geq 1}$ by an homogenized process. The limit can be written as an integral of limits of log-likelihoods of homogeneous HMMs, hence the name *integrated log-likelihood*.

4.C.1 Convergence of the log-likelihood to the integrated log-likelihood

Assume **(Aerg)**, **(Amax)**, **(Amin)**, **(Aint)** and **(Areg)**. In this section we shall prove Theorem 4.4. Let $M > 0$.

The normalized log-likelihood associated with the HMM $(Y_t, B_t)_{t\geq 1}$ can be written as

$$\begin{aligned}\frac{1}{n}\ell_n^{(Y,B)}(\theta) &= \frac{1}{n}\log \sum_{\substack{x_1^n \text{ s.t.} \\ \forall t,\, \mathbf{b}^\theta(x_t)=B_t}} \pi^\theta(x_1)Q^\theta(x_1,x_2)\dots Q^\theta(x_{n-1},x_n)\prod_{t=1}^n \gamma^\theta_{x_t}(Y_t - T^\theta_{x_t}(t)) \\ &= \frac{1}{n}\log \sum_{\substack{x_1^n \text{ s.t.} \\ \forall t,\, \mathbf{b}^\theta(x_t)=B_t}} \pi^\theta(x_1)Q^\theta(x_1,x_2)\dots Q^\theta(x_{n-1},x_n)\prod_{t=1}^n \gamma^\theta_{x_t}\left(Z'_t - D^{\theta,n}_{x_t}\left(\frac{t}{n}\right)\right),\end{aligned} \tag{4.44}$$

with $D^{\theta,n}_x = S^{\theta,n}_x - S^{*,n}_{\mathbf{b}^\theta(x)}$ and $Z'_t := Y_t - T^*_{B_t}(t)$. Note that $Z'_t = Z_t + \Delta(X_t)$. Recall that $\mathcal{D}(M)$ is defined by Equation (4.12).

Theorem 4.5 implies that $\mathrm{Cl}(\mathcal{D}(M))$ is compact, where $\mathrm{Cl}(\cdot)$ denotes the closure with respect to the supremum norm topology, hence Proposition 4.1 holds. Together with Arzelà–Ascoli's theorem, this entails that $\mathcal{D}(M)$ is uniformly equicontinuous and uniformly bounded by M. Hence there exists a continuity modulus ν such that for all $\delta > 0$,

$$|s-t| \leq \delta \;\Rightarrow \sup_{n\geq 4K(d+1)} \sup_{\theta\in\Theta_n^{\mathrm{OK}}(M)} \sup_{x\in\mathcal{X}^\theta} |D^{\theta,n}_x(s) - D^{\theta,n}_x(t)| \leq \nu(\delta). \tag{4.45}$$

Définition 4.8 (Log-likelihood of the homogenized process)**.** *For each $\eta > 0$ and $\theta \in \Theta_n^{OK}(M)$, let*

$$\begin{aligned}\frac{1}{n}\ell_n^{(Y,B)}[\eta](\theta) := \frac{1}{n}\log \sum_{x_1^n \;s.t.\; \forall t,\, \mathbf{b}^\theta(x_t)=B_t} &\pi^\theta(x_1)Q^\theta(x_1,x_2)\dots Q^\theta(x_{n-1},x_n) \\ &\times \prod_{t=1}^n \gamma^\theta_{x_t}\left(Z'_t - D^{\theta,n}_{x_t}\left(\eta\left\lfloor\frac{t}{\eta n}\right\rfloor\right)\right),\end{aligned}$$

be the normalized log-likelihood of the process where each residual trend is made constant over segments of length η.

Remark. *This quantity is indeed a log-likelihood : one has* $\frac{1}{n}\ell_n^{(Y,B)}[\eta](\theta) = \frac{1}{n}\ell_n^{(Y,B)}(\theta[N,n])$, *where the parameter* $\theta[N,n]$ *is defined by* $\pi^{\theta[N,n]} = \pi^\theta$, $Q^{\theta[N,n]} = Q^\theta$, $\gamma^{\theta[N,n]} = \gamma^\theta$, $\mathbf{b}^{\theta[N,n]} = \mathbf{b}^\theta$ *and*

$$\forall x \in \mathcal{X}^\theta, \quad T_x^{\theta[N,n]}(t) = T^*_{\mathbf{b}^\theta(x)}(t) + D_x^{\theta,n}(\lfloor N\frac{t}{n}\rfloor). \tag{4.46}$$

However, $\theta[N,n]$ *has piecewise polynomial trends instead of polynomial trends, so that it does not belong to* Θ.

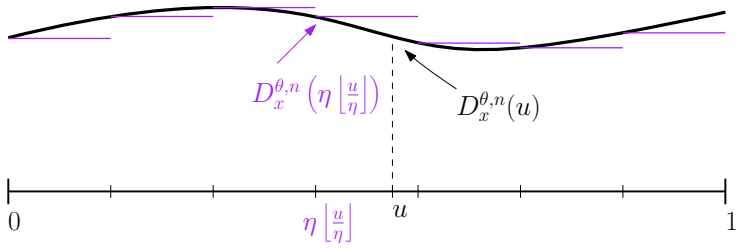

FIGURE 4.8 – Construction of the trends of the homogenized process.

$\frac{1}{n}\ell_n^{(Y,B)}[\eta](\theta)$ is an approximation of the log-likelihood of equation (4.44). Assumption **(Areg)** together with Equation (4.45) ensure that for all $\delta > 0$, $n \geq 1$, $\theta \in \Theta_n^{\mathrm{OK}}(M)$, $x \in \mathcal{X}^\theta$ and $t \in \{1,\dots,n\}$,

$$\gamma^\theta_{x_t}\left(Z'_t - D^{\theta,n}_{x_t}\left(\frac{t}{n}\right)\right) \in \left[e^{-L(|Z'_t|+M)\omega(\nu(\delta))}, e^{L(|Z'_t|+M)\omega(\nu(\delta))}\right]\gamma^\theta_{x_t}\left(Z'_t - D^{\theta,n}_{x_t}\left(\delta\left\lfloor\frac{t}{\delta n}\right\rfloor\right)\right),$$

hence, for all $\delta > 0$ and $n \geq 1$,

$$\begin{aligned}\sup_{\theta\in\Theta_n^{\mathrm{OK}}(M)}\left|\frac{1}{n}\ell_n^{(Y,B)}(\theta) - \frac{1}{n}\ell_n^{(Y,B)}[\delta](\theta)\right| &\leq \omega(\nu(\delta)) \times \frac{1}{n}\sum_{t=1}^n L(|Z'_t| + M)\\ &\leq \omega(\nu(\delta)) \times \frac{1}{n}\sum_{t=1}^n L(\|\Delta\|_\infty + M + |Z_t^{\max}|).\end{aligned} \tag{4.47}$$

Remark. *Under* ***(Areg)****, the law of large numbers entails that almost surely, for all* $N \geq 1$,

$$\begin{aligned}\limsup_{n\to+\infty}\sup_{\theta\in\Theta_n^{OK}(M)}&\left|\frac{1}{n}\ell_n^{(Y,B)}(\theta) - \frac{1}{n}\ell_n^{(Y,B)}\left[\frac{1}{N}\right](\theta)\right|\\ &\leq \omega\left(\nu\left(\frac{1}{N}\right)\right)\mathbb{E}^*[L(\|\Delta\|_\infty + M + |Z_1^{\max}|)].\end{aligned} \tag{4.48}$$

Recall the following notation. For all $K' \in \mathbb{N}^*$, for all K'-uple $\gamma = (\gamma_x)_{x\in[K']}$ of measurable functions and for all $\mathbf{D} = (D_x)_{x\in[K']} \in \mathbb{R}^{K'}$, let

$$\tau(\gamma, \mathbf{D}) := (z' \longmapsto \gamma_x(z' - D_x))_{x\in[K']}$$

the vector of functions γ translated by the vector $\mathbf{D}$.

Définition 4.9. *Let π be a probability measure on $[K']$, Q be a $K' \times K'$ transition matrix, γ be a vector of K' emission densities on $\mathbb{R}$ and $\mathbf{b}$ be a function $[K'] \longrightarrow \mathcal{B}^*$. Let $(X_t, (\tilde{Z}_t, \tilde{B}_t))_{t\geq 1}$ be a homogeneous HMM taking values in $[K'] \times (\mathbb{R} \times \mathcal{B}^*)$ with parameter $(\pi, Q, (\gamma_x \otimes \mathbf{1}_{\mathbf{b}(x)})_{x\in[K']})$. Denote by $\frac{1}{n}\ell_n^{hom}(\pi, Q, \gamma, \mathbf{b})\{(\tilde{z}, \tilde{b})_1^n\}$ (resp. $\ell^{hom}(Q, \gamma, \mathbf{b})$) the normalized log-likelihood of the parameter $(\pi, Q, \gamma, \mathbf{b})$ for the observations $(\tilde{z}, \tilde{b})_1^n$ (resp. the limit of the log-likelihood, if it exists), that is*

$$\frac{1}{n}\ell_n^{hom}(\pi, Q, \gamma, \mathbf{b})\{(\tilde{z}, \tilde{b})_1^n\} = \frac{1}{n}\log \sum_{x_1^n \in [K']^n} \pi(x_1)Q(x_1, x_2)\dots Q(x_{n-1}, x_n) \prod_{t=1}^{n} \gamma_{x_t}(\tilde{z}_t)\mathbf{1}_{\mathbf{b}(x_t)=\tilde{b}_t}$$

and

$$\ell^{hom}(Q, \gamma, \mathbf{b}) = \lim_{n\to\infty} \frac{1}{n}\ell_n^{hom}(\pi, Q, \gamma, \mathbf{b})\{(\tilde{Z}, \tilde{B})_1^n\}. \tag{4.49}$$

The following Lemma ensures the existence of the limit of the normalized log-likelihood in Definition 4.9 as well as its uniform continuity with respect to the parameter. It is a consequence of a result concerning homogeneous HMM stated in Douc et al. (2004).

Lemma 4.13. *Assume **(Amax)**, **(Amin)**, **(Aint)** and **(Areg)**. Let $K' \in \mathbb{N}^*$. Then, the following points hold.*

— *Almost surely, for all $Q \in \Sigma_{K'}^{\sigma^-}$, $\gamma \in \Gamma^{K'}$, $\mathbf{D} \in \mathbb{R}^{K'}$ and $\mathbf{b} : [K'] \longrightarrow \mathcal{B}^*$, the quantity*

$$\ell^{hom}(Q^\theta, \tau(\gamma^\theta, \mathbf{D}), \mathbf{b})$$

from Equation (4.49) exists and is finite almost surely under $\mathbb{P}^*$ when $(\tilde{Z}_t, \tilde{B}_t)_t = (Z'_t, B_t)_t$.

— *For all $K' \in \mathbb{N}^*$, the mapping*

$$(Q^\theta, \gamma^\theta, \mathfrak{D}, u, \mathbf{b}) \in \Sigma_{K'}^{\sigma^-} \times \Gamma^{K'} \times Cl(\mathcal{D}(M))^{K'} \times [0,1] \times (\mathcal{B}^*)^{K'} \longmapsto \ell^{hom}(Q^\theta, \tau(\gamma^\theta, \mathfrak{D}(u)), \mathbf{b}) \tag{4.50}$$

is continuous and its domain is compact, so that it is uniformly continuous.

— *Almost surely, for all $N \in \mathbb{N}^*$,*

$$\sup_{(\pi,Q,\gamma,\mathfrak{D},u,\mathbf{b})} \sup_{s\in\{0,\dots,(N-1)n\}} \left| \frac{1}{n}\ell_n^{hom}(\pi, Q, \tau(\gamma, \mathfrak{D}(u)), \mathbf{b})\{(Z', B)_{s+1}^{s+n}\} - \ell^{hom}(Q, \tau(\gamma, \mathfrak{D}(u)), \mathbf{b}) \right| \underset{n\to\infty}{\longrightarrow} 0$$

where the supremum is taken for $(\pi, Q, \gamma, \mathfrak{D}, u, \mathbf{b}) \in \Delta_{K'} \times \Sigma_{K'}^{\sigma^-} \times \Gamma^{K'} \times Cl(\mathcal{D}(M))^{K'} \times [0,1] \times (\mathcal{B}^*)^{K'}$.

Proof. Proof in Section 4.C.3 □

As a consequence, the family of functions

$$\bigcup_{K'=1}^{K} \{u \in [0,1] \longmapsto \ell^{\text{hom}}(Q, \tau(\gamma, \mathfrak{D}(u)), \mathbf{b})\}_{Q\in\Sigma_{K'}^{\sigma^-}, \gamma\in\Gamma^{K'}, \mathfrak{D}\in\mathcal{D}(M)^{K'}, \mathbf{b}\in(\mathcal{B}^*)^{K'}}$$

is uniformly equicontinuous, which ensures the following result.

Corollary 4.3 (Riemann approximation of the integral)**.** *The quantity*

$$R_N := \sup_{n\geq n_1(M)} \sup_{\theta\in\Theta_n^{OK}(M)} \left| \frac{1}{N}\sum_{i=0}^{N-1} \ell^{hom}\left(Q^\theta, \tau\left(\gamma^\theta, \mathfrak{D}^{\theta,n}\left(\frac{i}{N}\right)\right), \mathbf{b}^\theta\right) - \int_0^1 \ell^{hom}\left(Q^\theta, \tau\left(\gamma^\theta, \mathfrak{D}^{\theta,n}(u)\right), \mathbf{b}^\theta\right) du \right| \quad (4.51)$$

satisfies

$$R_N \underset{N\to+\infty}{\longrightarrow} 0.$$

The integrated log-likelihood ℓ^{int} from Definition 4.3 is continuous by uniform continuity of ℓ^{hom}. We may now prove the main result of this section, that is the convergence of the normalized log-likelihood to the integrated log-likelihood.

Proof of Theorem 4.4 By the triangle inequality and using Equations (4.51) and (4.47), for all $n \geq n_1(M)$ and $N \in \mathbb{N}^*$,

$$\begin{aligned}
&\sup_{\theta\in\Theta_n^{\text{OK}}(M)} \left| \frac{1}{n}\ell_n^{(Y,B)}(\theta) - \ell^{\text{int}}(Q^\theta, \gamma^\theta, \mathfrak{D}^{\theta,n}, \mathbf{b}^\theta)\right| \\
&\qquad \leq \omega\left(\nu\left(\frac{1}{N}\right)\right)\frac{1}{n}\sum_{t=1}^{n} L(\|\Delta\|_\infty + M + |Z_t^{\max}|) + R_N \\
&\qquad + \sup_{\theta\in\Theta_n^{\text{OK}}(M)} \left| \frac{1}{n}\ell_n^{(Y,B)}\left[\frac{1}{N}\right](\theta) - \frac{1}{N}\sum_{i=0}^{N-1} \ell^{\text{hom}}\left(Q^\theta, \tau\left(\gamma^\theta, \mathfrak{D}^{\theta,n}\left(\frac{i}{N}\right)\right), \mathbf{b}^\theta\right)\right|.
\end{aligned}$$

For the sake of simplicity, assume that $\frac{n}{N}$ is an integer. By Equation (4.46), for all $\theta \in \Theta_n^{\text{OK}}(M)$, there exists $\theta[N,n]$ such that

$$\frac{1}{n}\ell_n^{(Y,B)}\left[\frac{1}{N}\right](\theta) = \frac{1}{n}\ell_n^{(Y,B)}(\theta[N,n]),$$

so that

$$\begin{aligned}
\frac{1}{n}\ell_n^{(Y,B)}\left[\frac{1}{N}\right](\theta) &= \frac{1}{n}\sum_{i=0}^{N-1} \log p^{\theta[N,n]}\left((Y,B)_{1+i\frac{n}{N}}^{\frac{n}{N}+i\frac{n}{N}} \mid (Y,B)_1^{i\frac{n}{N}}\right) \\
&= \frac{1}{N}\sum_{i=0}^{N-1} \frac{1}{\frac{n}{N}}\ell_{\frac{n}{N}}^{\text{hom}}\left(\pi_{i\frac{n}{N}}^{\theta[N,n]}, Q^\theta, \tau\left(\gamma^\theta, \mathfrak{D}^{\theta,n}\left(\frac{i}{N}\right)\right), \mathbf{b}^\theta\right)\left\{(Z',B)_{1+i\frac{n}{N}}^{\frac{n}{N}+i\frac{n}{N}}\right\},
\end{aligned}$$

where $\pi_{i\frac{n}{N}}^{\theta[N,n]}$ is defined as the distribution of $X_{1+i\frac{n}{N}}$ conditionally to $(Y,B)_{1+i\frac{n}{N}}^{\frac{n}{N}+i\frac{n}{N}}$ under the parameter $\theta[N,n]$. Hence, Lemma 4.13 implies that almost surely,

$$\begin{aligned}
&\limsup_{n\to+\infty} \sup_{\theta\in\Theta_n^{\text{OK}}(M)} \left| \frac{1}{n}\ell_n^{(Y,B)}(\theta) - \ell^{\text{int}}(Q^\theta, \gamma^\theta, \mathfrak{D}^{\theta,n}, \mathbf{b}^\theta)\right| \\
&\qquad \leq \inf_{N\in\mathbb{N}^*}\left[\omega\left(\nu\left(\frac{1}{N}\right)\right)\mathbb{E}^*[L(\|\Delta\|_\infty + M + |Z_1^{\max}|)] + R_N\right] \\
&\qquad = 0.
\end{aligned}$$

The conclusion follows from Theorem 4.2.

4.C.2 Maximizers of the integrated log-likelihood and identifiability

In this section we prove Proposition 4.2. Assume **(Amax)**, **(Amin)**, **(Areg)**, **(Aid)** and **(Acentering)** and assume that $K^* = K$ is known. We identify $\mathcal{X}^\theta$, $\mathcal{X}^*$ and $[K]$ for all $\theta \in \Theta_K$.

Remark. *Assumptions **(Areg)**, **(Amax)** and **(Amin)** can be replaced by*

$$\begin{cases} \forall \theta \in \Theta, \quad \forall x \in \mathcal{X}^*, \quad z \in \mathbb{R} \longmapsto \gamma^\theta(z) \text{ is continuous,} \\ \forall x \in \mathcal{X}^*, \quad \gamma_x^*(z) \underset{|z| \to +\infty}{\longrightarrow} 0 \\ \forall x \in \mathcal{X}^*, \quad \forall z \in \mathbb{R}, \quad \gamma_x^*(z) > 0. \end{cases}$$

The maximum of ℓ^{int} is reached at $(Q, \gamma, \mathfrak{D} = (D_x)_{x \in \mathcal{X}^*}, \mathbf{b})$ if and only if the integrand is maximal for almost every $u \in [0,1]$, which means under **(Aid)** that

$$\Big(Q, \big(\gamma(\cdot - D_x(u)) \otimes \mathbf{1}_{\mathbf{b}(x)}\big)_{x \in \mathcal{X}^*}\Big) = \Big(Q^*, \big(\gamma_x^*(\cdot - \Delta(x)) \otimes \mathbf{1}_{\mathbf{b}^*(x)}\big)_{x \in \mathcal{X}^*}\Big)$$

up to permutation of the hidden states for all $u \in [0,1]$.
Let us assume that the permutation is not constant at u. Since there are only a finite number of possible permutations of $\mathcal{X}^*$, there exist two sequences $(u_i)_{i \geq 1}$ and $(v_i)_{i \geq 1}$ converging to u, one corresponding to a permutation p and the other to a permutation $p' \neq p$, that is

$$\forall i \geq 1, \quad \forall x \in \mathcal{X}^*, \quad \begin{cases} \gamma_x(\cdot - D_x(u_i)) = \gamma_{p(x)}^*(\cdot - \Delta(p(x))) \quad \text{and} \quad \mathbf{b}(x) = \mathbf{b}^*(p(x)) \\ \gamma_x(\cdot - D_x(v_i)) = \gamma_{p'(x)}^*(\cdot - \Delta(p'(x))) \quad \text{and} \quad \mathbf{b}(x) = \mathbf{b}^*(p'(x)) \end{cases}$$

Therefore, by continuity, for all $x \in \mathcal{X}^*$

$$(\gamma_{p(x)}^*(\cdot - \Delta(p(x))), \mathbf{b}^*(p(x))) = (\gamma_{p'(x)}^*(\cdot - \Delta(p'(x))), \mathbf{b}^*(p'(x))),$$

so that $p = p'$ according to **(Aid)**, which contradicts the assumption that the permutation is not constant in u. Therefore, the permutation does not depend on u.
One may assume without loss of generality that the permutation is the identity, in other words $Q = Q^*$, $\mathbf{b} = \mathbf{b}^*$ and

$$\forall u \in [0,1], \quad \forall x \in \mathcal{X}^*, \quad \gamma_x(\cdot - D_x(u)) = \gamma_x^*(\cdot - \Delta(x)).$$

Here, we took u in the whole segment $[0,1]$ instead of a subset with measure 1 because the mapping $u \in [0,1] \longmapsto \gamma_x(\cdot - D_x(u))$ is continuous under **(Areg)**. If D_x is not constant at some $x \in \mathcal{X}^*$, this entails that γ_x^* is invariant by translation, so that it is constant, which contradicts **(Amax)**. Therefore, $\mathfrak{D}$ is constant.
Finally,

$$\forall x \in \mathcal{X}^*, \quad \frac{1}{2} = \int_{z \leq D_x} \gamma_x(z - D_x) dz = \int_{z \leq D_x} \gamma_x^*(z - \Delta(x)) dz$$

using **(Acentering)**, so that D_x is a median of γ_x^*. To conclude, note that under **(Amin)** and **(Acentering)**, $\Delta(x)$ is the only median of γ_x^*.

4.C.3 Uniform convergence of the homogeneous log-likelihood

Let us prove Lemma 4.13. The following theorem is a reformulation of Proposition 2 of Douc et al. (2004). Note that their proof also works when the space of parameters is not parametric.

Theorem 4.6. *Let $\mathcal{V}$ be a Polish space and write $\mathcal{D}(\mathcal{V})$ the set of nonnegative functions of $\mathcal{V}$. Let $K \in \mathbb{N}^*$. Let $\mathcal{Q}_K$ be the set of transition matrices of size K and Δ_K the set of probability measures on $[K]$. Let $(V_t)_{t\geq 1}$ be an ergodic and stationary process taking values in $\mathcal{V}$ with distribution $\mathbb{P}^*$. Consider a compact metric space Ω and mappings $\omega \longmapsto Q^\omega \in \mathcal{Q}_K$ and $\omega \longmapsto \gamma^\omega \in \mathcal{D}(\mathcal{V})^K$. Assume that $\omega \longmapsto Q^\omega$ is continuous and for all $v \in \mathcal{V}$, the mapping $\omega \longmapsto \gamma^\omega(v) \in \mathbb{R}_+^K$ is continuous. Finally, assume that there exists a constant $\sigma_- > 0$ such that*

$$\inf_{\omega\in\Omega} \inf_{x,x'\in[K']} Q^\omega(x,x') \geq \sigma_-, \tag{4.52}$$

$$\sup_{\omega\in\Omega} \sup_{x\in[K']} \sup_{v\in\mathcal{V}} \gamma_x^\omega(v) < \infty, \tag{4.53}$$

$$\mathbb{E}^* \left[\sup_{\omega\in\Omega} \left(\log \sum_{x\in[K']} \gamma_x^\omega(V_1) \right)_- \right] < \infty. \tag{4.54}$$

For all $\pi \in \Delta_K$, $\omega \in \Omega$ and $v_1^n \in \mathcal{V}^n$, let

$$\frac{1}{n} l_n(\pi, Q^\omega, \gamma^\omega)\{v_1^n\} := \frac{1}{n} \log \sum_{x_1^n\in[K']^n} \pi(x_1) Q^\omega(x_1,x_2)\dots Q^\omega(x_{n-1},x_n) \prod_{t=1}^n \gamma_{x_t}^\omega(v_t)$$

be the log-likelihood corresponding to the HMM with parameters $(\pi, Q^\omega, \gamma^\omega)$ and to the observations v_1^n.
Then for all $\pi \in \Delta_K$ and $\omega \in \Omega$, there exists a finite $l(Q^\omega, \gamma^\omega)$ such that almost surely,

$$\frac{1}{n} l_n(\pi, Q^\omega, \gamma^\omega)\{V_1^n\} \underset{n\to\infty}{\longrightarrow} l(Q^\omega, \gamma^\omega).$$

In addition, for all $N \in \mathbb{N}^*$, the mapping $\omega \longmapsto l(Q^\omega, \gamma^\omega)$ is continuous and

$$\sup_{\omega\in\Omega} \sup_{\pi\in\Delta_{K'}} \left| \frac{1}{n} l_n(\pi, Q^\omega, \gamma^\omega)\{V_1^n\} - l(Q^\omega, \gamma^\omega) \right| \underset{n\to\infty}{\longrightarrow} 0$$

almost surely.

Let us check these assumptions. First, let $\mathcal{V} = \mathbb{R} \times \mathcal{B}^*$, $V_t = (Z_t', B_t)$, $K = K'$ and

$$\Omega = \left\{ \omega = (Q, \gamma, \mathfrak{D}, u, \mathbf{b}) \in \Sigma_{K'}^{\sigma_-} \times \Gamma^{K'} \times \mathrm{Cl}(\mathcal{D}(M))^{K'} \times [0,1] \times (\mathcal{B}^*)^{K'} \right\}.$$

By Proposition 4.1, Ω is compact. It is also metrizable under **(Areg)** : for instance, let $(x_i)_{i\geq 1}$ be a dense sequence in $\mathbb{R}$, endow Γ with the distance

$$d_\Gamma(\gamma, \gamma') = \sum_{i\geq 1} 2^{-i} (|\gamma(x_i) - \gamma'(x_i)| \wedge 1)$$

and thus Ω is metrizable as a product of metric spaces.
By the uniform continuity of $\mathrm{Cl}(\mathcal{D}(M))$, the mappings

$$\begin{cases} \omega = (Q, \gamma, \mathfrak{D}, u, \mathbf{b}) \in \Omega \longmapsto Q^\omega := Q \\ \omega = (Q, \gamma, \mathfrak{D}, u, \mathbf{b}) \in \Omega \longmapsto \gamma^\omega(z', b) := (\gamma_x(z' - D_x(u)) \mathbf{1}_{\mathbf{b}(x)}(b))_{x\in[K']} \end{cases}$$

are continuous for all $(z', b) \in \mathbb{R} \times \mathcal{B}^*$. The lower bound (4.52) on the transition matrices is ensured by **(Aerg)** and the upper bound (4.53) on the densities is implied by **(Amax)**. Finally, the

integrability condition (4.54) follows from the fact that for all $\omega = (Q, \gamma, \mathfrak{D} = (D_x)_{x\in[K']}, u, \mathbf{b}) \in \Omega$,

$$\sum_{x\in[K']} \gamma_x^\omega(Z_1', B_1) \geq \inf_{x\in[K']} \gamma_x(Z_1' - D_x(u)) \geq m(M + |Z_1^{\max}|)$$

by **(Amin)**, and $\mathbb{E}^*[-\log m(M + |Z_1^{\max}|)] < \infty$ by **(Aint)**.
Thus, the previous theorem holds, which shows that the application

$$\omega \longmapsto l(Q^\omega, \gamma^\omega) =: l^{\hom}(Q, \tau(\gamma, \mathfrak{D}(u)), \mathbf{b})$$

is continuous on Ω. For the uniform convergence, let π_U be the uniform distribution on $[K']$ and let

$$S_{s,n}(\omega) = \frac{1}{n} l_n(\pi_U, Q^\omega, \gamma^\omega)\{V_{s+1}^{s+n}\}$$

for all $s, n \in \mathbb{N}^*$ and $\omega \in \Omega$.
The theorem implies that almost surely,

$$\lim_{n\to\infty} \sup_{\omega\in\Omega} \left| \frac{1}{n} S_{0,n}(\omega) - l(Q^\omega, \gamma^\omega) \right| = 0.$$

Hence, for all $\epsilon > 0$, there exists a (random) $n(\epsilon)$ such that,

$$\forall n \geq n(\epsilon), \quad \sup_{\omega\in\Omega} \left| \frac{1}{n} S_{0,n}(\omega) - l(Q^\omega, \gamma^\omega) \right| \leq \epsilon. \tag{4.55}$$

The following Lemma is a reformulation of Lemma 2 of Douc et al. (2004) for compact nonparametric parameter spaces.

Lemma 4.14. *Under the same assumptions as the previous theorem, for all $v_1^n \in \mathcal{V}^n$,*

$$\sup_{\omega\in\Omega} \sup_{\pi\in\Delta_K} |l_n(\pi, Q^\omega, \gamma^\omega)\{v_1^n\} - l_n(\pi_U, Q^\omega, \gamma^\omega)\{v_1^n\}| \leq \frac{1}{\sigma_-^2}.$$

Therefore,

$$|nS_{s,n}(\omega) - l_n(\pi_{X_{s+1}|V_1^s, X_1\sim\pi_U}, Q^\omega, \gamma^\omega)\{V_{s+1}^{s+n}\}| \leq \frac{1}{\sigma_-^2}.$$

Note that

$$\begin{aligned} l_n(\pi_{X_{s+1}|V_1^s, X_1\sim\pi_U}, Q^\omega, \gamma^\omega)\{V_{s+1}^{s+n}\} &= l_{s+n}(\pi_U, Q^\omega, \gamma^\omega)\{V_1^{s+n}\} - l_s(\pi_U, Q^\omega, \gamma^\omega)\{V_1^s\} \\ &= (s+n)S_{0,s+n}(\omega) - sS_{0,s}(\omega), \end{aligned}$$

so that

$$|nS_{s,n}(\omega) - (s+n)S_{0,s+n}(\omega) - sS_{0,s}(\omega)| \leq \frac{1}{\sigma_-^2}.$$

Thus, equation (4.55) entails that for all $s \geq 1$, $n \geq n(\epsilon)$ and $\omega \in \Omega$,

$$|nS_{s,n}(\omega) - nl(Q^\omega, \gamma^\omega)| \leq (2s+n)\epsilon + \frac{1}{\sigma_-^2}.$$

Therefore, by Lemma 4.14, one has for all $n \geq n(\epsilon)$ and $s \in \{0, \ldots, (N-1)n\}$:

$$\sup_{\omega\in\Omega} \sup_{\pi\in\Delta_{K'}} \left| \frac{1}{n} l_n(\pi, Q^\omega, \gamma^\omega)\{V_{s+1}^{s+n}\} - l(Q^\omega, \gamma^\omega) \right| \leq (2N-1)\epsilon + \frac{2}{n\sigma_-^2},$$

which concludes the proof.

4.D Miscellaneous proofs

4.D.1 Proof of Lemma 4.8

Let us first state a Hoeffding inequality for uniformly ergodic Markov chains using **(Aerg)** (see e.g. Glynn and Ormoneit (2002)) : for all $\epsilon > 0$, $x_1 \in \mathcal{X}^*$ and $n \geq \frac{1}{2\epsilon\sigma_-}$,

$$\mathbb{P}\left(\mathbb{P}(X_1 = x^*) - \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{X_t = x^*} \geq \epsilon \;\middle|\; X_1 = x_1\right) \leq \exp\left(-\frac{\sigma_-^2\epsilon^2}{2}n\right).$$

The value of $\mathbb{P}(X_1 = x^*)$ in the inequality is the one corresponding to the stationary distribution, so it is bounded below by σ_- using **(Aerg)**. Thus, for all $\delta > 0$, $\epsilon > 0$, $n \geq \frac{1}{\delta\epsilon\sigma_-}$ and $x_1 \in \mathcal{X}^*$,

$$\mathbb{P}\left(\frac{2}{\delta n}\sum_{t=1}^{\delta n/2} \mathbf{1}_{X_t = x^*} \leq \sigma_- - \epsilon \;\middle|\; X_1 = x_1\right) \leq \exp\left(-\frac{\sigma_-^2\epsilon^2\delta}{4}n\right).$$

Assume $n \geq \frac{2}{\delta(\sigma_-)^2}$. Choose $\epsilon = \sigma_-/2$ and apply a union bound on a covering $\mathcal{R}$ of $\{1, \dots, n\}$ in at most $2n/\delta$ segments of size $\delta n/2$:

$$\mathbb{P}\left(\inf_{S\in\mathcal{R}} \frac{1}{n}\sum_{t\in S} \mathbf{1}_{X_t = x^*} \leq \frac{\delta\sigma_-}{4}\right) \leq \frac{2n}{\delta}\exp\left(-\frac{\sigma_-^4\delta}{16}n\right).$$

Borel-Cantelli's lemma yields the result.

4.D.2 Proof of Lemma 4.9

Without loss of generality, we assume $\delta_n \longrightarrow \delta$ almost surely (this is possible by replacing δ_n by $\delta_n \wedge \delta$ in the first statement and $\delta_n \vee \delta$ in the second).
Let us first show that the two statements are equivalent. Assume that the second one holds. Let $(U_t)_{t\geq 1}$ and $(\delta_n)_n$ as in the first statement. Apply the second one to the i.i.d sequence of non-negative integrable random variables $(-U_t)_{t\geq 1}$:

$$\begin{aligned}\limsup_{n\to\infty} \sup_{\substack{S\subset\{1,\dots,n\}\\ |S|\leq\delta_n n}} \frac{1}{n}\sum_{t\in S}(-U_t) &\leq \mathbb{E}[(-U_1)\mathbf{1}_{-U_1\geq -q_U(\delta)}] \\ &\leq -\mathbb{E}[U_1\mathbf{1}_{U_1\leq q_U(\delta)}].\end{aligned}$$

Add $\mathbb{E}[U_1]$ on each side and use the law of large numbers :

$$\limsup_{n\to\infty} \sup_{\substack{S\subset\{1,\dots,n\}\\ |S|\leq\delta_n n}} \frac{1}{n}\sum_{t\notin S} U_t \leq \mathbb{E}[U_1\mathbf{1}_{U_1 > q_U(\delta)}].$$

Finally, replace δ_n by $1-\delta_n$, δ by $1-\delta$ and the sets S by their complementary to obtain the first statement.

Let us now show the second statement (regarding non-negative random variables). Given a random vector $(V_1, \dots, V_n)$, we write $V_{(1)} \leq V_{(2)} \leq \cdots \leq V_{(n)}$ its order statistics. Let $\delta \in (0,1)$ and let $(\delta_n)_n$ be a non-increasing sequence of $[0,1]$-valued random variables whose limit is δ almost surely.

For all $\beta \in (0,1)$, write $\hat{q}_V(\beta) := V_{(\lfloor \beta n \rfloor)}$ the empirical β-quantile. Then

$$\sup_{\substack{S \subset \{1,\dots,n\} \\ |S| \leq \delta_n n}} \frac{1}{n} \sum_{t \in S} V_t = \frac{1}{n} \sum_{t=1}^{n} V_t \mathbf{1}_{V_t \geq \hat{q}_V(1-\delta_n)}.$$

Let us show that almost surely

$$\left| \frac{1}{n} \sum_{t=1}^{n} V_t \mathbf{1}_{V_t \geq \hat{q}_V(1-\delta_n)} - \frac{1}{n} \sum_{t=1}^{n} V_t \mathbf{1}_{V_t \geq q_V(1-\delta)} \right| \underset{n \to \infty}{\longrightarrow} 0$$

and the result will follow by the law of large numbers. Thus we have to show that

$$\frac{1}{n} \sum_{t=1}^{n} V_t \left(\mathbf{1}_{q_V(1-\delta) \leq V_t \leq \hat{q}_V(1-\delta_n)} + \mathbf{1}_{\hat{q}_V(1-\delta_n) \leq V_t \leq q_V(1-\delta)} \right)$$

goes to 0 almost surely. Using Hoeffding's inequality, for all $\beta \in (0,1)$,

$$\mathbb{P}\left(\left| \frac{|\{t \in \{1,\dots,n\} \text{ s.t. } V_t \geq q_V(\beta)\}|}{n} - \mathbb{P}(V_1 \geq q_V(\beta)) \right| \geq \sqrt{\frac{\log n}{n}} \right) \leq 2n^{-2}.$$

In particular, taking $\beta = 1-\delta$, Borel-Cantelli's lemma shows that almost surely, for large enough n,

$$\hat{q}_V\left(1 - \delta - \sqrt{\frac{\log n}{n}}\right) \leq q_V(1-\delta) \leq \hat{q}_V\left(1 - \delta + \sqrt{\frac{\log n}{n}}\right),$$

so that there are at most $\sqrt{n \log n}$ terms between $q_V(1-\delta)$ and $\hat{q}_V(1-\delta)$. Hence there are at most $\sqrt{n \log n} + (\delta_n - \delta)n$ terms between $q_V(1-\delta)$ and $\hat{q}_V(1-\delta_n)$. As $(\delta_n)_n$ is non-increasing, this yields

$$\hat{q}_V(1-\delta_n) \leq \hat{q}_V(1-\delta) \leq q_V\left(1 - \delta + \sqrt{\frac{\log n}{n}}\right),$$

which ensures that these terms are bounded above by $q_V(1 - \delta + \sqrt{\frac{\log n}{n}})$. Thus, almost surely, for n large enough,

$$\frac{1}{n} \sum_{t=1}^{n} V_t \left(\mathbf{1}_{q_V(1-\delta) \leq V_t \leq \hat{q}_V(1-\delta_n)} + \mathbf{1}_{\hat{q}_V(1-\delta_n) \leq V_t \leq q_V(1-\delta)} \right) \leq \left[\sqrt{\frac{\log n}{n}} + (\delta_n - \delta) \right] q_V(1 - \delta/2),$$

which indeed converges to 0.

Deuxième partie

Application à la modélisation de variables météorologiques

Chapitre 5

Modélisation bivariée température et précipitations

Bivariate modeling of temperature and precipitations

5.1 Introduction

Historically, the management and planning of electricity demand and generation has involved long lasting observed or synthetic temperature time series, because temperature is the main driver of electricity demand. Then the centralized generation facilities are managed to match the anticipated demand. With a growing part of less manageable renewable generation based on wind and solar, the need for the same type of meteorological information, but not restricted to temperature anymore, emerges. The necessity for the system to be robust to as many different meteorological situations as possible involves a need for large samples of consistent evolutions of many meteorological variables, such as temperature, wind speed, solar radiation and rainfall for example. Since observation or reanalysis products are all available over quite limited time periods, stochastic weather generators are valuable tools to enrich the samples. For example, stochastic generators for temperature are commonly used as part of pricing derivatives, in relation with energy prices (Campbell and Diebold (2005), Mraoua (2007), Benth and Šaltytė Benth (2011)).

Single site multivariate models have been studied for several decades. The most widely cited model for weather variables has been proposed by Richardson (Richardson, 1981) in the framework of crop development, and lots of models have then been developed on the same basis (see Wilks and Wilby (1999) for a review). These models condition the evolution of the non-precipitation variables on two states based on occurrence and nonoccurrence of rainfall. Then the simulation of the non-precipitation variables is obtained through a multivariate autoregressive process, mostly using Gaussian distributions. In some cases, the autoregressive parameters depend on weather types. Flecher et al. (2010) extend this concept by using more weather types and skew normal distributions. The weather types are identified through classifications of the rainy and non-rainy days separately for each season and the number of weather types is chosen according to the BIC

criterion. Vrac et al. (2007) define a model used for precipitation downscaling based on weather types identified a priori through classifications either of the precipitation data or of exogenous atmospheric variables. However, such a priori definitions of the weather types may not be optimal to infer the stochastic properties of the variable to generate.

Hidden Markov Models (HMM) introduce the weather types as latent variables. In theses models, the states form a latent Markov chain and conditionally to the states, the observations are independent. Although simple, they are very flexible :

- the determination of the states is data driven instead of depending on arbitrarily chosen exogeneous variables,
- they allow non-parametric state-dependent distributions,
- using few parameters, they are able to model complex time dependence for the observations.

Homogeneous HMM are generally used for multisite generation either of rainfall occurrences (Zucchini and Guttorp, 1991) or of the whole rainfall field. Kirshner (2005) proposes an overview and tests different options for the multivariate emissions, from conditional independence to complex dependence structures, going through tree structures. Ailliot et al. (2015a) offers a more recent overview of the weather type based stochastic weather generators, including HMM. Extensions to nonhomogenous HMM are also proposed in order to introduce a diurnal cycle (Ailliot and Monbet, 2012) or to let the probability of a hidden state depend on the value of an external input variable (Hughes and Guttorp (1994), Hughes et al. (1999)).

Recently, new ways of generating meteorological variable have been studied. As an example, Peleg et al. (2017) designed a model mixing physically and stochastically based features in order to generate gridded climate variables at high spatial and temporal resolution.

Our contribution In this chapter, we introduce a nonhomogeneous HMM for the single site generation of temperature and rainfall at different locations in Europe presenting different climatic conditions. The model is here designed for a single site generation, because electricity load and generation balance is more and more studied at a very local scale in relation to the decentralization of electricity generation based on renewables. Furthermore, global balance is generally studied on the basis of geographical (possibly weighted) averages of the demand and generation respectively. The proposed HMM is nonhomogeneous because the seasonality is introduced in the transition matrix between the hidden states, as well as in the state-dependent distributions. Most of stochastic weather generators in the literature elude the problem of the non-stationarity of weather variables by defining a different model for each season or month independently, assuming local stationarity inside each block. For example, Lennartsson et al. (2008) consider blocks of lengths one, two or three months. This approach has several drawbacks :

- the local stationarity assumption may be difficult to check,
- the data used to fit each model is obtained as a concatenation of data that do not belong to the same year. This is a problem if our data exhibits a strong time dependence,
- a stochastic weather generator should be able to simulate long times series using only one model.

Our model, in contrast, allows the generation of synthetic climate variables, without splitting the data, and without any preprocessing. Furthermore, the generation of temperature implies handling the warming trend. Whereas Flecher et al. (2010) proposed a standardization of the temperature and radiation fields beforehand, the choice has been made here to explicitly introduce a trend in the temperature generation.

One of the main issues when dealing with multivariate modeling is being able to capture the possibly complex dependence structure between the variables. We introduced the state-dependent

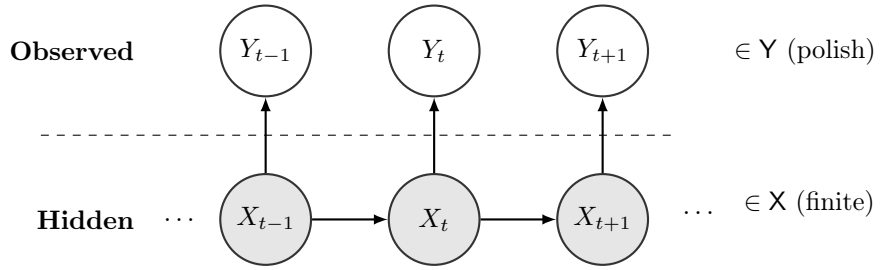

FIGURE 5.1 – The dynamic of a hidden Markov model.

distributions of our HMM as mixtures of tensor products (see equation (5.4)). This allows us, provided that the number of components in the mixture is large enough and that the marginals are well chosen, to approximate any state-dependent bivariate distribution for temperature and precipitation, without any a-priori on their dependence structure. Besides, it makes it easy to generalize the model to a larger number of variables without changing its global definition.

Outline A general description of our non homogeneous HMM is given in Section 5.2, as well as a reminder of the existing theoretical results linked to our model. Section 5.3 deals with the application of the model to precipitation and temperature observations. In this section, we present the data used, then we describe more precisely the modeling framework by giving the parametrization of our model that is specific to this application, and we discuss the results for the different locations. We finally go to the main conclusions and perspectives in the last section.

5.2 Model

In this section, we describe the mathematical framework in which we developed our model. We first give a general definition of nonhomogeneous hidden Markov models, before addressing the topic of theoretical results regarding these models. The full details of the parametrization of our model are given in Section 5.3.2, along with its application to climate data.

General formulation Let us first recall the definition of a finite state space hidden Markov model (HMM). Let K be a positive integer and $\mathsf{X} = \{1, \dots, K\}$. Let Y be a Polish space equiped with its Borel σ-algebra $\mathcal{Y}$. A (nonhomogeneous) hidden Markov model with state space X and observation space Y is a $\mathsf{X} \times \mathsf{Y}$-valued stochastic process $(X_t, Y_t)_{t\geq 1}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that

— $(X_t)_{t\geq 1}$ is a Markov chain with state space X.

— For all $t \geq 1$, the distribution of Y_t given $(X_s)_{s\geq 1}$ only depends on t and X_t, and conditional on $(X_s)_{s\geq 1}$, the $(Y_t)_{t\geq 1}$ are independent. Figure 5.1 summarizes this dynamic.

The key point here is that, from a statistical point of view, we do not observe the state sequence $(X_t)_{t\geq 1}$ but only $(Y_t)_{1\leq t\leq n}$. Hence the X_t are called the *hidden states*. The law of the Markov chain $(X_t)_{t\geq 1}$ is determined by its initial distribution π such that, for $k \in \mathsf{X}$, $\pi_k = \mathbb{P}(X_1 = k)$, and its transition matrices $(Q(t))_{t\geq 1}$ defined, for $i, j \in \mathsf{X}$ and $t \geq 1$, by $Q(t)_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$. The conditional distributions of Y_t given X_t are called the *emission distribution*. For $t \geq 1$ and $k \in \mathsf{X}$, we shall denote by $\nu_k(t)$ the distribution of Y_t given $X_t = k$. Formally, $Y_t \mid \{X_t = k\} \sim \nu_k(t)$. We assume that for any $t \geq 1$ and $k \in \mathsf{X}$, $\nu_k(t)$ is absolutely continuous with respect to some dominating measure μ defined on $\mathcal{Y}$ and we denote by $f_{k,t} : \mathsf{Y} \longrightarrow \mathbb{R}_+$ the corresponding density,

which will be refered to as the *emission density*. When the transition matrices and emission distributions are constant over time, the corresponding model is sometimes called *homogeneous hidden Markov model*, or just *hidden Markov model* when there is no ambiguity.

Parametric framework Assume that the transition matrices and the emission distributions depend on a parameter $\theta \in \Theta$, where Θ is a compact subset of $\mathbb{R}^q$, for some $q \geq 1$. We now precise the way they depend on θ.

- The function $t \mapsto Q(t)$ belongs to a known parametric family indexed by an unknown parameter β to be estimated.
- For any $t \geq 1$, the emission distributions $\nu_k(t)$ belong to a known (not depending on t) parametric family (e.g. gaussian) indexed by a parameter $\theta^Y(t)$. In addition, we assume that the function $t \mapsto \theta^Y(t)$ itself belongs to a known parametric family (e.g. affine functions), with parameter δ.
- Thus we can write $\theta = (\beta, \delta)$. We will denote by $Q^\theta(t)$ the transition matrix at time t when the parameter is θ and, for $1 \leq k \leq K$, $f^\theta_{k,t}$ the k-th emission density at time t when the parameter is θ.

Note that we consider that the number of states K is known, although this is not the case in practice. We discuss this issue in section 5.3.3. In this framework we proceed to the estimation of the parameters using maximum likelihood inference. Having observed $(Y_1, \dots, Y_n)$, the likelihood of the model when the parameter is θ and the initial distribution is π is given by

$$p^{\theta,\pi}(Y_1, \dots, Y_n) = \sum_{x_1,\dots,x_n} \pi_{x_1} Q^\theta(1)_{x_1 x_2} f^\theta_{x_1,1}(Y_1) Q^\theta(2)_{x_2 x_3} f^\theta_{x_2,2}(Y_2) \dots Q^\theta(n-1)_{x_{n-1} x_n} f^\theta_{x_n,n}(Y_n).$$

Then we define the maximum likelihood estimator (MLE) :

$$\hat{\theta}_{\pi,n} = \arg\max_{\theta \in \Theta} p^{\theta,\pi}(Y_1, \dots, Y_n).$$

The practical computation of the MLE can be performed using the well-known Expectation Maximization (EM) algorithm. See section 5.3.3 for the details of this algorithm in our framework.

Theoretical guarantees The statistical properties of hidden Markov models have been studied extensively since the 1960's. However, general identifiability conditions have only been proved recently. Following Allman et al. (2009), the authors of Gassiat et al. (2016a) proved that stationary non parametric HMM are identifiable from the law of three consecutive observations (up to permutation of the states), provided that the emission distribution are linearly independent and that the transition matrix has full rank. Alexandrovich et al. (2016) prove a similar result with slightly weaker assumptions. The properties of the maximum likelihood estimator in homogeneous HMM are now well-known. Its strong consistency has first been established by Baum and Petrie (1966) in the case where both the state space and the observation space are finite. This result has been generalized to continuous observation spaces by Leroux (1992). See also Douc et al. (2011) and references therein. The literature is less abundant when it comes to nonhomogeneous hidden Markov models. In Ailliot and Pene (2015) and Pouzo et al. (2016), the authors consider models where the observation distribution depends not only on the current state, but also on previous observations, and where the transition matrices depend on the previous observations. They prove the consistency of the maximum likelihood estimator in this framework. More recently, Diehn et al. (2018) study the case of nonhomogeneous hidden Markov models that can be asymptotically approximated by an homogeneous one. They introduce a quasi-maximum likelihood estimator and prove its consistency. None of the previously stated results apply to the model we introduce in this chapter.

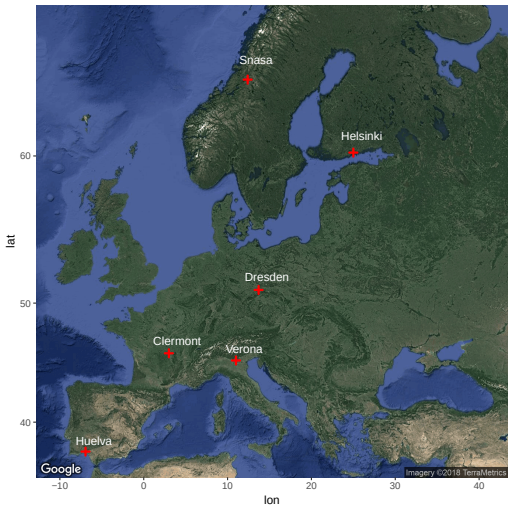

FIGURE 5.2 – Stations locations

5.3 Application to precipitation and temperature

5.3.1 Data

Description of the data We used data from the *European Climate Assessment and Datasets* (ECA&D) project [1]. Six weather stations were considered : Helsinki (Finland), Dresden (Germany), Verona (Italy), Huelva (Spain), Clermont-Ferrand (France) and Snåsa (Norway). For each station, the data consists of mean daily temperature and daily precipitation from 1954/01/01 to 2014/12/31. Figure 5.2 presents the locations of the weather stations, whereas Figure 5.3 shows that the distribution of temperature in the 6 stations differ in many ways : mean, variance, range, skewness, seasonal behaviour...

Table 5.1 presents some basic statistics concerning precipitations at the 6 stations under study. Mean yearly precipitations range from 501.5 mm in Huelva (Spain) to 964.2 mm in Snåsa (Norway), where the precipitation frequency is 0.62 compared to only 0.19 in Huelva. Also, the maximum observed daily precipitation in Snåsa is 65.9 mm, compared to 198 mm in Verona. The mean value of (non-zero) daily precipitations ranges from 3.4 mm in Helsinki to 7.3 mm in Huelva. We chose on purpose weather stations where climate strongly differs in order to test the robustness of our model when applied to different climates.

	Clermont	Huelva	Dresden	Verona	Helsinki	Snasa
Mean yearly precip. (mm)	584.2	501.5	665.3	803.3	638.9	964.2
Max. observed precip. (mm)	75.3	160.0	158.0	198.0	79.3	65.9
Precip. frequency	0.40	0.19	0.49	0.34	0.51	0.62
Mean positive precip. (mm)	4.0	7.3	3.7	6.5	3.4	4.2

TABLE 5.1 – Basic statistics for precipitation.

1. Data freely available at https://www.ecad.eu//dailydata/index.php

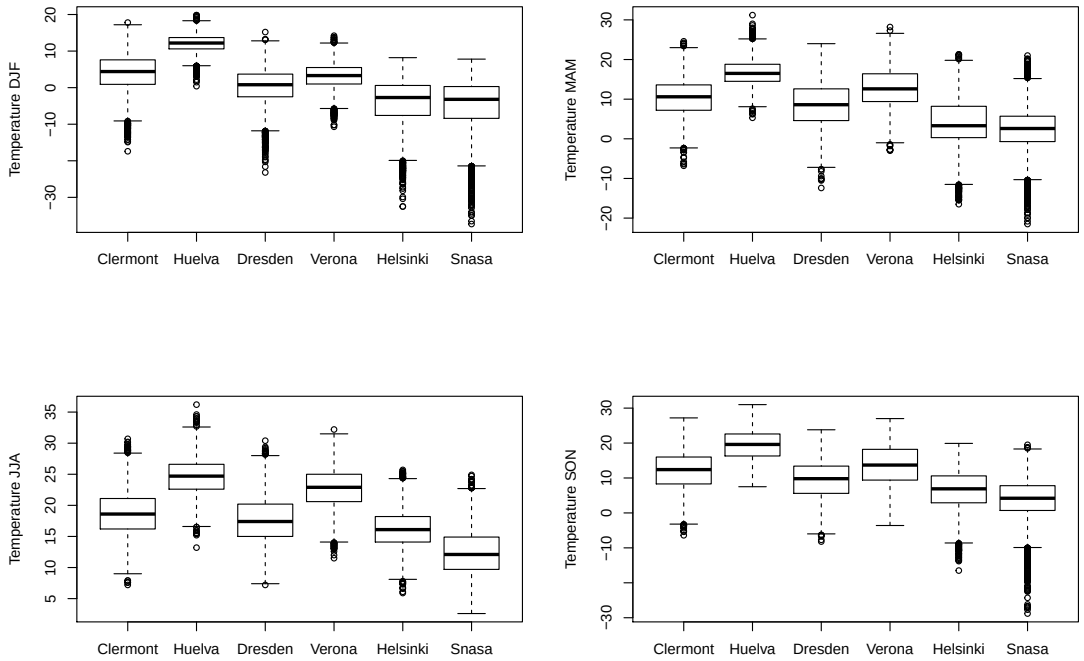

FIGURE 5.3 – Boxplots of the temperature at the different locations for each season.

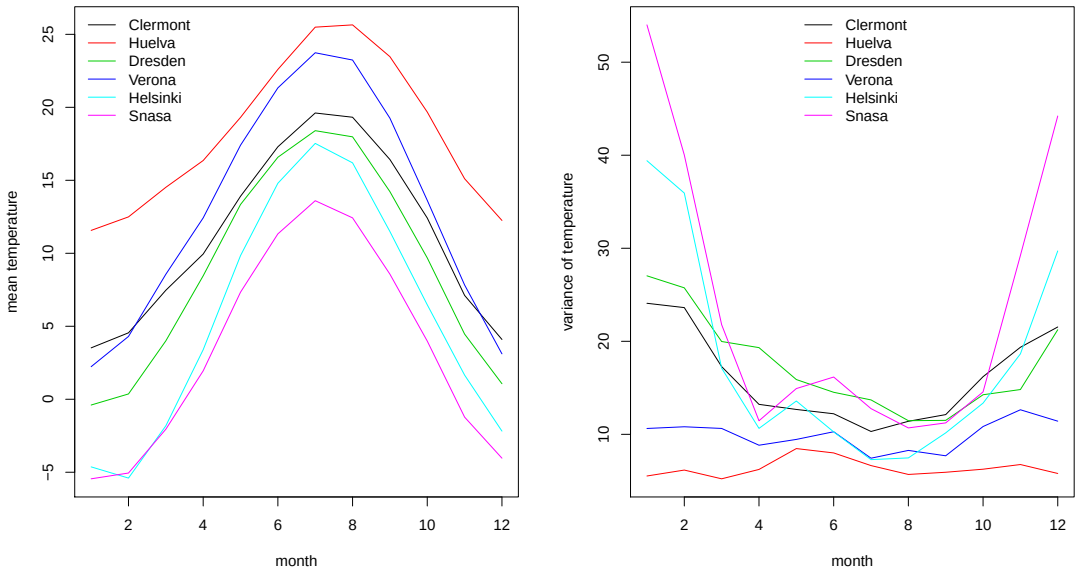

FIGURE 5.4 – Monthly mean and variance of temperature.

Seasonalities and trends When observed at a daily time step temperature and precipitation times series are not stationary, in that their distribution varies through time. Temperature obviously exhibits a seasonal cycle, as shown on the left panel of Figure 5.4. The right panel of Figure 5.4 shows the seasonality of the variance of temperature, which is rather flat in the two stations of Southern Europe (Verona and Huelva). We also notice that these stations have the lowest variances. On the opposite, the variance of temperature displays a very clear seasonality in the stations of Northern Europe (Snåsa and Helsinki), the variability being much higher in the winter months.

The non-stationarity of temperature is also caused by the existence of a trend, corresponding to global warming. The black lines in Figure 5.5 are the yearly mean temperature for each site. All sites except Huelva seem to exhibit an increasing trend. However, the shape and the slope of this trend differ among the different sites. For stations like Helsinki or Dresden it seems reasonable to consider a linear trend over the whole observed period, whereas Verona or Clermont exhibit a change point in the warming rate. Thus the modeling of the trend should be site-specific. Figure 5.5 suggests that a simple parametric form for the trends could be linear or piecewise linear with two pieces. Then for each site, two questions arise :

— is there a breaking point (i.e. a change in the slope of the trend) ?
— if yes, where is it ?

To answer these questions, we can use the yearly mean temperatures $(\bar{Y}_a)_{a\in\{1954,\dots,2014\}}$. The linear regression is then given by

$$(\hat{\alpha}^{LM}, \hat{\beta}^{LM}) = \underset{\alpha,\beta\in\mathbb{R}}{\arg\min} \sum_{a=1954}^{2014} (\bar{Y}_a - \alpha a - \beta)^2$$

In a piecewise linear model, the optimal breaking point can be found by computing

$$\hat{\tau} = \underset{1954\leq\tau\leq 2014}{\arg\min} \min_{\alpha,\beta,\gamma\in\mathbb{R}} \sum_{a=1954}^{2014} (\bar{Y}_a - \alpha a - \beta - \gamma(a-\tau)\mathbf{1}_{t>\tau})^2.$$

Then the corresponding piecewise linear regression is given by

$$(\hat{\alpha}^{PLM}, \hat{\beta}^{PLM}, \hat{\gamma}^{PLM}) = \underset{\alpha,\beta,\gamma\in\mathbb{R}}{\arg\min} \sum_{a=1954}^{2014} (\bar{Y}_a - \alpha a - \beta - \gamma(a-\hat{\tau})\mathbf{1}_{t>\hat{\tau}})^2.$$

In order to test for the significance of the breaking point, we perform a likelihood ratio test : we compute the test statistic

$$\Lambda = 2\left(\log \mathcal{L}(\hat{\alpha}^{PLM}, \hat{\beta}^{PLM}, \hat{\gamma}^{PLM}) - \log \mathcal{L}(\hat{\alpha}^{LM}, \hat{\beta}^{LM})\right),$$

with $\mathcal{L}$ the likelihood function, where we considered gaussian residuals (this assumption was tested with a Kolmogorov-Smirnov test). Then Λ is compared to a quantile of the χ^2 distribution with one degree of freedom.

Station	**Test result**	**p-value**	τ
Helsinki	PL	0.049	1980
Dresden	L	0.16	-
Verona	PL	3.10^{-6}	1987
Huelva	PL	0.026	1961
Clermont	PL	0.012	1978
Snåsa	PL	0.047	1980

TABLE 5.2 – Test of the parametric form of the trends.

Table 5.2 gives the results of this procedure for the six sites. In the "test result" column, PL means that the test rejected the linear model at the risk level 0.05 (which means that the trend is piecewise linear), and L means that the test did not reject the linear model. The last column τ is the year of the change in the slope of the trend, when there is such a change. Thus, the test rejects the simple linear trend for all sites but Dresden. For piecewise linear trends, the breaking point is in the decade 1978-1987, which is consistent with climatology, except for the site of Huelva (1961), surprisingly. The optimal trends are depicted in Figure 5.5. We see that as far as Huelva is concerned, despite the result of the test, the piecewise linear trend cannot be considered significant. First, the change point in 1961 is not consistant with climatology, as it would correspond to a warming stopping in 1961. Then, we see that the procedure described above was misleaded by the unusually cold year 1956. For these reasons, we choose a simple linear trend for this station.

Furthermore, climate change does not affect summer and winter the same way. For each location, we computed the mean temperature of summer months (June, July, August) and we centered it to remove the shift between the stations. We then applied a rolling mean with a window width of 15 years in order to smooth the effect of interannual variability. Thus the value corresponding to the year 1968 is the mean over the period 1954-1968. We performed the same operation considering winter months (December, January, February). The results can be seen in Figure 5.6. In summer, all the stations show an increasing trend. However, the shape of the trend may differ according to the stations. Also, the amplitude of the warming is higher in Verona (2.5°C) than in Helsinki and Snåsa (about 1°C). In winter, the amplitude of the warming is higher in Northern Europe

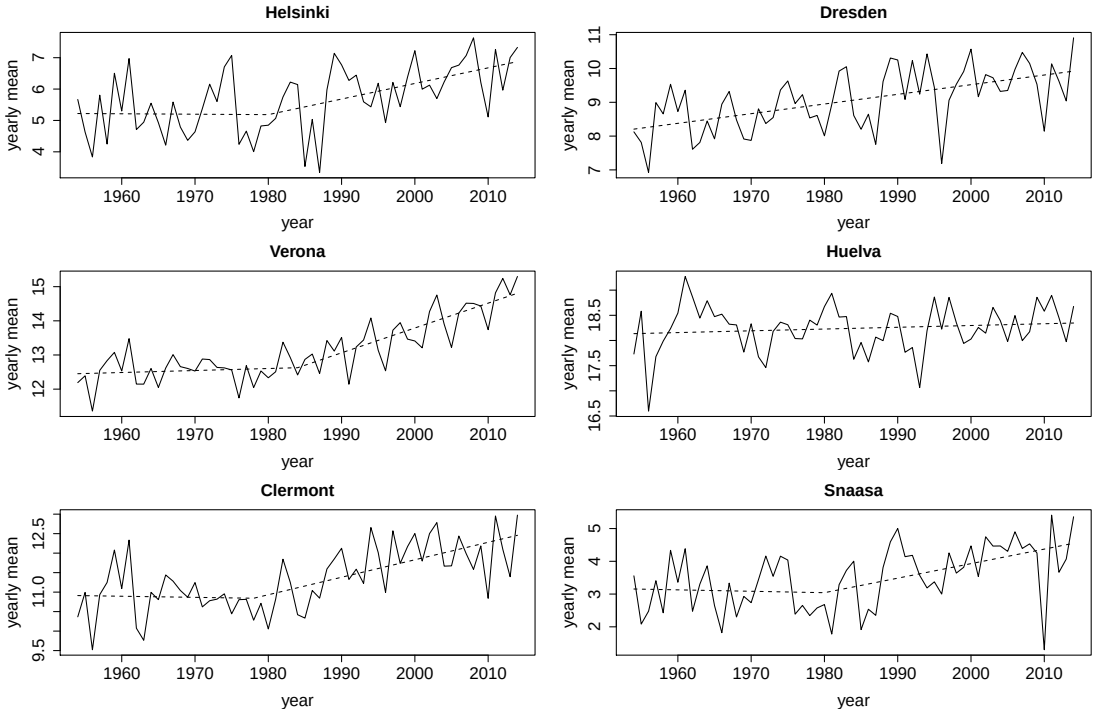

FIGURE 5.5 – Yearly means of temperature (black) and associated optimal trends (blue).

(Helsinki, Snåsa) than in the South (Verona, Huelva). We actually do not notice any warming in Huelva. Our model will deal with this phenomenon using seasonal transition probabilities between the states, and state-dependent trends (see Section 5.3.2).

As for precipitation, both the occurrence process (frequency of precipitation) and the intensity process have a seasonal behaviour, as shown in Figure 5.7. The left panel displays the frequency of precipitation, month by month, for each station. Here again, we observe that the shape as well as the amplitude of this seasonality depend on the location. In Helsinki for example, the frequency of precipitation is at its lowest in spring and at its highest in winter. It's the opposite in Verona. The station of Huelva becomes very dry in summer and its maximum precipitation frequency is reached in January (only 0.3). The right panel in Figure 5.7 displays the mean value of non-zero precipitation, month by month, for every site. Observe that this quantity exhibits a strong seasonal behaviour (except in Snåsa) that is different from the one of the occurrence process. Once again, this phenomenon requires some modeling effort that we describe in Section 5.3.2.

Figures 5.8 to 5.10 show the annual, summer and winter precipitation amounts respectively, as well as the associated regression lines. The most notable results here are the increasing trends in the two stations of Northern Europe, especially in winter, and also a slightly decreasing trend in Clermont-Ferrand in winter.

The existence of a trend in the yearly precipitation amounts can be caused either by a trend in the occurrence process (precipitations become more rare or more frequent) or in this intensity process (the mean value of positive precipitations changes). A slight decrease of the frequency of winter precipitations can be observed in Clermont, whereas the increasing trend in winter in

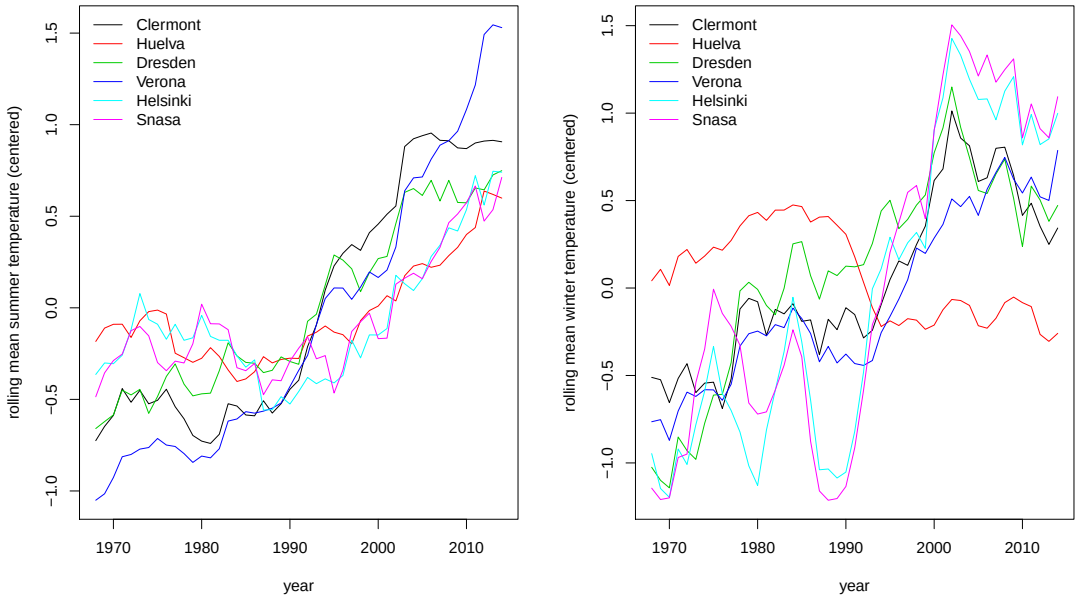

FIGURE 5.6 – Temperature trend in summer (left) and winter (right).

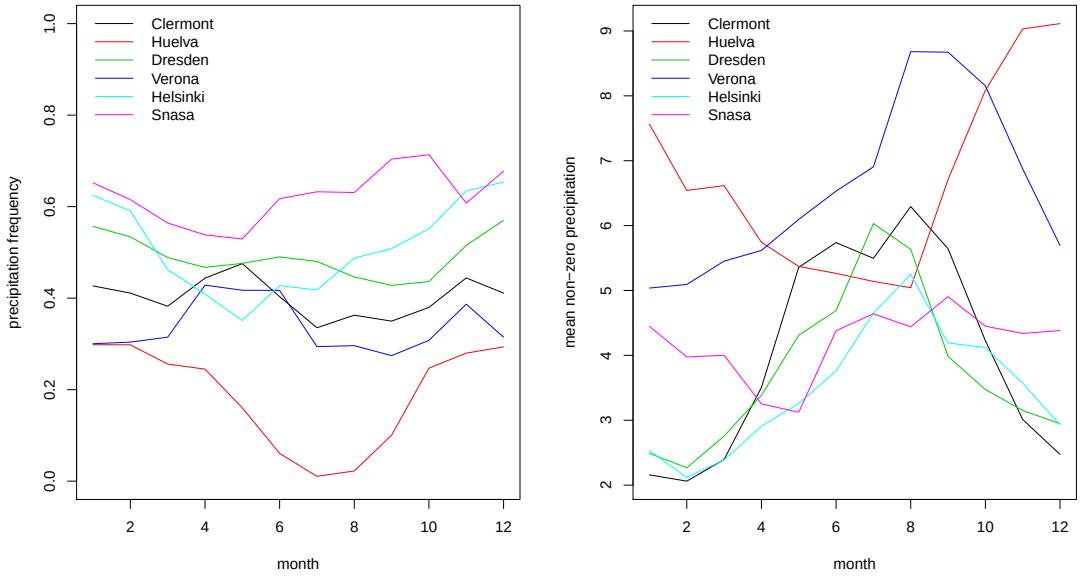

FIGURE 5.7 – Seasonality of precipitation.

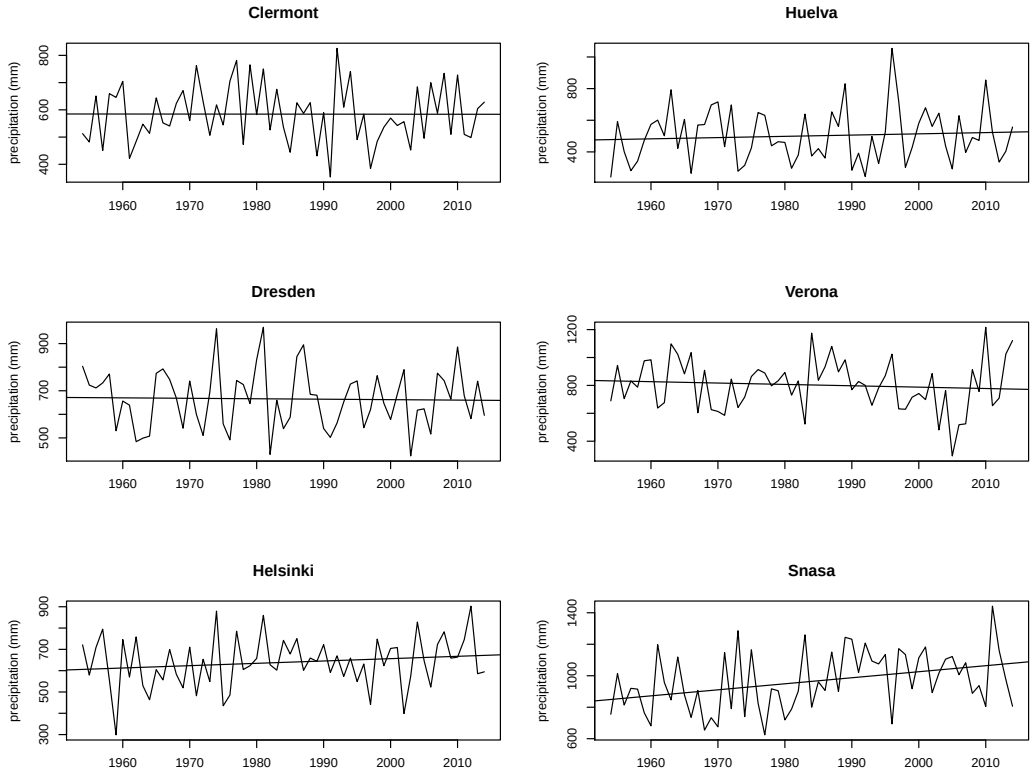

FIGURE 5.8 – Trend of precipitation.

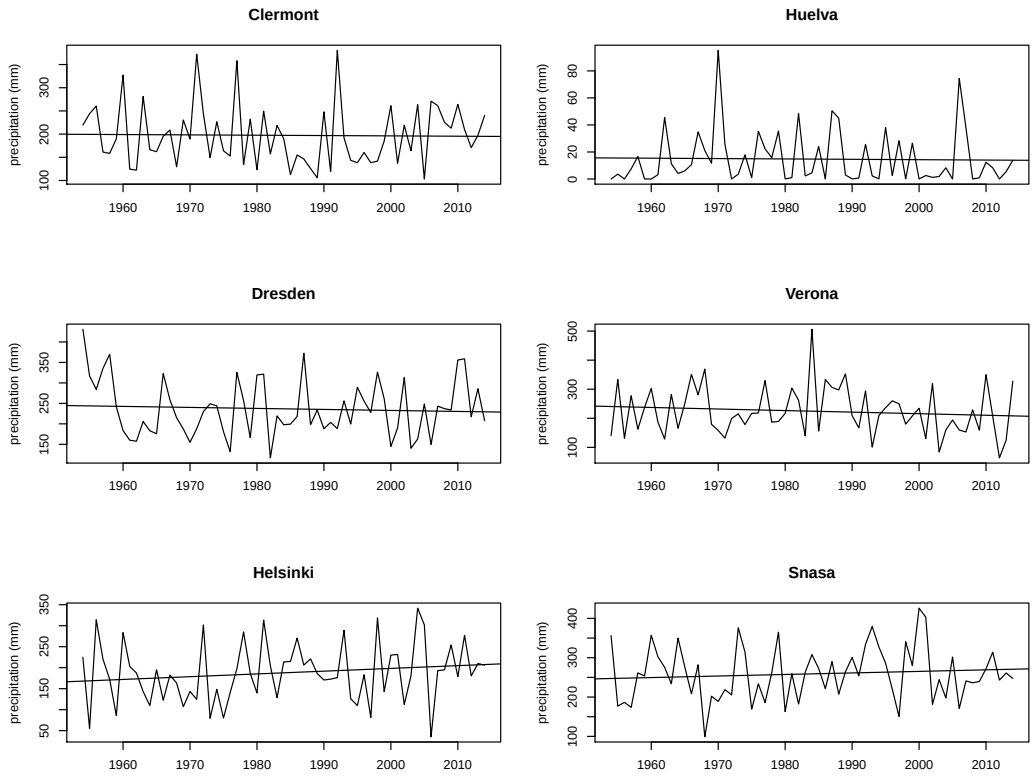

FIGURE 5.9 – Trend of precipitation in summer.

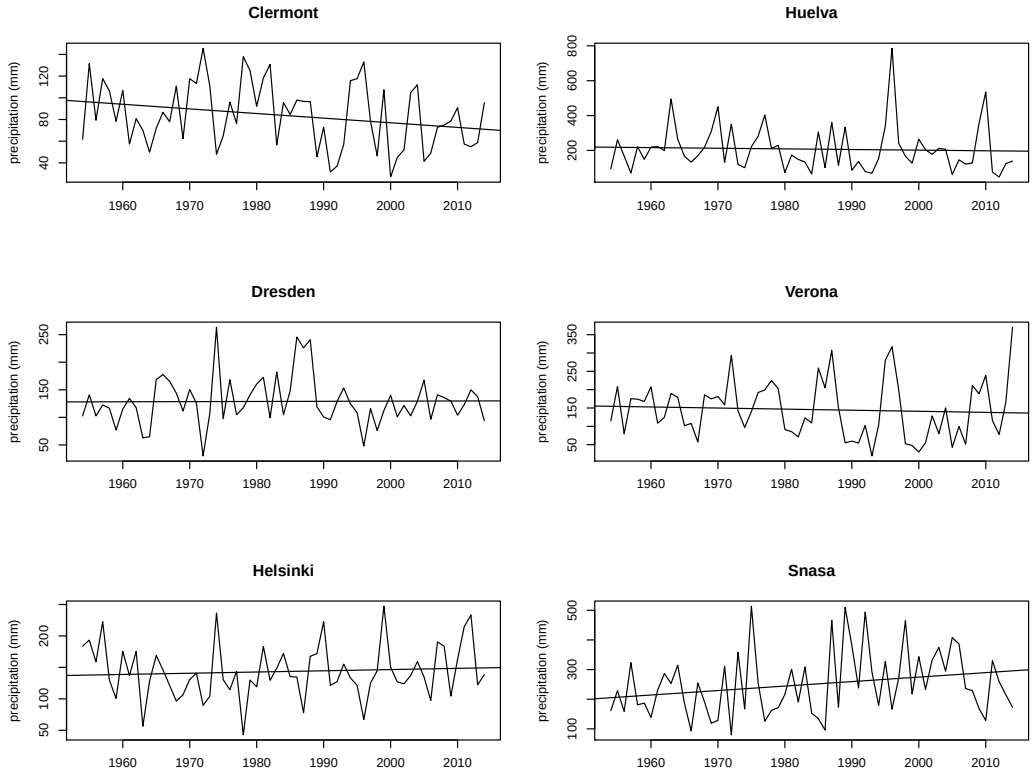

FIGURE 5.10 – Trend of precipitation in winter.

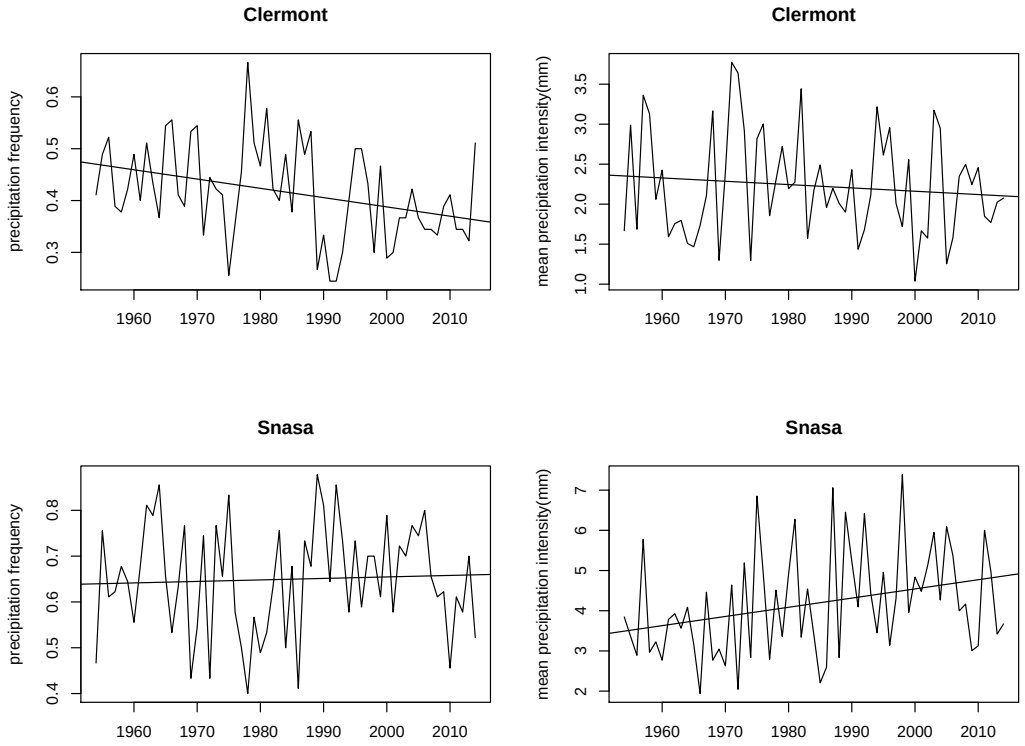

FIGURE 5.11 – Trends in the occurrence (left) and intensity (right) of precipitations in winter for the stations of Clermont and Snåsa

Snåsa is mostly the result of an increasing precipitation intensity (Figure 5.11). As these trends are light and concern only two stations, we chose not to include them in the model, which does not degrade the results.

Finally, we highlight the fact that there is also a seasonality in the dependence structure between temperature and precipitation. To see this, we can plot the monthly correlations between these two variables, as in Figure 5.12. All stations but Clermont share a similar shape for this curve, with a negative minimum of correlation in summer, and a positive maximum in winter. This reflects the fact the precipitations mostly occur when the temperature is moderate rather than extreme.

5.3.2 Specification of the model

In section 5.2, we gave a very general description of our model. In this section, we get more into details as we give the specific forms of the transition structure and the emission distributions. As we want to model simultaneously precipitations and temperature, the observation space is $\mathsf{Y} = \mathbb{R}_+ \times \mathbb{R}$ and $Y_t = \left(Y_t^{(1)}, Y_t^{(2)}\right)$. The superscript (1) refers to precipitation and (2) refers to temperature.

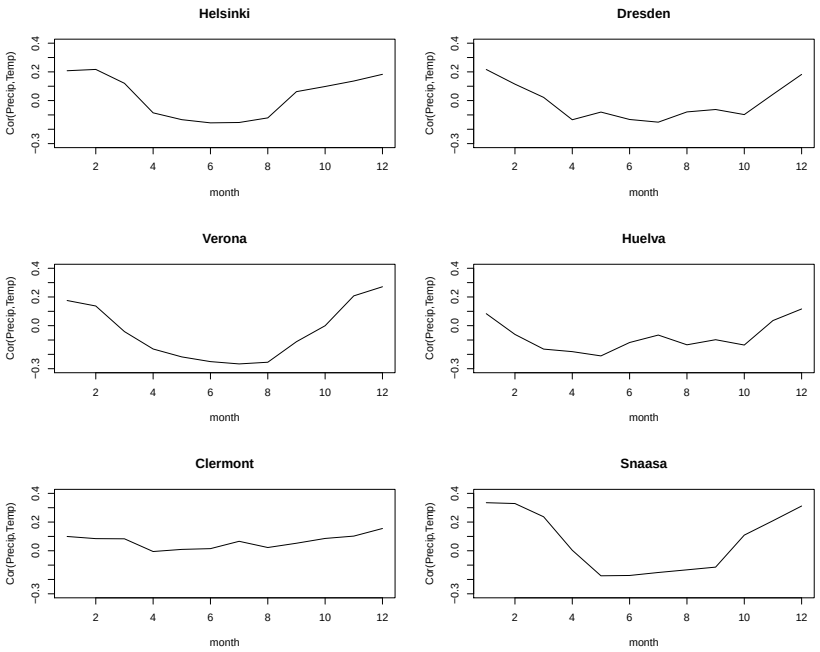

FIGURE 5.12 – Monthly correlations between temperature and precipitations

Transition probabilities The transition matrix at time t in our model is given by

$$Q(t)_{ij} = \frac{\exp(P_{ij}(t))}{1 + \sum_{l=1}^{K-1} \exp(P_{il}(t))}, 1 \leq j \leq K-1, \quad 1 \leq i \leq K \tag{5.1}$$

$$Q(t)_{iK} = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(P_{il}(t))}, 1 \leq i \leq K. \tag{5.2}$$

This is indeed a stochastic matrix, as for all $1 \leq i \leq K$, $\sum_{j=1}^{K} Q(t)_{ij} = 1$ and $Q(t)_{ij} > 0$. For $1 \leq i \leq K$ and $1 \leq j \leq K-1$, P_{ij} is a trigonometric polynomial with (known) degree d and period T. More precisely,

$$P_{ij}(t) = \beta_{ij0} + \sum_{l=1}^{d} \left(\beta_{ij,(2l-1)} \cos\left(\frac{2\pi}{T} lt\right) + \beta_{ij,2l} \sin\left(\frac{2\pi}{T} lt\right) \right) \tag{5.3}$$

Hence the transition probabilities of the hidden Markov chain are periodic functions of time. Therefore, the relative frequencies of the states will vary through time in a periodic manner. This allows the model to reproduce some of the seasonal behaviours of climate variables.

Emission distributions In order to allow for flexibility, we choose mixtures as emission distributions. More precisely, the conditional distribution of Y_t given $X_t = k$ is

$$\nu_k(t) = \sum_{m=1}^{M} p_{km} \nu_{k,m}^{(1)}(t) \otimes \nu_{k,m}^{(2)}(t) \tag{5.4}$$

The p_{km} are the weights of the mixture, satisfying $p_{km} \geq 0$ and $\sum_{m=1}^{M} p_{km} = 1$. The component corresponding to precipitations is defined by

$$\nu_{km,t}^{(1)} = \begin{cases} \delta_0 & ,1 \leq m \leq M_1 \\ \mathcal{E}\left(\frac{\lambda_{km}}{1+\sigma_k(t)}\right) & ,M_1+1 \leq m \leq M \end{cases}, \tag{5.5}$$

where δ_0 denotes the Dirac mass at 0, the λ_{km} are positive parameters depending on both the state k and the population m in the mixture, and $\sigma_k(\cdot)$ is a state-dependent trigonometric polynomial, modeling the seasonal variations of the intensity of precipitations. Thus the marginal distribution of precipitations in state k at time t is a mixture of a Dirac mass with weight

$$p_{k0} := \sum_{m=1}^{M_1} p_{km}$$

and exponential distributions with parameters $\frac{\lambda_{km}}{1+\sigma_k(t)}$ and weights $(p_{km})_{M_1+1 \leq m \leq M}$. When p_{k0} is close to 1, the state k is considered as *dry* whereas it is considered as *wet* when p_{k0} is close to 0. As the frequency of a given state varies across the year thanks to the seasonal transition probabilities, this will allow the model to capture the seasonal behaviour of the frequency of precipitations.

Regarding temperature,

$$\nu_{km,t}^{(2)} = \mathcal{N}\left(T_k(t) + S_k(t) + \mu_{km}, \sigma_{km}^2\right), \tag{5.6}$$

so that the marginal distribution of temperature in state k at time t is a mixture of gaussian distributions, with variances σ_{km}^2.

- μ_{km} is a mean parameter that depends on both the state k and the component m.
- $T_k(\cdot)$ is a function corresponding to the temperature trend in state k. As shown in Section 5.3.1, the parametric form of the trends depends on the sites. We follow the conclusions of Section 5.3.1 and we choose linear trends for Huelva and Dresden, and piecewise linear trends (with site-specific change points) for the other stations.
- $S_k(\cdot)$ is a trigonometric polynomial with degree d corresponding to the seasonal cycle of temperature in state k.

Note that we allow both the trend and the seasonality of temperature to depend on the hidden state, hence the subscript k. Equation (5.4) shows that in each state and each component of the mixture, precipitations and temperature are independent, but of course they are not globally independent. The choices of K, M, M_1 and d are discussed in the next section.

5.3.3 Inference of the parameters

The EM algorithm The computation of the maximum likelihood estimator is done using the EM algorithm (Dempster et al., 1977), which is a generic approach to perform maximum likelihood inference in latent variables models. The details of the algorithm in our particular framework can be found in Touron (2018). Recall that the EM algorithm does not guarantee to find the global maximum of the likelihood function but only a local maximum, depending on its initial point. To overcome this drawback, we launch the algorithm multiple times, using randomly chosen initial points. See also Biernacki et al. (2003) where the authors compare several initialization procedures for the EM algorithm.

Model selection Our model requires to specify several hyper-parameters :

— K the number of hidden states,
— d the degree of the trigonometric polynomials, which sets the complexity of the seasonality,
— M and M_1 which correspond to the complexity of the emission distributions.

As the dimensionality of the parameter θ is a quadratic function of K and a linear function of both d and M, the larger these hyper-parameters, the more complex the model is and the better we capture the statistical properties of the data. However, we cannot use too large hyper-parameters, for the following reasons :

— We may overfit the data.
— The M step of the EM algorithm requires to solve an optimization problem. As its solution admits no closed form, this is done using a numerical optimization algorithm, which can be difficult and time-consuming if the number of parameters is too large.
— The likelihood function of models with a large number of parameters may have many sub-optimal local maxima.
— A large number of hidden states leads to a loss of interpretability of the states, which may be a problem for some practitionners.

Therefore, the complexity of the model, especially the number of hidden states, must be chosen carefully. To do so, a standard approach is to use information criteria such as Akaike Information Criterion (AIC), Integrated Completed Likelihood (ICL, see Biernacki et al. (2000)) or Bayesian Information Criterion (BIC, see Schwarz et al. (1978)). The latter is very popular in applications of HMM, although not justified in theory. The idea of AIC and BIC is to penalize the models with a large number of parameters, in order to realize a trade-off between goodness-of-fit and complexity. If we have to choose a model among a collection $\mathcal{M}$, we minimize over $m \in \mathcal{M}$ the criterion $-\mathcal{L}(\hat{\theta}_m) + \text{pen}(m)$, where $\mathcal{L}(\hat{\theta}_m)$ is the maximum likelihood in the model m, and $\text{pen}(m)$ is the penalty associated to the model m. For example, in the case of BIC, $\text{pen}(m) = -\frac{p(m)}{2}\log(n)$ where $p(m)$ is the number of parameters of the model m and n is the number of observations. Another approach is cross-validated likelihood (Celeux and Durand, 2008), even though it is computationally intensive. In Lehéricy (2019), the author introduces a penalized least square estimator for the order of a nonparametric HMM and proves its consistency. However, when dealing with real world data, other considerations should be taken into account, such as interpretability of the states, computing time, or the ability of the model to reproduce some behaviour of the data, as explained in Bellone et al. (2000). Indeed, according to Pohle et al. (2017), the popular AIC (Akaike Information Criterion) and BIC, as well as other penalized criteria, tend to overestimate the number of states as soon as the data generating process differs from a HMM (e.g. the presence of a conditional dependence), which is often the case in practice. Hence it is advised to use such a criterion as a guide, without following it blindly. Keeping in mind these considerations, we chose to use the BIC criterion to select the number of states, which is by far the most important hyper-parameter because the number of parameters of the model is quadratic in K. We found $K = 6$ or $K = 7$ (depending on the stations) to be good choices. We used our previous experience on univariate models to select $d = 2$, $M = 4$ and $M_1 = 2$.

5.3.4 Results

Estimated parameters

As we use a hidden Markov model, we do not need to define the states *a priori*, thus we do not need to give them an interpretation before estimating the parameters (e.g. wet or dry state). The determination of the states is data driven and this is one of the perks of HMM. Besides, as

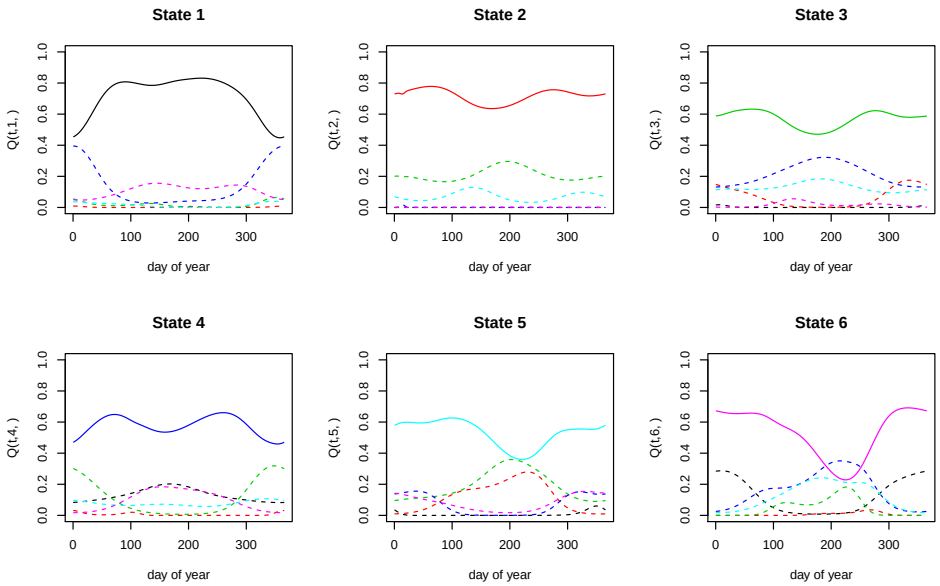

FIGURE 5.13 – Estimated transition probabilities, Verona. Each panel corresponds to a row of the transition matrices. The solid lines represent the diagonal coefficients, i.e. the functions $\hat{Q}_{ii}(\cdot)$ whereas the dashed lines represent the other coefficients. Each color corresponds to a different state.

our purpose is to produce realistic time series of weather variables, we only need to investigate on the simulations produced by the model, not its parameters. However, it is interesting to take a look at the estimated parameters themselves, thus interpreting the states *a posteriori*. Indeed, interpretability of the states gives credit to the model as it provides a first indication of whether or not it manages to capture some specific meteorological behaviours. We will not provide all the estimated parameters for all of the six stations. Rather, we will give some examples to shed light on the way the different parameters of our model can be physically interpreted. More precisely, the following elements are interesting to look at :

- Transitions (see Equations (5.1) to (5.3)) :
  - transition matrices, specifically the functions $t \mapsto \hat{Q}_{ij}(t)$ for $1 \leq i, j \leq K$. The estimated transition probabilities for the station of Verona are depicted in Figure 5.13. This provides information on the stability of the states. For example, State 1 is rather stable in summer, (the probability of remaining in that state is approximately 0.8), whereas it is much less stable in winter. On the opposite, State 6 is unstable in summer and more stable in winter.
  - the relative frequencies of the states, i.e. the probabilities $\mathbb{P}^{\hat{\theta}}(X_t = k)$ for $1 \leq k \leq K$, $1 \leq t \leq T$. These quantities can be directly obtained through the transition matrices. We represented them for the station of Verona in Figure 5.14. We see that they vary throughout the year, the most frequent states in summer being 1 and 4, whereas it is State 3 in winter.

- Temperature parameters (see Equation (5.6)) :
  - the trends $\hat{T}_k(\cdot)$ for $1 \leq k \leq K$. Recall that the trends are modeled as linear for the sites of Huelva and Dresden, and piecewise linear for the others (with site-dependent

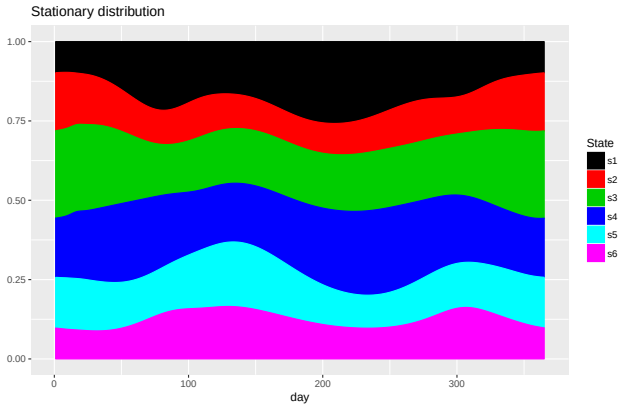

FIGURE 5.14 – Estimated relative frequencies of the states, Verona.

breaking points) and that we estimate a different trend for each state. The results for the six stations are presented in Figure 5.15 and reflect an increase in mean temperature ranging from 0.6°C in Huelva to $\simeq 2$°C (depending on the state) in Verona.

— the state-dependent seasonalities $\hat{S}_k(\cdot)$, corresponding to the yearly cycle of temperature (see Figure 5.16).

— the random noise, i.e. the centered gaussian mixture that remains, in each state, when we removed the state-dependent trend and the state-dependent seasonal component. Recall that we chose $M = 4$, so that there are 4 components in each gaussian mixture. We can see in Figure 5.17 some peaks in the probability density functions (especially in Dresden). These are caused by estimators $\hat{\sigma}^2_{km}$ that are close to zero. They must be understood as numerical issues in the estimation process (EM algorithm) and have no physical intepretation. Furthermore, they have no negative impact on the quality of the simulated process, as the associated weights $\hat{p}_{km}$ are also close to zero. Figure 5.17 also shows that some states have heavier tails (e.g. the state corresponding to the black line in Helsinki). These states can be interpreted as *extreme values states*. Indeed, they induce a larger probability for large deviations from the mean temperature.

— Precipitation parameters (see Equation (5.5)) :

— the estimated seasonal components $1 + \hat{\sigma}_k(t)$: the larger they are, the heavier the precipitations, if any. They are depicted in Figure 5.18. One can see that for all the considered sites, some of the states clearly exhibit a seasonal behaviour regarding the intensity of precipitations, in accordance with the right panel of Figure 5.7.

— the weights of the Dirac masses, i.e. $(\hat{p}_{k0})_{1\leq k\leq K} = \left(\sum_{m=1}^{M_1} \hat{p}_{km}\right)_{1\leq k\leq K}$: how dry or wet are the states. This can be compared to the left panel of Figure 5.7. For example, for the station of Verona,

$$(\hat{p}_{k0})_{1\leq k\leq K} = (0.88, 0.68, 0.93, 0.95, 0.09, 0.09),$$

so that the states 1, 3 and 4 are mostly dry, whereas the states 5 and 6 are rainy.

— the estimated parameters of the exponential distributions involved in each state, i.e.

$$\left(\hat{\lambda}_{km}\right)_{\substack{1\leq k\leq K \\ M_1+1\leq m\leq M}}.$$

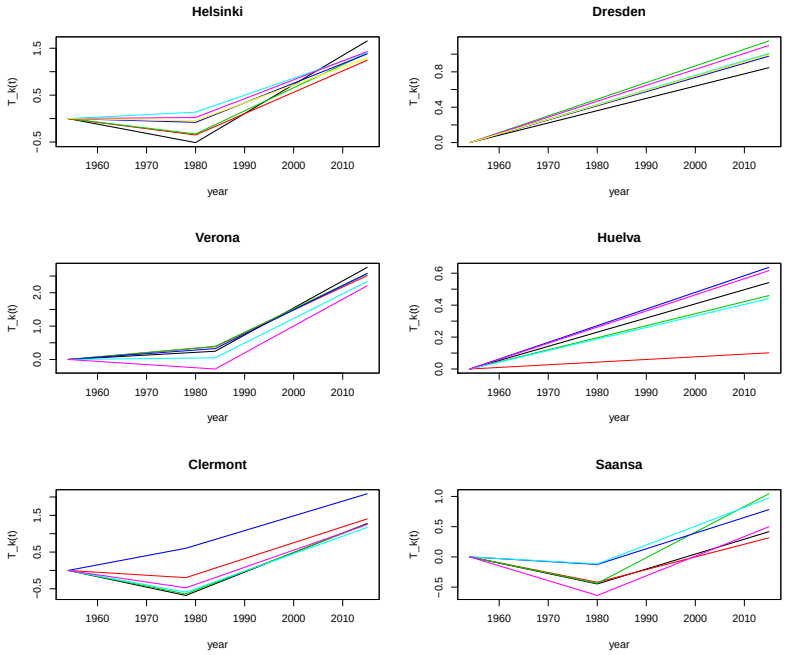

FIGURE 5.15 – Estimated trends (one color per state)

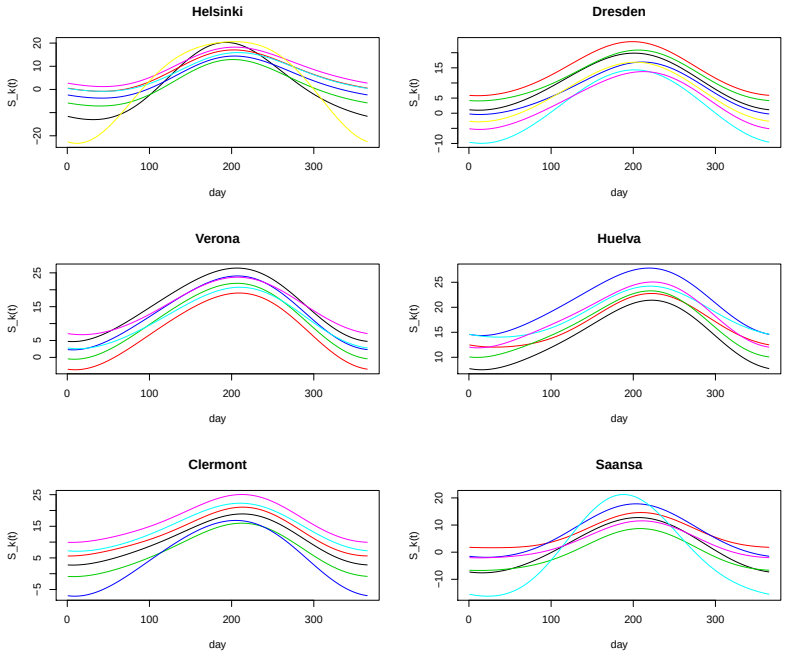

FIGURE 5.16 – Estimated seasonalities of temperature (one color per state)

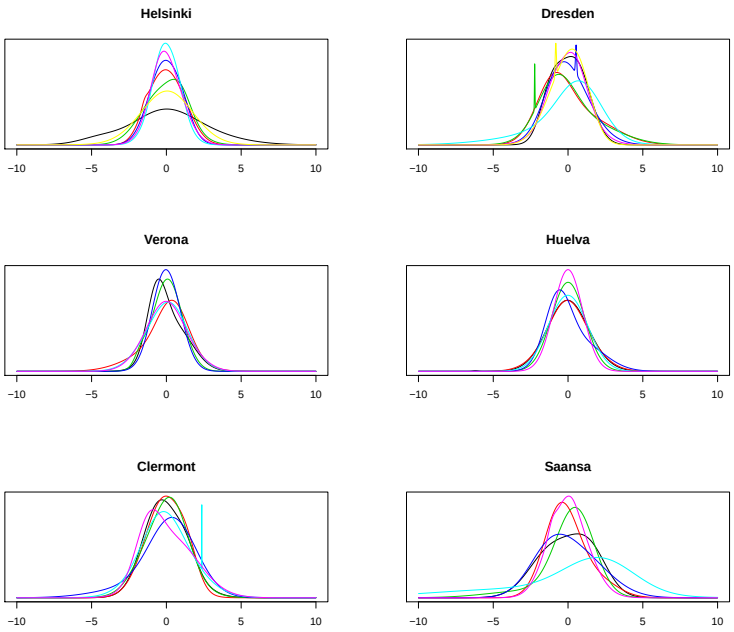

FIGURE 5.17 – Probability density functions of the random noises of the temperature (one color per state)

They represent the baseline intensity of rainfall (independently from the seasonal variations).

Until now, we have not tested the model using simulations, but the estimated parameters are consistent with what we would expect from precipitations and temperatures. The validation of the model through simulations is the purpose of the next paragraph.

Validation of the model

Simulation At this stage, for each site, we have ran the EM algorithm using our precipitation and temperature data as inputs, and we have at our disposal a vector of estimated parameters $\hat{\theta}$. This vector of parameters can be used to produce synthetic time series of temperature and precipitations $\left(Y_t^{\text{sim}}\right)_{1\leq t\leq n} = \left(Y_t^{(1),\text{sim}}, Y_t^{(2),\text{sim}}\right)_{1\leq t\leq n}$, where n is the length of our observed time series. The simulation procedure is the following.

1. Using the estimated transition matrices $\left(\hat{Q}(t)\right)_{1\leq t\leq n}$, simulate the states $\left(X_t^{\text{sim}}\right)_{1\leq t\leq n}$. The initial state X_1^{sim} can be chosen arbitrarily or drawn according to the stationary distribution of $\hat{Q}(1)$.
2. Given $\left(X_t^{\text{sim}}\right)_{1\leq t\leq n}$, simulate $\left(Y_t^{\text{sim}}\right)_{1\leq t\leq n}$. At time t, if $X_t^{\text{sim}} = k$,
   (a) Pick a component $m \in \{1, \ldots, M\}$ according to the probability vector $(\hat{p}_{km})_{1\leq m\leq M}$.
   (b) Take $Y_t^{(2),\text{sim}}$ as a realization of a $\mathcal{N}\left(\hat{T}_k(t) + \hat{S}_k(t) + \hat{\mu}_{km}, \hat{\sigma}_{km}^2\right)$ distribution. This is our simulated temperature at time t.

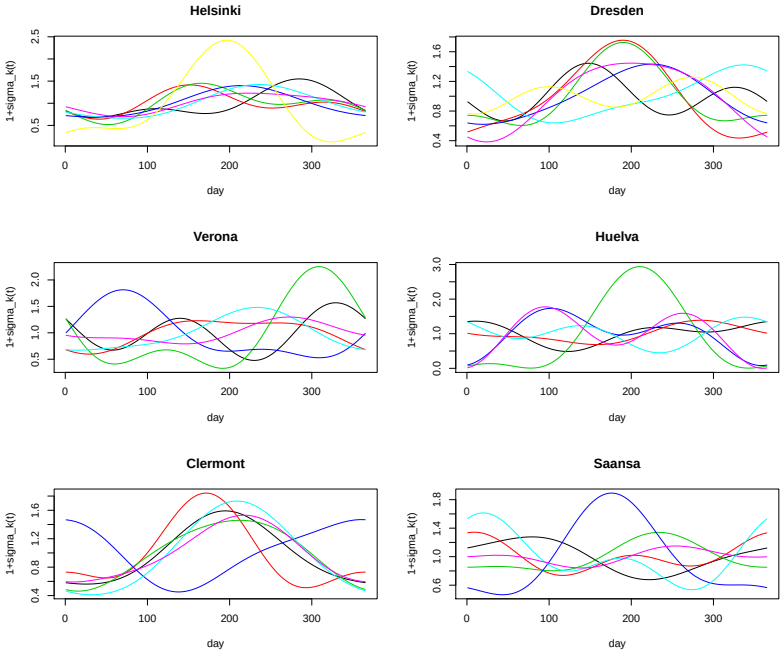

FIGURE 5.18 – Seasonalities in the precipitation intensity (one color per state).

(c) If $1 \leq m \leq M_1$, then $Y_t^{(1),\text{sim}} = 0$, else take $Y_t^{(1),\text{sim}}$ as a realization of the exponential distribution $\mathcal{E}\left(\frac{\hat{\lambda}_{km}}{1+\hat{\sigma}_k(t)}\right)$. This is our simulated precipitation at time t.

Note that the simulation algorithm gives insight into the reason why precipitation and temperature are dependent in our model : they always share the same state k, and the same mixture component m. This is crucial because we obviously do not want precipitation and temperature to be simulated independently.

Thus, using the above algorithm, we are able to simulate very easily and quickly a large number of synthetic time series of the same length as the observed ones. Using a standard laptop, we simulated $N_{\text{sim}} = 1000$ independent trajectories of length $n = 22265$ in a few minutes, for each site. Now our goal is to compare these simulations to the observed times series, in order to check if our simulations are realistic, with regard to several criteria. Our validation procedure is divided into three parts, aiming to answer the three following questions :

— is the distribution of the precipitation process $(Y_t^{(1)})_{1\leq t\leq n}$ well reproduced ?
— is the distribution of the temperature process $(Y_t^{(2)})_{1\leq t\leq n}$ well reproduced ?
— is the dependence structure between precipitation and temperature well reproduced ?

To this aim, we shall consider several statistics, or criteria of validation. These statistics will be computed from the data and from each of the N_{sim} simulated trajectory, thus providing a Monte-Carlo estimate of the distribution of each statistic under the model.

Temperature Forgetting about the temporal aspect of the temperature time series, we can start by looking at the overall distribution of the temperatures. Figure 5.19 shows the quantile-quantile plot (QQ-plot) of observed versus simulated temperatures, for each site. Here we mix

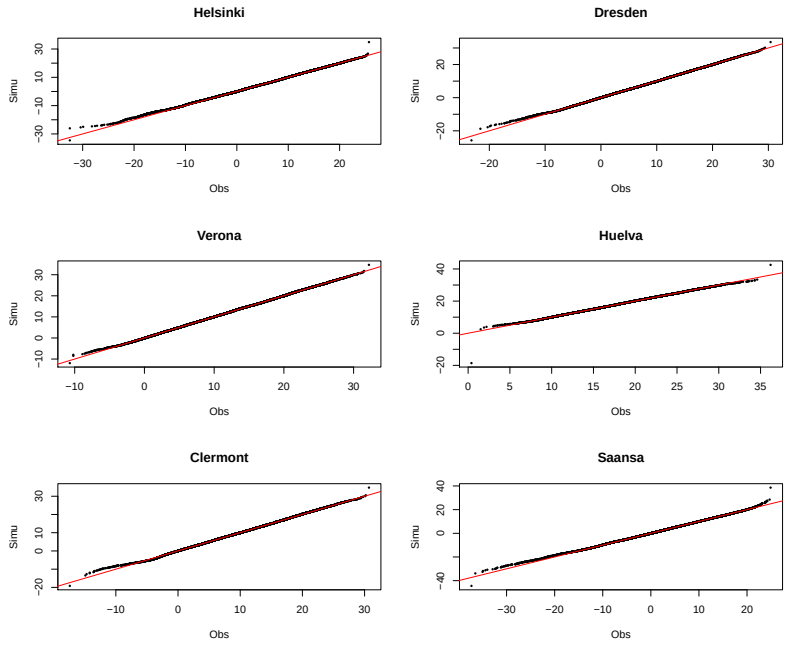

FIGURE 5.19 – Quantile-quantile plots of observed vs simulated temperatures

all the simulated values from the 1000 trajectories of length $n = 61 \times 365 = 22265$. We see that the model can generate values that are more extreme than those observed in the data, both in the left and the right tail of the distribution. In the station of Huelva, the model has generated unrealistically cold values : 61 values are below -5 whereas the minimum observed value is 0.4. However, these are rare, considering the large number $N_{\text{sim}} \times n$ of simulated values. Apart from that, the overall distribution is well reproduced.

Another way to look at the overall distribution of temperatures is to compute its observed quantiles and to compare them with the distributions of the quantiles generated by the model. For $\alpha \in (0, 1)$, we can compute the observed empirical α-th quantile, denoted by q_α^{obs}. Then we perform the same calculation for each simulation $s \in \{1, \dots, N_{\text{sim}}\}$ and we obtain the quantiles q_α^s. Finally, q_α^{obs} is compared to the distribution of $(q_\alpha^s)_{1 \le s \le N_{\text{sim}}}$ (estimated using a kernel density estimator). Figure 5.20 shows the example of the station of Clermont.

Now let us have a look at some daily statistics, beginning with daily moments of the temperature distribution. For a day of year $t \in \{1, \dots, 365\}$ (e.g. January 1st), we compute the empirical mean temperature at day t by averaging over the years the temperatures observed this day :

$$\bar{Y}_t^{(2)} := \frac{1}{N_{\text{year}}} \sum_{i=1}^{N_{\text{year}}} Y_{t+365(i-1)}^{(2)},$$

where N_{year} is the number of observed years (here $N_{\text{year}} = 61$). We perform the same calculation for each simulated scenario $s \in \{1, \dots, N_{\text{sim}}\}$, that is

$$\bar{Y}_{t,s}^{(2)} := \frac{1}{N_{\text{year}}} \sum_{i=1}^{N_{\text{year}}} Y_{t+365(i-1),s}^{(2)},$$

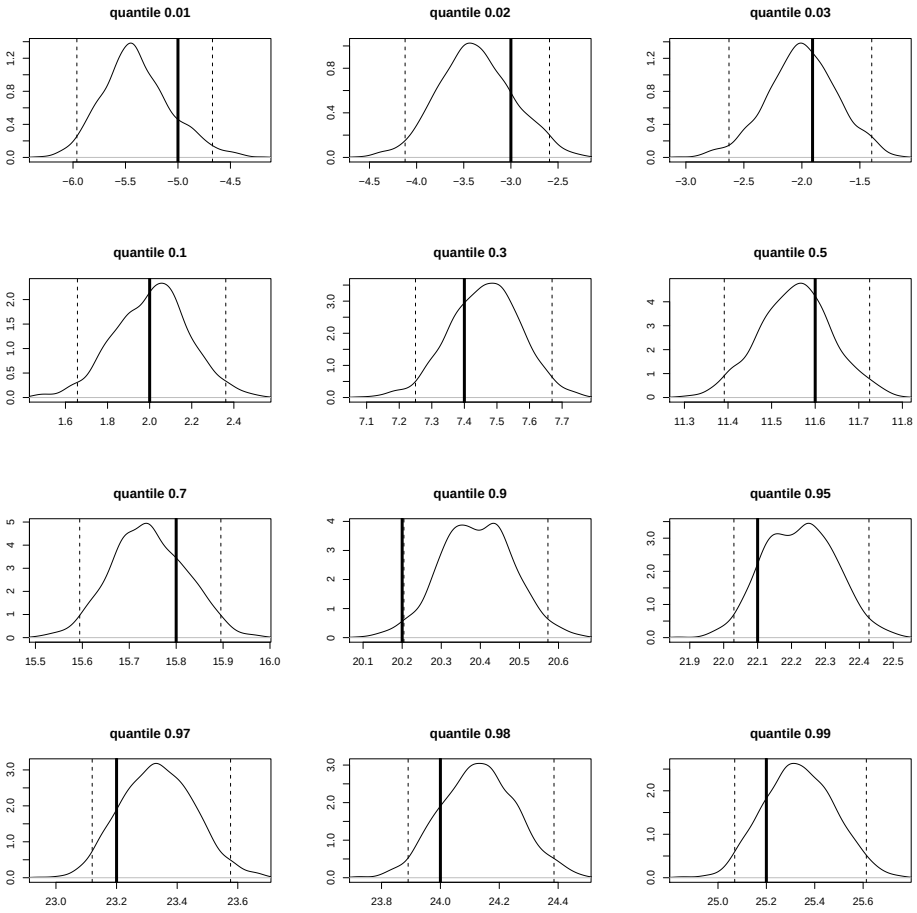

FIGURE 5.20 – Observed (vertical lines) superimposed to the distribution of simulated (curve) quantiles of temperature for the station of Clermont. The dashed vertical lines are the bounds of a 95% confidence interval.

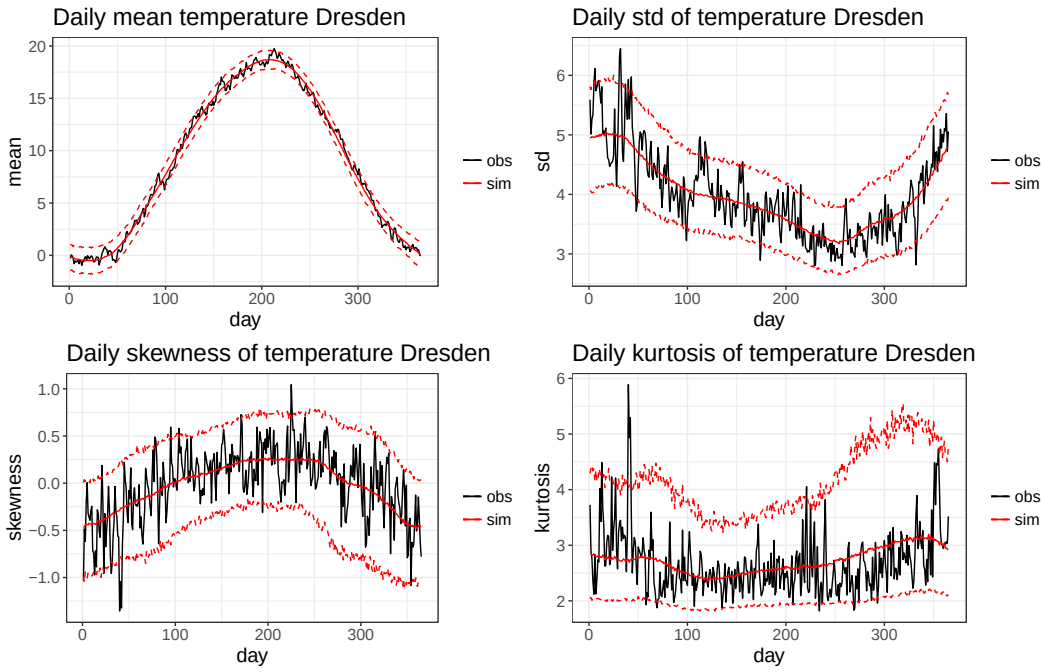

FIGURE 5.21 – Daily moments of temperature for Dresden : observations (black line), mean of simulations (red solid line), 2.5% and 97.5% quantiles of simulations (red dashed lines).

where $Y_{t,s}^{(2)}$ is the simulated temperature at time t in the s-th simulation. Then, using the simulated means $\left(\bar{Y}_{t,s}^{(2)}\right)_{1\leq s\leq N_{\text{sim}}}$, we can estimate the distribution of the mean temperature at day t under the model. To be specific, we compute the mean and a 95% confidence interval based on quantiles from the values $\left(\bar{Y}_{t,s}^{(2)}\right)_{1\leq s\leq N_{\text{sim}}}$. Finally, the same computations are performed for all $t \in \{1,\dots,365\}$, and for the next 3 moments : standard deviation, skewness (asymmetry coefficient) and kurtosis, as shown in Figure 5.21 for the station of Dresden and Figure 5.22 for the station of Huelva. The first daily moments are well reproduced by the model. Figure 5.21 highlights the seasonality in the variability of temperature, as the standard deviation is maximal in winter, then decreases until it reaches its minimum at the end of summer, before increasing again. The shape of this seasonality is common to all the stations we studied, except Huelva (see Figure 5.22). This is consistent with what was observed in Figure 5.4. Another interesting observation is the asymmetry of the distribution of temperatures, measured by the third moment (skewness). Recall that a negative (resp. positive) skewness means that the distribution is skewed to the left (resp. right) whereas a skewness of 0 means that the distribution is symmetric. The temperatures in Dresden clearly exhibit a seasonal behaviour in the asymmetry : the skewness is negative in winter and positive in summer. This reflects the presence of cold extremes in winter and hot extremes in summer. Using gaussian mixtures as emission distributions instead of simple gaussian distributions allows the model to reproduce this asymmetry. As Figure 5.22 shows, the station of Huelva, whose climate strongly differs from the climate of Dresden, does not exhibit the same seasonal behaviour, as the skewness curve is rather flat.

Besides the moments, it is interesting to pay attention to the interannual minimum and maximum

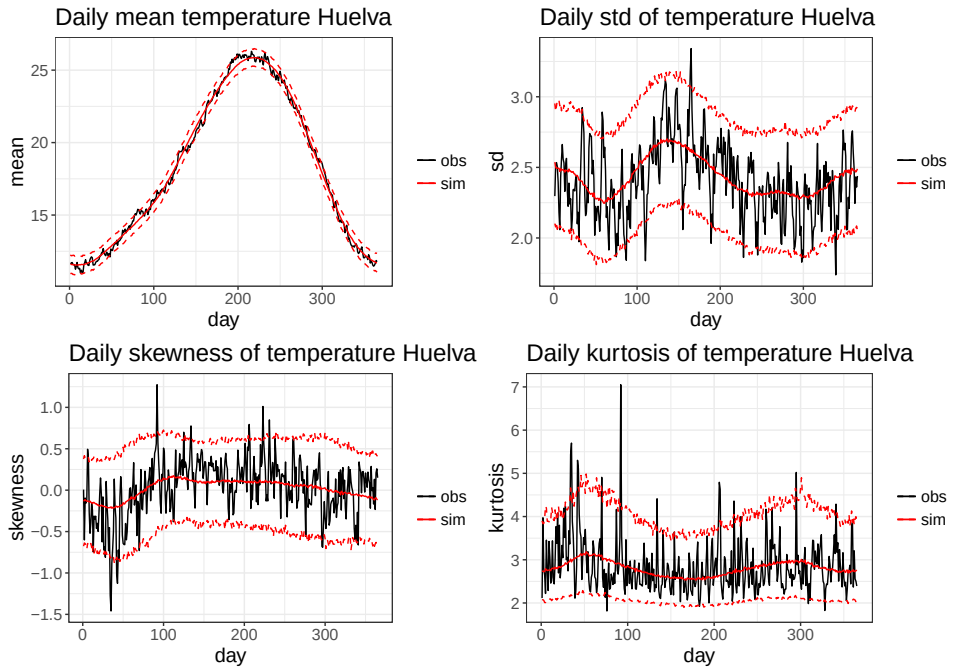

FIGURE 5.22 – Daily moments of temperature for Huelva.

of daily mean temperature for each day. Precisely, for a day of year $t \in \{1, \dots, 365\}$, the observed maximum is $\max_{1 \leq i \leq N_{\text{year}}} Y^{(2)}_{t+365(i-1)}$ and we estimate the distribution of this quantity under the model using the simulations, in the same way as we did for the moments. A similar computation is performed for the daily minima. This statistic is well reproduced by the model. As an example, Figure 5.23 shows the results for Clermont, Helsinki and Huelva.

The temperatures do not form an independent process, as it is strongly autocorrelated. In our model, autocorrelation is introduced through the hidden Markov chain : even though the observations are generated independently conditionally to the states, they are not independent because the state process is autocorrelated. Hence the next criterion to be considered is the empirical autocorrelation of temperature with lags 1, 2 and 3 days. Figure 5.24 shows that the model slightly underestimates the autocorrelations.

The last criterion that we are interested in regarding temperature is its persistence in extreme values. We fix some threshold u (e.g. the α-th quantile of temperature, with α close to 1) and we consider the durations of the episodes exceeding u. Let us give an example. Assume that, for some $t \geq 2$, $Y^{(2)}_{t-1} \leq u$, $Y^{(2)}_t > u$, $Y^{(2)}_{t+1} > u$, $Y^{(2)}_{t+2} > u$ and $Y^{(2)}_{t+3} \leq u$. In such a case, we say that $\left(Y^{(2)}_t, Y^{(2)}_{t+1}, Y^{(2)}_{t+2}\right)$ is a *hot* cluster of length 3 because the temperature remains for 3 days above the threshold u. Thus the length of a cluster is a positive integer, possibly 1. Similarly, we can define *cold* clusters by considering the times when the temperature drops below some low threshold. The results for hot clusters are shown by Figure 5.25. Here the threshold u is the 95-th percentile (hence it varies according to the site). Clearly, for all sites, the model produces too many clusters of length 1, therefore not enough longer clusters. We tried various thresholds between the 90-th and the 99-th quantiles and the same conclusion can be drawn. Thus the persistence in extreme values is underestimated by the model. This is not surprising, considering

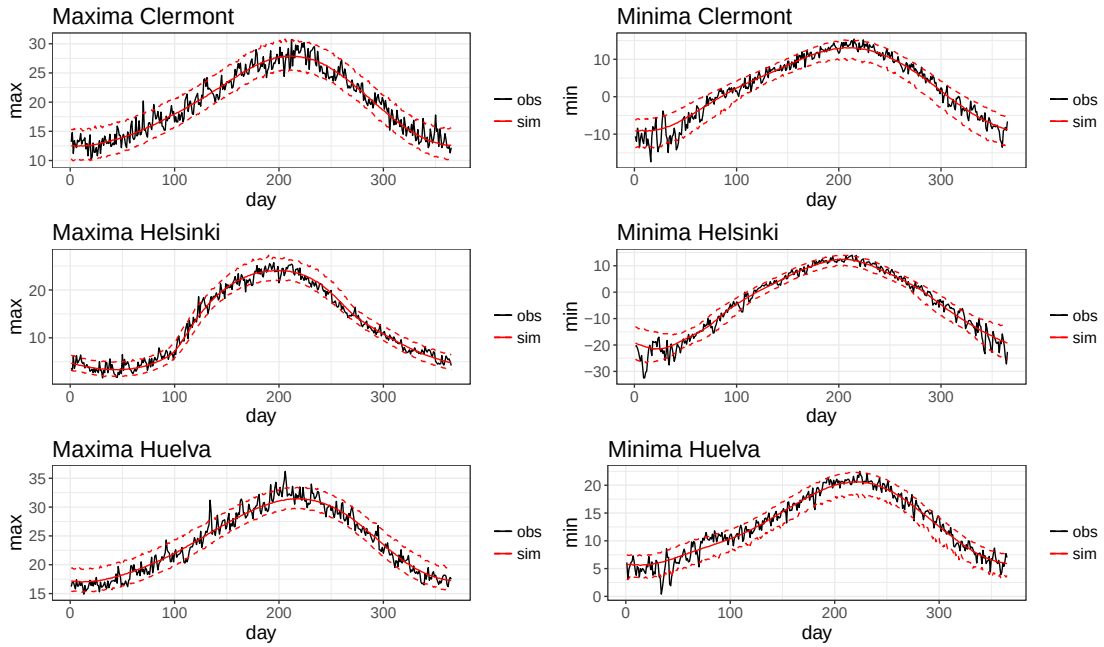

FIGURE 5.23 – Interannual minimum and maximum of daily mean temperature for Clermont, Helsinki and Huelva.

that the autocorrelations of temperature are slightly underestimated too (see Figure 5.24). The same issue appears when we consider cold clusters, as we can see in Figure 5.26.

Precipitation Figure 5.27 shows the quantile-quantile plots of precipitation, for each of the six sites. To be specific, we plotted the α-quantiles of observed precipitations versus the corresponding simulated α-quantiles, for $\alpha \in \{\frac{i}{1000}, 1 \leq i \leq 999\}$. As the distribution of precipitations is very asymmetric, we zoom-in to the α greater than 0.9 (see Figure 5.28). As we can see, the overall distribution of precipitations, including its tail, is well reproduced. It is also interesting to note that our model is able to simulate precipitation values that are larger than all the observed values. As an example, the maximum observd value of daily precipitation in Helsinki is 79.3 mm, but 14 of the 1000 simulated trajectories include a larger value, the maximum being 110.7 mm. This is one of the assets of model compared to models based on resampling. However, the exponential distribution being unbounded, performing a large number of simulations sometimes leads to unrealistic precipitations values.

We can also check the distributions of the simulated quantiles. As an example, Figure 5.29 shows the distributions of some upper quantiles for the station of Snåsa, and their observed counterparts.

As we did for temperature, we estimate the daily moments of precipitations. We also estimate the daily frequencies of precipitations by computing, for $t \in \{1, \ldots, T\}$, an estimate of $\mathbb{P}(Y_t^{(1)} > 0)$ as

$$\frac{1}{N_{\text{year}}} \sum_{i=1}^{N_{\text{year}}} \mathbf{1}_{Y_{t+365(i-1)}^{(1)} > 0},$$

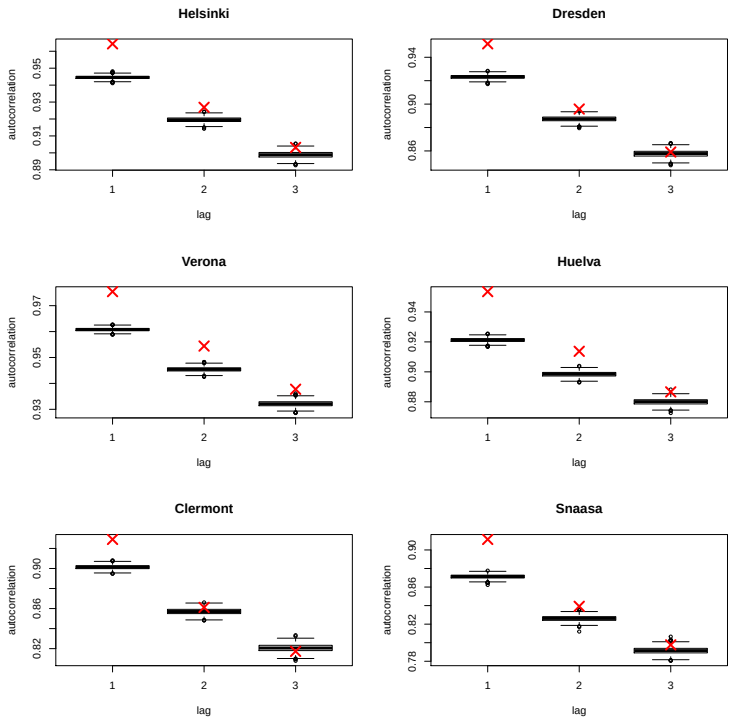

FIGURE 5.24 – Autocorrelation of temperatures : the red crosses are the observed values and the boxplots are related to the different simulations.

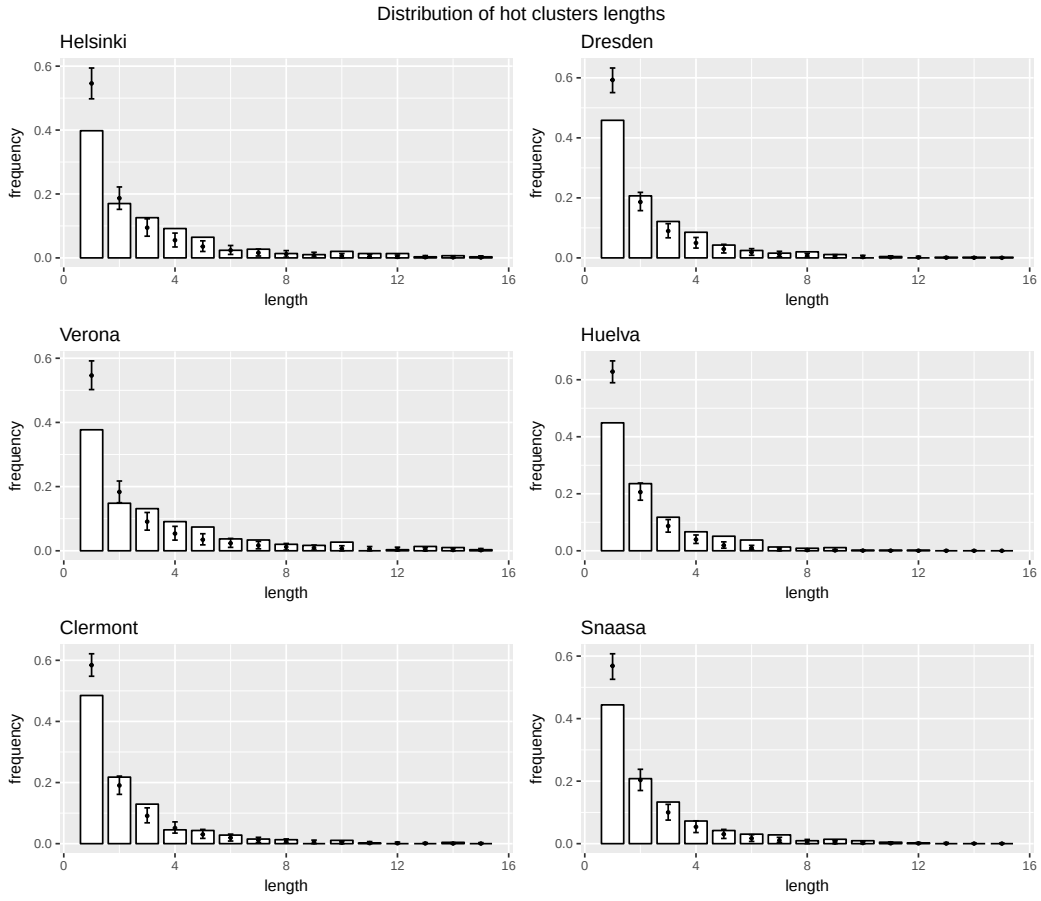

FIGURE 5.25 – Distribution of the lengths of the hot clusters. The threshold is the 95-th percentile. White bars : observed values. Errorbars : 95% confidence interval based on the simulations.

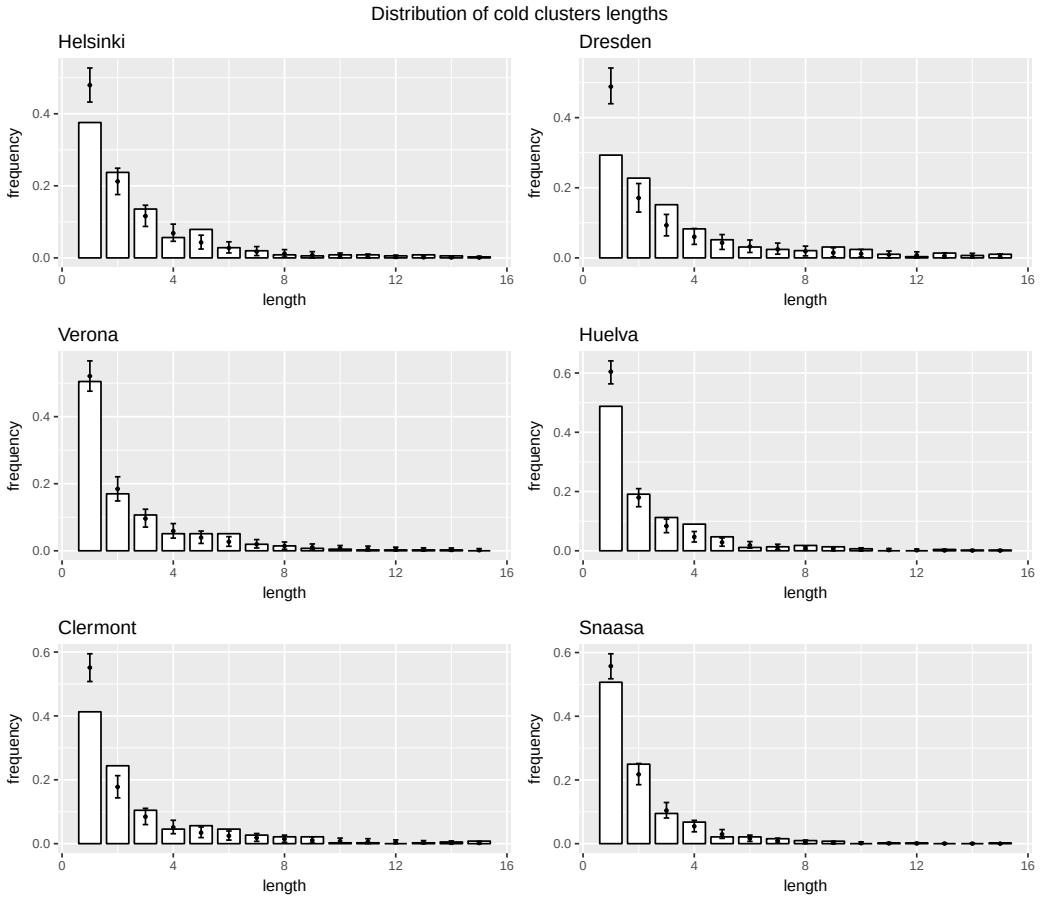

FIGURE 5.26 – Distribution of the lengths of the cold clusters. The threshold is the 5-th percentile. White bars : observed values. Errorbars : 95% confidence interval based on the simulations.

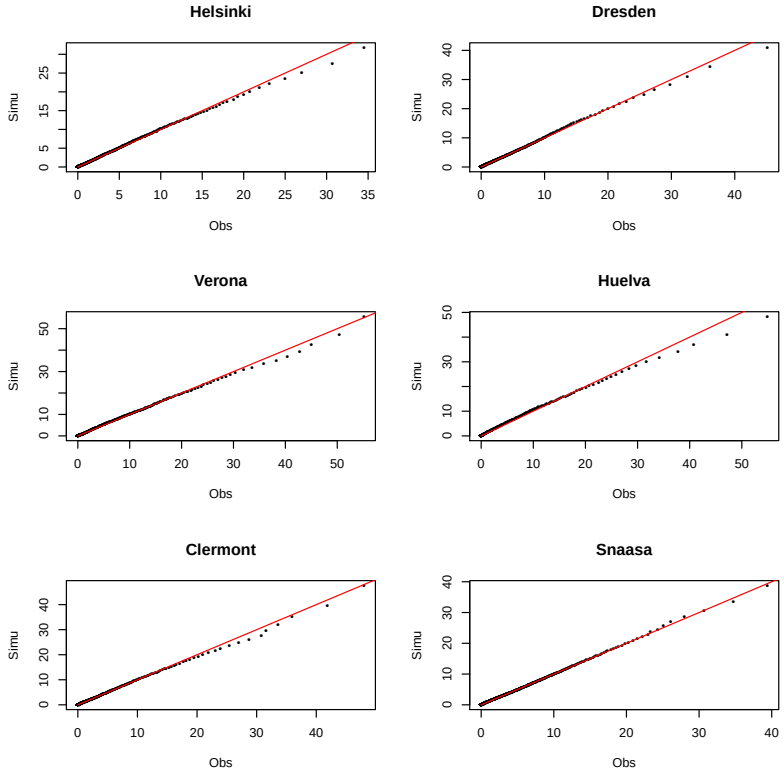

FIGURE 5.27 – Precipitation quantile-quantile plots.

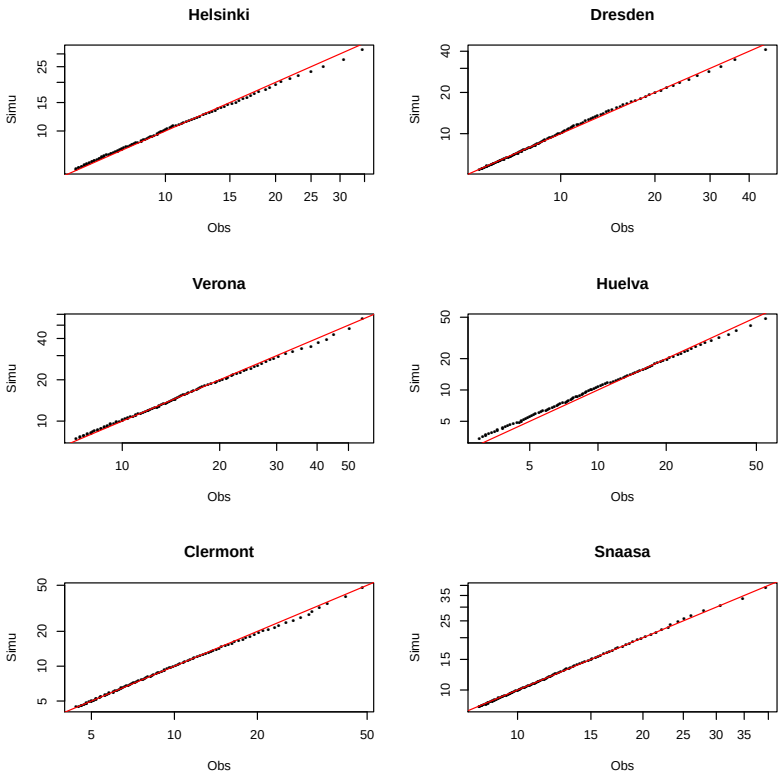

FIGURE 5.28 – Precipitation quantile-quantile plots, tail of distribution (logarithmic scale).

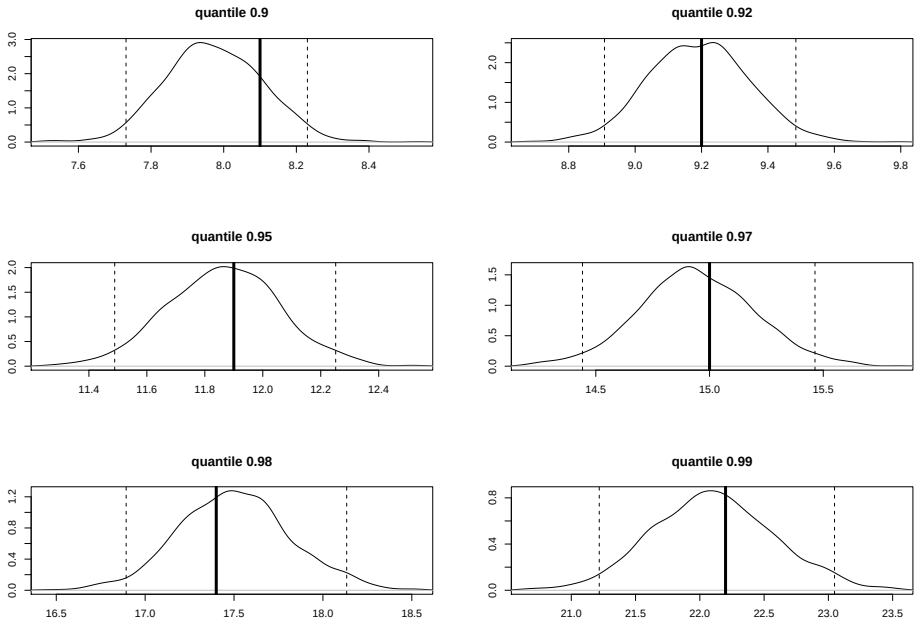

FIGURE 5.29 – Observed (vertical lines) superimposed to the distribution of simulated (curve) quantiles of precipitation for the station of Snåsa. The dashed vertical lines are the bounds of a 95% confidence interval.

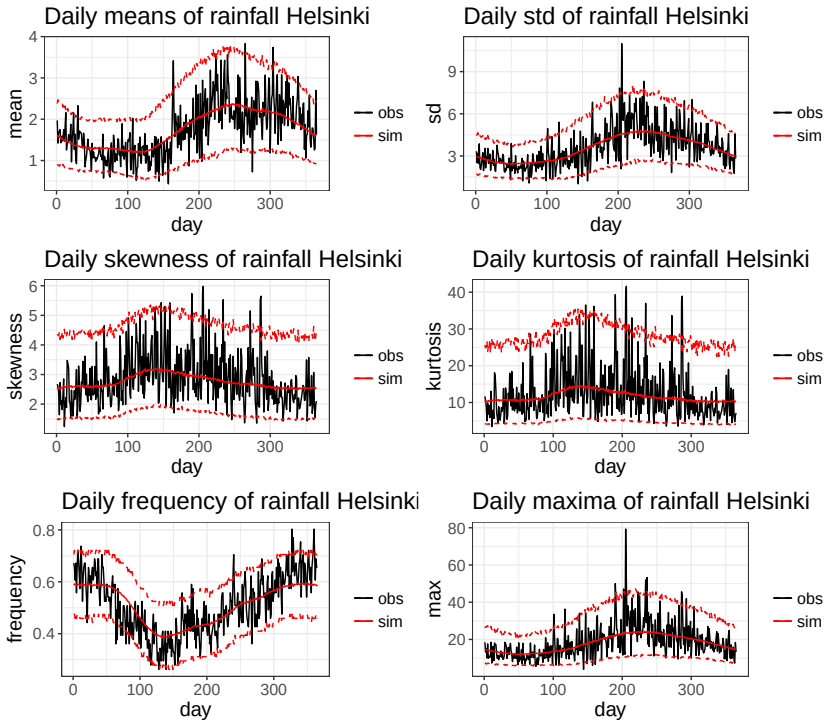

FIGURE 5.30 – Daily moments of precipitations

and the maxima of daily precipitations totals :

$$\max_{1\leq i\leq N_{\text{year}}} Y^{(1)}_{t+365(i-1)}, \quad 1 \leq t \leq 365$$

Using the simulations, we estimate the distribution of these statistics under the model. The results are presented in Figure 5.30 for the station of Helsinki, together with the first four daily moments. The seasonalities in the intensity and in the frequency of precipitations are well reproduced by the model.

A dry spell is a period of time during which it does not rain. When there are d consecutive days without rain and non-zero precipitations on the $(d+1)$-th day, this constitutes a dry spell of length d. Similarly, we define wet spells as consecutive days with non-zero precipitations. Figures 5.31 and 5.32 show the observed and simulated distributions of dry and wet spells. The lengths of dry spells are quite well reproduced by the model but for some stations (e.g. Clermont), the model clearly underestimates the number of wet spells longer than one day.“
Stochastic precipitations generators often underestimate the interannual variability of precipitations (Katz and Parlange, 1998). Thus we focus on yearly rainfall and we look at its interannual variability. The histograms in Figure 5.33 are the observed distributions of yearly precipitations (thus each histogram has been computed with 61 observations). The lines are the kernel density estimations of simulated yearly precipitations. We have performed the same computations with monthly precipitations (see Figure 5.34 for the station of Clermont). Our model does not underestimate interannual variability, as it is able to generate rainy as well as dry years or months.

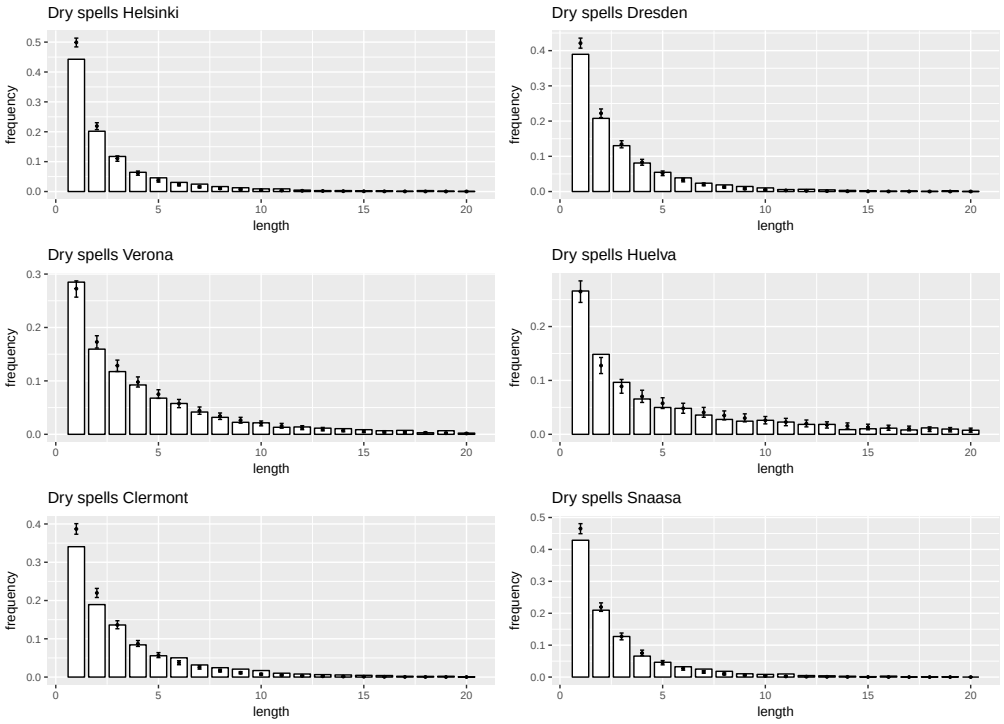

FIGURE 5.31 – Distribution of the length of dry spells. White bars : observed values. Errorbars : 95% confidence interval based on the simulations.

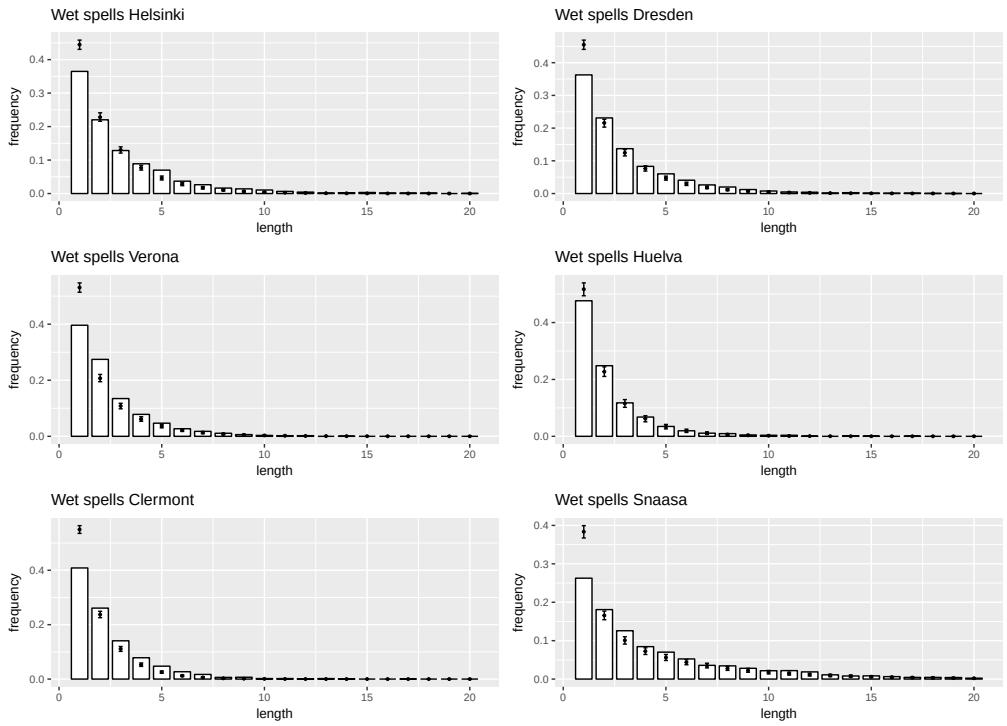

FIGURE 5.32 – Distribution of the length of wet spells. White bars : observed values. Errorbars : 95% confidence interval based on the simulations.

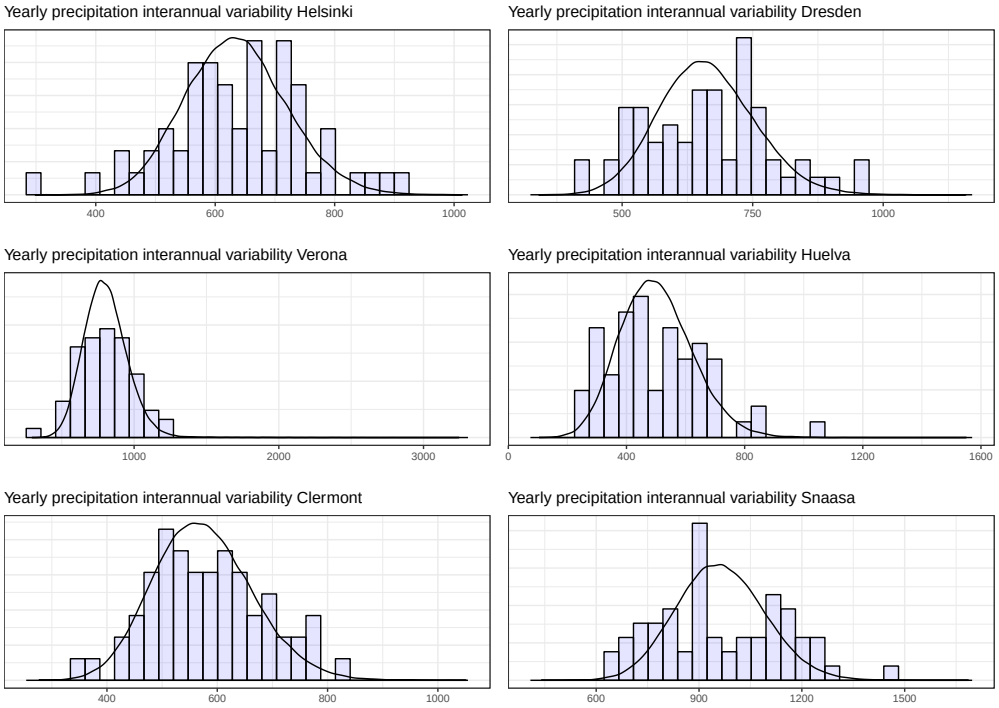

FIGURE 5.33 – Interannual variability of yearly precipitations.

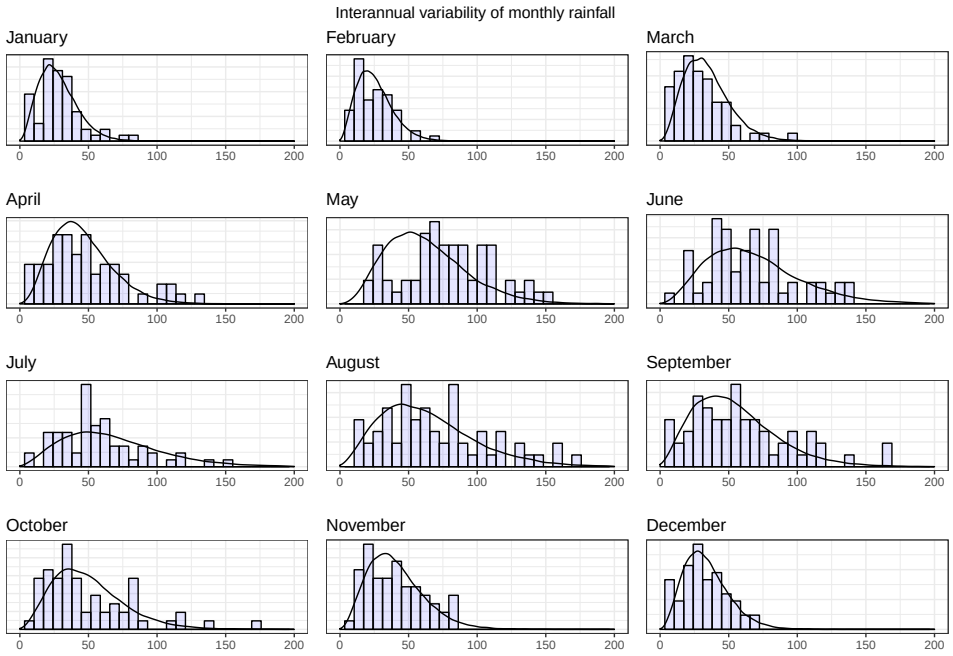

FIGURE 5.34 – Interannual variability of monthly precipitations.

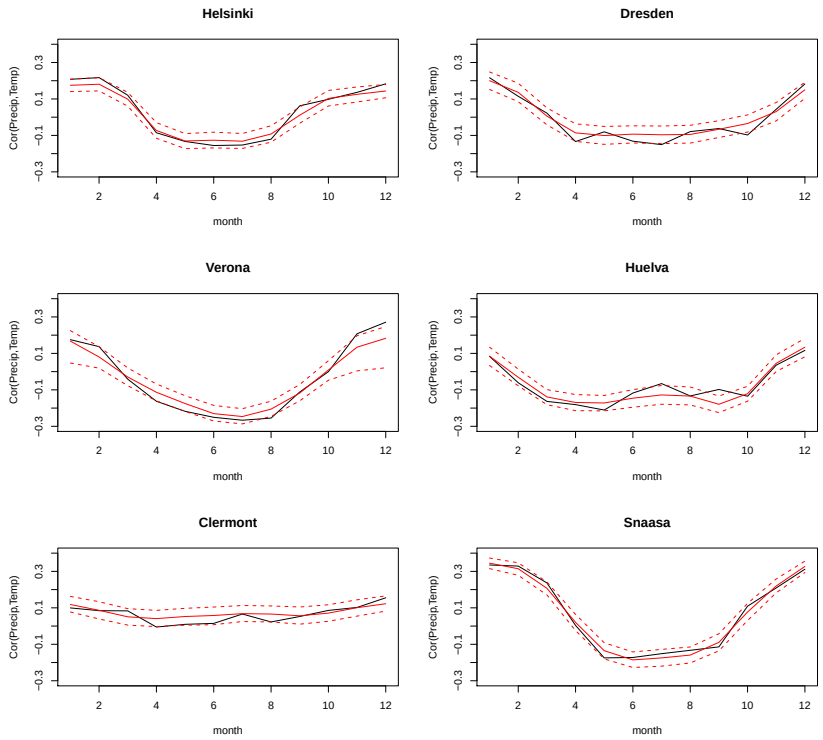

FIGURE 5.35 – Monthly correlations between temperature and precipitations. Black solid line : observed. Red solid line : mean of simulations. Red dashed lines : 95% confidence interval from simulations.

Temperature and precipitation coupling At this stage, we have assessed the performance of our model for temperature and precipitations separately. We shall now concentrate on the relationship between these two variables, as they are not independent. Figure 5.35 shows that the model provides realistic monthly correlations between temperature and precipitations.

However, a better representation of the relationship between temperature and precipitations can be obtained. For example, the probability of observing precipitations varies with temperature : for $u, v \in \mathbb{R}$, in general,

$$\mathbb{P}\left(Y_t^{(1)} > 0 \mid Y_t^{(2)} = u\right) \neq \mathbb{P}\left(Y_t^{(1)} > 0 \mid Y_t^{(2)} = v\right).$$

Similarly, the expected value of non-zero precipitation depends on temperature : for $u, v \in \mathbb{R}$, in general,

$$\mathbb{E}\left[Y_t^{(1)} \mid Y_t^{(1)} > 0, Y_t^{(2)} = u\right] \neq \mathbb{E}\left[Y_t^{(1)} \mid Y_t^{(1)} > 0, Y_t^{(2)} = v\right].$$

Note that the quantities $\mathbb{P}\left(Y_t^{(1)} > 0 \mid Y_t^{(2)} = u\right)$ and $\mathbb{E}\left[Y_t^{(1)} \mid Y_t^{(1)} > 0, Y_t^{(2)} = u\right]$ depend not only on u but also on t as the process $\left(Y_t^{(1)}, Y_t^{(2)}\right)_{t\geq 1}$ is not stationary.

Let $\mathcal{K}$ be the gaussian kernel defined by $\mathcal{K}(x) = \exp\left(-\frac{x^2}{2}\right)$. For $y \in \mathbb{R}$ and a bandwidth $h > 0$, we consider the following statistics.

$$r(y) := \frac{\sum_{t=1}^{n} \mathcal{K}\left(\frac{Y_t^{(2)}-y}{h}\right) \mathbf{1}_{Y_t^{(1)}>0}}{\sum_{t=1}^{n} \mathcal{K}\left(\frac{Y_t^{(2)}-y}{h}\right)}$$

$$R(y) := \frac{\sum_{t=1}^{n} \mathcal{K}\left(\frac{Y_t^{(2)}-y}{h}\right) Y_t^{(1)}}{\sum_{t=1}^{n} \mathcal{K}\left(\frac{Y_t^{(2)}-y}{h}\right) \mathbf{1}_{Y_t^{(1)}>0}}$$

If we had a sample $\left(Y_t^{(1)}, Y_t^{(2)}\right)_{1\leq t\leq n}$ of i.i.d. copies of $\left(Y^{(1)}, Y^{(2)}\right)$, then $r(y)$ would be an estimator of $\mathbb{P}\left(Y^{(1)} > 0 \mid Y^{(2)} = y\right)$ and $R(y)$ would be an estimator of $\mathbb{E}\left[Y^{(1)} \mid Y^{(1)} > 0, Y^{(2)} = y\right]$. Although this is not the case, these statistics still provide some information on the dependence between precipitation occurrence and temperature and it is interesting to see how the model behaves with respect to $r(y)$ and $R(y)$. Therefore, the functions $y \mapsto r(y)$ and $y \mapsto R(y)$ are computed from the observations and from each of the 1000 bivariate simulations. Figures 5.36 and 5.37 show the results for the six stations (with $h = 2$). Here again, the model is performing well with regard to these statistics.

For $t \in \{1, \dots n\}$, we denote by $\bar{t} \in \{1, \dots, 365\}$ its representative modulo 365, that is the day of year. For $t, s \in \{1, \dots, n\}$, let

$$\|t - s\| := \min(|\bar{t} - \bar{s}|, 365 - |\bar{t} - \bar{s}|)$$

be the cyclic distance between days t and s, that is the number of days between the corresponding days of year. For $h_1, h_2 > 0$, $y \in \mathbb{R}$ and $t \in \{1, \dots, 365\}$, let us define

$$r(t, y) := \frac{\sum_{s=1}^{n} \mathcal{K}\left(\frac{\|t-s\|}{h_1}\right) \mathcal{K}\left(\frac{y-Y_s^{(2)}}{h_2}\right) \mathbf{1}_{Y_s^{(1)}>0}}{\sum_{s=1}^{n} \mathcal{K}\left(\frac{\|t-s\|}{h_1}\right) \mathcal{K}\left(\frac{y-Y_s^{(2)}}{h_2}\right)},$$

and

$$R(t, y) := \frac{\sum_{s=1}^{n} \mathcal{K}\left(\frac{\|t-s\|}{h_1}\right) \mathcal{K}\left(\frac{y-Y_s^{(2)}}{h_2}\right) Y_s^{(1)}}{\sum_{s=1}^{n} \mathcal{K}\left(\frac{\|t-s\|}{h_1}\right) \mathcal{K}\left(\frac{y-Y_s^{(2)}}{h_2}\right) \mathbf{1}_{Y_s^{(1)}>0}}.$$

Then $r(t, y)$ is a proxy for $\mathbb{P}\left(Y_t^{(1)} > 0 \mid Y_t^{(2)} = y\right)$ and $R(t, y)$ is a proxy for the conditional expectation $\mathbb{E}\left[Y_t^{(1)} \mid Y_t^{(1)} > 0, Y_t^{(2)} = u\right]$. Hence we use these statistics as validation criteria. Figures 5.38 and 5.39 show the results for the station of Verona for four different days of year. Results are in general satisfying and demonstrate a realistic coupling between the two variables.

5.4 Conclusion

We introduced a seasonal hidden Markov model for the joint modeling of daily temperature and precipitations. The nonhomogeneity of the underlying Markov chain allows the model to account for the complex seasonal features of these weather variables, as well as climate change, in a unified framework, without resorting to pre-processing the data or fitting multiple models. Our model

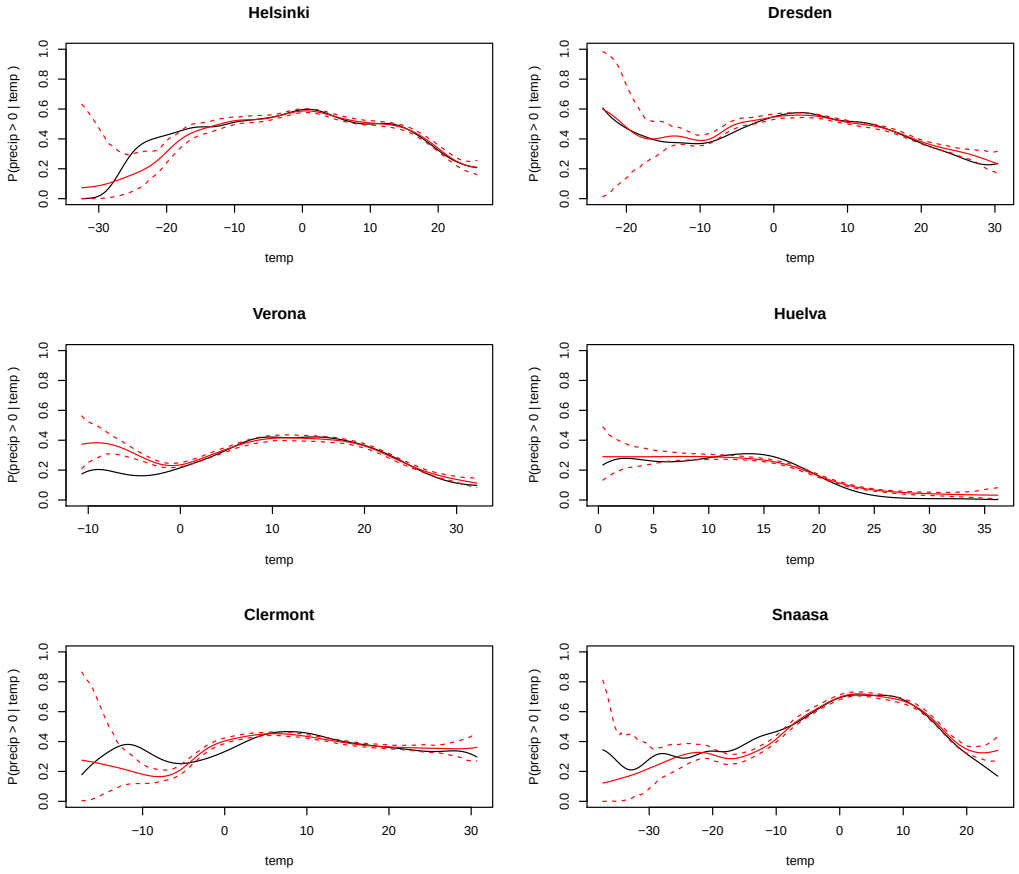

FIGURE 5.36 – Graph of the function $y \mapsto r(y)$. Solid black curve : computation on the observations. Solid red curve : mean of the simulated values. Dashed red curves : 95% confidence interval based on simulations.

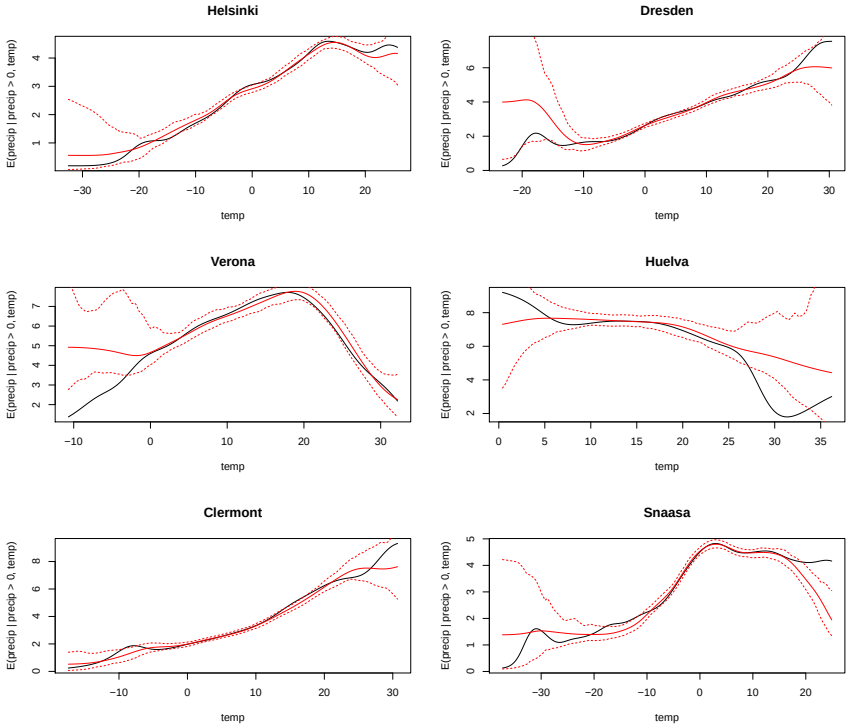

FIGURE 5.37 – Graph of the function $y \mapsto R(y)$. Solid black curve : computation on the observations. Solid red curve : mean of the simulated values. Dashed red curves : 95% confidence interval based on simulations.

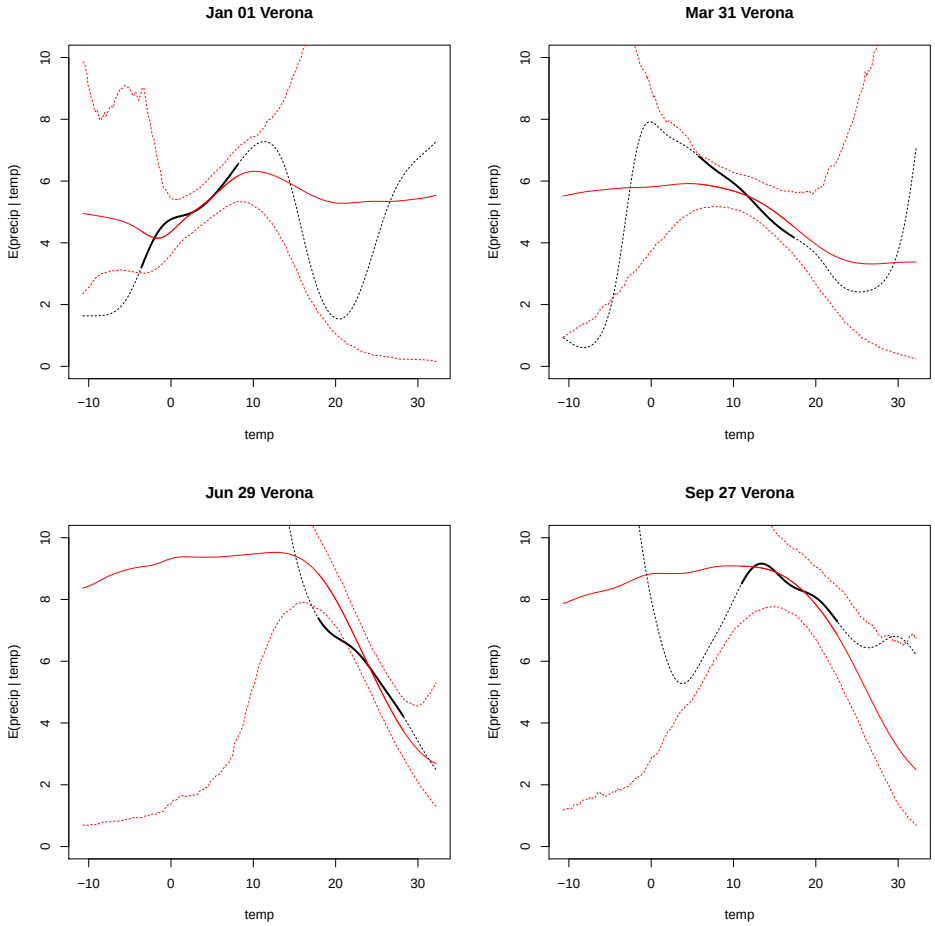

FIGURE 5.38 – Graphs of the function $y \mapsto r(t, y)$ for different values of t (day of year) in Verona. Solid black curve : computation on the observations. Solid red curve : mean of the simulated values. Dashed red curves : 95% confidence interval based on simulations. The dashed black curve is an extrapolation of $y \mapsto r(t, y)$ outside the observed range of temperature at day of year t : although the computation is feasible, it is not relevant.

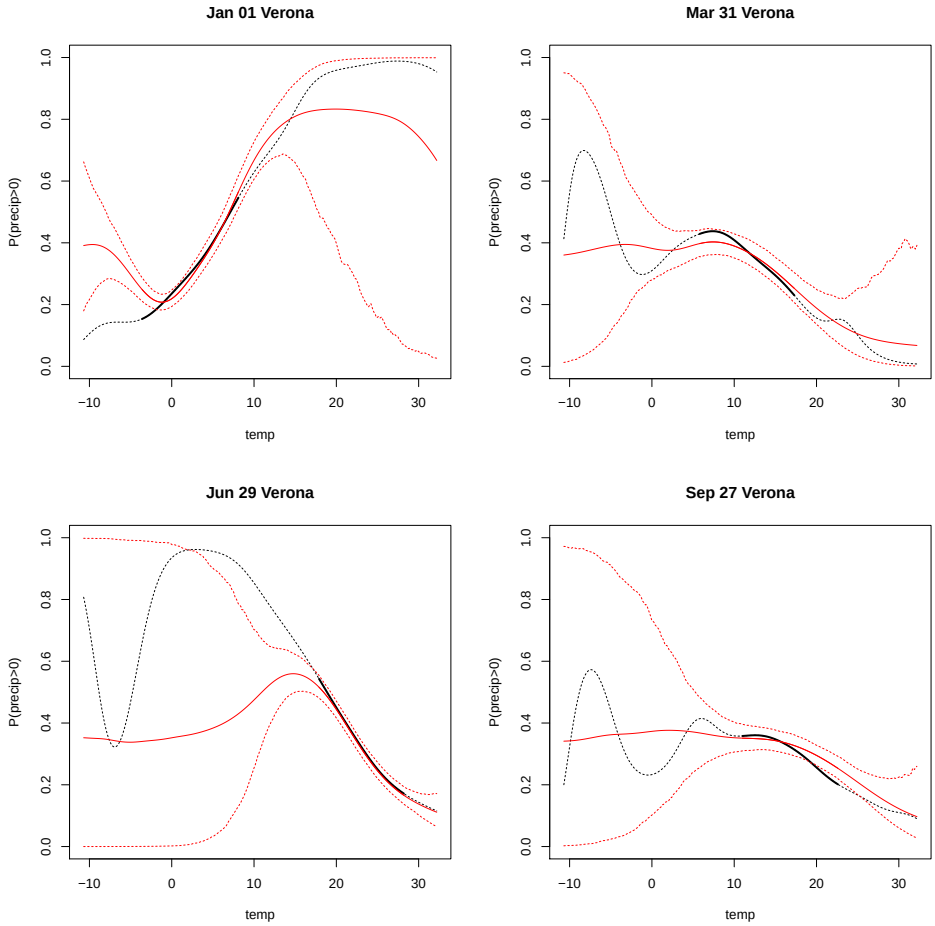

FIGURE 5.39 – Graphs of the function $y \mapsto R(t, y)$ for different values of t (day of year) in Verona. Solid black curve : computation on the observations. Solid red curve : mean of the simulated values. Dashed red curves : 95% confidence interval based on simulations. The dashed black curve is an extrapolation of $y \mapsto r(t, y)$ outside the observed range of temperature at day of year t : although the computation is feasible, it is not relevant.

can be used as a stochastic weather generator, as it can quickly generate realistic synthetic time series of temperature and precipitations, at a given site. Considering many criteria of interest, we showed that these simulations closely reproduce the behaviour of the data, be it the marginal distributions of the two variables or their dependence relationships. Furthermore, we showed that investigating the estimated parameters of the model leads to giving *a posteriori* a physical intepretation to the hidden states, thus avoiding the pitfall of a *black box* model. We also proved the robustness of our model by testing it on different sites with various climates.

Several extensions of our model can be considered and be the subject of future works. First, we noted that in many cases, it fails to reproduce correctly the extreme heat or cold episodes, and the dry and rainy spells. This flaw can be caused by a lack of autocorrelation and could be adressed by adding autoregression in the process. Using our notations, the distribution of Y_t could depend on t, Y_{t-1} and X_t instead of just being a function of t and X_t. Then, extreme values of temperature and precipitations can be investigated more closely. We did not focus on this particular point but some applications need a fine modeling of extremes. To this aim, it may be necessary to choose other emission distributions, even though we showed that the upper quantiles of temperature and precipitations were well reproduced. In order to apply the model to sites where there is a sensible trend in the distibution of precipitations, it would have to be modified. Finally, the structure of our model can easily be extended to more variables (e.g. wind speed), the main difficulty being the choice of the emission distributions.

Chapitre 6

Application hydrologique

6.1 Introduction

Dans ce chapitre, nous présentons un travail réalisé en collaboration avec la Direction Technique Générale (DTG) d'EDF. Ce travail constitue un exemple concret d'application de notre générateur bivarié pluie/température. Un *modèle hydrologique* permet à la DTG de modéliser le fonctionnement d'un bassin versant (voir Garçon (1996)). Schématiquement, les entrées du modèle hydrologique sont des chroniques de température et de précipitations sur le bassin versant considéré, et il fournit en sortie une chronique de débit. Plusieurs variables intermédiaires sont calculées, comme l'enneigement ou l'évapotranspiration (évaporation de l'eau du sol et transpiration des plantes). Les paramètres d'un tel modèle sont calibrés statistiquement sur chaque bassin versant. Ainsi, on peut utiliser des simulations de précipitations et température obtenues grâce à un générateur de temps en entrée d'un modèle hydrologique, pour obtenir des simulations de débit. (voir Figure 6.1).

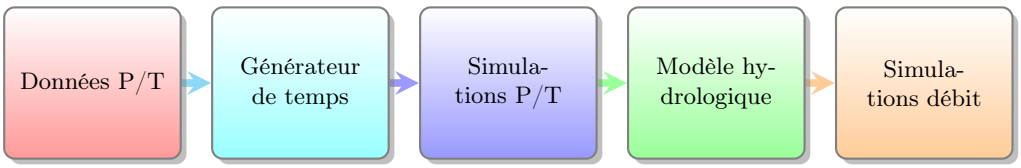

FIGURE 6.1 – Simulations de débits

Nous ne nous intéressons pas ici au modèle hydrologique, que l'on considèrera comme une "boîte noire", mais dans la Section 6.2, nous allons comparer les simulations de précipitations et température obtenues par le générateur de la DTG d'une part, et par notre générateur bivarié introduit dans le Chapitre 5 d'autre part, ainsi que les simulations de débit correspondantes. Dans la Section 6.3, nous utilisons notre générateur, couplé au modèle hydrologique, pour estimer, sur différentes périodes, la distribution bivariée des températures et des débits hivernaux, au pas de temps annuel. Enfin, dans la Section 6.4, nous estimons les états cachés du HMM et nous tentons de les intepréter en les liant à des variables météorologiques à plus grande échelle.

Méthode des analogues Le générateur de temps de la DTG est basé sur une méthode de rééchantillonage appelée méthode des analogues. Celle-ci s'appuie sur les champs géopotentiels à 700 et 1000 hPa [1] à 0h et 24h sur une grille de 110 points centrée sur le Sud-Est de la France, comme expliqué dans Garavaglia et al. (2010). A chaque jour est donc associé un point dans l'espace $\mathbb{R}^{440}$ des géopotentiels. Cet espace est muni d'une distance, appelée *score de Teweles-Wobus* (Teweles Jr and Wobus, 1954), optimisée pour la comparaison de champs géopotentiels. Lorsque l'on veut générer précipitations et température pour un jour donné t, on choisit les précipitations et la température d'un jour tiré aléatoirement parmi les *analogues* du jour t, c'est-à-dire les 10 jours les plus proches de t au sens de la distance de Teweles-Wobus. Pour respecter la saisonnalité, les analogues sont sélectionnés uniquement dans une fenêtre temporelle de 60 jours autour de t. La simple méthode des analogues telle que décrite ici ne permet pas de prendre en compte le changement climatique. En effet, on pourrait par exemple simuler une température de 2010 en utilisant un analogue de 1960, ou inversement, ce qui pose problème compte tenu du fait que la température possède une tendance croissante. Il est donc nécessaire de procéder à un ajustement. On utilise pour cela une température "grande échelle" T^{GE} comme témoin du réchauffement climatique. Cette température grande échelle est obtenue à partir de l'épaisseur de la couche d'air entre les champs géopotentiels à 700 et 1000 hPa. Supposons que l'on souhaite simuler la température T_t^{sim} à la date t à partir de la date analogue s. On a alors

$$T_t^{\mathrm{sim}} = T_s + (T_t^{\mathrm{GE}} - T_s^{\mathrm{GE}}).$$

Cette méthode est à rapprocher de la méthode ANATEM écrite dans Kuentz et al. (2015).

Nous venons bien de décrire un générateur de temps, puisque cette procédure permet de générer aléatoirement des chroniques de température et précipitations.

Données Nous avons travaillé sur les données de température, précipitations et débit de deux bassins versants dans les Alpes : Dorinet@Hauteluce (24 km^2) et Souloise@Infernet (163 km^2). La température est en fait une moyenne pondérée des températures relevées en différents points du bassin versant. Le débit est le débit moyen journalier [m^3/s] de la rivière (Dorinet ou Souloise) au point considéré (Hauteluce ou Infernet). D'autre part, notons que notre modèle est destiné à la modélisation des températures et précipitations ponctuelles. Ici la situation est légèrement différente puisque l'on considère les précipitations et températures sur une étendue géographique un peu plus grande, mais ramenées à une grandeur scalaire. Néanmoins, les deux bassins versants sont suffisamment petits pour que l'on puisse supposer que notre modèle "local" s'applique. C'est en tout cas ce que l'on se propose de vérifier dans ce chapitre. Outre leur faible étendue géographique, d'autres critères ont guidé le choix de ces deux bassins versants :

- Longueur suffisante pour la série de débits observés.
- Bonne performance du modèle hydrologique.
- Possibilié d'évaluer la bonne simulation de l'enneigement et des séquences sèches.

La période d'observation pour les trois variables est 1964-2010 pour Hauteluce et 1969-2013 pour Infernet. Cependant pour les précipitations et la température, les observations commencent à 1948, ce qui explique que dans les sections suivantes, certains graphes démarrent en 1948.

6.2 Comparaison de deux générateurs

Dans les résultats qui vont suivre, ANA2 désigne le générateur de la DTG basé sur la méthode des analogues et R&D désigne notre générateur basé sur les HMM. Pour chacun des deux bas-

1. Le géopotentiel à 1000 hPa en un point donné est l'altitude, à la verticale de ce point, à laquelle la pression atmosphérique est de 1000 hPa.

sins versants, nous allons comparer les performances de ces générateurs selon différents critères concernant les précipitations, la température, l'enneigement et le débit. A chaque fois, 50 trajectoires pluie/température ont été simulées.

6.2.1 Précipitations

Les critères suivants ont été considérés :

- Cumuls annuels
- Cumuls mensuels
- Précipitations journalières
- Maxima annuels des précipitations sur trois jours (séquences pluvieuses) et minima annuels des précipitations sur 30 jours (séquences sèches).

La Figure 6.2 montre les cumuls annuels de précipitations observés et simulés (la période d'observation et simulation démarre à 1948 pour la méthode des analogues) pour Infernet. Les résultats sont similaires pour Hauteluce. Dans les deux cas, les trajectoires simulées sont réalistes mais on constate que toutes les trajectoires simulées par la méthode ANA2 ressemblent à la trajectoire observée, tandis que le générateur HMM permet davantage de variabilité. Cela s'explique facilement. Dans la méthode des analogues, chaque valeur est tirée au hasard parmi un ensemble restreint d'observations, cet ensemble dépendant de la situation atmosphérique (champs géopotentiels). Les simulations sont donc fortement contraintes par l'historique des géopotentiels, qui est le même pour chaque trajectoire simulée. La méthode HMM présente quant à elle davantage d'aléa puisqu'à chaque simulation, la séquence des états est tirée aléatoirement, et conditionnellement à cette séquence d'états, les variables à simuler sont tirées aléatoirement selon des lois paramétriques, et non parmi un ensemble d'observations. Cela a aussi pour conséquence que la méthode ANA2 est limitée dans les extrêmes, puisqu'elle ne peut générer que des valeurs observées. Ce problème n'existe pas pour les HMM. On peut le voir sur les deux graphes du bas de la Figure 6.2 : la fonction de répartition empirique des simulations HMM s'étend davantage dans les extrêmes. Ces constatations sont générales et ne concernent pas uniquement les cumuls annuels de précipitations. Ce manque de variabilité est le défaut principal de la méthode ANA2, et plus généralement des méthodes de rééchantillonage.

Les précipitations mensuelles sont correctement reproduites par les deux modèles pour Hauteluce. Pour Infernet, les résultats sont présentés par la Figure 6.3. Sur cette station, on observe une transition rapide entre de faibles cumuls mensuels en août à des cumuls beaucoup plus élevés dans les mois qui suivent. La méthode ANA2 reproduit correctement cette transition tandis que le HMM ne l'a pas bien détectée.

Sur les deux stations, les deux générateurs reproduisent correctement les précipitations moyennes journalières (moyennes des précipitations pour chaque jour calendaire) et la distribution globale des précipitations, malgré une légère sous-estimation des précipitations extrêmes par le HMM sur la station d'Infernet. Ces résultats sont représentés sur la Figure 6.4. On peut formuler les mêmes remarques que pour les cumuls annuels concernant le manque de variabilité de la méthode ANA2 et la limitation dans les extrêmes.

Le VCX3 annuel est le maximum, sur une année, de la moyenne glissante sur 3 jours des précipitations. On s'intéresse à la distribution de VCX3. Celle-ci donne une indication sur l'intensité des séquences pluvieuses. De même on définit le VCN30 comme le minimum annuel de la moyenne glissante sur 30 jours. Il caractérise les séquences sèches. Sur les deux stations, on constate une surestimation des VCN30 et une sous-estimation des VCX3 les plus élevés par le HMM (voir la Figure 6.5 pour les résultats d'Infernet). Cela est la conséquence d'un manque de dépendance

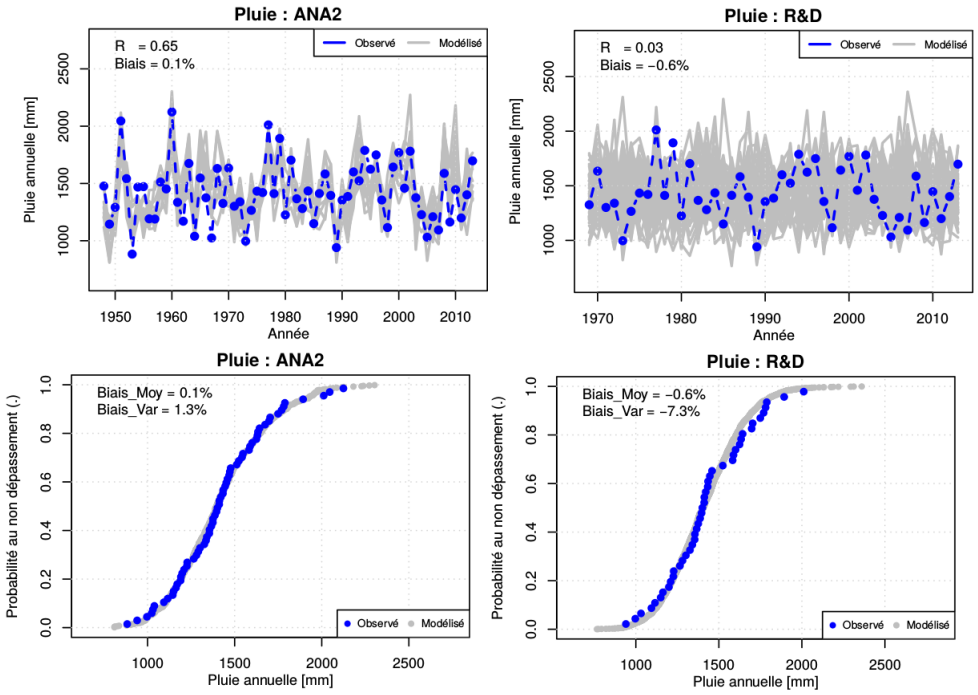

FIGURE 6.2 – Cumuls annuels des précipitations Infernet : analogues (gauche) et HMM (droite). En haut : les trajectoires observées et simulées. En bas : les fonctions de répartition empiriques.

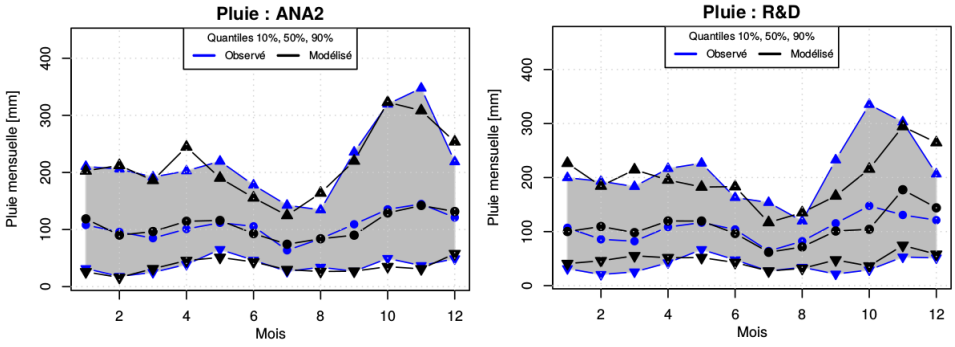

FIGURE 6.3 – Cumuls mensuels des précipitations Infernet : analogues (gauche) et HMM (droite).

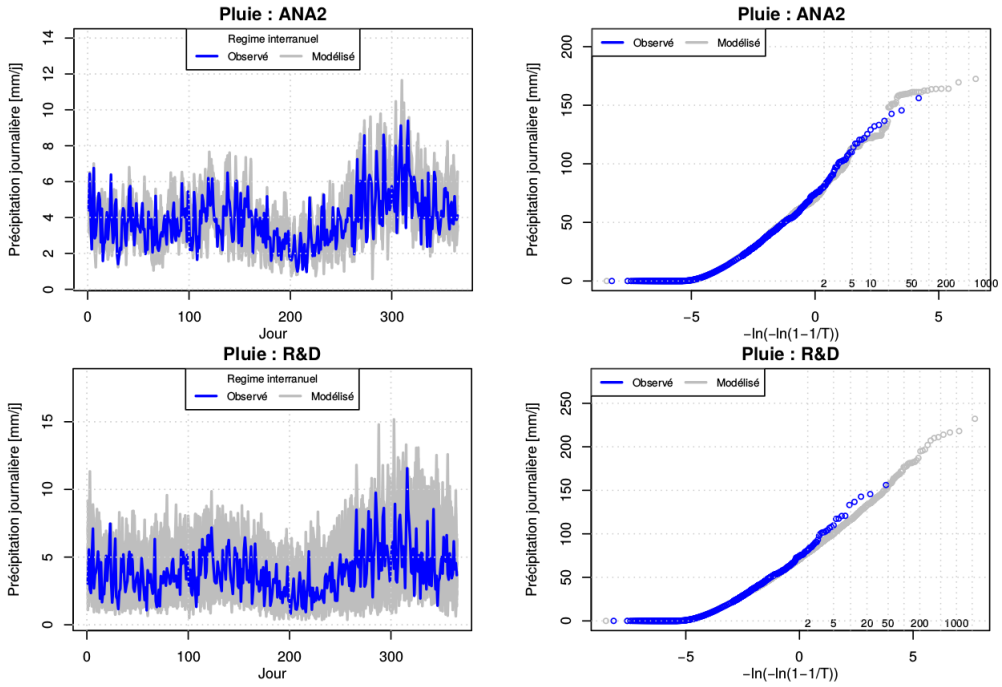

FIGURE 6.4 – Précipitations moyennes journalières Infernet : analogues (haut) et HMM (bas). A gauche : les précipitations moyennes pour chaque jour de l'année. A droite, la distribution des précipitations journalières exprimées en termes de niveaux de retour (T est la période de retour en années).

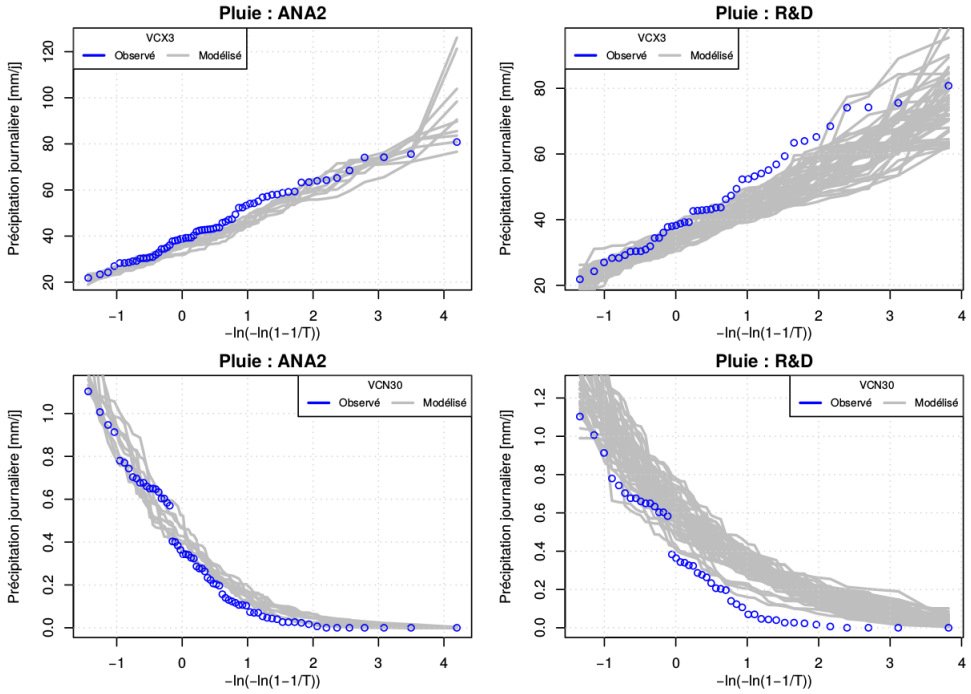

FIGURE 6.5 – Séquences pluvieuses (VCN30) et sèches (VCX3) Infernet

temporelle dans le HMM, défaut que nous avons déjà identifié. En effet le HMM produit trop peu de longues séquences sèches, ce qui a pour effet de surestimer le VCN30, et trop peu de séquences pluvieuses de plus d'un jour, ce qui a pour effet de sous-estimer le VCN3. La méthode ANA2 ne présente pas ce problème.

6.2.2 Température

Les critères suivants ont été considérés :

- Moyennes annuelles
- Moyennes mensuelles
- Températures journalières
- Maxima annuels des températures moyennes sur trois jours (séquences chaudes) et minima annuels des températures moyennes sur 30 jours (séquences froides).

Sur les deux stations, les températures moyennes annuelles (Figure 6.6 pour Hauteluce) et mensuelles (Figure 6.7) sont bien reproduites par les deux modèles, y compris le réchauffement climatique. Comme pour les précipitations, on observe davantage de variabilité interannuelle avec le HMM et une plus grande capacité à générer des valeurs éloignées de la moyenne.

Les températures journalières sont correctement modélisées avec les deux modèles et sur les deux stations. On observe néanmoins (voir Figure 6.8) une légère sous-estimation des températures les plus élevées sur la station de Hauteluce avec le modèle ANA2.

La Figure 6.9 présente les distributions respectives de VCX3 (maximum annuel des températures moyennes sur 3 jours) et VCN30 (minimum annuel des températures moyennes sur 30 jours) pour

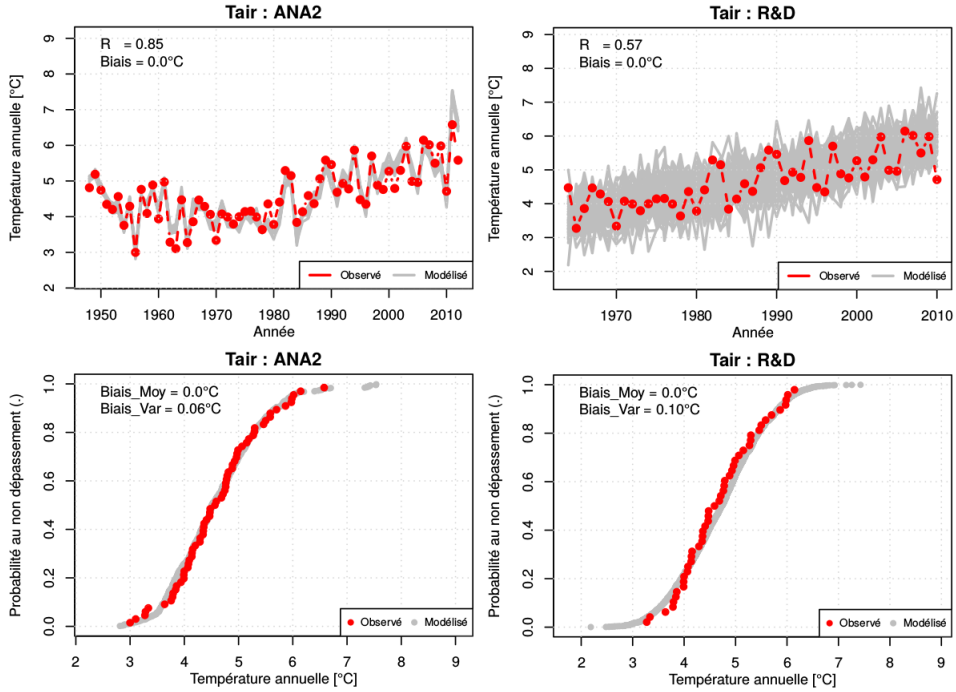

FIGURE 6.6 – Températures moyennes annuelles Hauteluce : analogues (gauche) et HMM (droite). En haut : les trajectoires observées et simulées. En bas : les fonctions de répartition empiriques.

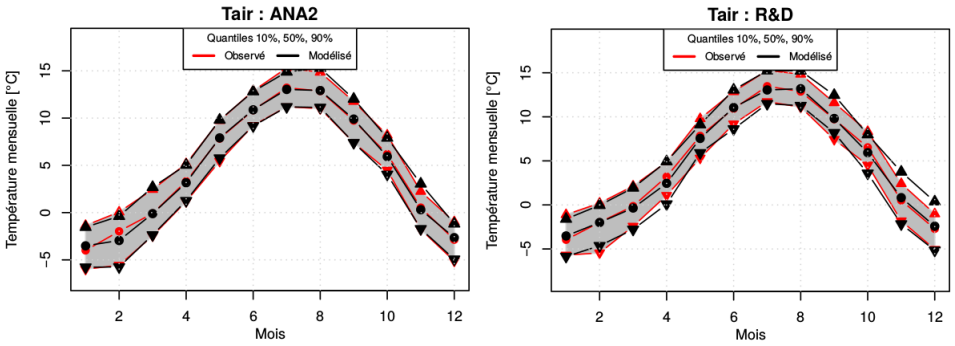

FIGURE 6.7 – Températures moyennes mensuelles Hauteluce : analogues (gauche) et HMM (droite).

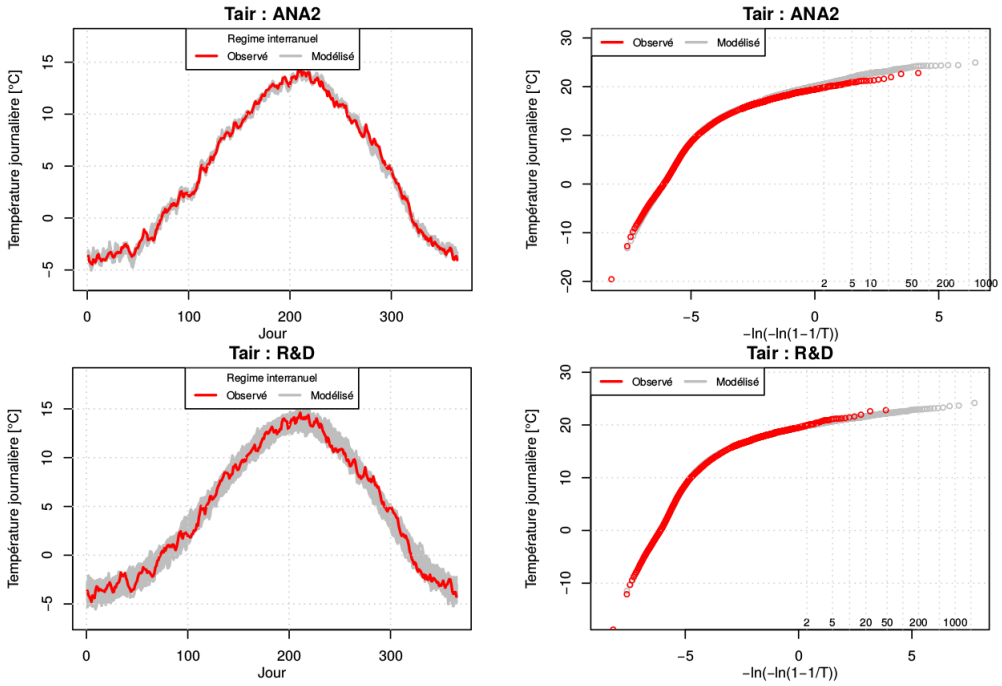

FIGURE 6.8 – Températures moyennes journalières (à gauche) et distribution des températures (à droite) pour Hauteluce : analogues (en haut) et HMM (en bas).

les deux modèles sur la station de Hauteluce. Sur les deux stations, le modèle ANA2 surestime légèrement VCX3 et VCN30. Ce problème n'apparaît pas avec la modélisation HMM.

6.2.3 Enneigement

L'enneigement "observé" n'est pas issu de mesures mais correspond à l'enneigement simulé par le modèle hydrologique à partir des températures et précipitations observées. On le compare avec l'enneigement simulé par le modèle hydrologique à partir des précipitations et températures simulées par les générateurs de temps. Bien que l'enneigement ne soit qu'une variable intermédiaire dans le modèle hydrologique, on s'y intéresse car une bonne modélisation de l'enneigement indique un bon couplage entre la température et les précipitations, ce que l'on cherche à obtenir. D'autre part, une bonne modélisation de l'enneigement est nécessaire pour obtenir une bonne modélisation du débit, la variable d'intérêt dans cette application.

L'enneigement annuel moyen est correctement simulé par les deux modèles pour la station d'Infernet. En revanche, il est sous-estimé par les deux modèles sur la station de Hauteluce, comme le montre la Figure 6.10. Cette sous-estimation concerne toute la distribution avec la méthode des analogues, et seulement la queue de distribution supérieure avec le HMM. Notons également que l'on observe, en particulier sur la sation de Hauteluce, une tendance à la baisse de l'enneigement annuel, corollaire immédiat de la hausse des températures moyennes liée au changement climatique.

L'enneigement mensuel moyen est présenté Figure 6.11 pour les deux stations. Sur la station de Hauteluce, l'enneigement mensuel moyen est sous-estimé par les deux modèles, tous les mois où

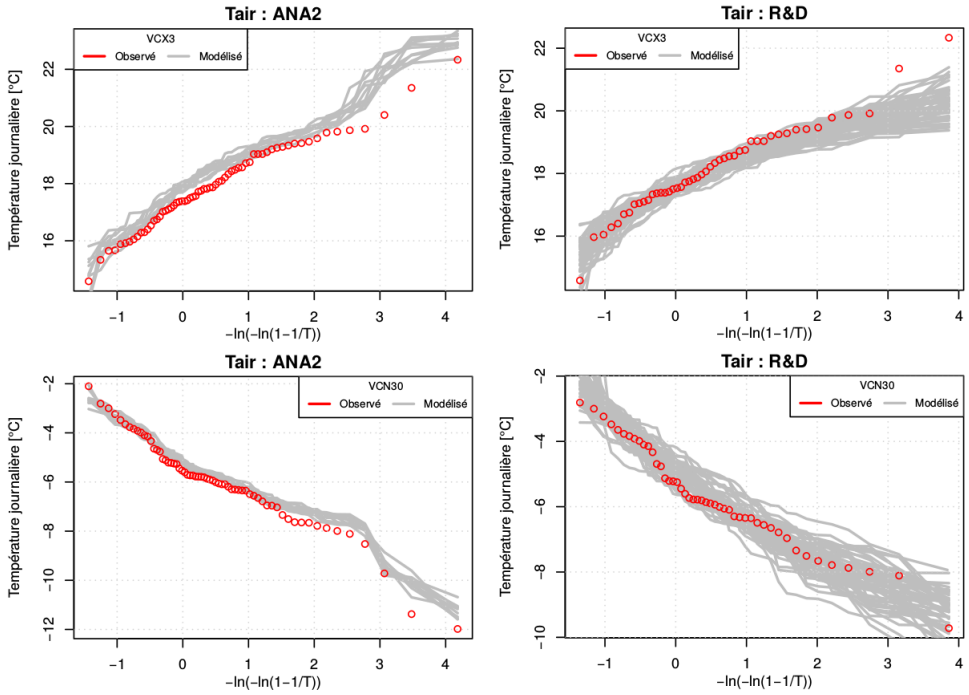

FIGURE 6.9 – Séquences chaudes (VCX3) et froides (VCN30) pour Hauteluce : analogues (à gauche) et HMM (à droite).

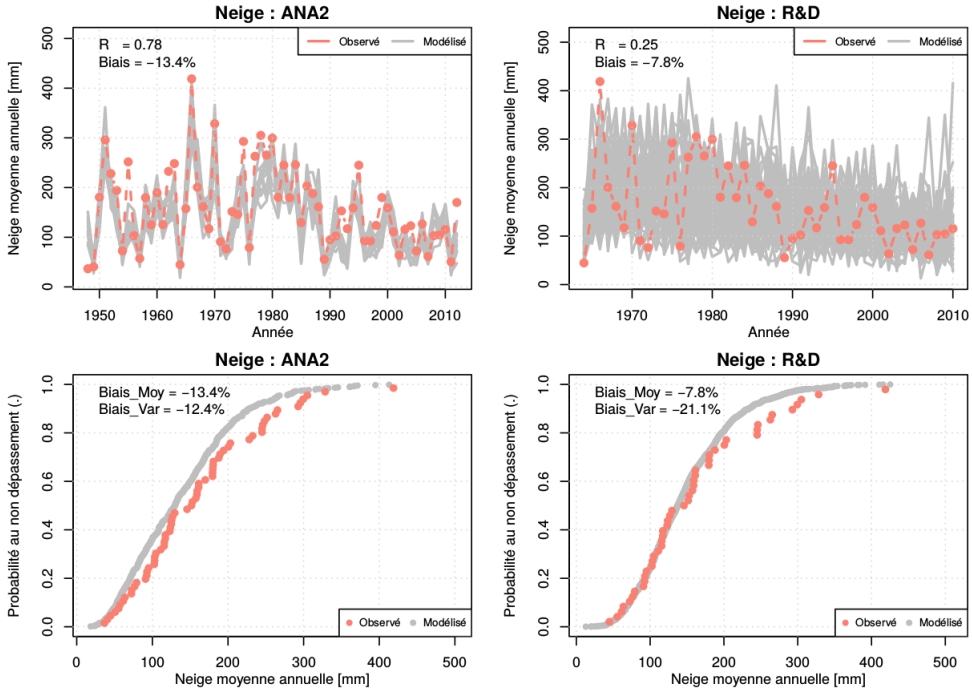

FIGURE 6.10 – Enneigement moyen annuel Hauteluce : analogues (gauche) et HMM (droite). En haut : les trajectoires observées et simulées. En bas : les fonctions de répartition empiriques.

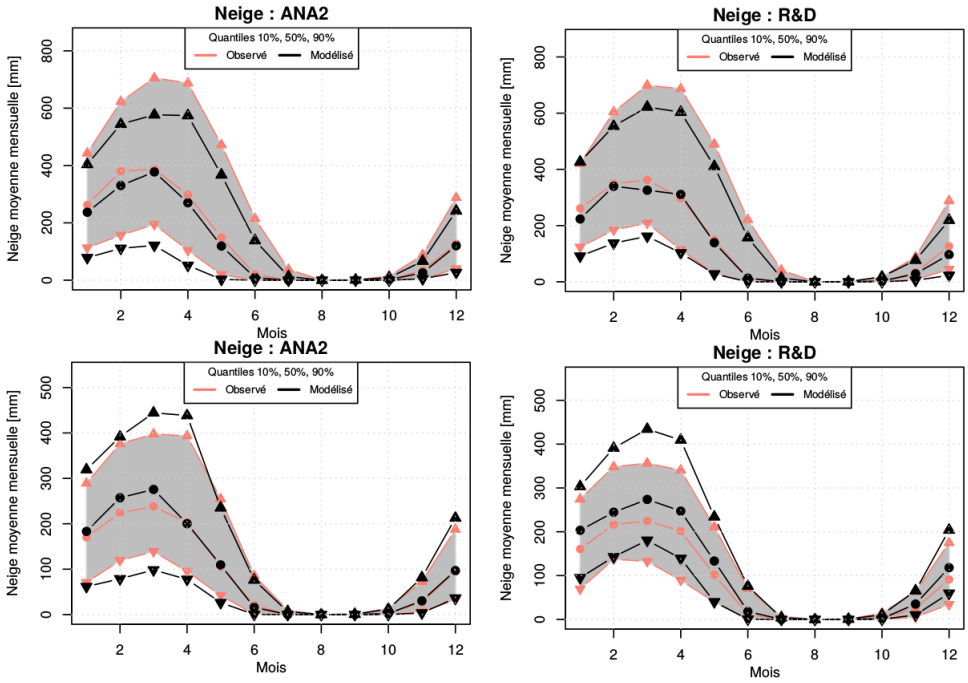

FIGURE 6.11 – Enneigement moyen mensuel : analogues (gauche) et HMM (droite) pour Hauteluce (en haut) et Infernet (en bas).

l'enneigement n'est pas nul. Cette sous-estimation est un peu moins forte avec le modèle HMM. A l'inverse, l'enneigement mensuel moyen est sur-estimé par le HMM sur la station d'Infernet.

La Figure 6.12 présente les résultats pour l'enneigement moyen journalier, sur les deux stations et les deux modèles. Sur la station de Hauteluce, on observe une sous-estimation assez nette des chutes de neige en hiver et au printemps avec le modèle ANA2. Ce défaut n'apparaît pas sur la station d'Infernet. De ce point de vue, le modèle HMM se comporte mieux, même si on peut également détecter une légère sous-estimation de l'enneigement journalier au printemps sur la station de Hauteluce.

6.2.4 Débit

Pour les deux stations étudiées, on dispose des chroniques journalières des débits mesurés. Le modèle hydrologique simule des débits de façon déterministe, soit à partir des chroniques de précipitations et température observées, soit à partir des chroniques simulées par un générateur (analogues ou HMM). Ainsi trois types de comparaisons sont possibles :

— Comparer le débit observé avec le débit simulé par le modèle hydrologique à partir des précipitations et températures observées permet d'évaluer la performance du modèle hydrologique.
— Comparer le débit simulé à partir des précipitations et températures observées avec le débit simulé à partir des précipitations et températures simulées permet d'évaluer la performance du générateur de temps vis-à-vis du débit sans inclure les éventuels défauts du modèle hydrologique.

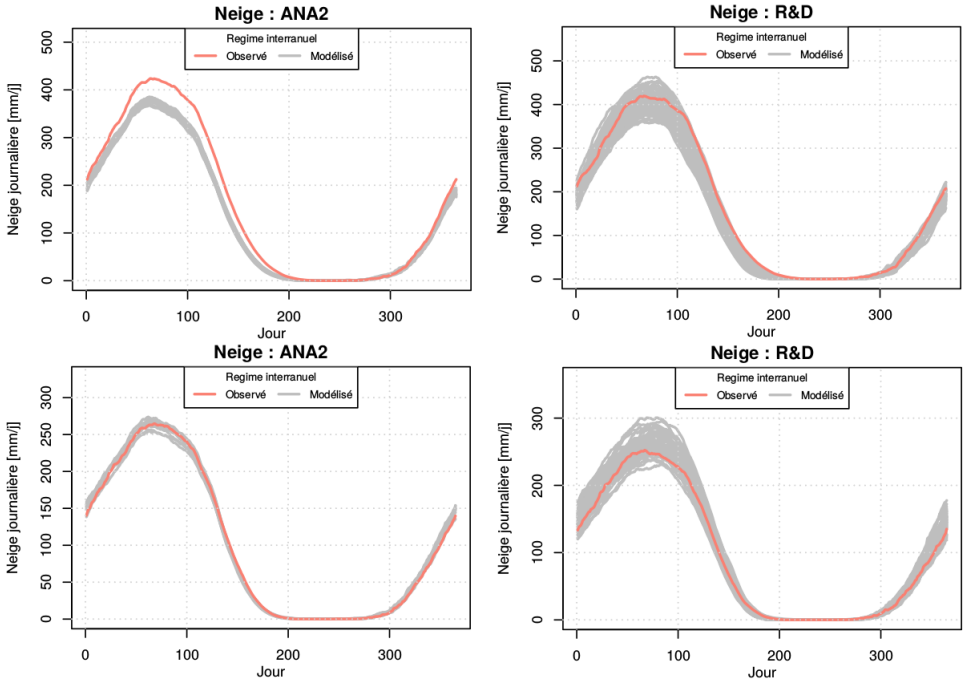

FIGURE 6.12 – Enneigement moyen journalier : analogues (gauche) et HMM (droite) pour Hauteluce (en haut) et Infernet (en bas).

— Comparer le débit observé au débit simulé à partir des précipitations et températures simulées permet d'évaluer la capacité du système *générateur de temps+modèle hydrologique* à générer des débits réalistes.

Nous allons présenter les résultats de la seconde comparaison puisque nous souhaitons évaluer la performance de notre générateur de temps, sans tenir compte de la qualité du modèle hydrologique. Nous allons voir que les défauts du HMM relativement aux précipitations sont répercutés sur les simulations de débit.

La distribution du débit annuel moyen, représentée sur la Figure 6.13 pour Infernet, est correctement simulée par les deux modèles. Remarquons l'importante variabilité interannuelle du débit annuel moyen, résultat des variabilités respectives des précipitations et des températures, ainsi que de leur couplage, dont dépend l'enneigement. Cette variabilité est bien reproduite par le HMM, dont les simulations peuvent conduire à des débits annuels moyens de 2 à 8 m^3/s pour Infernet, par exemple. La méthode ANA2 reproduit aussi cette variabilité, mais les trajectoires ont toutes tendance à ressembler à la trajectoire observée.

Les résultats sur les débits mensuels moyens sont satisfaisants sur la station de Hauteluce. Sur la station d'Infernet (Figure 6.14), on constate une sous-estimation par le HMM du débit à l'automne, ce défaut n'apparaissant pas avec le modèle ANA2. C'est la conséquence de la sous-estimation des précipitations automnales qui apparaît sur la Figure 6.3.

Sur la station d'Infernet, les débits moyens journaliers sont correctement simulés par les deux modèles, mais quelques défauts apparaissent sur la station de Hauteluce (voir Figure 6.15). Le modèle ANA2 sur-estime légèrement le débit hivernal et sous-estime le débit en juin/juillet.

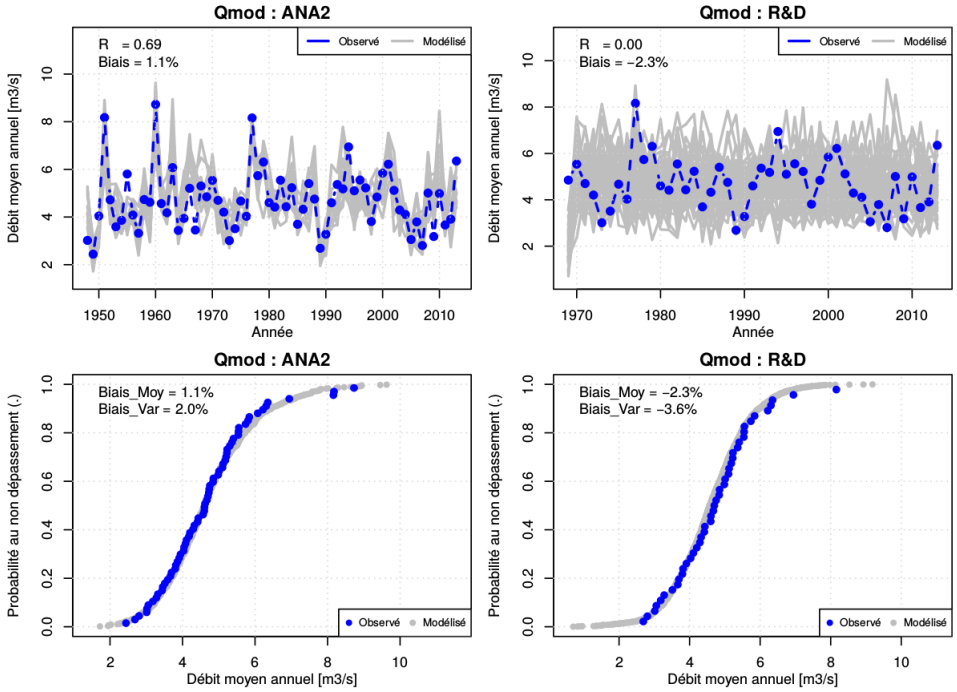

FIGURE 6.13 – Débit annuel moyen Infernet : analogues (gauche) et HMM (droite). En haut : les trajectoires observées et simulées. En bas : les fonctions de répartition empiriques.

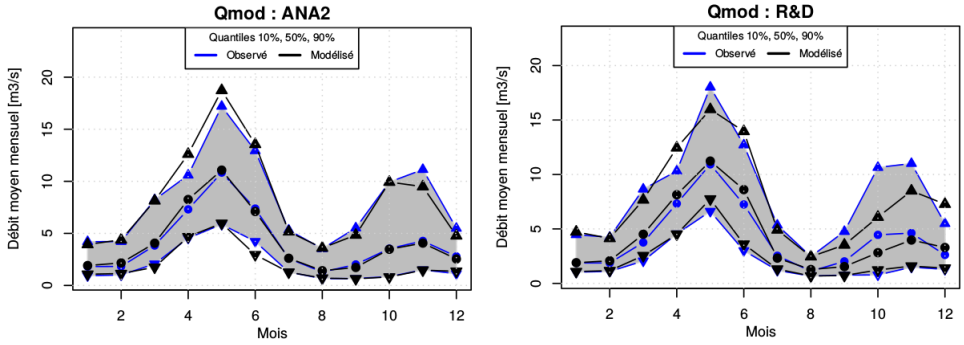

FIGURE 6.14 – Débit mensuel moyen Infernet : analogues (gauche) et HMM (droite).

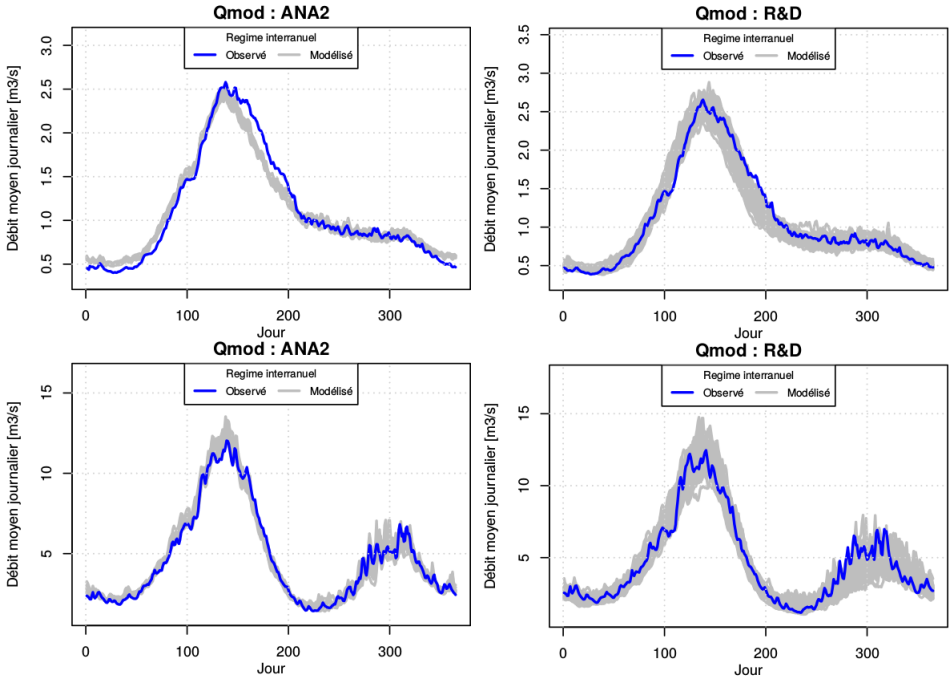

FIGURE 6.15 – Débit moyen journalier : analogues (gauche) et HMM (droite), pour Hauteluce (en haut) et Infernet (en bas).

C'est la conséquence de la sous-estimation des chutes de neige en hiver (Figure 6.12). On note également, concernant Hauteluce, une légère sous-estimation du débit estival sur les simulations issues du HMM.

La Figure 6.16 représente les distributions des débits journaliers. Cette distribution est correctement simulée pour la station de Hauteluce. En revanche, sur la station d'Infernet, la queue de distribution est mal simulée par les deux modèles : sur-estimation par ANA2 et sous-estimation par HMM. Dans le cas du HMM, cela peut provenir de la sous-estimation des précipitations les plus élevées (voir Figure 6.4).

La Figure 6.17 présente les résultats concernant la distribution du VCX3 (maximum annuel de la moyenne glissante sur 3 jours du débit. Le HMM sous-estime le VCX3 sur la station d'Infernet. Cela est cohérent avec le fait que le VCX3 des précipitations est également sous-estimé par le HMM sur cette station (Figure 6.5).

La Figure 6.18 présente les résultats concernant la distribution du VCN30 (minimum annuel de la moyenne glissante sur 30 jours du débit). Le HMM a tendance à surestimer les débits minimaux, en particulier sur Infernet. Sur la station de Hauteluce, le VCN30 du débit est assez nettement sous-estimé, tandis que cette statistique est bien modélisée pour Infernet. Ces observations sont à mettre en relation avec le VCN30 des précipitations (Figure 6.5).

La comparaison des débits simulés à partir des précipitations et températures simulées avec les débits réellement observés fournit des résultats similaires. Cela n'est pas surprenant dans la mesure où les deux bassins versants ont été sélectionnés pour la bonne qualité du modèle hydrologique associé.

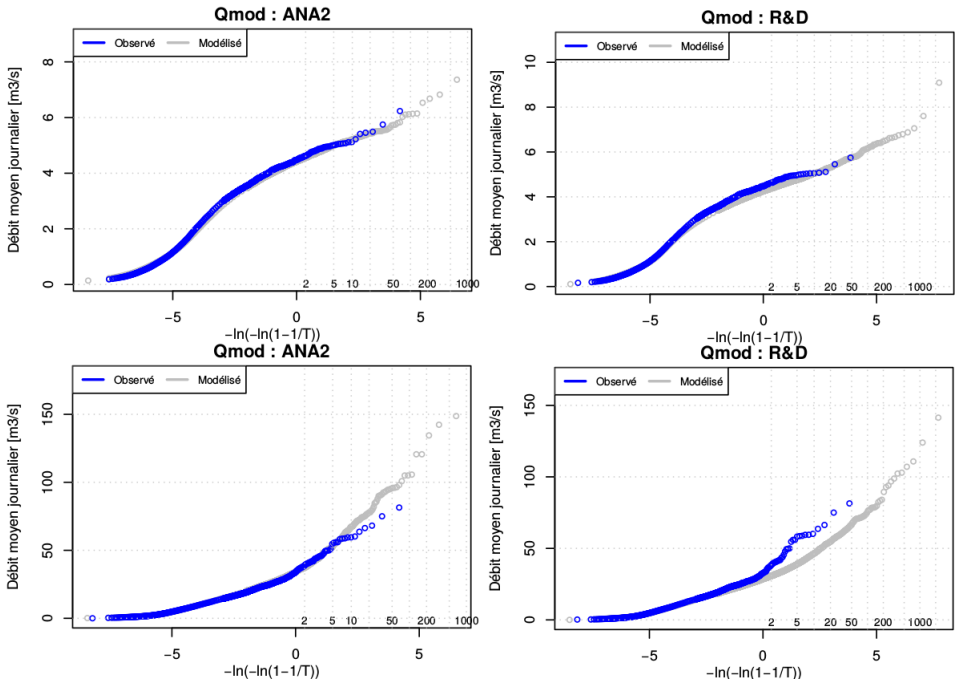

FIGURE 6.16 – Distribution des débits journaliers : analogues (gauche) et HMM (droite), pour Hauteluce (en haut) et Infernet (en bas).

6.3 Application : hivers froids et faible hydraulicité

La production hydroélectrique dite "fatale" est celle des centrales "au fil de l'eau" qui ne possèdent pas de retenue d'eau et qui ne fonctionnent que grâce au débit naturel des cours d'eau. Ce moyen de production est donc entièrement dépendant du débit instantané, de la même manière que la production éolienne dépend de la vitesse du vent. Du point de vue de ce mode de production, une situation critique pour l'équilibre entre la production et la consommation est celle d'un hiver froid, ce qui implique une consommation élevée, et à faible hydraulicité (débit plus faible que la normale), ce qui implique une faible production hydraulique fatale. Une telle configuration n'est pas rare : en hiver une situation anticyclonique correspond souvent à un froid sec. D'autre part si les précipitations tombent sous forme de neige, elles n'alimentent les cours d'eau que lorsque la neige fond. Dans cette section, nous allons montrer que notre générateur de temps, associé à un modèle hydrologique, peut servir à quantifier la probabilité d'occurrence d'un tel hiver.

On considère le bassin versant de Dorinet@Hauteluce, et les hivers (décembre, janvier, février) de la période 1965-2010 (46 hivers). Pour chaque hiver on calcule à partir des observations journalières le débit moyen Q et la température moyenne T. On s'intéresse à la probabilité pour que cette température soit inférieure à son 1er décile, noté $T10$ et que dans le même temps, le débit moyen soit inférieur à son 1er décile, noté $Q10$. Ces deux événements étant positivement corrélés, on s'attend à ce que cette probabilité soit supérieure à 1%, qui correspondrait à la situation d'indépendance.

Sur la Figure 6.19 (à gauche), les points sont les 46 observations dans le plan (T, Q). A partir

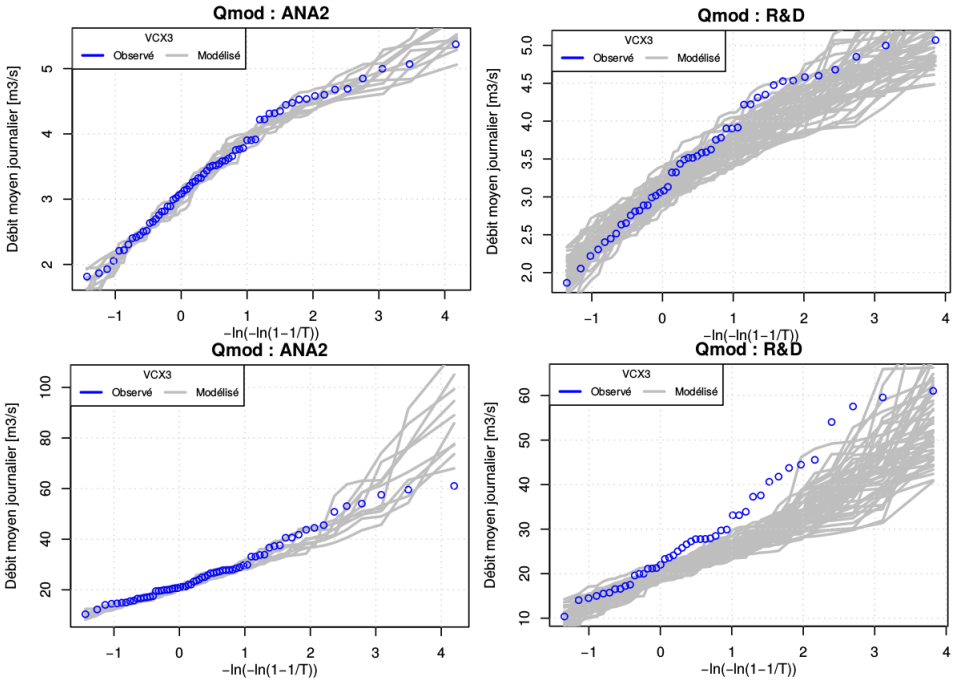

FIGURE 6.17 – Distribution des maxima annuels du débit moyen sur 3 jours (VCX3) pour Hauteluce (en haut) et Infernet (en bas).

de ces points, nous avons calculé un estimateur à noyau (gaussien) de la densité de probabilité du couple (T, Q), représentée en couleurs et par ses lignes de niveau. Cette estimation de densité permet d'obtenir une estimation de la probabilité cherchée $\mathbb{P}(T < T10, Q < Q10)$: c'est l'intégrale de la densité estimée sur le quadrant inférieur gauche. Nous obtenons 7.3%. Notons que cette valeur est bien supérieure à 1%. Nous avons constaté que, le nombre d'observations étant faible, cette valeur est sensible au choix du paramètre de lissage dans l'estimation de la densité.

Une autre façon d'estimer cette probabilité est d'utiliser les simulations issues de notre générateur. Nous disposons de 1000 chroniques simulées de température et de débit. On peut donc remplacer les 46 observations par 46000 pseudo-observations. Celles-ci peuvent alors être utilisées pour estimer la densité de probabilité de (T, Q), puis la probabilité cherchée. Les résultats obtenus sont présentés sur la Figure 6.19 (à droite). Nous obtenons une valeur légèrement inférieure à celle obtenue à partir des seules observations : 4.6%. Cet écart peut s'expliquer par une sous-estimation de la variabilité interannuelle des débits hivernaux par le simulateur. On peut aussi arguer du fait qu'avec seulement 46 observations, il est difficile d'estimer correctement une densitée de probabilité bivariée, et donc la valeur de 6.9% est à prendre avec précaution. Notons que la forme de la densité estimée est différente de celle obtenue avec les observations. Cette dernière était ellipsoïdale, alors qu'en utilisant le simulateur, on voit apparaître une asymétrie due au fait que le débit est borné inférieurement par 0.

La Figure 6.19 concerne la totalité de la période 1965-2010. Nous avons également restreint le calcul à plusieurs sous-périodes : 1965-1979, 1980-1994, 1995-2010 et 2008-2010. Sur chaque sous-période, nous avons estimé, en utilisant le simulateur, la probabilité d'intérêt. Nous obtenons respectivement 9.7%, 3.4%, 1% et 0.5%. (voir la Figure 6.20). Cette décroissance est liée au

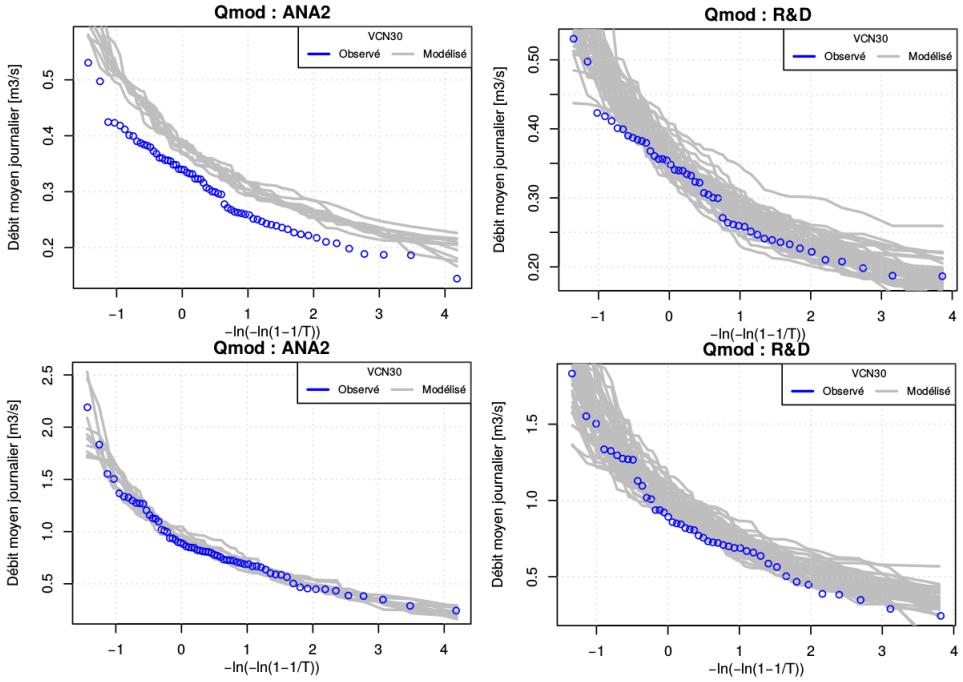

FIGURE 6.18 – Distribution des minima annuels du débit moyen sur 30 jours (VCN30) pour Hauteluce (en haut) et Infernet (en bas).

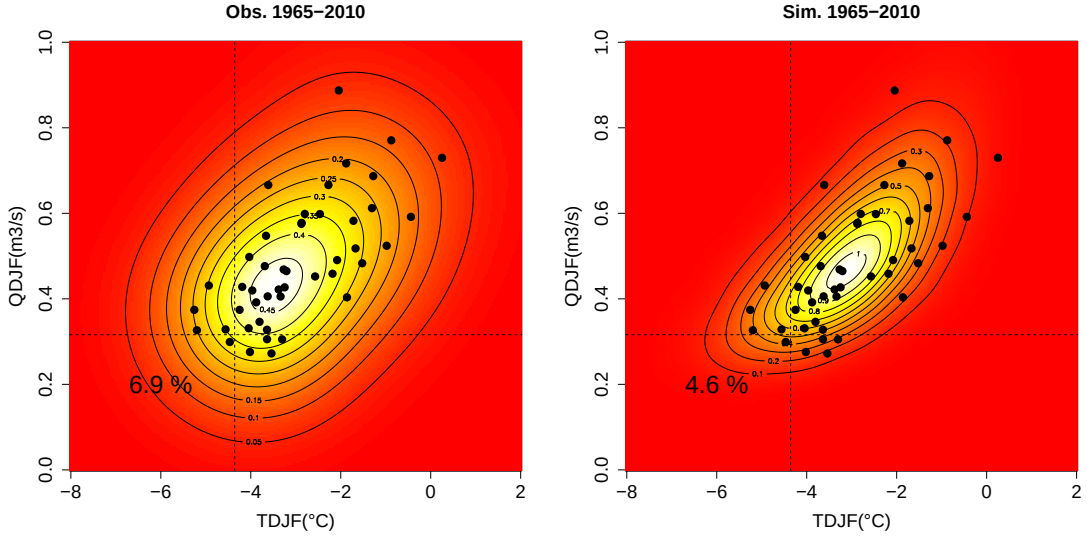

FIGURE 6.19 – Observations de la température moyenne et du débit moyen de chaque hiver (points noirs). Les lignes pointillées matérialisent $T10$ et $Q10$ et les lignes de niveau sont celles de l'estimateur de la densité de (T, Q) obtenue à partir des observations (à gauche) et des simulations (à droite).

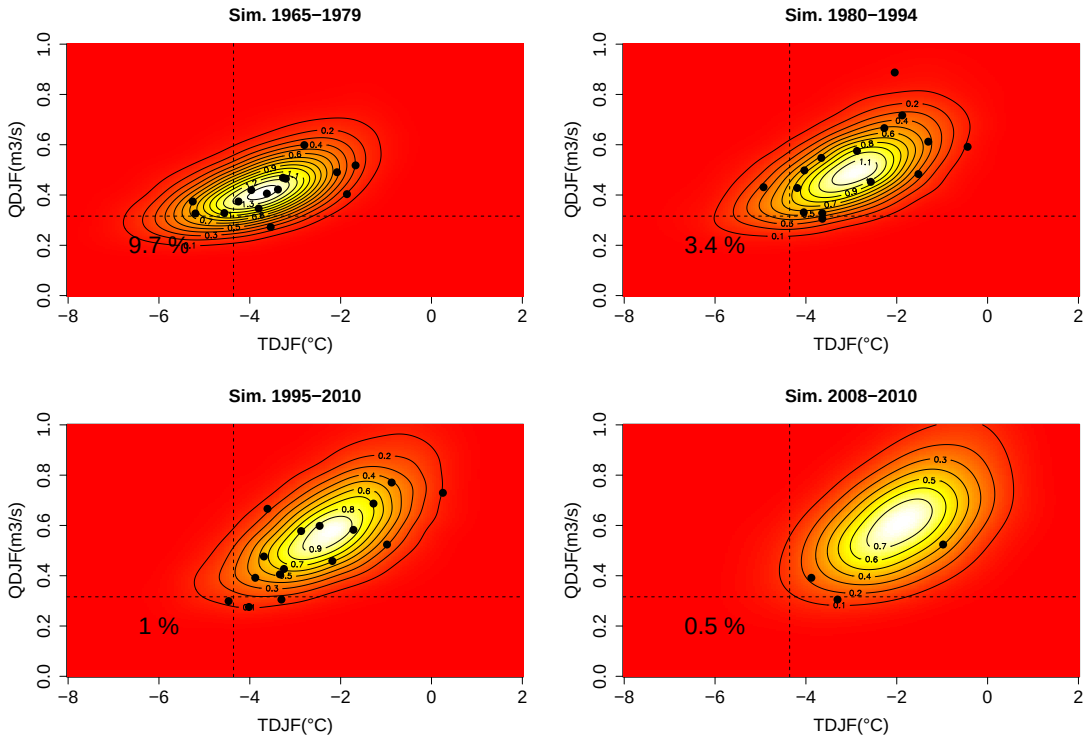

FIGURE 6.20 – Observations (points noirs) et densités estimées à partir des simulations, sur différentes sous-périodes.

réchauffement climatique : comme les températures sont en moyenne plus élevées, la probabilité d'observer un hiver dont la température moyenne est inférieure à $T10$ (quantile calculé sur toute la période) est de plus en plus faible. Ainsi le risque de conjonction d'un hiver froid et d'une faible hydraulicité diminue.

En particulier, nous nous sommes intéressés à la problématique "temps présent" : compte tenu du climat "actuel" (vu de 2010), comment estimer la probabilité d'observer un hiver à risque ? Pour cela nous nous sommes restreints à la période 2008-2010. Sur cette période, il est exclu d'utiliser les observations puisqu'on ne peut pas estimer quoi que ce soit avec seulement trois observations. Le simulateur est alors d'un grand secours puisqu'il permet de générer un nombre quelconque de pseudo-observations et ainsi d'obtenir une estimation de la densité du couple (T, Q) pour les hivers "actuels". Le résultat est représenté sur la Figure 6.20 (en bas à gauche) : on obtient une probabilité de 0.5%. Il est intéressant de comparer la densité estimée sur la période 1965-1979 avec celle estimée sur 2008-2010 pour voir l'effet du changement climatique sur le couple (T, Q) en hiver.

Dans cette section, nous avons utilisé les simulations issues du HMM couplé au modèle hydrologique MORDOR pour obtenir une estimation de la probabilité d'intérêt. En principe, un autre générateur de temps pourrait être utilisé. La méthode des analogues n'est cependant pas adaptée à cet usage : comme nous l'avons déjà remarqué, les simulations qu'elle produit ont tendance à être proches de la chronique observée. Il est donc difficile de l'utiliser pour estimer des probabilités d'événements rarement observés.

Remarque En réalité, pour la prévision de consommation, EDF utilise la "Température France", qui est une moyenne des températures de 32 villes, pondérée par leur consommation. De même, pour la production hydroélectrique fatale, c'est un débit agrégé qui est utilisé, appelé "Fil France".

6.4 Interprétation des états et classification en types de temps

Dans le Chapitre 5, nous avons vu qu'en examinant les estimateurs obtenus par maximisation de la vraisemblance, il était possible de donner une interprétation aux différents états. Dans cette section, nous proposons une autre approche pour interprérer les états, basée sur l'estimation de la séquence d'états la plus probable. Rappelons que si $(X_t, Y_t)_{t\geq 1}$ est un HMM paramétré par θ, on peut calculer par l'algorithme de Viterbi (voir le Chapitre 2) la séquence d'états la plus probable, définie, sous le paramètre θ, par

$$(\hat{X}_1, \ldots, \hat{X}_n) \in \underset{(x_1,\ldots,x_n)\in \mathsf{X}^n}{\arg\max}\ \mathbb{P}^\theta\left(X_1 = x_1, \ldots, X_n = x_n \mid Y_{1:n}\right). \tag{6.1}$$

Nous avons calculé cette suite d'états avec $\theta = \hat{\theta}_n$ l'estimateur du maximum de vraisemblance. On obtient ainsi une partition des jours de la période d'observation :

$$\{1, \ldots, n\} = \bigsqcup_{1\leq k\leq K} \{t \in \{1, \ldots, n\},\ \hat{X}_t = k\}.$$

Notons que même dans le cas d'un modèle bien identifié (i.e. les observations sont vraiment des réalisations du modèle) et sous le vrai paramètre, la séquence d'états la plus probable n'est pas nécessairement celle qui s'est réalisée. La classification ainsi obtenue n'est donc pas parfaite. Nous traiterons ici l'exemple du bassin versant de Souloise@Infernet. De la classification obtenue via l'algorithme de Viterbi, on peut immédiatement déduire quelques statistiques élémentaires sur les différentes états (voir Table 6.1) : fréquence des précipitations, température moyenne et précipitations moyennes (lorsqu'il y en a).

Etat	nb occurrences	fréquence précip	temp moyenne (°C)	précip moyenne (mm)
1	2095	0.35	9.96	6.51
2	2301	0.11	2.58	0.83
3	2479	0.97	2.83	10.78
4	2255	0.98	8.17	12.81
5	2789	0.13	8.15	1.13
6	1879	0.47	-2.29	4.59
7	2627	0.15	5.95	1.22

TABLE 6.1 – Statistiques élémentaires état par état, pour Souloise@Infernet

Outre le fait que les états sont distribués de façon relativement homogène, on voit que deux états (numérotés 3 et 4) sont presque exclusivement pluvieux, avec des précipitations intenses. Ils se différencient cependant par la température, celle-ci étant en moyenne plus élevée dans l'état 4. A l'inverse, les états 2, 5 et 7 sont secs, avec des précipitations faibles. Ils se distinguent également par leurs températures. Enfin, les états 1 et 6 sont modérément pluvieux. L'état 6 est

particulièrement froid tandis que l'état 1 est particulièrement chaud. Cette interprétation des états est obtenue uniquement à partir des variables locales, et elles rejoignent celles qui peuvent être obtenues en examinant l'estimateur du maximum de vraisemblance. Mais peut-on établir en lien entre ces états "locaux", obtenus par le HMM à partir des variables mesurées in-situ, et des variables météorologiques à plus grande échelle ? Nous avons examiné les éléments suivants :

— Comparaison avec une autre classification.
— Champ géopotentiel à 1000hPa.
— Champ de précipitations.
— Champ de température.

1. **Comparaison avec une autre classification** On veut comparer la classification en types de temps issue du HMM avec la classification construite par Garavaglia et al. (2010). Cette dernière est obtenue à partir des champs géopotentiels à 700hPa et 1000hPa, et de 54 séries de précipitations, dans une zone centrée sur le Sud-Est de la France. Cette classification n'est donc pas directement basée sur les températures (mais seulement indirectement puisque la température et le géopotentiel sont liés). En utilisant un algorithme de classification hiérarchique, huit classes sont déterminées. Pour simplifier l'analyse, celles-ci peuvent être regroupées en 4 classes (voir Figure 3 dans Garavaglia et al. (2010)) : O (circulation d'Ouest), E (circulation d'Est), S (circulation du Sud) et A (anticyclone). La Table 6.2 donne pour chaque état la ventilation sur les différents types de temps. La similarité entre deux partitions peut être mesurée par l'Adjusted Rand Index (ARI, voir Hubert and Arabie (1985)). Une valeur de 0 signifie que les partitions sont indépendantes, tandis qu'une valeur de 1 signifie qu'elles sont identiques. Ici, ARI = 0.28. Les états 3 et 4, les plus pluvieux, sont principalement associés à une circulation d'Ouest, puis d'Est pour l'état 3 et de Sud pour l'état 4. Sans surprise, ils sont très peu associés à l'anticyclone. Les états secs 2 et 7 sont principalement associés à l'anticyclone, puis à la circulation d'Ouest. L'état sec 5, le plus chaud, est très peu associé à la circulation d'Est, mais se répartit équitablement entre les autres types de temps. L'état 1, chaud est associé aux circulations d'Ouest et de Sud, et très peu à la circulation d'Est. L'état 6, froid, est très rarement associé à la circulation de Sud, mais peut concerner les trois autres classes à peu près à parts égales. Les deux classifications présentent ainsi des caractéristiques communes, même si elles sont différentes, l'une étant construite à partir des précipitations sur une zone géographique étendue et des champs géopotentiels, l'autre étant obtenue par les données locales.

Etat	O	S	E	A
1	0.38	0.42	0.03	0.17
2	0.24	0.06	0.16	0.54
3	0.58	0.15	0.23	0.05
4	0.49	0.28	0.13	0.10
5	0.31	0.30	0.05	0.34
6	0.30	0.02	0.32	0.36
7	0.29	0.13	0.11	0.47

TABLE 6.2 – Comparaison des deux classifications : répartition des types de temps dans chaque état.

2. **Champ géopotentiel à 1000hPa** Nous avons utilisé les données de réanalyses ERA5 [2] pour le géopotentiel à 1000hPa à 12h sur une zone comprise entre 30°O et 30°E en longitude,

2. https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

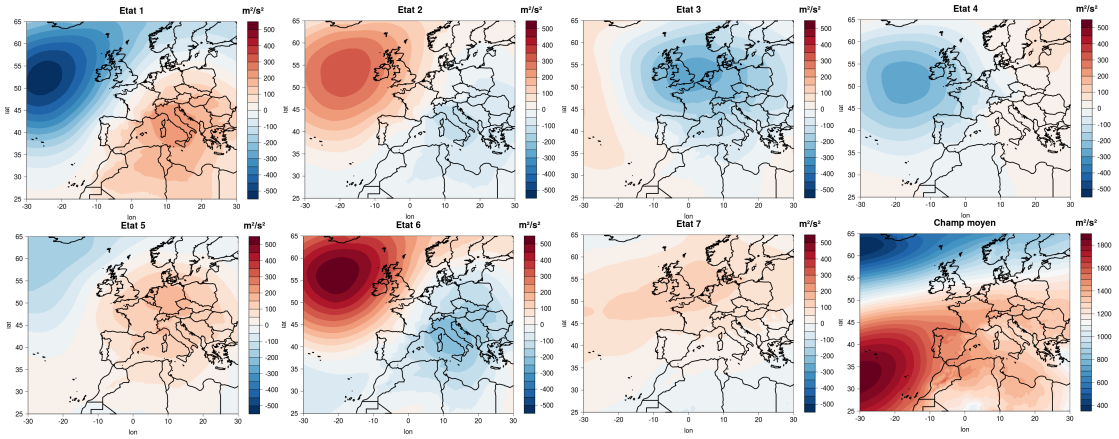

FIGURE 6.21 – Anomalies du champ géopotentiel à 1000hPa dans chaque état. Rouge : anomalie positive. Bleu : anomalie négative.

et entre 25°N et 65°N en latitude. La résolution spatiale est de 31km. Nous avons calculé le champ moyen état par état (pour chaque état k, cellule par cellule, on moyenne le géopotentiel sur les jours affectés à l'état k). Sur la Figure 6.21, nous avons représenté les anomalies de géopotentiel dans chaque état, c'est-à-dire le champ moyen dans l'état auquel on soustrait le champ moyen tous états confondus (en bas à droite sur la Figure 6.21). Les deux états pluvieux, 3 et 4, présentent des champs géopotentiels similaires avec une anomalie négative sur l'Ouest de l'Europe qui donne lieu à une circulation d'Ouest. Dans l'état 3, la dépression est située un peu plus à l'Est que dans l'état 4. Ces deux états présentent un fort contraste avec les états 2 et 6 : les deux "pôles" d'anomalies sont inversés. Notons cependant que les anomalies sont plus marquées dans l'état 6 que dans l'état 2 (en particulier l'anomalie positive centrée au sud de l'Islande). Cela peut expliquer le fait que l'état 6 soit davantage pluvieux que l'état 2. Les états 1 et 5 sont également similaires, avec une anomalie positive centrée sur l'Italie ou sur l'Allemagne, mais les contrastes sont plus marqués dans l'état 1 que dans l'état 5, qui reste proche du champ moyen. Enfin, l'état 7 est lui aussi proche du champ moyen, avec néanmoins une anomalie positive centrée sur les îles britanniques. Ainsi, lorsque l'on reporte les états déterminés par le HMM sur le champ de géopotentiel à 1000hPa, on obtient dans certains états des motifs à grande échelle très contrastés, ce qui montre le lien fort entre ces états locaux et un champ à l'échelle continentale, mais aussi d'autres états pour lesquels les anomalies sont faibles en valeur absolue. D'autre part, plusieurs états HMM correspondent à des champs géopotentiels similaires. La classification HMM est donc plus fine. Ce n'est pas surprenant car d'une part elle a été déterminée à partir des observations locales, et d'autre part, nous n'avons utilisé qu'un seul descripteur grande échelle, ce qui est certainement insuffisant.

3. **Champ de précipitations** On peut aussi s'intéresser au lien entre les états HMM et la distribution à plus grande échelle des variables à modéliser. Commençons par les précipitations. Nous avons utilisé les données grillées E-OBS [3] (cumuls journaliers, résolution spatiale de 0.25°). La Figure 6.22 représente, état par état, le rapport entre le champ moyen de l'état et le champ moyen des précipitations tous états confondus. Comme pour les géopotentiels, les états sont très contrastés. Les deux états qui ont définis comme pluvieux à l'échelle locale

3. https://www.ecad.eu/download/ensembles/download.php

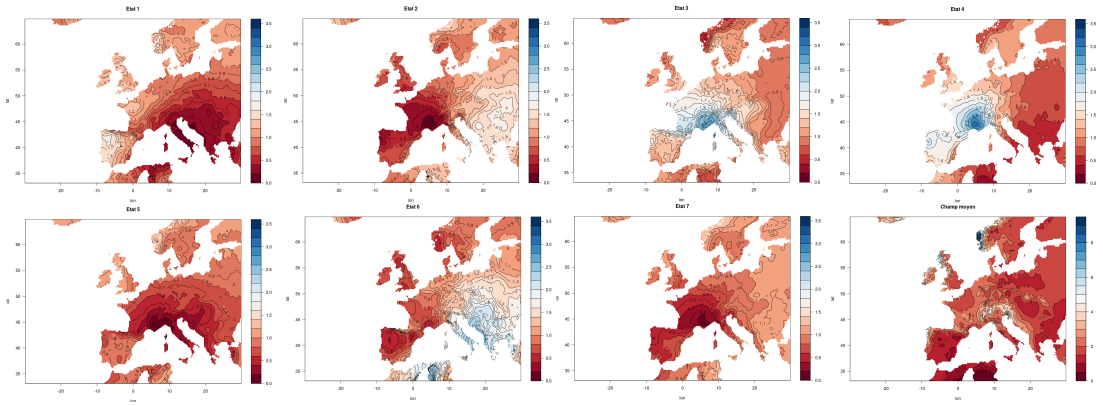

FIGURE 6.22 – Champs de précipitation état par état : anomalies par rapport au champ moyen. Bleu (resp. rouge) signifie davantage (resp. moins) de précipitations qu'en moyenne.

(états 3 et 4) correspondent en fait à des précipitations plus élevées que la moyenne sur la France, en particulier la partie Sud. De même, les états identifiés comme secs (2, 5 et 7) correspondent à des précipitations plus faibles que la moyenne sur le Sud de la France. Ces états diffèrent cependant par le champ de précipitations dans le reste de l'Europe. Par exemple en Italie, l'état 5 est sec tandis que l'état 2 est pluvieux. Notons que le bassin versant étudié est situé dans les Alpes, qui est souvent à l'interface entre deux régions homogènes en termes de précipitations.

4. **Champ de température** Enfin, nous nous sommes intéressés au champ de température (moyenne journalière) issu des données E-OBS. Nous avons traité séparément l'été (juin juillet, août) et l'hiver (décembre, janvier, février) et avons considéré les anomalies de température par rapport au champ moyen. Les résultats sont représentés sur les Figures 6.23 et 6.24. Comme pour les géopotentiels, certains états présentent des anomalies significatives (états 1, 2 et 6 en hiver, états 1, 3 et 6 en été) tandis que d'autres (état 3 en hiver, état 7 en été par exemple) sont plus proches du champ de température moyen. Remarquons qu'un même état peut correspondre à une anomalie froide en été, et chaud en hiver (état 6). Nous voyons donc ici l'intérêt d'une classification locale, qui va déterminer des états de manière optimale pour la station considérée, et donc capturer des comportements qui ne peuvent pas l'être en observant les variables d'intérêt à l'échelle continentale. Les états du HMM ne sont néanmoins pas indépendants des variables à grande échelle, les variables locales étant évidemment liées à la situation météorologique globale.

6.5 Conclusion

Dans ce chapitre, nous avons présenté une application hydrologique de notre modèle bivarié précipitations/température. En utilisant un modèle hydrologique, nous avons simulé des chroniques d'enneigement et de débit pour deux bassins versants dans les Alpes. De telles simulations sont utilisées par la DTG d'EDF dans le cadre de la production hydroélectrique. Nous avons donc comparé, du point de vue des précipitations, des températures, de l'enneigement et du débit, les résultats fournis par notre générateur HMM à ceux fournis par le générateur utilisé par la DTG, qui s'appuie sur la méthode des analogues.

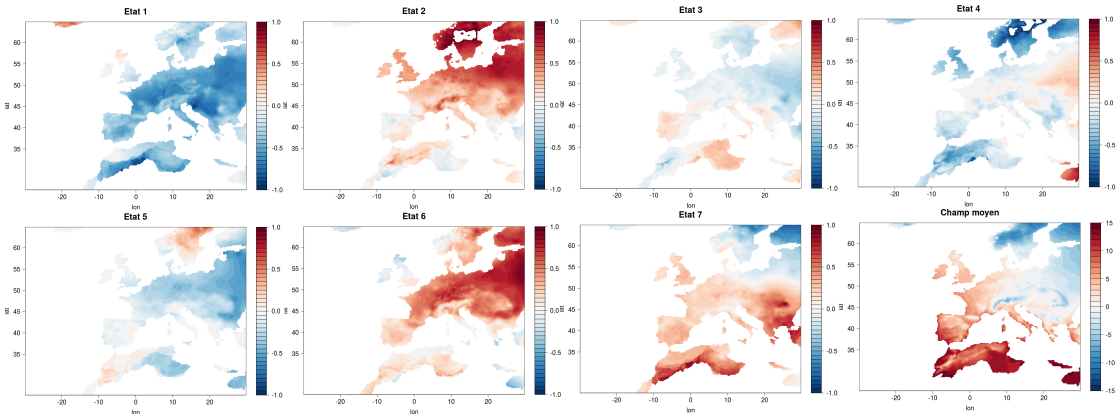

FIGURE 6.23 – Anomalies de température état par état en hiver.

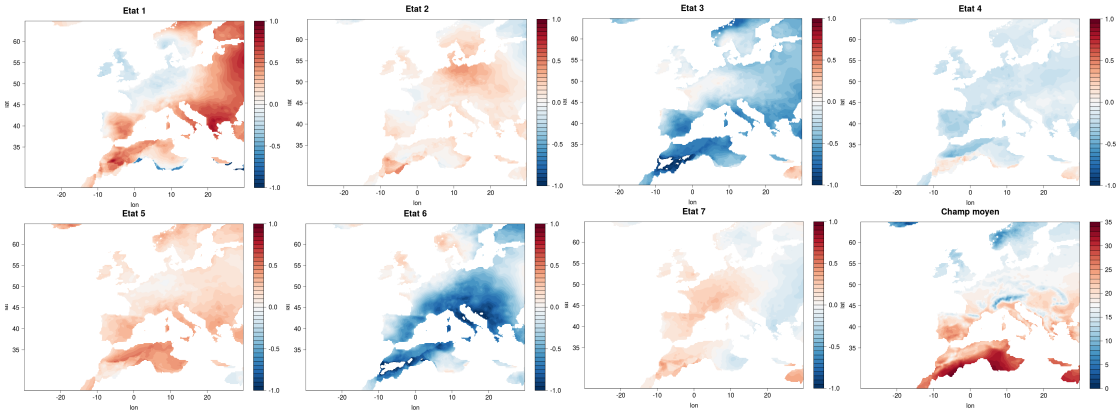

FIGURE 6.24 – Anomalies de température état par état en été.

Les deux approches sont très différentes. La méthode des analogues est non paramétrique et consiste en un rééchantillonage de l'historique des observations. Elle est guidée par des considérations physiques puisque les analogues sont sélectionnés à partir de champs géopotentiels. A l'inverse le HMM conduit à l'ajustement aux observations d'un modèle paramétrique, sans variables exogènes. L'avantage du HMM par rapport à la méthode de rééchantillonage est sa capacité à générer des séries réalistes mais néanmoins possiblement très différentes des observations, tandis que la méthode des analogues aura tendance à générer des séries proches des observations, en particulier lorsque l'on s'intéresse à des statistiques annuelles.

Globalement, le HMM donne de bons résultats malgré quelques défauts. En particulier, même si les distributions marginales des précipitations à différents pas de temps (journalier, mensuel, annuel) sont bien reproduites, on constate un défaut d'auto-corrélation qui affecte la qualité des séquences sèches (VNC30) ou pluvieuses (VCX3). D'autre part, un changement brutal dans l'intensité des précpitations en automne peut ne pas être détecté. Ce problème pourrait éventuellement être résolu en augmentant le degré des polynômes trigonométriques intervenant dans les transitions et les lois d'émission. Sur les températures, le HMM donne de très bons résultats, et certains défauts de la méthode des analogues (notamment sous-estimation des VCX et VCN) sont corrigés. Même si sa modélisation n'est pas parfaite, l'enneigement est bien simulé par le HMM. Les résultats de ce point de vue sont meilleurs que ceux du modèle ANA2 sur les deux stations considérées. Cette bonne modélisation de l'enneigement démontre que la dépendance entre la température et les précipitations est bien reproduite par le HMM, ce qui est l'une des qualités que nous cherchons à obtenir. Les défauts constatés sur les précipitations sont répercutés sur le débit, notamment les VCX3 (périodes de fort débit). Cependant, pour la plupart des critères considérés, les simulations de débit obtenues via le HMM sont réalistes.

Nous avons utilisé le générateur HMM pour estimer la densité bivariée, au pas de temps annuel, de la température et du débit, pour obtenir une estimation de la probabilité d'observer un hiver froid à faible hydraulicité. Nous avons mis en évidence sur un exemple l'impact du changement climatique sur cette densité, ce qu'il est difficile de faire en utilisant les seules observations.

Enfin, nous avons estimé les états cachés du HMM grâce à l'algorithme de Viterbi, obtenant ainsi une classification "locale" en type de temps. Nous avons comparé cette classification à une autre classification obtenue à partir des précipitations et du champ géopotentiel à plus grande échelle, puis nous avons montré que les états du HMM correspondent à des configurations météorologiques particulières, en considérant les champs de géopotentiels, de précipitations et de températures. Au passage, nous avons illustré la possibilité d'utiliser une modélisation HMM comme méthode de classification non supervisée.

Chapitre 7

Modélisation conjointe de la température, des précipitations et de la vitesse du vent

Dans les chapitres 5 et 6, nous avons introduit et appliqué un modèle bivarié pour simuler des séries de températures et de précipitations. Rappelons que ce modèle est un HMM non-homogène, dont les probabilités de transitions sont périodiques, et dont les lois d'émission sont des mélanges de la forme

$$\nu_k(t) = \sum_{m=1}^{M} p_{km} \nu_{km}^{\text{precip}}(t) \otimes \nu_{km}^{\text{temp}}(t).$$

Dans ce chapitre, nous allons étendre ce modèle de façon à inclure la vitesse de vent. La façon la plus naturelle de le faire est de conserver la même structure de transition et de considérer les lois d'émission

$$\nu_k(t) = \sum_{m=1}^{M} p_{km} \nu_{km}^{\text{precip}}(t) \otimes \nu_{km}^{\text{temp}}(t) \otimes \nu_{km}^{\text{vent}}(t). \tag{7.1}$$

La modélisation trivariée se résume alors au choix d'une forme paramétrique pour les $\nu_{km}^{\text{vent}}(t)$.

7.1 Données

Pour tester notre modèle, nous avons travaillé sur les données du projet *European Climate Assessment & Dataset*[1]. Ces données sont des mesures in-situ de plusieurs variables météorologiques sur plusieurs milliers de stations en Europe et sur le pourtour méditerrannéen (toutes les variables ne sont pas disponibles pour toutes les stations). Les variables que nous avons considérées sont la température moyenne, le cumul des précipitations et la vitesse moyenne du vent (au pas journalier). Pour chacune de ces variables, nous avons sélectionné des stations selon les critères suivants :

1. Au moins 20000 observations.
2. Moins de 1000 valeurs manquantes.
3. Moins de 30 valeurs manquantes consécutives.

1. https://www.ecad.eu//dailydata/index.php

Par intersection des trois ensembles de stations ainsi obtenus, on détermine l'ensemble des stations valides pour étudier le modèle trivarié. Parmi ces stations, nous en avons sélectionné 6 que nous avons effectivement étudiées : Toulouse, Vlissingen, Rennes, Nice, Dublin et Madrid (voir Figure 7.1). Comme dans le chapitre 5, nous avons choisi des stations avec des climats variés pour tester la robustesse du modèle.

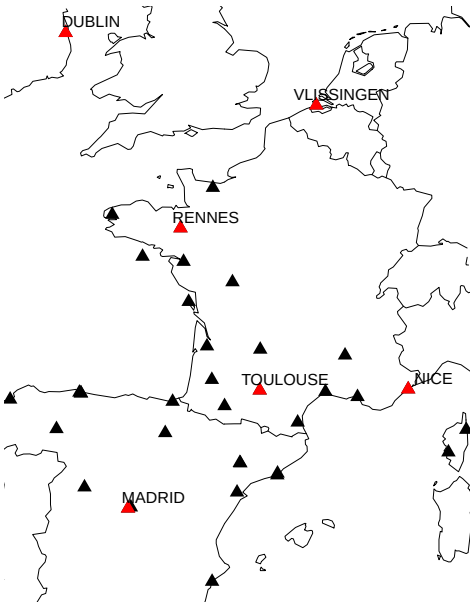

FIGURE 7.1 – Stations valides pour le modèle trivarié. En rouge, les stations effectivement étudiées.

7.2 Préambule : modélisation univariée de la vitesse du vent

Pour choisir une forme paramétrique pour les $\nu_{km}^{\text{vent}}(t)$ dans l'équation (7.2), nous avons commencé par l'étude univariée de la vitesse du vent. On introduit donc un HMM pour la modélisation du vent seul. Rappelons que nous considérons des moyennes journalières de la vitesse du vent. Une loi classique pour la modélisation de la vitesse du vent dans le contexte de l'énergie éolienne est la *loi de Weibull* à deux paramètres (Carta et al., 2009). On dit qu'une variable aléatoire positive X suit une loi de Weibull $W(a,b)$ lorsque pour tout $x > 0$,

$$\mathbb{P}(X > x) = e^{-\left(\frac{x}{b}\right)^a},$$

où $a > 0$ est un paramètre de forme et $b > 0$ est un paramètre d'échelle. Remarquons que si $a = 1$, on retrouve la loi exponentielle. Lorsque $0 < a \leq 1$, la densité de la loi $W(a,b)$ est strictement décroissante, et donc son mode est 0. Cela ne correspond pas à ce qu'on l'on observe sur la vitesse du vent, dont le mode est strictement positif. On aura donc en général $a > 1$.

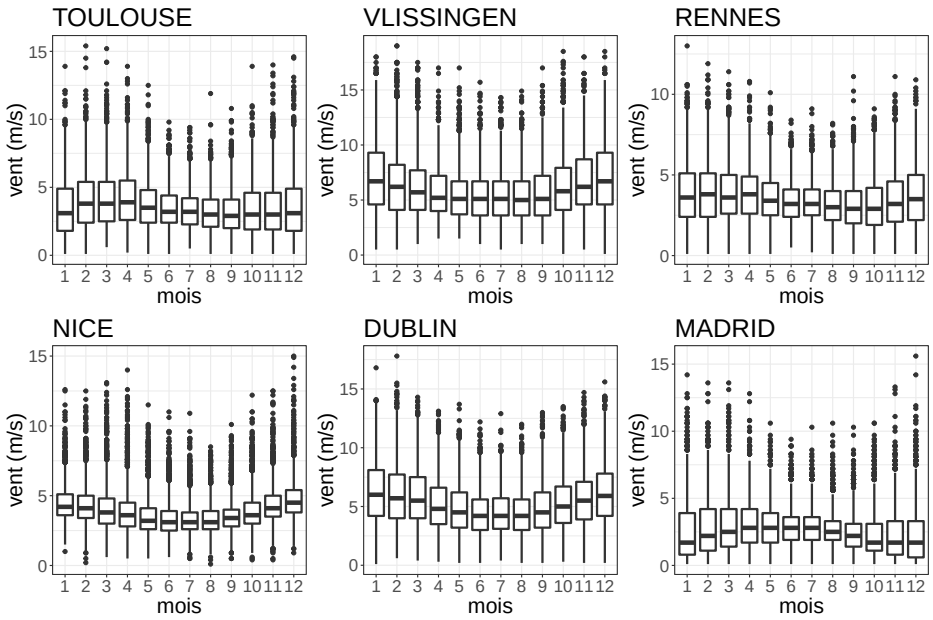

FIGURE 7.2 – Boxplots des distributions mensuelles du vent pour les six stations étudiées.

Notons que si $X \sim W(a, b)$, alors pour tout $\alpha > 0$, $\alpha X \sim W(a, \alpha b)$. Cette propriété nous servira pour introduire une saisonnalité multiplicative.

Comme toutes les variables climatiques, la vitesse du vent possède un comportement saisonnier, à la fois dans la moyenne et dans la variance (cf par exemple Smith (1983)). Sur la Figure 7.2, nous avons représenté sous forme de boxplots les distributions mensuelles de la vitesse du vent, pour les six stations étudiées. On constate sur cet exemple que la forme de la saisonnalité est variable en fonction de la station. Notons qu'il existe aussi un cycle diurne (de période 24h) mais celui-ci ne nous concerne pas puisque nous considérons des moyennes journalières.

La Figure 7.3 présente, pour chaque mois, l'histogramme de la vitesse du vent, et son ajustement à une densité de Weibull pour la station de Vlissingen. Même si l'ajustement n'est pas satisfaisant pour tous les mois, l'utilisation d'un HMM implique que la loi marginale de la vitesse du vent sera un mélange de lois de Weibull de la forme $\sum_{k=1}^{K} \mathbb{P}(X_t = k) W(a_k, b_k)$, ce qui offre davantage de flexibilité.

Pour chaque mois, l'ajustement à la loi de Weibull a nécessité l'estimation (par maximum de vraisemblance) d'un paramètre de forme $\hat{a}$ et d'un paramètre d'échelle $\hat{b}$. Ceux-ci sont représentés sur la Figure 7.4. On retrouve la saisonnalité dans ces estimateurs.

Dans le cadre d'une modélisation par HMM, si l'on choisit des lois d'émission Weibull, la saisonnalité peut être introduite à différents niveaux :

— dans les transitions entre les états : $\mathbb{P}(X_t = j \mid X_{t-1} = i) = Q_{ij}(t)$.
— dans le paramètre de forme a : la loi d'émission dans l'état k est alors une $W(a_k(t), b_k)$.
— dans le paramètre d'échelle b : la loi d'émission dans l'état k est alors une $W(a_k, b_k(t))$.

En excluant le cas où l'on n'inclut aucune saisonnalité, cela donne 7 modèles potentiels, les plus simples étant ceux où la saisonnalité est introduite dans un seul des paramètres, et le plus

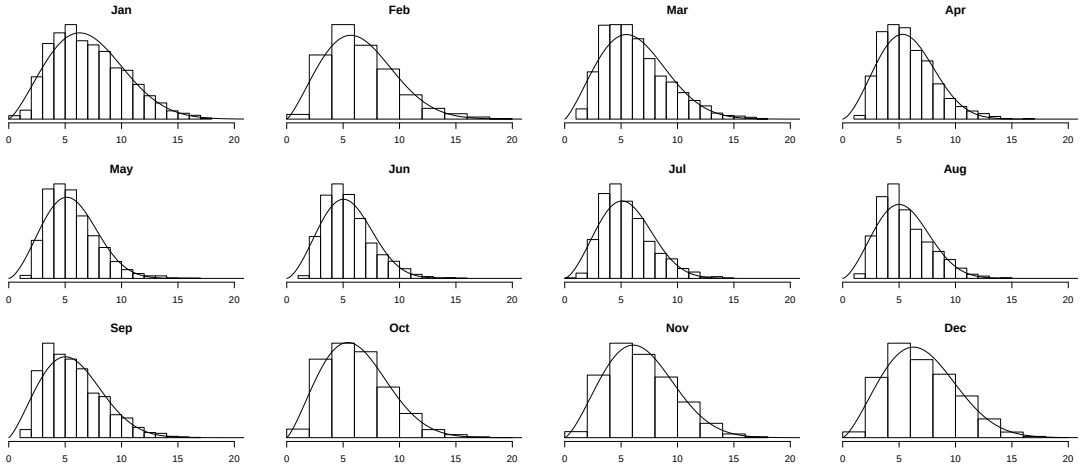

FIGURE 7.3 – Histogrammes des distributions mensuelles de la vitesse du vent et leurs ajustements à des densités de Weibull (Vlissingen)

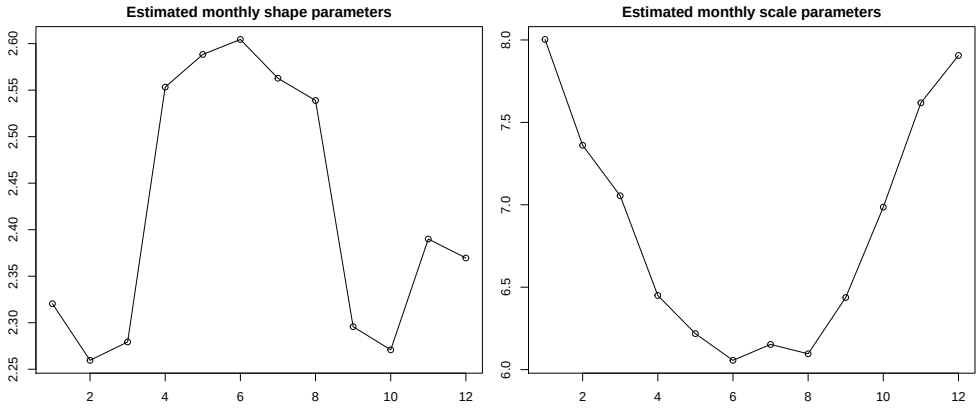

FIGURE 7.4 – Paramètres estimées des lois de Weibull mensuelles (Vlissingen)

Modèle	Saisonnalité dans...		
	Q	a	b
1	✓	✗	✗
2	✗	✗	✓
3	✗	✓	✗
4	✓	✗	✓
5	✓	✓	✗
6	✗	✓	✓
7	✓	✓	✓

TABLE 7.1 – Différents HMM pour la modélisation univariée du vent

complexe celui où la saisonnalité figure dans la matrice de transition et dans les deux paramètres des lois de Weibull (voir la table 7.1).

Parmi ces modèles, nous avons choisi de tester sur les données des 6 stations définies dans la Section 7.1 deux des modèles les plus simples : celui où les lois d'émission sont constantes (modèle 1), et celui où seul le paramètre d'échelle est périodique (modèle 2). Le premier a l'avantage de s'intégrer facilement dans un modèle multivarié (voir 7.3), et le second permet de facilement prendre en compte la saisonnalité.

7.2.1 Modèle 1 : transitions périodiques, lois d'émission constantes

Dans ce premier modèle, les probabilités de transition de la chaîne de Markov cachée $(X_t)_{t\geq 1}$ sont périodiques, de période 365, de la même forme que dans le modèle bivarié présenté dans le chapitre 5 (voir Equation (5.1)). Les lois d'émission sont des lois de Weibull constantes dans le temps : pour $k \in \{1, \dots, K\}$, $Y_t \mid \{X_t = k\} \sim W(a_k, b_k)$. Ainsi dans ce modèle, le comportement saisonnier de la vitesse du vent résulte uniquement de la saisonnalité des transitions entre états. Supposons par exemple que $K = 2$ (deux états) et que les deux lois d'émission correspondent à un état de "vent faible" (état 1) et un état de "vent fort" (état 2). Dans un tel cas, si le vent est en moyenne plus fort en hiver qu'en été, les probabilités de transition seront telles que l'état 2 sera plus fréquent en hiver qu'en été, et inversement pour l'état 1. C'est donc la fréquence des états qui induit la saisonnalité. Si l'on note d le degré des polynômes trigonométriques qui définissent les matrices de transition et K le nombre d'états, alors le nombre de paramètres du modèle est $K(K-1)(2d+1) + 2K$. En pratique, $d = 2$ et K est choisi pour chaque station en utilisant le critère BIC et d'autres considérations (cf Section 2.6). Typiquement, $K = 5$ et le modèle comporte alors 110 paramètres. L'algorithme EM (cf Section 2.3.3) est utilisé pour estimer les paramètres par maximum de vraisemblance.

La Figure 7.5 représente les densités d'émission Weibull estimées dans le modèle 1, pour chaque station. Remarquons que nous avons choisi $K = 5$ états pour la plupart des stations (6 états pour Nice) et que les états sont caractérisés par la force du vent qu'ils génèrent. Par exemple, pour la station de Vlissingen, un vent fort de 15 m/s (rappelons qu'il s'agit de moyennes journalières) ne peut être généré que dans l'état noir, un vent de 10 m/s peut être généré dans les états noir ou rouge, et un vent faible de 1 m/s peut être généré dans l'état vert ou, plus rarement, dans l'état bleu.

Ayant obtenu, pour $1 \leq t \leq 365$, un estimateur $\hat{Q}(t)$ de la matrice de transition à l'instant t, on peut en déduire une estimation de la distribution des états à chaque instant : voir la Figure 7.6. On constate que les fréquences relatives des états ne sont pas constantes au cours de l'année.

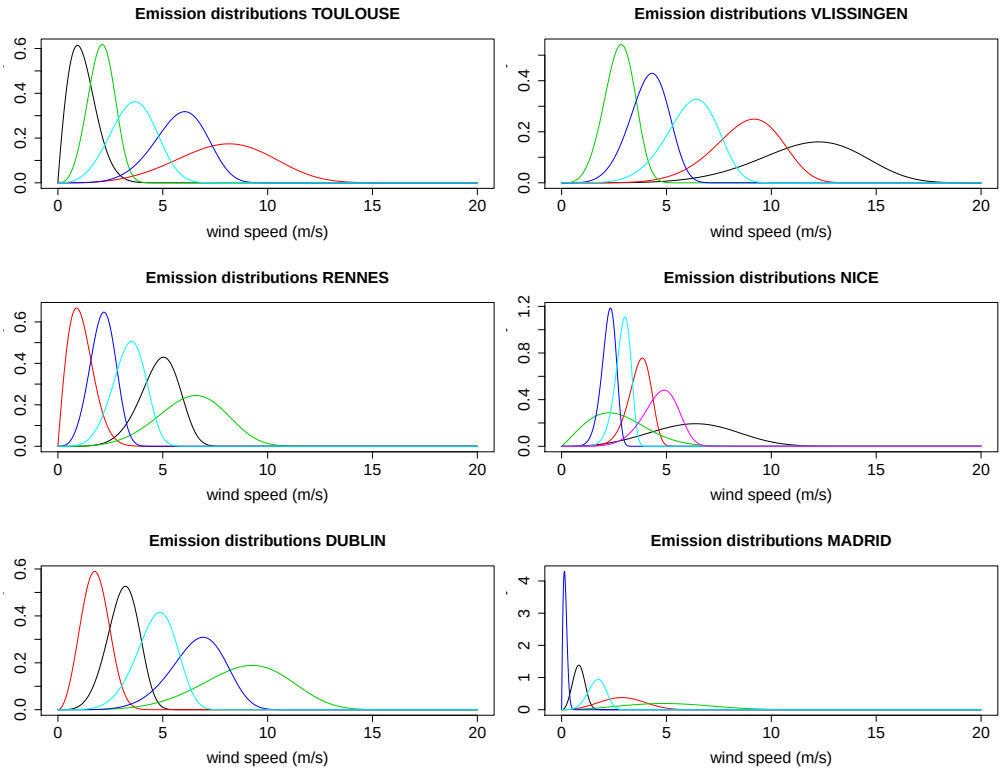

FIGURE 7.5 – Lois d'émission estimées dans le modèle 1. Chaque couleur correspond à un état.

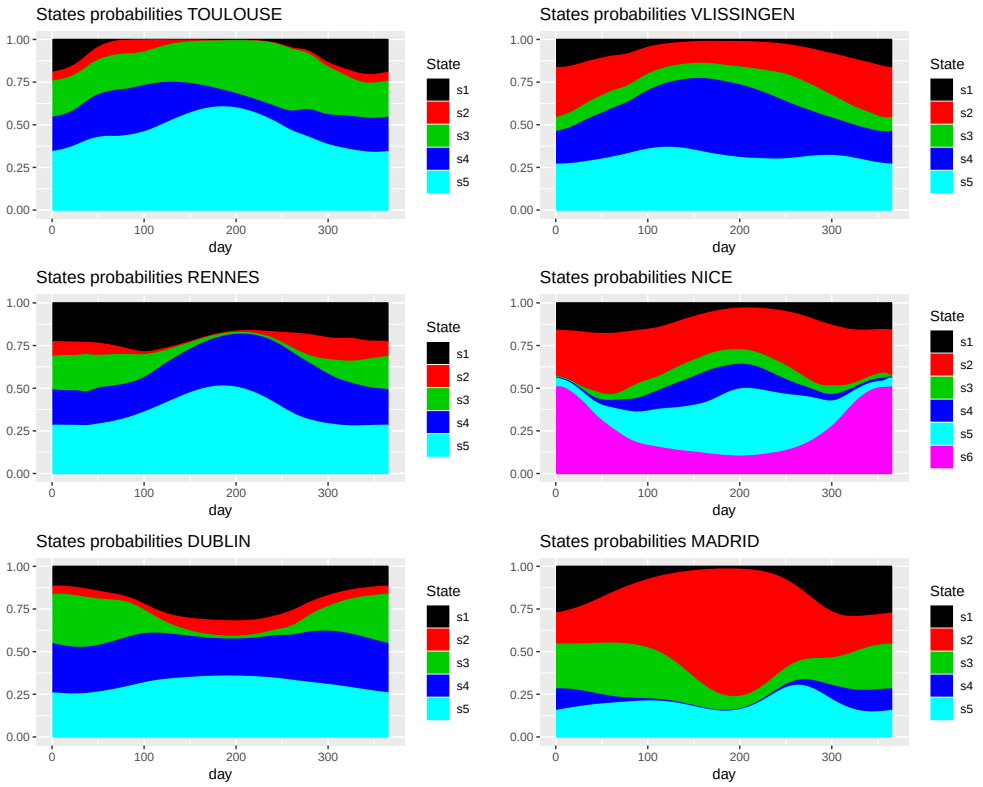

FIGURE 7.6 – Estimation des fréquences relatives des états, pour $1 \leq t \leq 365$.

Pour interpréter ces résultats, il convient de les mettre en relation avec ceux de la Figure 7.5. Prenons l'exemple de la station de Nice. Les états 1 et 6 (respectivement noir et rose) sont les deux états correspondant aux vents les plus forts. On observe sur la Figure 7.6 que ces deux états sont prépondérants en hiver, tandis qu'ils deviennent rares en été. A l'inverse, l'état 4 (bleu) qui correspond à un vent faible atteint sa fréquence maximale en été, et devient presque inexistant en hiver. Ainsi on retrouve bien le phénomène saisonnier que l'on observe sur la vitesse du vent sur cette station, avec un vent globalement plus fort en hiver qu'en été. Une analyse semblable peut être menée pour les autres stations et l'on peut déjà supposer, par simple inspection des paramètres estimés, que le modèle reproduit bien la saisonnalité du vent.

Pour nous en assurer, nous avons mis en place une procédure de validation du modèle basée sur des simulations, comme dans la Section 5.3.4 pour le modèle bivarié température et précipitations. Pour chaque station, nous avons donc simulé $N_{\text{sim}} = 1000$ trajectoires indépendantes de la même longueur que la série d'observations, en utilisant les paramètres estimés. Nous commençons par vérifier que la distribution globale de la vitesse du vent est bien respectée par le modèle. Nous avons pour cela calculé des estimateurs à noyau de la densité de la vitesse du vent, sur les observations puis sur les simulations issues du modèle. La Figure 7.7 montre que les simulations ont une distribution proche de celle des observations. Cela n'est cependant pas suffisant pour conclure quant à la qualité du modèle, puisque le processus des observations n'est pas i.i.d., ni même stationnaire. On s'intéresse donc aussi à des statistiques journalières.

Les premiers moments journaliers sont correctement reproduits par le modèle. Les résultats

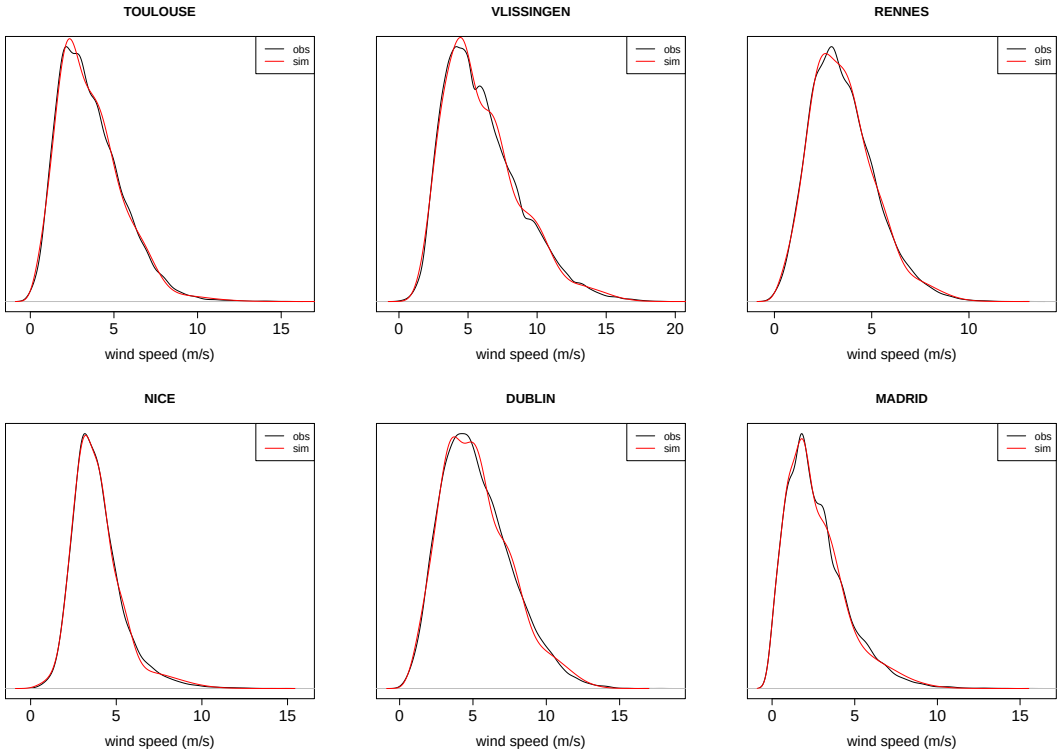

FIGURE 7.7 – Estimation de la densité de la vitesse du vent basée sur les observations (en noir) et sur les simulations (en rouge).

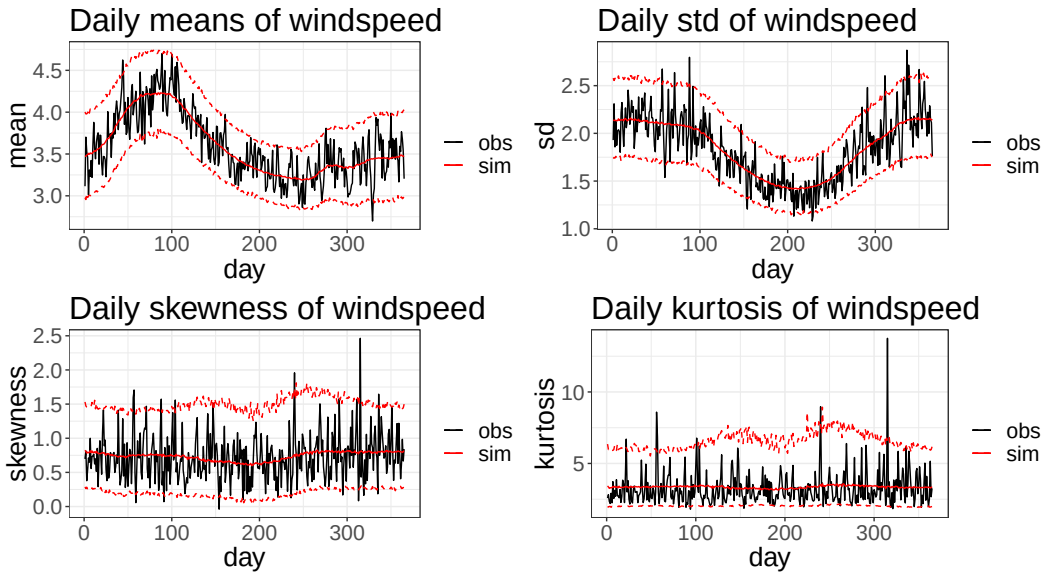

FIGURE 7.8 – Moments journaliers de la vitesse du vent (Toulouse). Les valeurs calculées sur les observations sont en noir. Le trait plein rouge est la moyenne sur les 1000 trajectoires et les pointillés rouges forment un intervalle de confiance à 95% obtenus à partir des simulations.

sont représentés sur la Figure 7.8 pour la station de Toulouse. On retrouve bien, sur toutes les stations, la saisonnalité des deux premiers moments. Les moments d'ordre supérieur ne présentent pas de saisonnalité nette. Notons que le coefficient d'asymétrie (skewness) positif indique une distribution asymétrique à droite.

Outre les moments, on peut aussi s'intéresser aux maxima de la vitesse du vent, pour chaque jour calendaire. Ceux-ci sont représentés sur la Figure 7.9 pour les six stations.

Observée au pas de temps journalier, la vitesse du vent présente une dépendance temporelle : la vitesse du vent d'un jour donné n'est pas indépendante de la vitesse du vent la veille. Nous voulons nous assurer que le modèle reproduit bien cette caractéristique. Pour cela nous avons calculé les auto-corrélations d'ordre 1, 2 et 3, c'est-à-dire, respectivement, la corrélation entre Y_t et Y_{t-1}, Y_t et Y_{t-2}, et Y_t et Y_{t-3}, sur les observations et sur les trajectoires simulées. Les résultats sont présentés sur la Figure 7.10. On voit que pour toutes les stations étudiées, le modèle sous-estime légèrement l'auto-corrélation d'ordre 1. Les auto-corrélations d'ordre 2 et 3 sont légèrement sur-estimées, sauf à Nice et Madrid. On constate au passage que les auto-corrélations varient selon les stations.

La dépendance temporelle peut aussi être mesurée en considérant la distribution des durées des périodes de vent faible ou fort. Nous appelons période de vent faible (resp. fort) une période de un ou plusieurs jours pendant laquelle la vitesse du vent est inférieure à son 1er quartile (resp. supérieure à son 95ème percentile). Sur la Figure 7.11 (resp. Figure 7.12), nous avons représenté en blanc la distribution observée des longueurs des périodes de vent faible (resp. fort). Les barres d'erreur noires donnent un intervalle de confiance à 95% sous le modèle de ces quantités, obtenu par simulation. Les points noirs figurent la moyenne sur toutes les simulations. Sur la plupart des stations, pour le vent faible comme pour le vent fort, on observe une légère sur-estimation par le modèle de la fréquence des épisodes de longueur 1 (et donc une sous-estimation de la

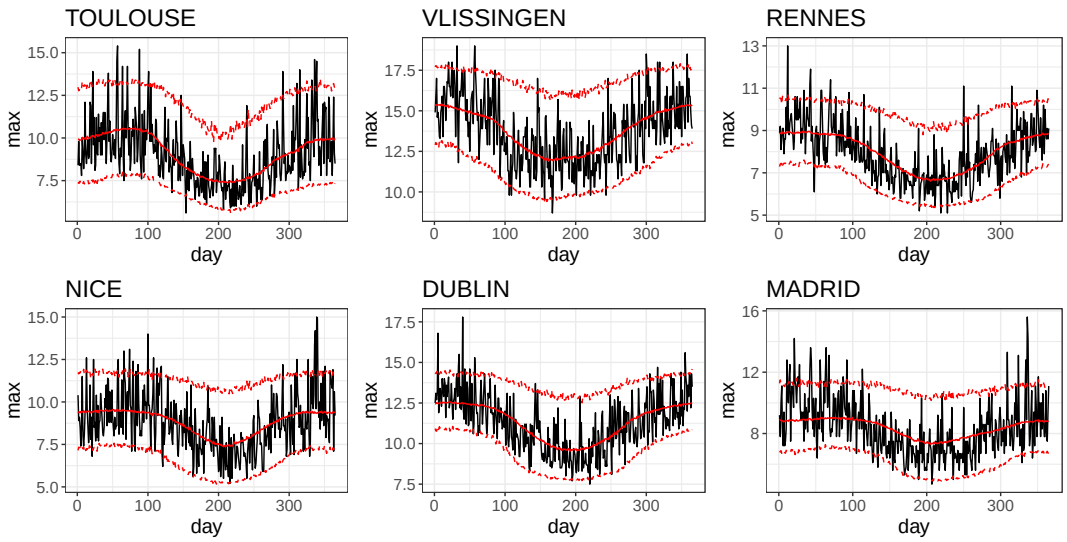

FIGURE 7.9 – Maxima de la vitesse du vent sur les observations (en noir) et sur les simulations (en rouge).

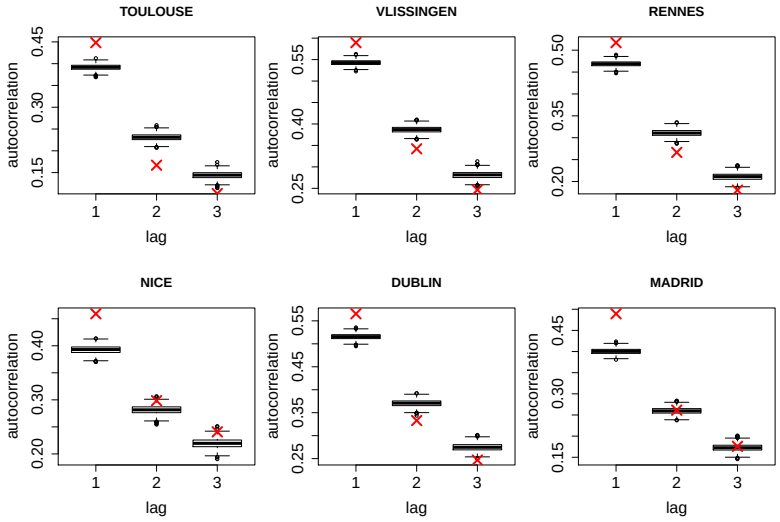

FIGURE 7.10 – Auto-corrélations d'ordre 1 à 3. En rouge, les valeurs calculées sur les observations. Les boxplots ont été obtenus à partir des 1000 simulations.

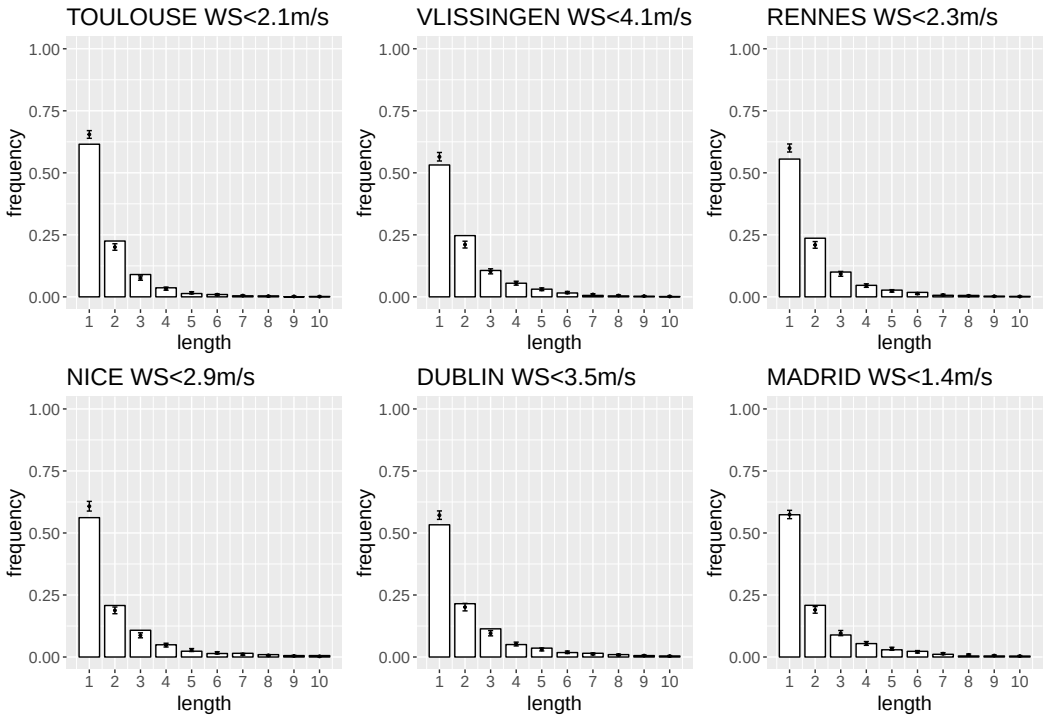

FIGURE 7.11 – Longueurs des périodes de vent faible.

fréquence des épisodes plus longs). Pour le vent fort, on fait le même constat si l'on remplace le 95ème percentile par un autre quantile extrême. Cela traduit un léger défaut de dépendance temporelle du modèle : ces résultats sont cohérents avec ceux de la Figure 7.10. Malgré ce léger défaut, la distribution des durées des périodes de vent faible ou fort est correctement reproduite : la dépendance markovienne d'ordre 1 est suffisante pour des données de vent journalières. Si le pas de temps était horaire, la dépendance serait beaucoup plus forte.

La vitesse du vent possède une variabilité interannuelle : certaines années sont plus venteuses que d'autres. Sur la Figure 7.13, nous avons représenté les moyennes annuelles de la vitesse du vent observée et simulée. On peut observer, notamment sur les données de Toulouse ou Nice, des périodes de plusieurs années consécutives ou le vent est plus fort (ou plus faible) que la moyenne. Bien entendu, cette caractéristique intéresse particulièrement les producteurs d'énergie éolienne puisqu'elle est synonyme d'incertitude sur la quantité d'énergie qui pourra être produite par les installations éoliennes sur une période donnée (voir Lee et al. (2018)). On constate sur les Figures 7.15 et 7.16 que la variabilité de la vitesse moyenne annuelle du vent est sous-estimée par le modèle.

Il est également possible de le voir en estimant les états cachés grâce à l'algorithme de Viterbi (cf 2.3.2) et en calculant les fréquences relatives de ces états estimés pour chaque année. Une telle opération appliquée aux observations montre de fortes variations dans la fréquence des états d'une année sur l'autre. Les résultats pour les données de Vlissingen sont présentés sur la Figure 7.14.
C'est un problème classique pour les générateurs de temps. En revanche, la variabilité interannuelle de la vitesse moyenne mensuelle du vent est mieux reproduite par le modèle, comme le

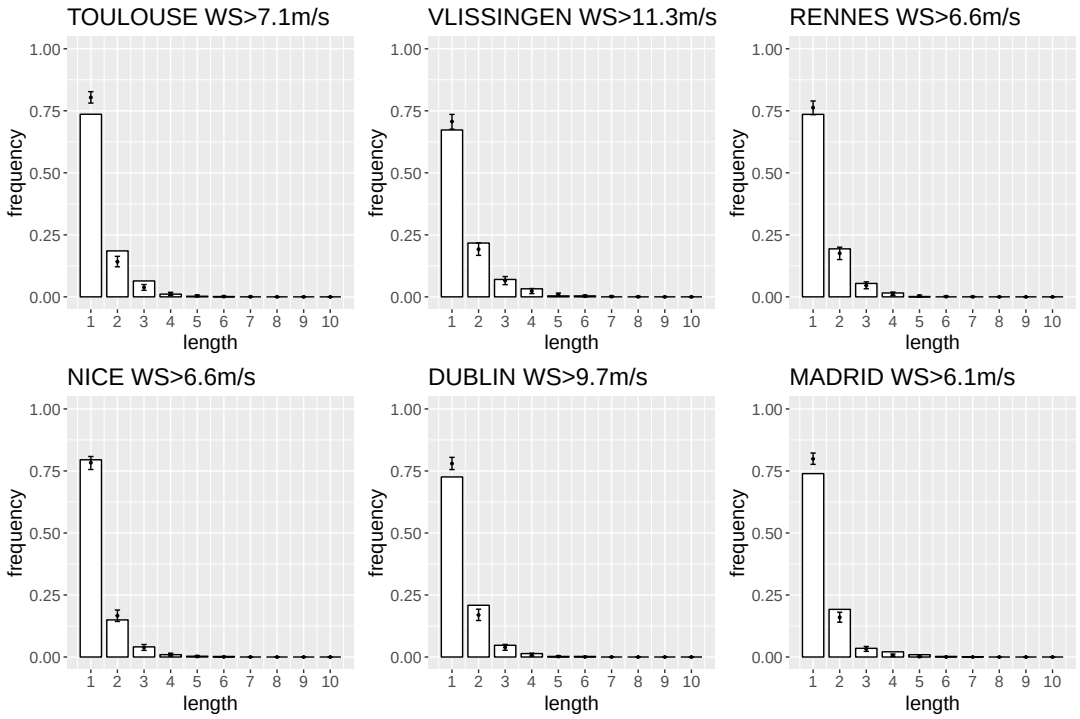

FIGURE 7.12 – Longueurs des périodes de vent fort.

montrent les Figures 7.17 et 7.18. Ainsi, le problème semble venir de l'incapacité du modèle à générer plusieurs mois consécutifs où la vitesse du vent est en moyenne élevée (ou faible).

La variabilité interannuelle peut être la résultante de cycles climatiques à grande échelle spatiale et temporelle. Par exemple, dans Ailliot and Monbet (2012), les auteurs constatent une corrélation positive entre les états cachés de leur modèle (un HMM auto-régressif) et l'indice AMO (Atlantic Multidecadal Oscillation), une oscillation de la température de surface de l'océan Atlantique, dont la période est d'approximativement 60 ans. Ils proposent d'introduire une tendance dans leur matrice de transition pour obtenir ces variations interannuelles. En l'absence d'une telle tendance, notre modèle est cyclo-stationnaire au sens où la distribution de chaque année est identique, ce qui peut expliquer la trop faible variabilité interannuelle. Une solution pourrait être d'identifier des variables exogènes permettant d'expliquer cette variabilité et de les intégrer au modèle.

7.2.2 Modèle 2 : transitions constantes, paramètre d'échelle périodique

Dans le second HMM que nous proposons pour la modélisation de la vitesse du vent, la chaîne de Markov cachée est homogène : les probabilités de transition ne varient pas au cours du temps. En revanche les lois d'émission sont périodiques, de période 365. Rappelons que si $X \sim W(a, b)$, alors pour tout $\alpha > 0$, $\alpha X \sim W(a, \alpha b)$, et donc l'espérance et l'écart-type d'une loi de Weibull sont proportionnels à son paramètre d'échelle. Ainsi on peut facilement introduire une saisonnalité multiplicative en multipliant le paramètre d'échelle b par une fonction périodique. On propose

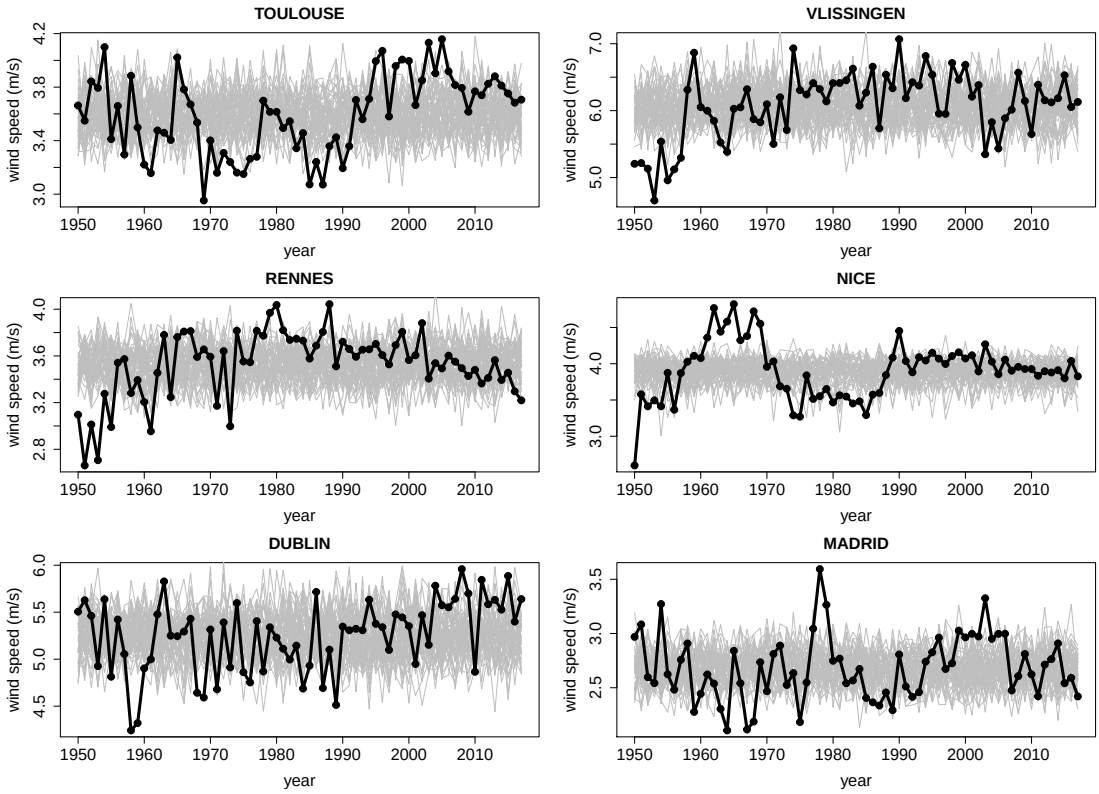

FIGURE 7.13 – Moyennes annuelles de la vitesse du vent observée (en noir) et simulée (50 trajectoires, en gris)

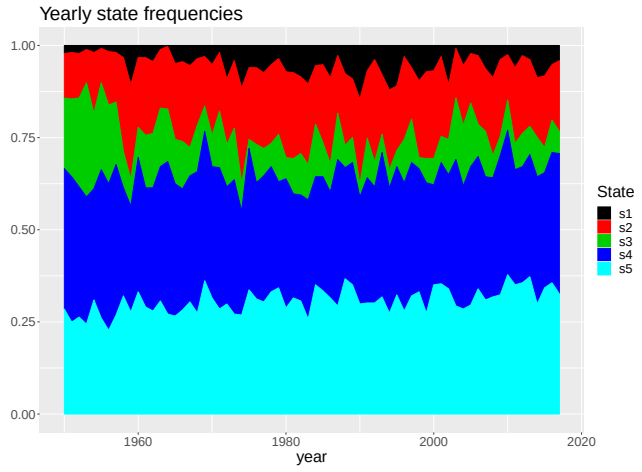

FIGURE 7.14 – Fréquences annuelles des états cachés estimés par l'algorithme de Viterbi (Vlissingen). L'état de vent faible est l'état 3 (vert) et les états de vent fort sont les états 1 (noir) et 2 (rouge).

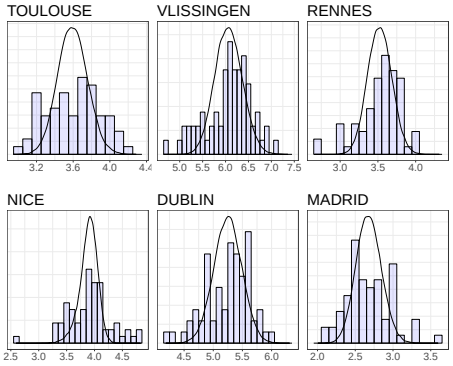

FIGURE 7.15 – Vitesse moyenne annuelle du vent. Histogramme : observations. La courbe noire est une estimation de la densité de probabilité de la vitesse moyenne annuelle du vent sous le modèle, obtenue par simulations.

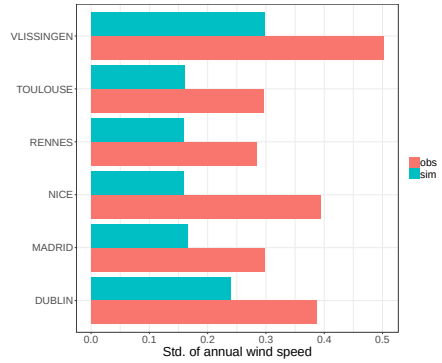

FIGURE 7.16 – Ecart-type de la moyenne annuelle de la vitesse du vent : observé versus simulé.

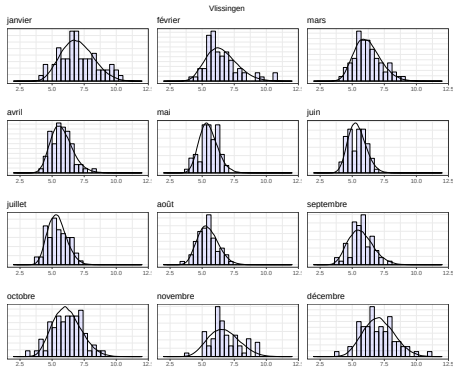

FIGURE 7.17 – Vitesse moyenne mensuelle du vent, station de Vlissingen. Histogramme : observations. La courbe noire est une estimation de la densité de probabilité la vitesse moyenne mensuelle du vent sous le modèle, obtenue par simulations.

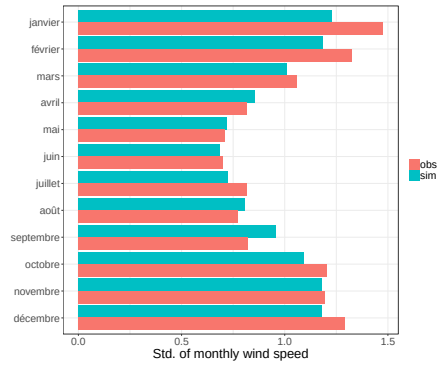

FIGURE 7.18 – Ecart-type de la moyenne mensuelle de la vitesse du vent : observé versus simulé.

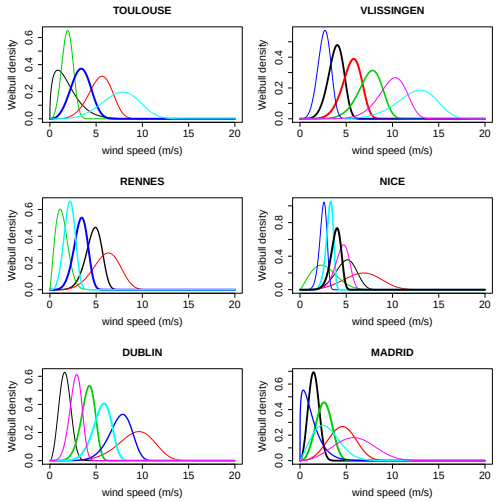

FIGURE 7.19 – Densités d'émission des $W(a_k, c_k)$. Chaque couleur correspond à un état et l'épaisseur du trait est proportionnelle à la fréquence des états.

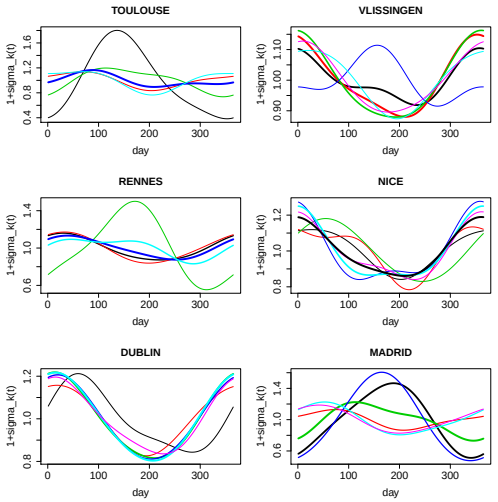

FIGURE 7.20 – Saisonnalités $1 + \sigma_k(t)$. Chaque couleur correspond à un état et l'épaisseur du trait est proportionnelle à la fréquence des états.

donc des lois d'émission de la forme

$$Y_t \mid \{X_t = k\} \sim W\left(a_k, b_k(t)\right),$$

où $b_k(t) = c_k\left(1 + \sigma_k(t)\right)$, avec $c_k > 0$ et σ_k est un polynôme trigonométrique sans terme constant tel que pour tout t, $\sigma_k(t) > -1$. Le nombre de paramètres d'un tel modèle est $K(K-1) + 2K + 2dK$, si d est le degré des polynômes σ_k. Typiquement, $d = 2$ et $K = 6$, et donc le modèle a 66 paramètres. Comme pour le premier modèle, ceux-ci sont estimés par maximum de vraisemblance par l'algorithme EM.

Sur la Figure 7.19, nous avons représenté les densités d'émission sans le facteur multiplicatif, c'est-à-dire les densités des $W(a_k, c_k)$. Celles-ci sont multipliées par les facteurs $1 + \sigma_k(t)$ représentés sur la Figure 7.20. Le comportement saisonnier de la vitesse du vent est bien illustré sur les stations de Vlissingen, Nice et Dublin, où σ_k est positif en hiver et négatif en été, dans la plupart des états. Par conséquent, pour ces stations, le vent est en moyenne (et en variance) plus fort en hiver qu'en été. Remarquons également qu'à Toulouse, Vlissingen, Rennes et Madrid, le sens de la saisonnalité de l'état associé au vent faible (exemple : états noir et bleu pour Madrid) est inversé par rapport aux autres états. Cela signifie que sur ces stations, le vent faible tend à l'être un peu moins en été.

Nous avons mené les mêmes tests de validation que pour le modèle précédent. Nous ne détaillons pas les résultats ici car ils sont très similaires à ceux obtenus sur le Modèle 1, et les mêmes défauts peuvent être identifiés.

7.2.3 Conclusion

En utilisant des lois d'émission Weibull et en introduisant de la saisonnalité soit dans les probabilités de transition, soit dans les lois d'émission, nous obtenons deux modèles de Markov cachés

qui reproduisent correctement de nombreuses caractéristiques de la vitesse moyenne journalière du vent. Nous notons cependant un léger défaut dans l'auto-corrélation, et dans la variabilité interannuelle. Le second modèle, dont la matrice de transition est constante, est plus parcimonieux que le premier, même s'il nécessite parfois un nombre légèrement plus élevé d'états cachés. Si l'objectif est la simulation de chroniques univariées de vent, il est donc préférable de choisir le second modèle, qui est aussi plus facilement interprétable. En revanche, pour l'intégration à un modèle trivarié, nous choisirons le premier modèle, puisque le modèle bivarié température/précipitations contient déjà une structure de transition périodique, commune à toutes les variables. La section suivante est consacrée à l'étude d'un tel modèle trivarié.

7.3 Modèle trivarié

7.3.1 Description du modèle

Pour la modélisation trivariée précipitations/température/vent, nous considérons un modèle de Markov caché $(X_t, Y_t)_{t\geq 1}$, où X_t est une chaîne de Markov à valeurs dans $\{1, \dots, K\}$, pour K un entier strictement positif, et

$$Y_t = \left(Y_t^{(1)}, Y_t^{(2)}, Y_t^{(3)}\right) \in \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+.$$

$Y_t^{(1)}$ correspond aux précipitations, $Y_t^{(2)}$ à la température, et $Y_t^{(3)}$ à la vitesse du vent. Suivant la conclusion de la section précédente, nous étendons le modèle bivarié précipitations/température (voir le paragraphe 5.3.2) en utilisant une loi de Weibull (constante de le temps) pour la composante correspondant au vent. Ainsi les probabilités de transition de la chaîne de Markov $(X_t)_{t\geq 1}$ sont les $Q_{ij}(t) := \mathbb{P}(X_{t+1} = j \mid X_t = i)$ donnés par :

$$\begin{aligned} Q_{ij}(t) &= \frac{\exp(P_{ij}(t))}{1 + \sum_{l=1}^{K-1} \exp(P_{il}(t))}, 1 \leq j \leq K-1, \quad 1 \leq i \leq K \\ Q_{iK}(t) &= \frac{1}{1 + \sum_{l=1}^{K-1} \exp(P_{il}(t))}, 1 \leq i \leq K, \end{aligned}$$

où les P_{ij} sont des polynômes trigonométriques de degré d et de période $T = 365$. Les lois d'émission sont de la forme, pour $t \geq 1$ et $k \in \{1, \dots, K\}$:

$$\nu_k(t) = \left(\sum_{m=1}^{M} p_{km} \nu_{km}^{(1)}(t) \otimes \nu_{km}^{(2)}(t) \right) \otimes \nu_k^{(3)} \tag{7.2}$$

Comme dans le modèle bivarié,

$$\nu_{km,t}^{(1)} = \begin{cases} \delta_0 & , 1 \leq m \leq M_1 \\ \mathcal{E}\left(\frac{\lambda_{km}}{1+\sigma_k^P(t)}\right) & , M_1 + 1 \leq m \leq M \end{cases},$$

pour les précipitations, et

$$\nu_{km,t}^{(2)} = \mathcal{N}\left(T_k(t) + S_k(t) + \mu_{km}, \sigma_{km}^2\right),$$

pour la température. Rappelons que $\mathcal{E}$ désigne la loi exponentielle, pour $1 \leq k \leq K$ et $M_1 \leq m \leq M$, $\lambda_{km} > 0$, σ_k^P et S_k sont des polynômes trigonométrique de degré d et de période 365, T_k est un terme de tendance linéaire ou linéaire par morceaux, $\mu_{km} \in \mathbb{R}$ et $\sigma_{km}^2 > 0$ pour tout

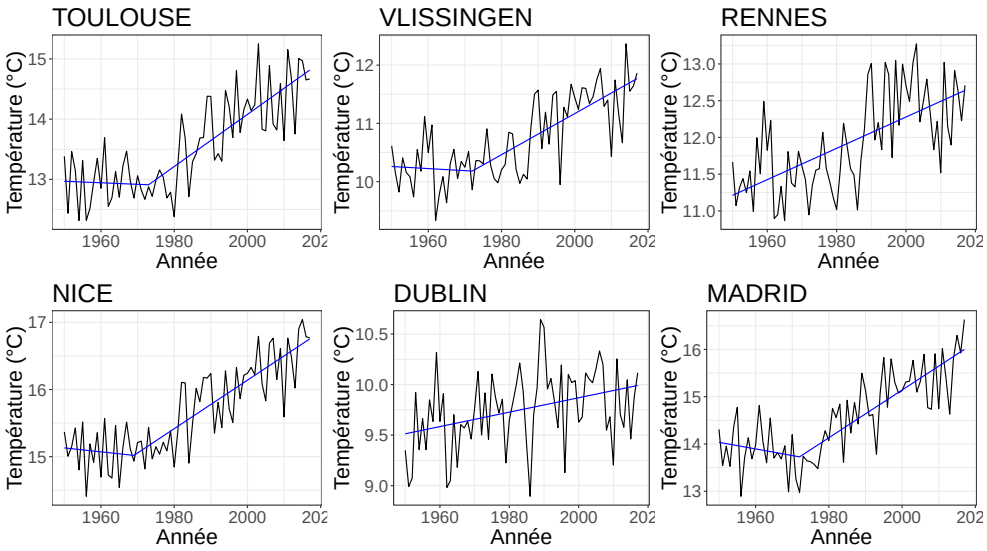

FIGURE 7.21 – Température moyenne annuelle et tendance linéaire ou linéaire par morceaux.

$m \in \{1, \dots, M\}$. La forme paramétrique de la tendance est obtenue comme dans le Chapitre 5 (voir la Figure 7.21). Enfin, pour le vent :

$$\nu_k^{(3)} = W(a_k, b_k),$$

avec $a_k, b_k > 0$, ne dépend pas du temps. Notons que dans cette modélisation, *conditionnellement à l'état*, la vitesse du vent est indépendante des deux autres variables.

7.3.2 Résultats

Nous avons appliqué ce modèle aux données présentées dans la Section 7.1. Pour les 6 stations considérées, nous avons déterminé le nombre optimal d'états. Nous avons obtenu $K = 8$ pour toutes les stations sauf Rennes, pour laquelle $K = 9$. Rappelons que pour le modèle bivarié température/précipitations (Chapitre 5), le nombre d'états était 6 ou 7. L'ajout du vent dans le modèle implique d'augmenter le nombre d'états si l'on veut décrire correctement le processus trivarié, mais l'augmentation reste raisonnable puisque le vent n'est pas indépendant des deux autres variables.

Paramètres estimés

Comme dans le Chapitre 5, commençons par examiner les paramètres estimés. A partir des estimations $\hat{Q}(t)$ des matrices de transition, on obtient des estimations des probabilités $\mathbb{P}(X_t = k)$ pour $1 \leq t \leq 365$ et $k \in \{1, \dots, K\}$. Rappelons que ces états sont communs aux trois variables, ils gouvernent donc toute la dynamique du modèle. Les résultats pour les 6 stations sont représentés sur la Figure 7.22. Certaines stations (par exemple Nice) présentent de fortes amplitudes dans la saisonnalité des fréquences relatives des états : certains états sont très fréquents en été et presque inexistants en hiver, ou inversement. Cela induit une saisonnalité marquée des variables simulées.

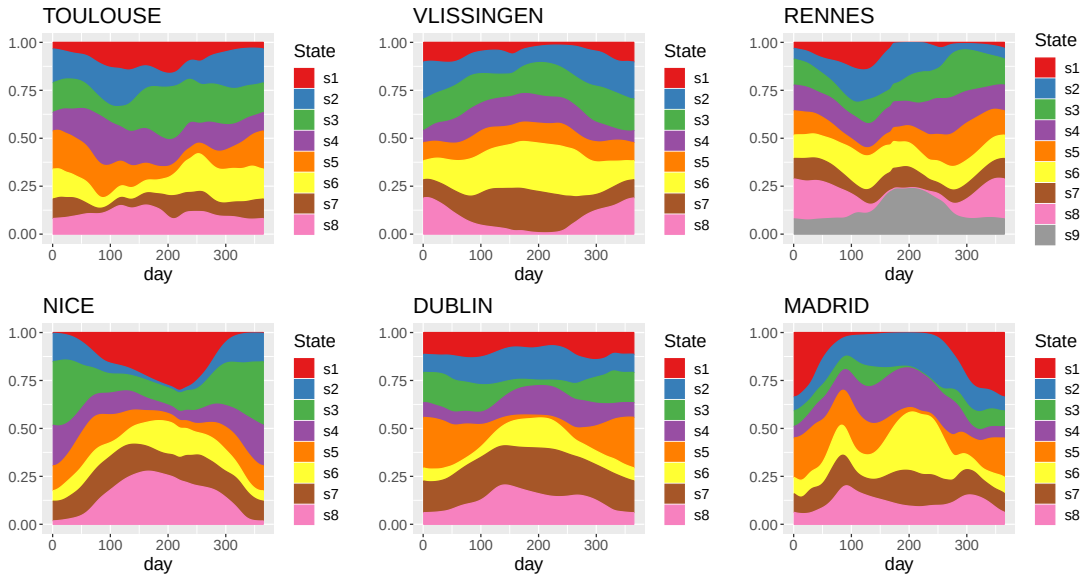

FIGURE 7.22 – Fréquences relatives des états sur une année

Sur la Figure 7.23, nous avons représenté sur les six graphes supérieurs les tendances estimées de la température, c'est-à-dire les $T_k(t)$ pour $1 \leq t \leq n$ et $k \in \{1, \ldots, K\}$ pour les six stations. Quelques remarques peuvent être faites à partir de ces graphes :

- Pour toutes les stations, on observe une tendance globale au réchauffement, ce qui est cohérent avec les graphes de la Figure 7.21. Néanmoins, l'amplitude du réchauffement n'est pas la même partout : elle est plus forte à Nice et Madrid (climat Méditerranéen), et moins forte à Dublin et Vlissingen (climat océanique).
- Le modèle permet à chaque état d'avoir sa propre tendance, mais dans de nombreux cas, elles ne diffèrent que par une constante.
- Pour les stations dont les tendances sont linéaires par morceaux, la première partie de certaines tendances est presque nulle, ou légèrement décroissante.
- Les stations de Dublin et Madrid présentent des tendances singulières, qui se différencient des autres tendances : même si globalement la température augmente, l'un ou plusieurs des états peuvent se caractériser par une trajectoire différente, voire décroissante.

Les six graphes inférieurs de la Figure 7.23 représentent, pour chaque station, le cycle annuel moyen des températures dans chaque état. Pour assurer la lisibilité, nous avons inclus la moyenne du bruit dans la saisonnalité, ce sont donc les $S_k(t) + \sum_{m=1}^{M} p_{km}\mu_{km}$ qui sont représentés sur ces graphes.

Les saisonnalités dans l'intensité des précipitations $1+\sigma_k^P(t)$ sont représentées sur la Figure 7.24.

Sur la Figure 7.25, nous avons tracé, pour chaque station et chaque état $k \in \{1, \ldots, K\}$, la composante des densités d'émission correspondant à la vitesse du vent, soit la densité Weibull $W(a_k, b_k)$. Il est clair que ces résultats ressemblent à ceux obtenus avec le modèle de vent univarié (voir la Figure 7.7). Le nombre d'états étant ici plus grand, il peut arriver que certaines densités liées à des états différents soient presque identiques. Cela ne signifie pas nécessairement que le nombre d'états est trop élevé car ces états, très proches en ce qui concerne le vent, peuvent être différents lorsque l'on considère les autres variables.

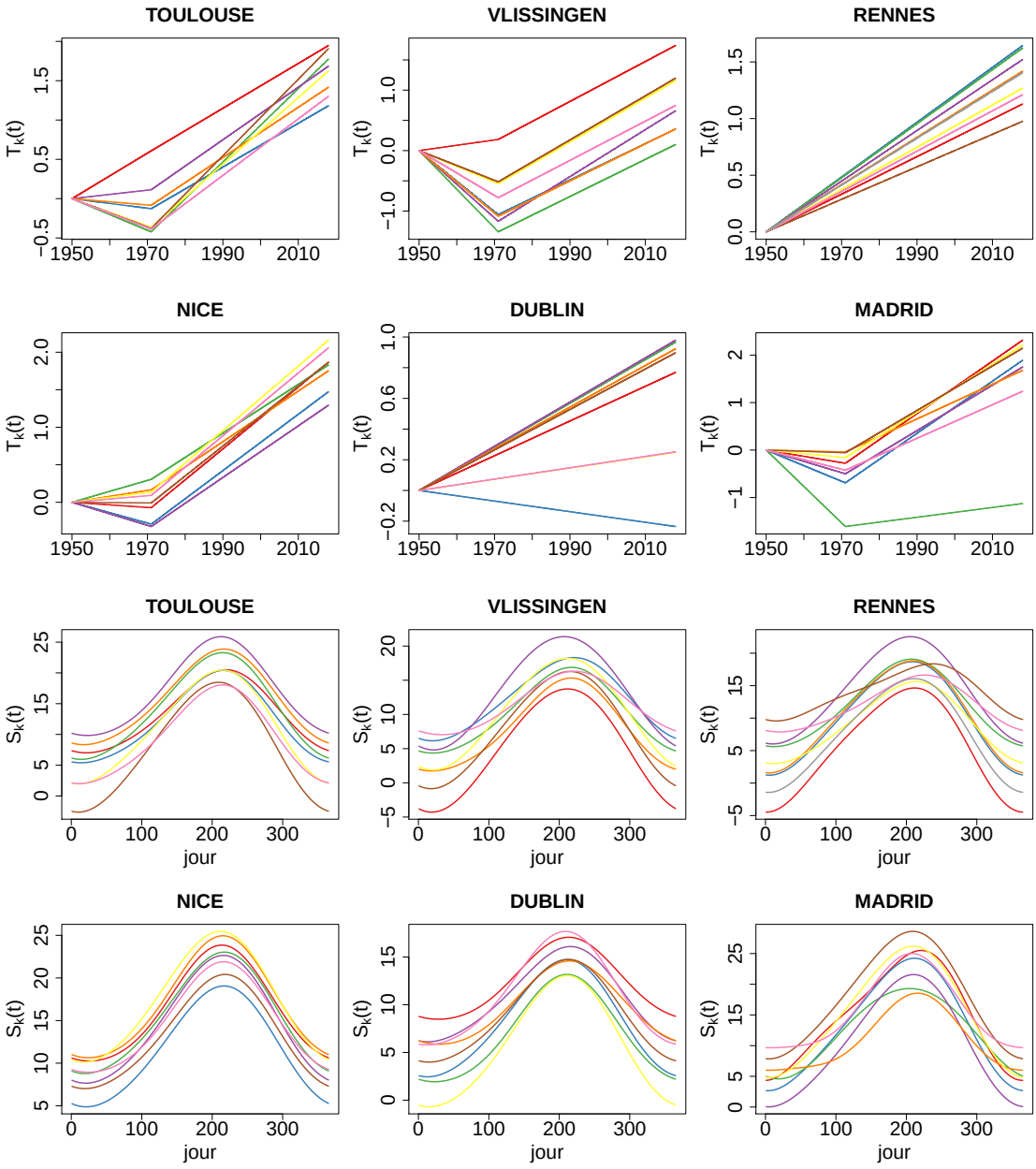

FIGURE 7.23 – Tendances et saisonnalités des températures.

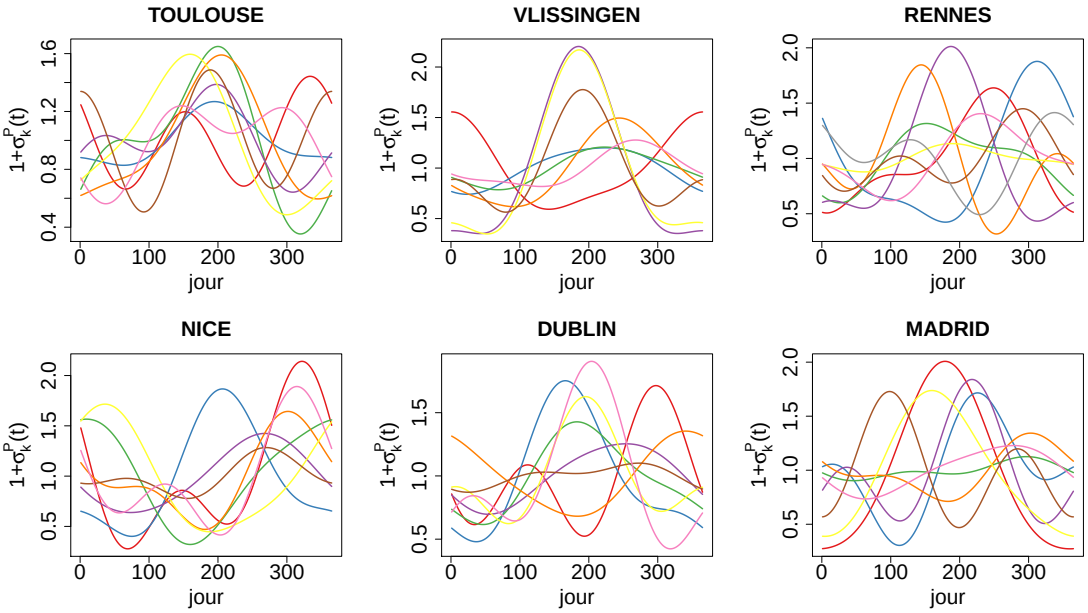

FIGURE 7.24 – Saisonnalité de l'intensité des précipitations dans chaque état

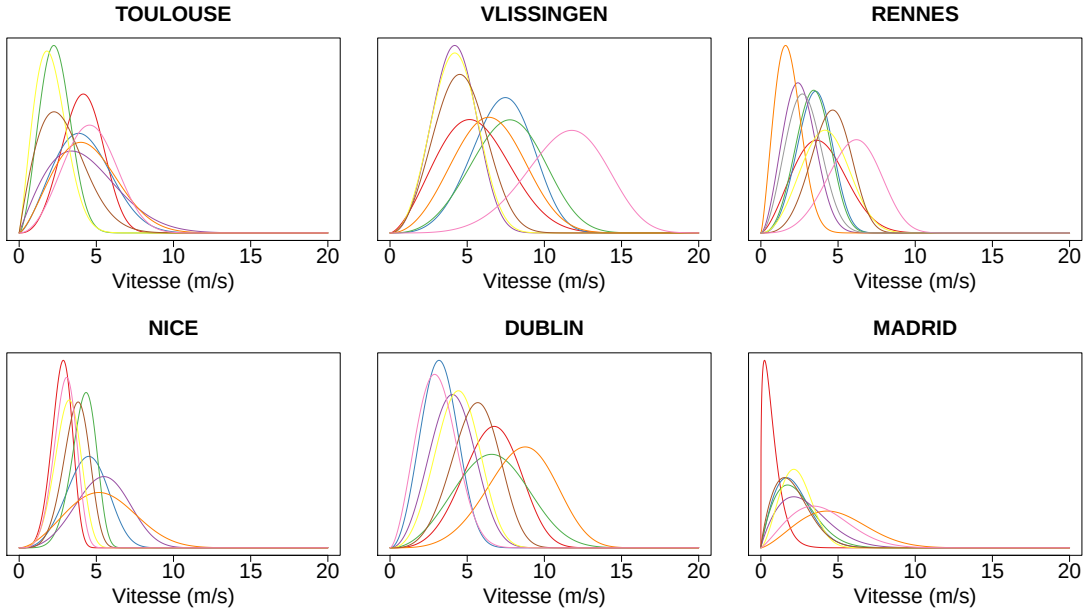

FIGURE 7.25 – Densités d'émission estimées pour la vitesse du vent : $W(a_k, b_k)$.

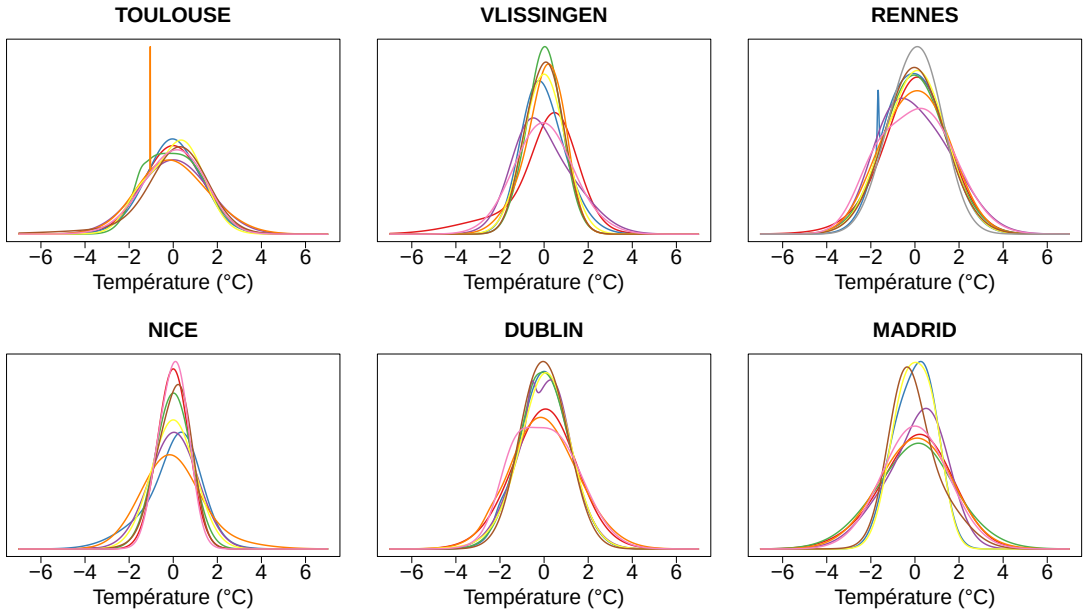

FIGURE 7.26 – Densités du "bruit" centré de la température

La Figure 7.26 représente les densités du bruit centré, c'est-à-dire, pour $k \in \{1, \dots, K\}$, la densité de la loi $\sum_{m=1}^{M} p_{km} \mathcal{N}(0, \sigma_{km}^2)$, et celles de la Figure 7.27 celles de l'intensité des précipitations, indépendamment de la saisonnalité :

$$\frac{1}{\sum_{m=M_1+1}^{M} p_{km}} \sum_{m=M_1+1}^{M} p_{km} \mathcal{E}\left(\lambda_{km}\right).$$

Enfin, la Figure 7.28 représente les $\left(\sum_{m=1}^{M} p_{km}\right)_{1 \leq k \leq K}$, soit la probabilité de générer des précipitations nulles dans l'état k.

L'analyse des paramètres estimés permet de donner une interprétation aux différents états sous le modèle et sous les paramètres estimés. Pour juger de la qualité du modèle, il est cependant nécessaire de confronter les simulations aux observations.

Tests de validation

Nous voulons vérifier que notre modèle permet de générer des séries synthétiques trivariées qui reproduisent raisonnablement bien les propriétés statistiques des séries observées. Dans cette optique, comme dans le Chapitre 5, nous avons généré pour chaque station, en utilisant le modèle, 1000 trajectoires indépendantes les unes des autres, de même longueur que les données. Dans la suite, nous nous intéressons à différentes statistiques univariées et bivariées et nous confrontons les simulations aux observations. Notre modèle trivarié étant une extension directe de notre modèle bivarié température/précipitations, on peut s'attendre à ce que les tests de validation concernant ces deux variables soit satisfaisants. Il est néanmoins nécessaire de le vérifier. En effet, les trois variables sont liées via les états, et l'ajout du vent dans le modèle pourrait ainsi avoir une action perturbatrice sur les matrices de transition, et donc indirectement sur les deux

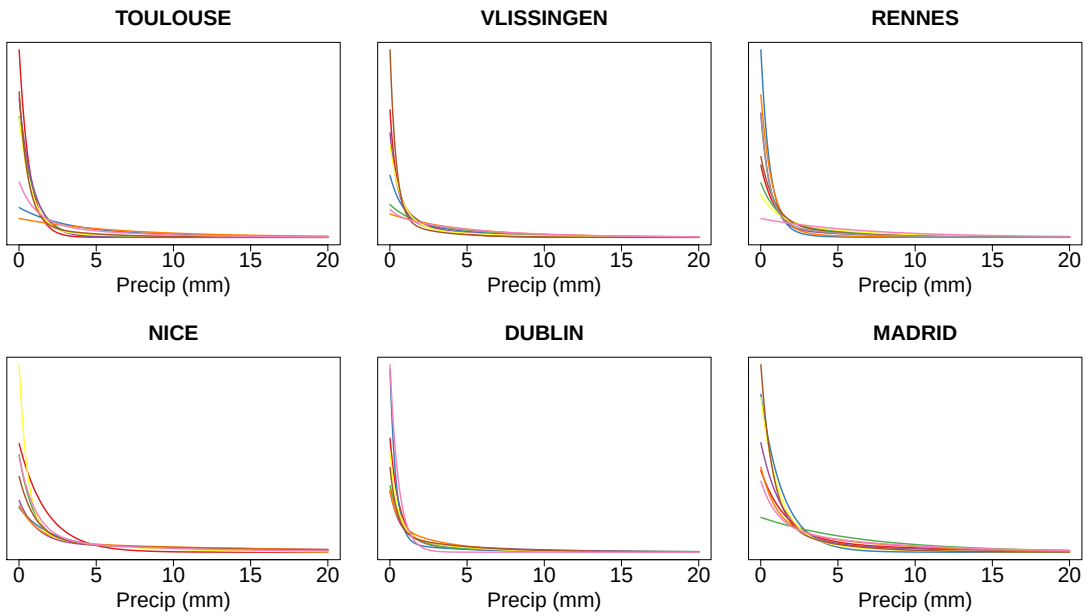

FIGURE 7.27 – Densités d'émission des précipitations pour les jours pluvieux

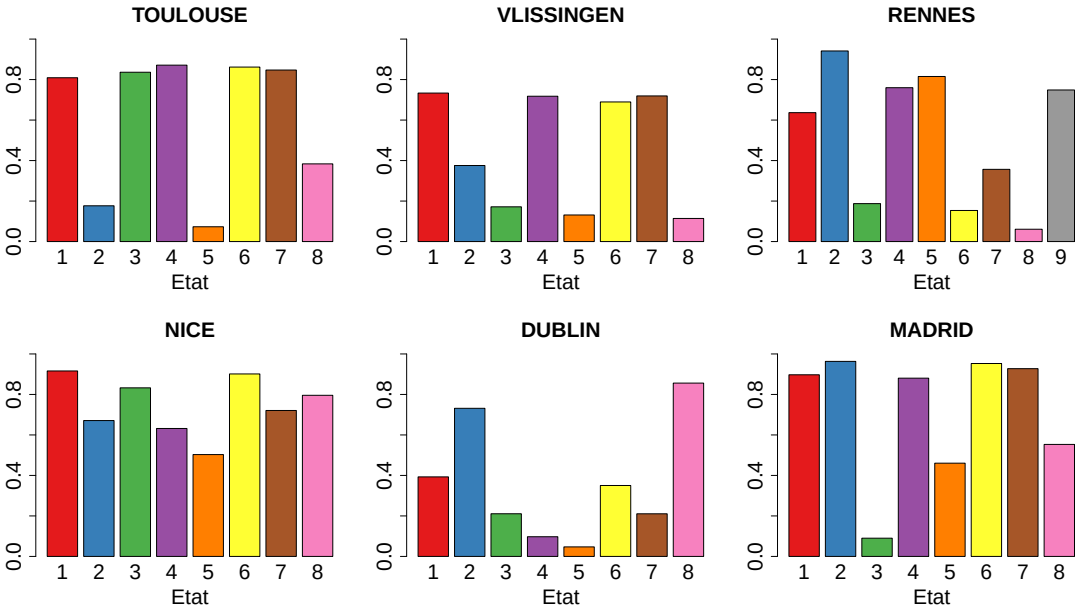

FIGURE 7.28 – Probabilité de jour sec dans chaque état

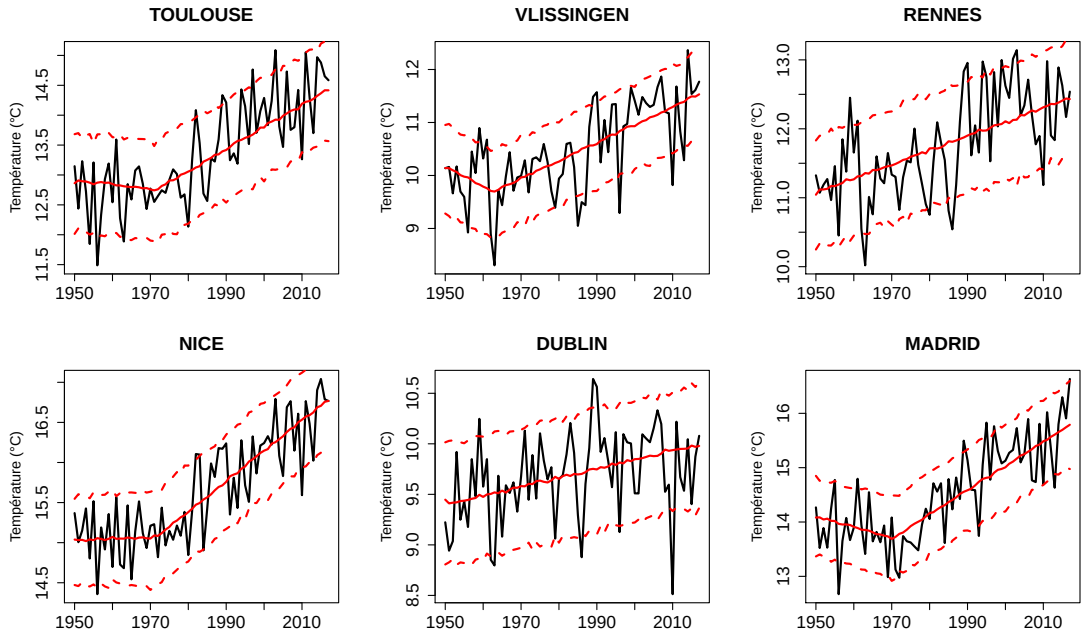

FIGURE 7.29 – Température moyenne annuelle. Les observations sont en noir, la moyenne sur les simulations en trait plein rouge, et l'intervalle entre les deux lignes pointillées contient 95% des simulations.

autres variables. De la même façon, les résultats corrects du modèle univarié pour le vent (voir la Section 7.2.1) ne sont pas nécessairement retrouvés au sein du modèle trivarié. Il est donc également nécessaire de vérifier que la vitesse du vent est correctement reproduite par le modèle trivarié, indépendamment des autres variables. Nous allons donc successivement examiner les variables suivantes :

— Température
— Précipitations
— Vent
— (Température,Précipitations)
— (Température,Vent)
— (Précipitations,Vent)

Température

Comme le montre la Figure 7.29, l'introduction de tendances dans les lois d'émission permet de correctement prendre en compte le changement climatique. D'autre part, l'intervalle de confiance montre que le simulations permettent d'obtenir une variabilité interannuelle du même ordre que celle que l'on observe sur la série historique : de 1 à 2°C autour de la tendance moyenne.

Sur la Figure 7.30, nous avons représenté par leur densité (ou plutôt par leur estimation à noyau) les distributions mensuelles des températures sur l'exemple de Vlissingen. Les résultats sont aussi bons pour les autres stations.

Les résultats sont aussi satisfaisants lorsque l'on considère des statistiques journalières. Nous

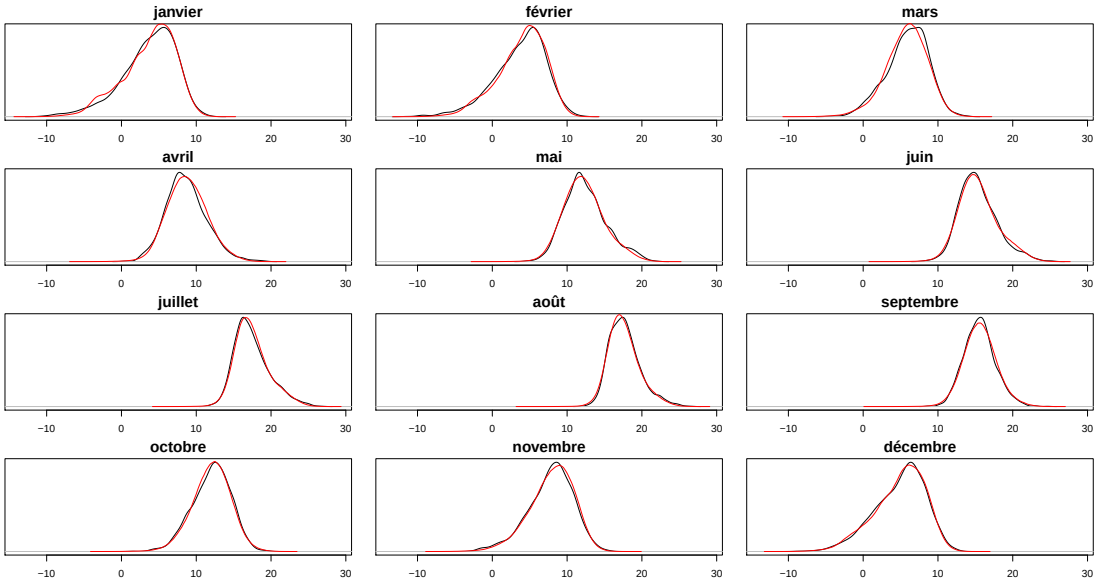

FIGURE 7.30 – Densité de probabilité de la température mois par mois, Vlissingen. En noir : observations, en rouge : simulations.

nous sommes intéressés aux quatre premiers moments, aux minima et aux maxima pour chaque jour calendaire : voir les Figures 7.31-7.32-7.33.

La Figure 7.34 (resp. 7.35) permet de comparer les distributions observée et simulée des maxima (resp. minima) annuels de la température. Les résultats sont corrects, même si l'on peut noter une légère surestimation par le modèle des maxima annuels sur certaines stations.

Comme dans le modèle bivarié, les autocorrélations des températures sont légèrement sous-estimées (Figure 7.36). Cela entraîne une sous-estimation des longueurs des vagues de chaud et de froid, comme nous l'avons déjà constaté dans le Chapitre 5 (voir la Figure 7.37 pour les vagues de chaleur).

Outre le comptage des dépassements de seuil, une autre manière de quantifier l'intensité des vagues de chaleur ou de froid est le calcul des VCXn (resp. VCNn) annuels. Ces mesures issues de l'hydrologie ont été introduite dans le Chapitre 6. Rappelons que le VCXn (resp. VCNn) d'une année pour une variable est le maximum (resp. minimum) sur l'année de la moyenne glissante sur n jours de cette variable. Dans le cas des précipitations, la moyenne est remplacée par le cumul. Les Figures 7.38 et 7.39 présentent respectivement les distributions du VCN15 et du VCX3 de la température. T est le temps de retour en années. On constate que les VCN15 et VCX3 sont relativement bien reproduits par le modèle. On constate toutefois dans la plupart des cas une sur-estimation des VCN15 les plus extrêmes (les plus froids) et une sous-estimation des VCX3 les plus extrêmes (les plus chauds). Cela est cohérent avec le fait déjà observé que le modèle a des difficultés à reproduire des vagues de chaleur ou de froid extrêmes.

Ces résultats permettent d'affirmer que l'ajout du vent au modèle bivarié n'altère pas la qualité des simulations en ce qui concerne la température. Les quelques défauts que nous avons identifiés existaient déjà dans le modèle bivarié. Nous allons vérifier que c'est aussi le cas des précipitations.

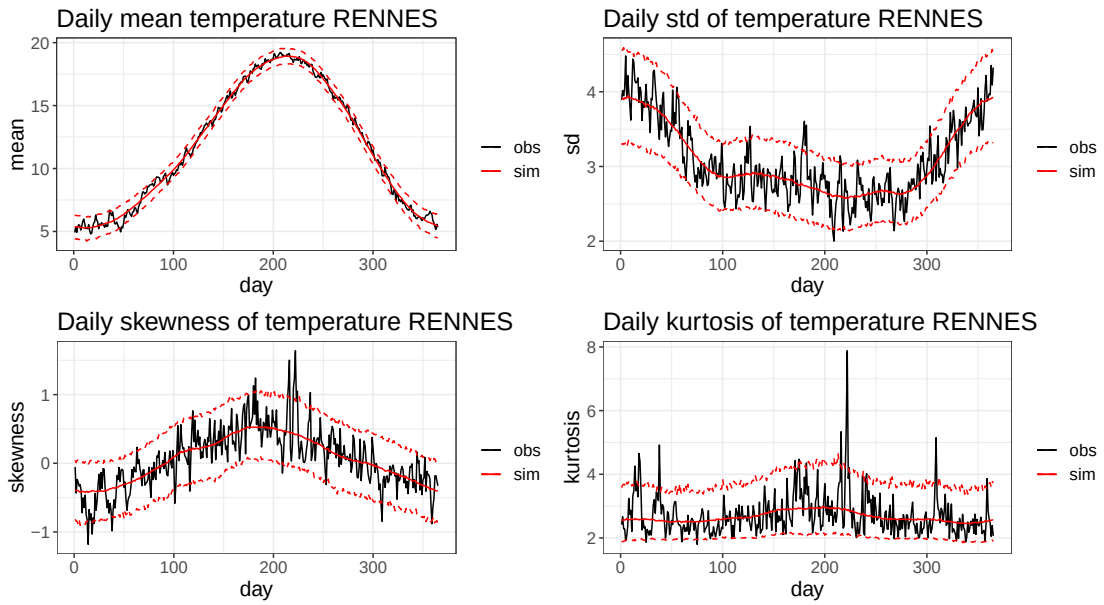

FIGURE 7.31 – Moments journaliers de la température, Rennes.

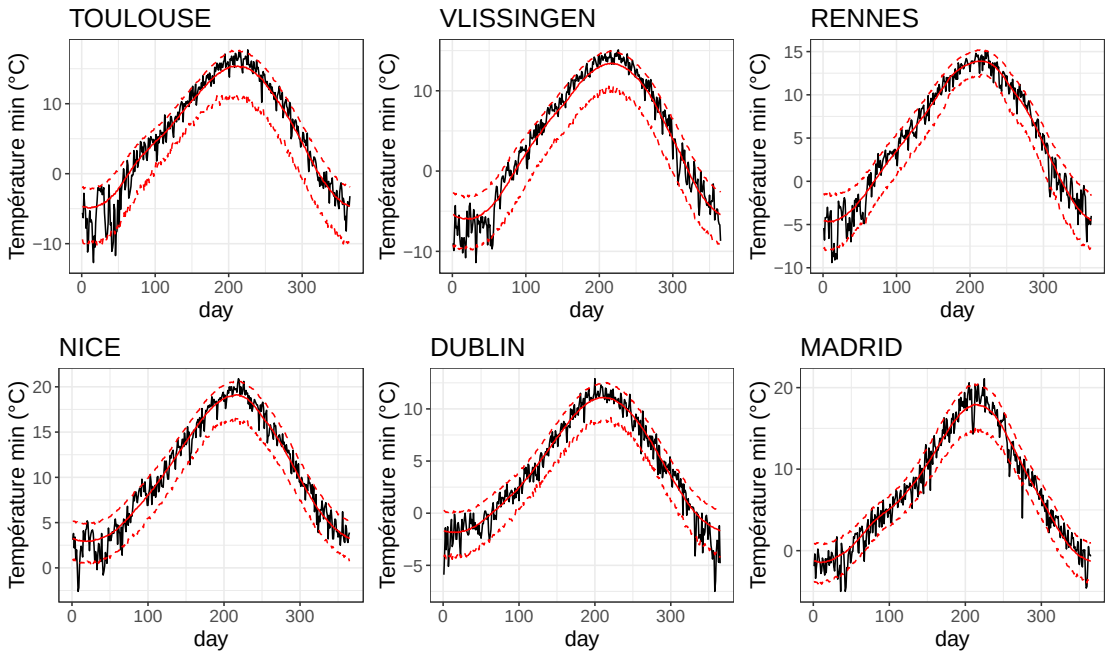

FIGURE 7.32 – Minima journaliers des températures.

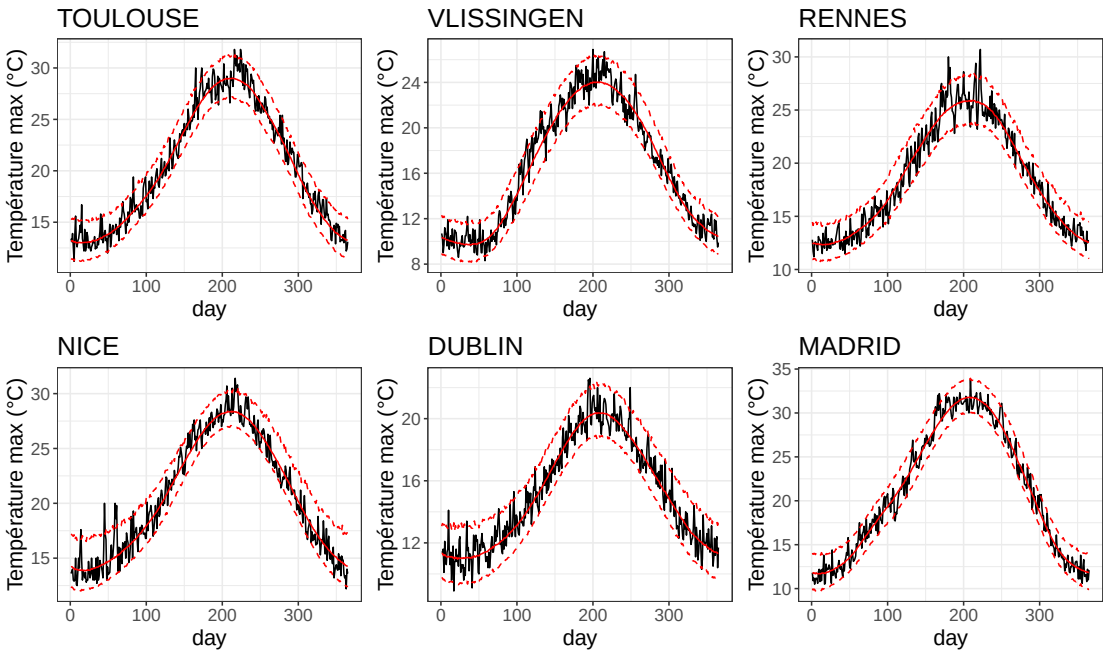

FIGURE 7.33 – Maxima journaliers des températures.

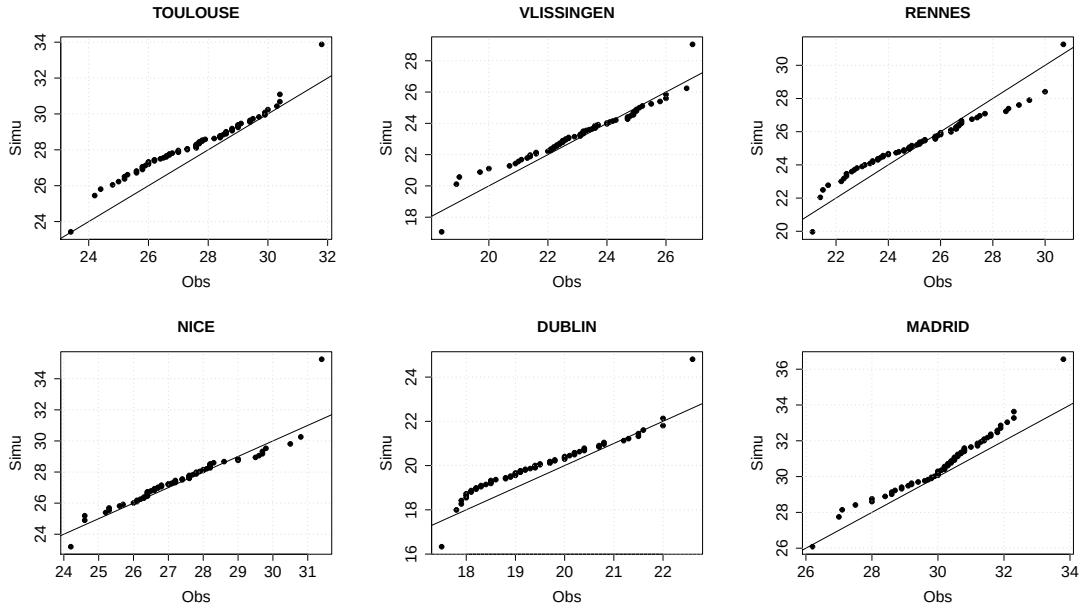

FIGURE 7.34 – QQ-plot des maxima annuels de la température.

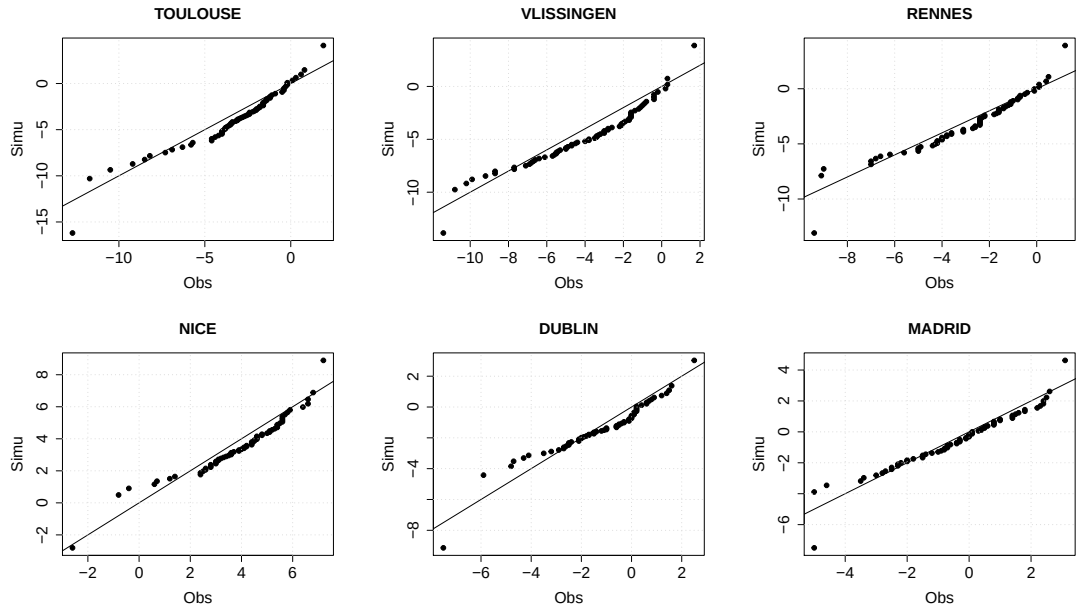

FIGURE 7.35 – QQ-plot des minima annuels de la température.

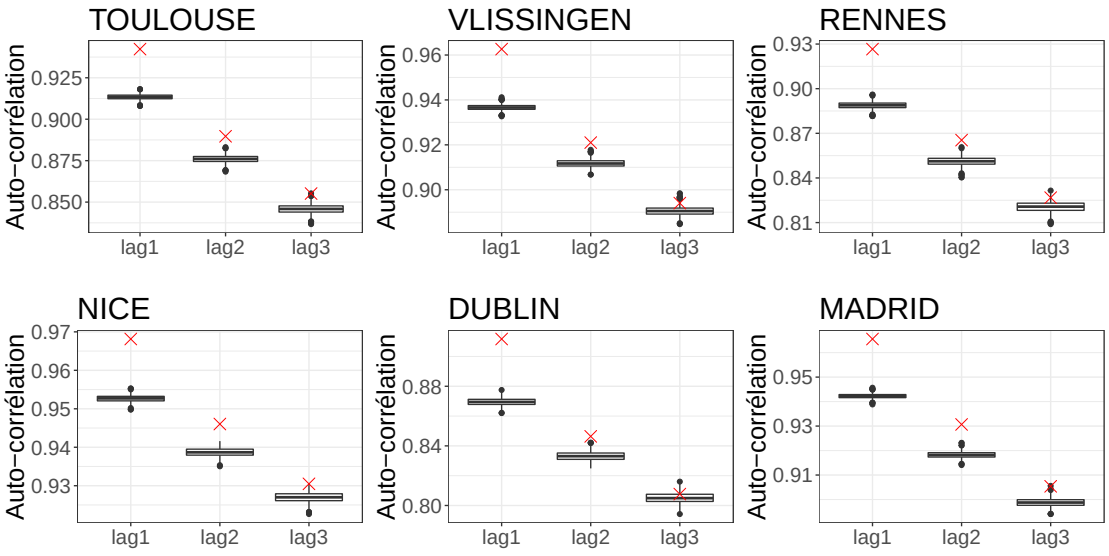

FIGURE 7.36 – Autocorrélations des températures. En rouge : les valeurs calculées sur les observations. En noir : les boxplots des distributions obtenues par les simulations.

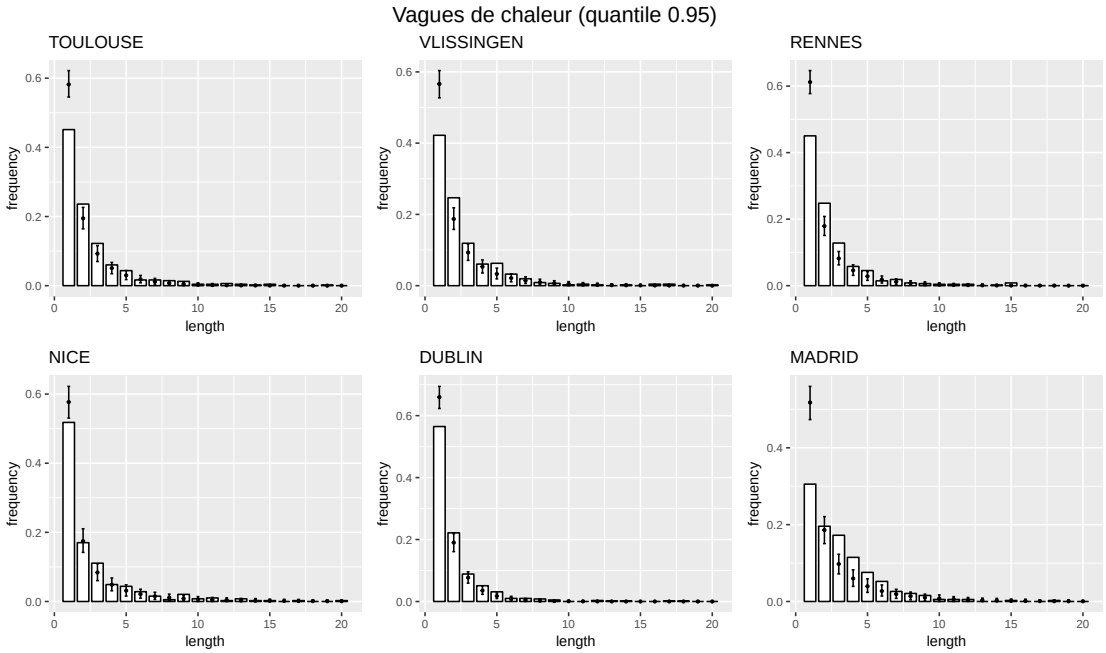

FIGURE 7.37 – Distributions des longueurs des clusters au-delà du 95ème percentile de la température. En blanc : les distributions obtenues sur les observations. En noir : un intervalle de confiance à 95% obtenu sur les simulations.

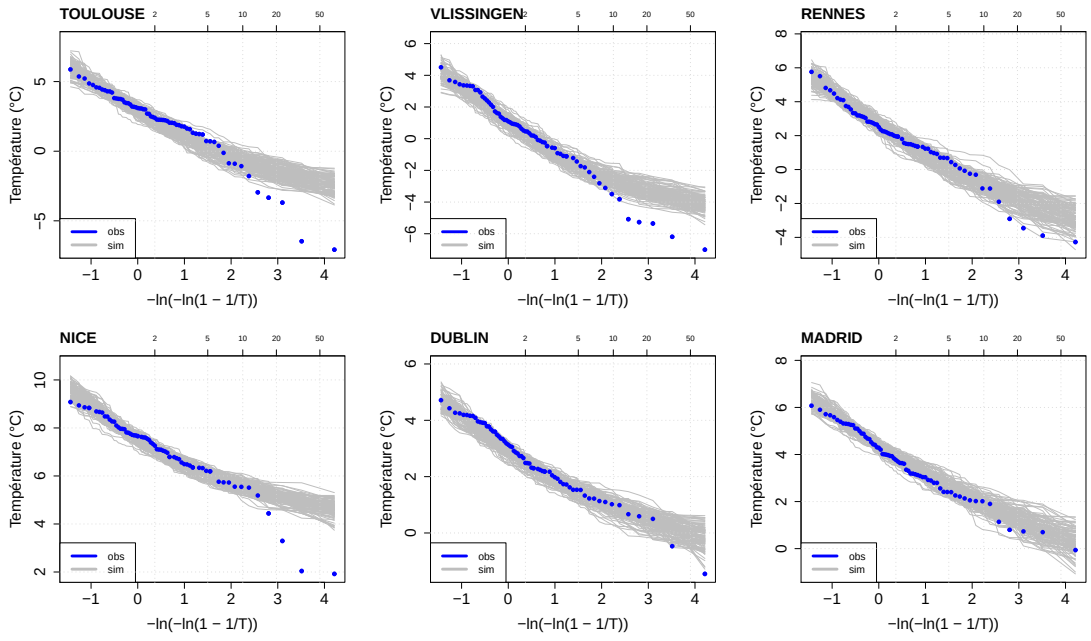

FIGURE 7.38 – Distribution du VCN15 de la température.

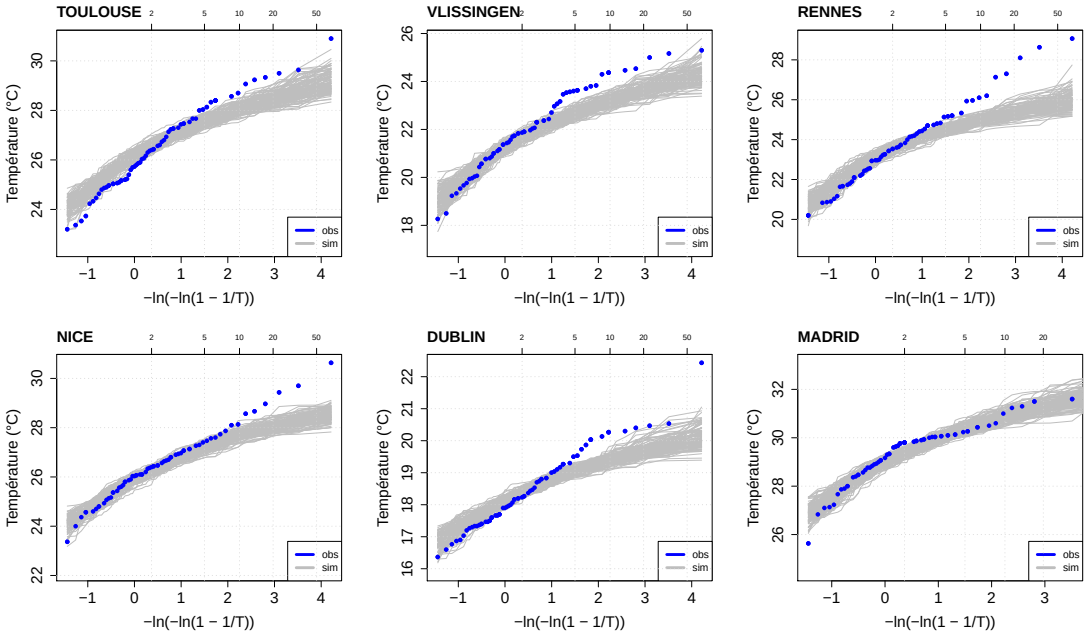

FIGURE 7.39 – Distribution du VCX3 de la température.

Précipitations

Pour chaque station, la distribution marginale des précipitations est correctement reproduite par le modèle (Figure 7.40), ainsi que le cumul annuel moyen (Figure 7.41).

La Figure 7.42 représente les quatre premiers moments journaliers, ainsi que les occurrences et maxima journaliers des précipitations, pour la station de Toulouse. On obtient des résultats similaires sur les autre stations.

Les distributions des longueurs des séquences sèches et pluvieuses sont représentées respectivement sur les Figures 7.43 et 7.44. Le modèle a tendance à sous-estimer légèrement la longueur des séquences sèches, et ce sur la plupart des stations. Le même problème est observé, de façon plus marqué, sur les séquences pluvieuses. Nous avons déjà observé ce phénomène sur le modèle bivarié.

Ce défaut de dépendance temporelle peut également s'observer à travers les VCN30 (minimum annuel du cumul des précipitations sur 30 jours) et VCX3 (maximum annuel du cumul des précipitations sur 3 jours) qui quantifient respectivement les périodes de sécheresse et les épisodes de fortes précipitations. Les distributions des VCN30 et VCX3 (observés et simulés) sont respectivement représentées sur les Figures 7.45 et 7.46. On constate que les VCN30 sont légèrement sur-estimés par le modèle, tandis que les VCX3 sont légèrement sous-estimés. Notons par exemple que la plupart des VCN30 observés sur les stations de Nice et Madrid sont nuls, ce qui signifie que presque tous les ans (surtout à Madrid), ces stations connaissent une période sans pluie d'au moins 30 jours. Le modèle ne parvient pas à reproduire cette situation : sur une période de 30 jours, il produit toujours quelques millimètres de pluie.

La distribution des maxima annuels des précipitations est en revanche bien reproduite par le modèle, comme le montre la Figure 7.47. Notons que pour chaque station, un point isolé se

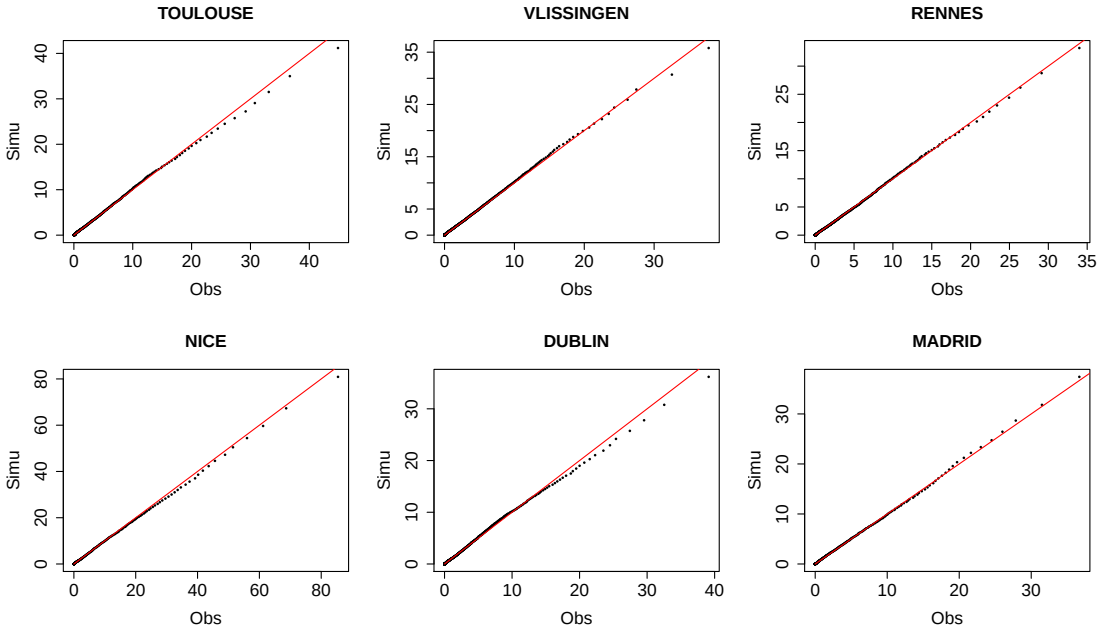

FIGURE 7.40 – Graphe quantile-quantile pour la distribution des précipitations.

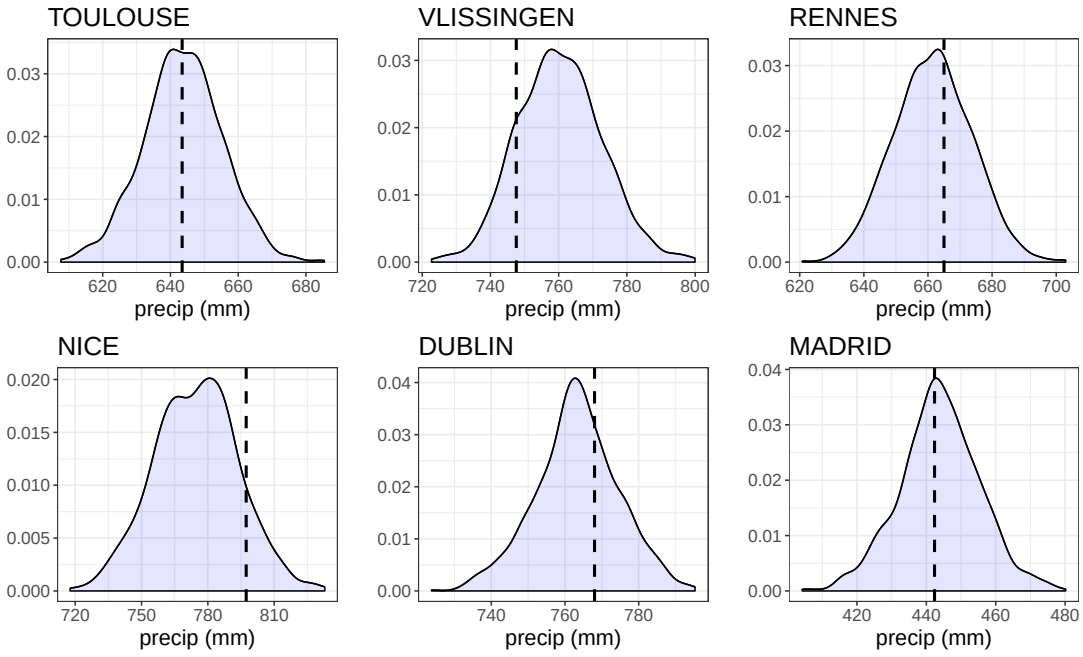

FIGURE 7.41 – Cumul annuel moyen des précipitations

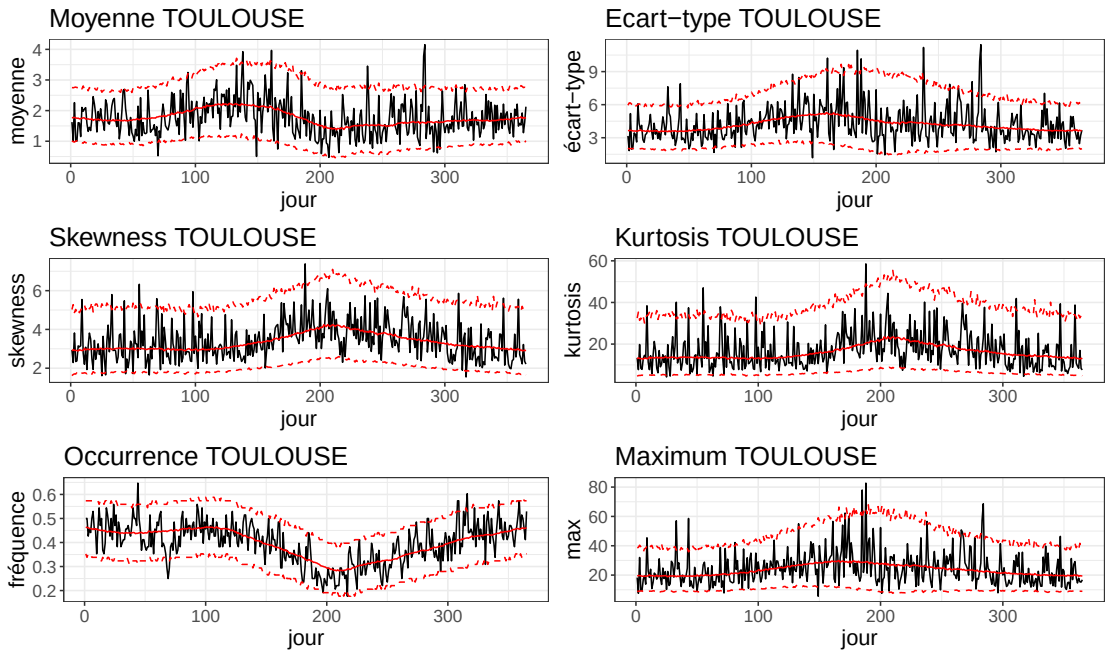

FIGURE 7.42 – Moments et maxima des précipitations journalières à Toulouse

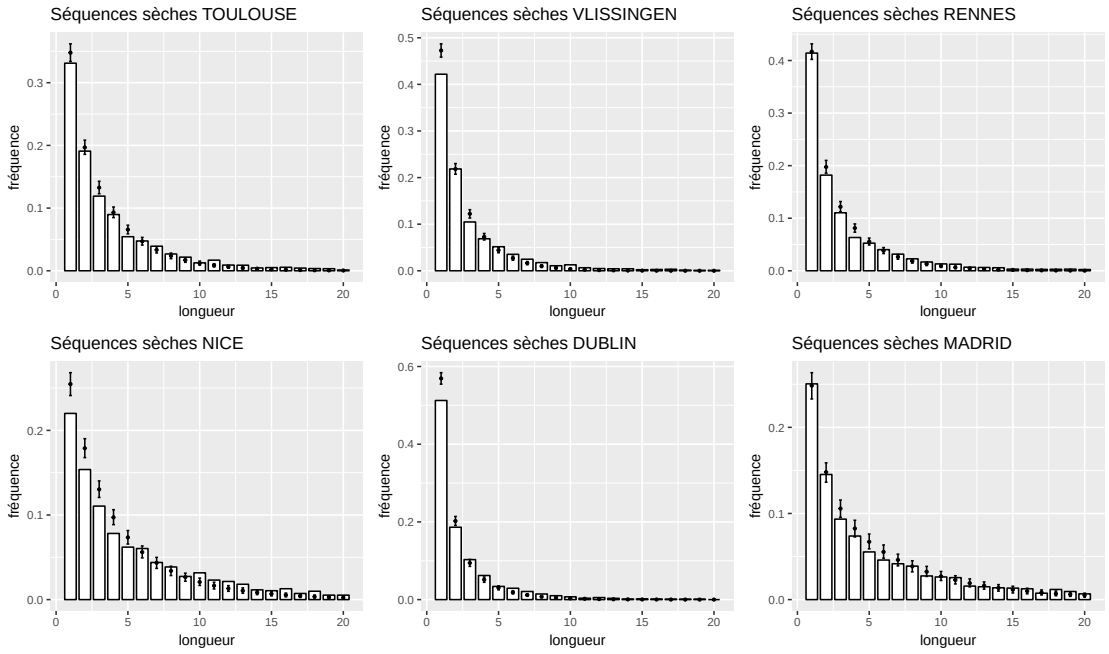

FIGURE 7.43 – Séquences sèches.

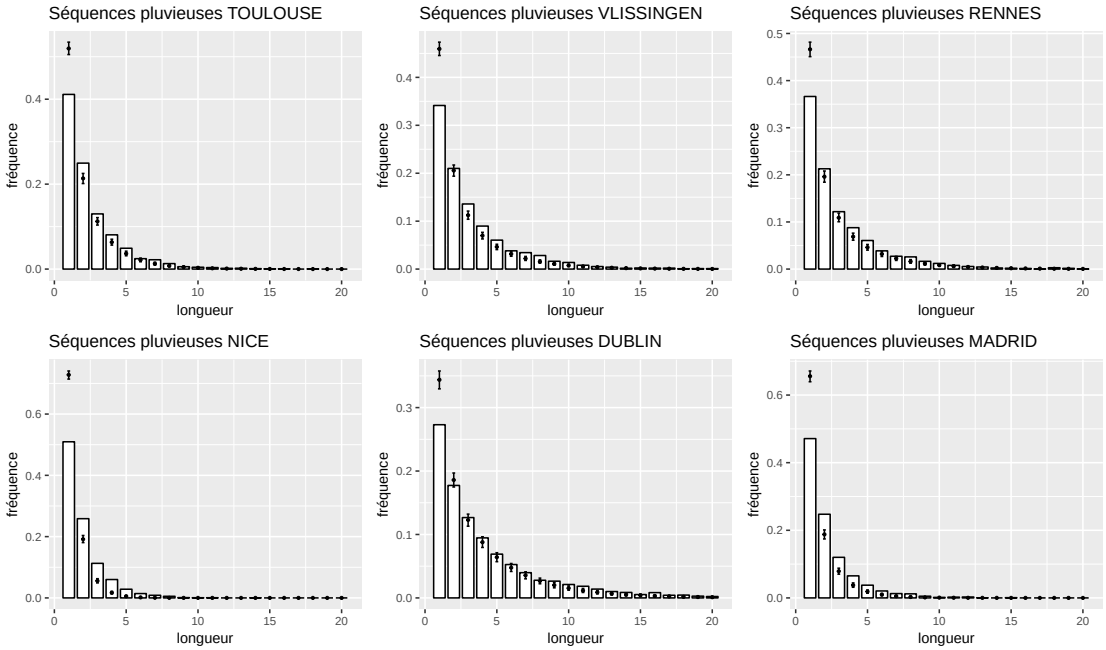

FIGURE 7.44 – Séquences pluvieuses.

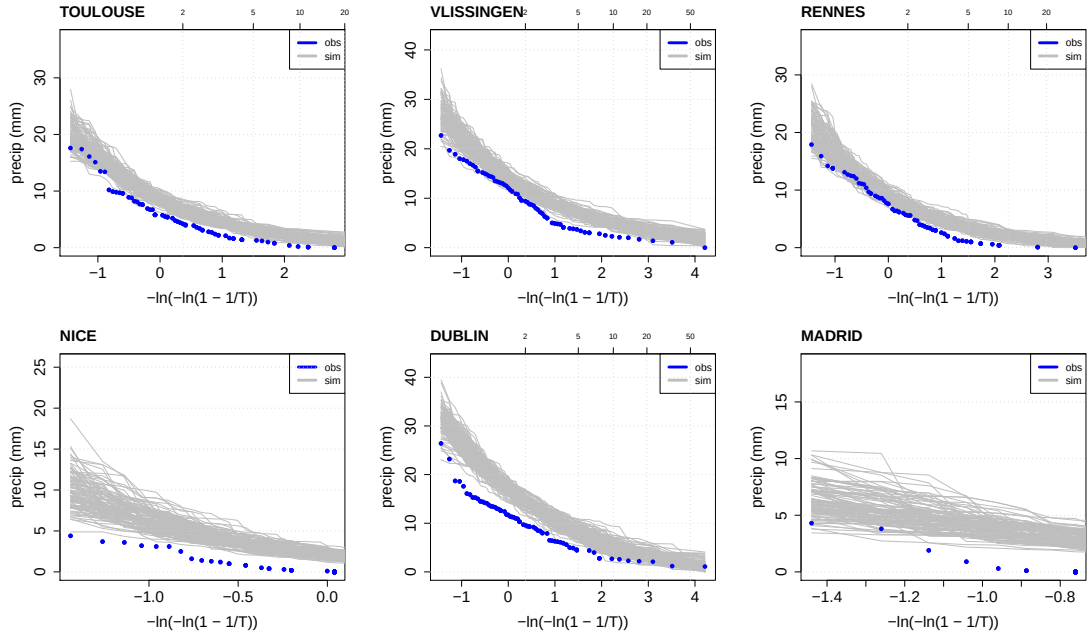

FIGURE 7.45 – VCN30 des précipitations.

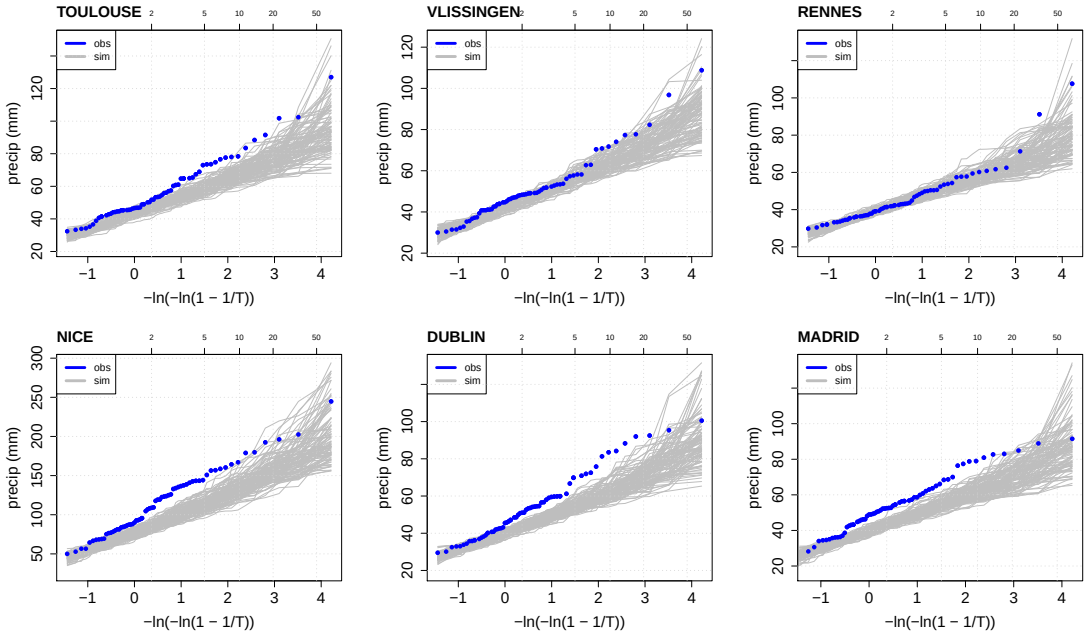

FIGURE 7.46 – VCX3 des précipitations.

trouve largement au-dessus de la première bissectrice. Cela signifie que le modèle a généré une valeur qui dépasse la valeur maximale observée. Par exemple, la valeur maximale observée à Vlissingen est 81mm, tandis que le modèle a généré pour cette station une valeur de 127mm (rappelons que 1000 trajectoires de longueur $n = 24820$ ont été générées). Notre simulateur n'est donc pas borné par les valeurs observées. C'est un avantage de ce type de simulateur par rapport à un simulateur basé sur une méthode de rééchantillonnage.

Enfin, la Figure 7.48 montre que la variabilité interannuelle du cumul annuel des précipitations est bien reproduite par le modèle.

On peut tirer la même conclusion que pour les températures : la modélisation des précipitations avec le modèle trivarié possède les mêmes qualités et défauts qu'avec le modèle bivarié. L'ajout du vent dans le modèle n'a donc pas perturbé la modélisation des précipitations. Dans la Section 7.2, nous avons présenté et testé deux modèles pour la modélisation univariée du vent. Nous allons maintenant voir si la modélisation du vent via le modèle trivarié est d'aussi bonne qualité que celle issue d'un modèle univarié.

Vent

La densité de probabilité de la vitesse du vent est bien reproduite pour chaque mois, comme le montre la Figure 7.49 pour la station de Dublin.

Les moments et les maxima journaliers de la vitesse du vent sont généralement bien reproduits par le modèle (voir la Figure 7.50). On observe néanmoins, pour certaines stations, un défaut de saisonnalité, comme à Toulouse pour la variance (cf la Figure 7.51). Ce problème n'était pas présent avec le modèle univarié pour le vent (voir la Figure 7.8).

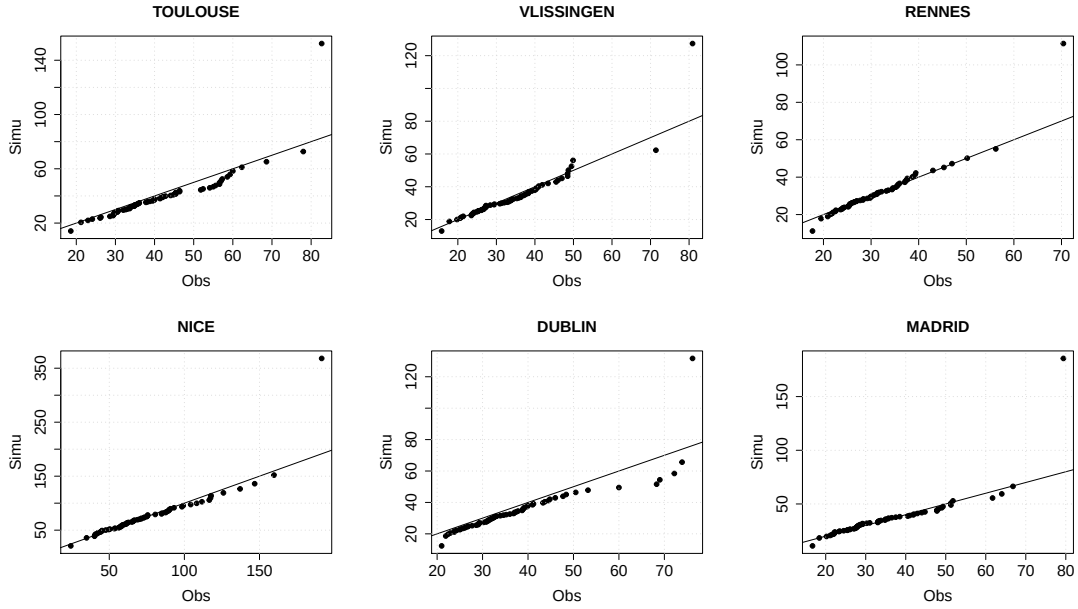

FIGURE 7.47 – QQ-plot des maxima annuels des précipitations.

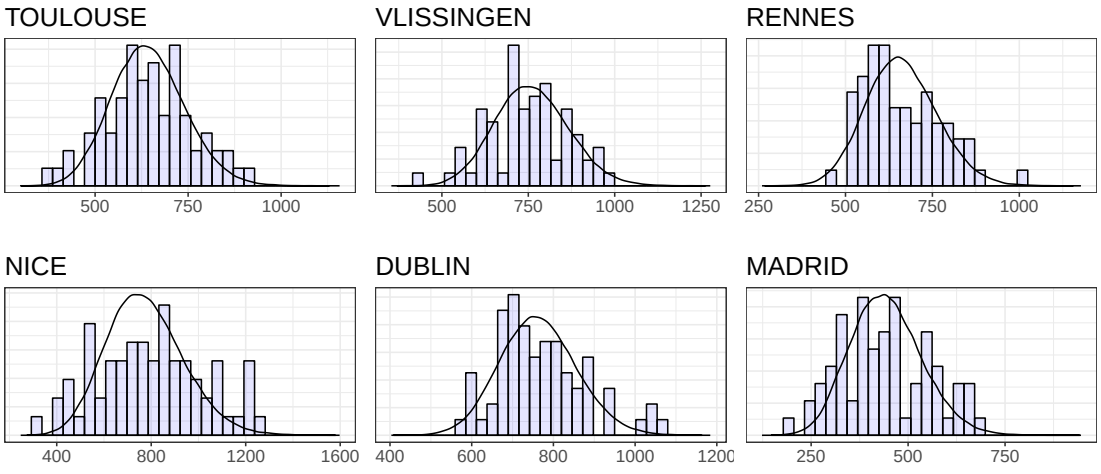

FIGURE 7.48 – Variabilité interannuelle des cumuls annuels des précipitations

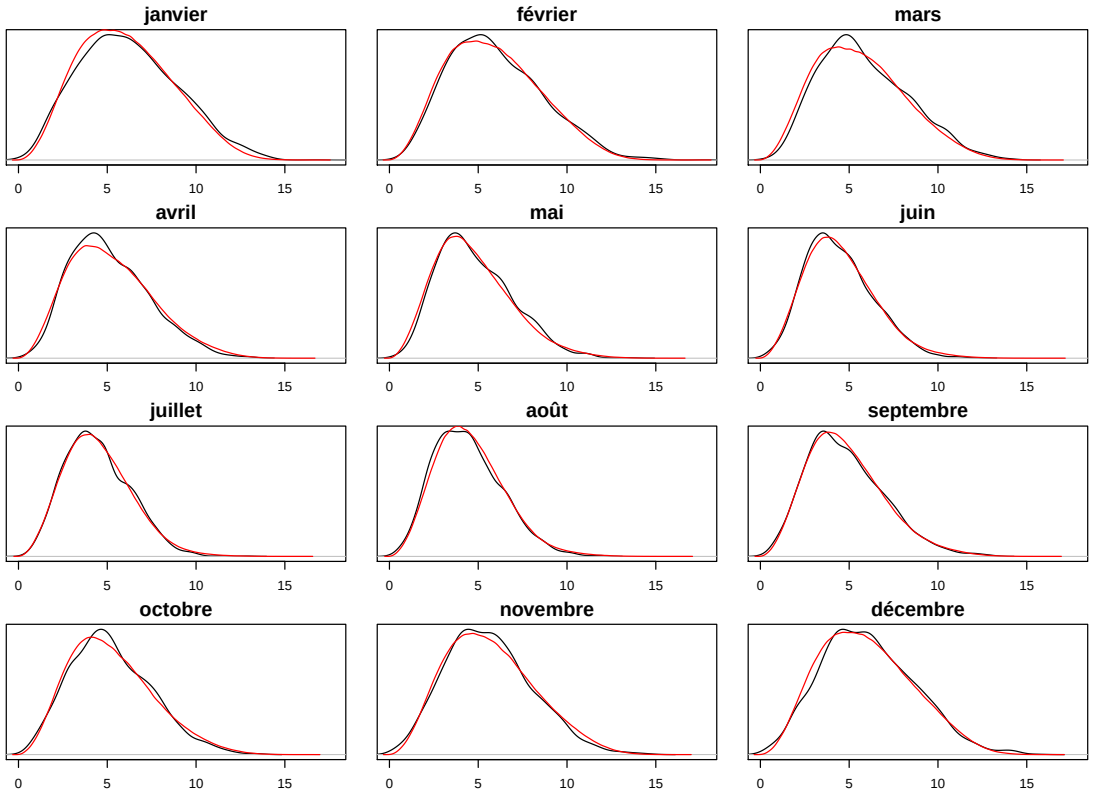

FIGURE 7.49 – Densité de probabilité de la vitesse du vent mois par mois, Dublin. En noir : observations, en rouge : simulations.

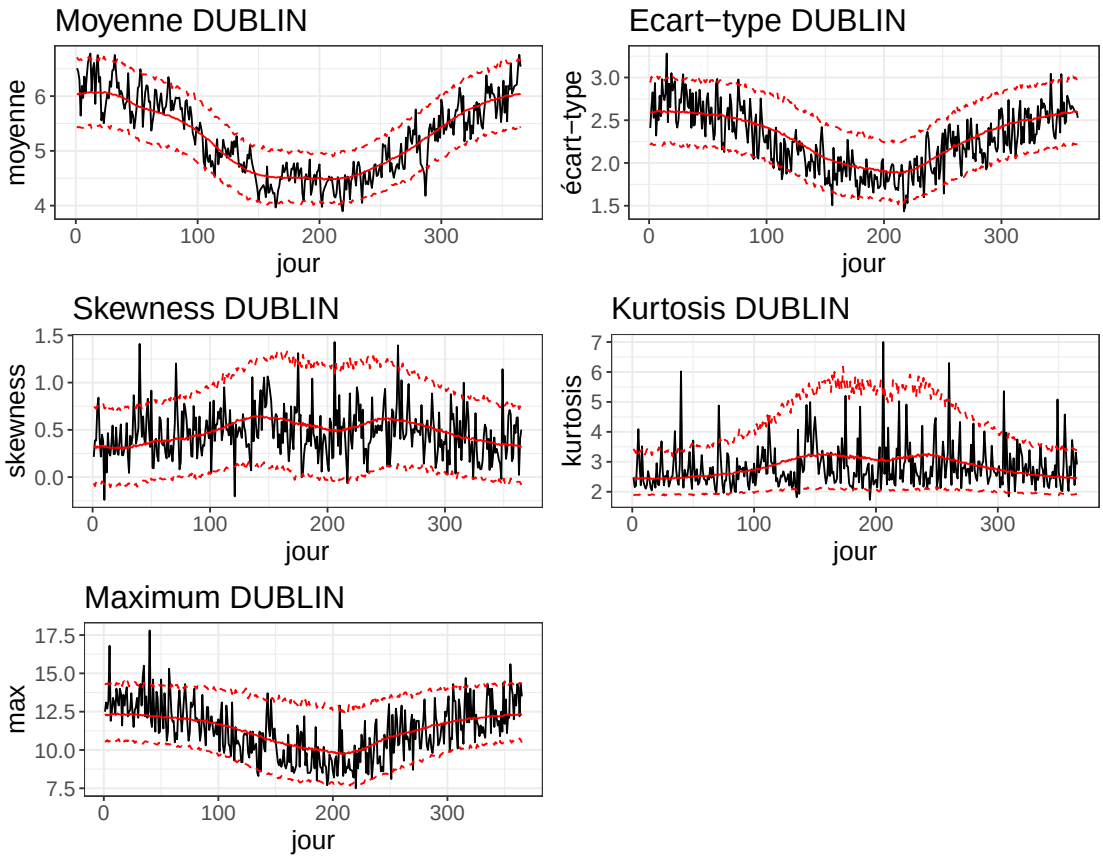

FIGURE 7.50 – Moments et maxima journaliers de la vitesse du vent, Dublin.

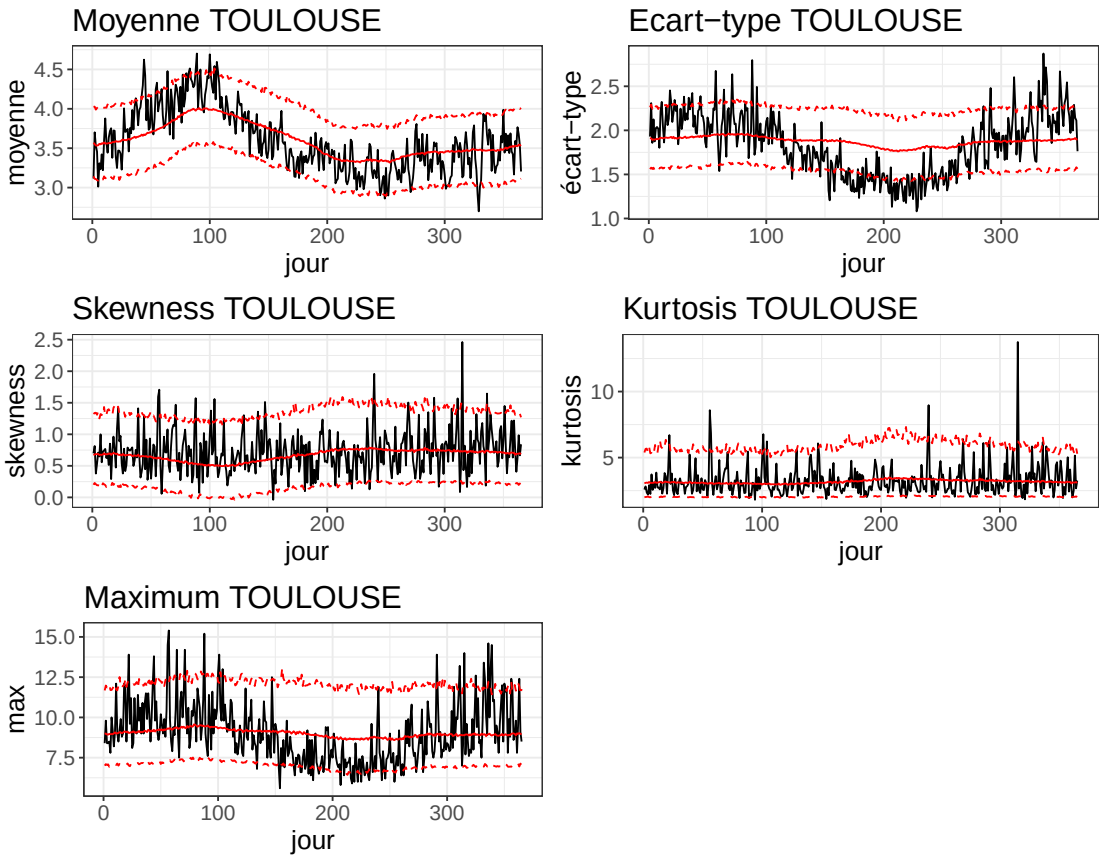

FIGURE 7.51 – Moments et maxima journaliers de la vitesse du vent, Toulouse

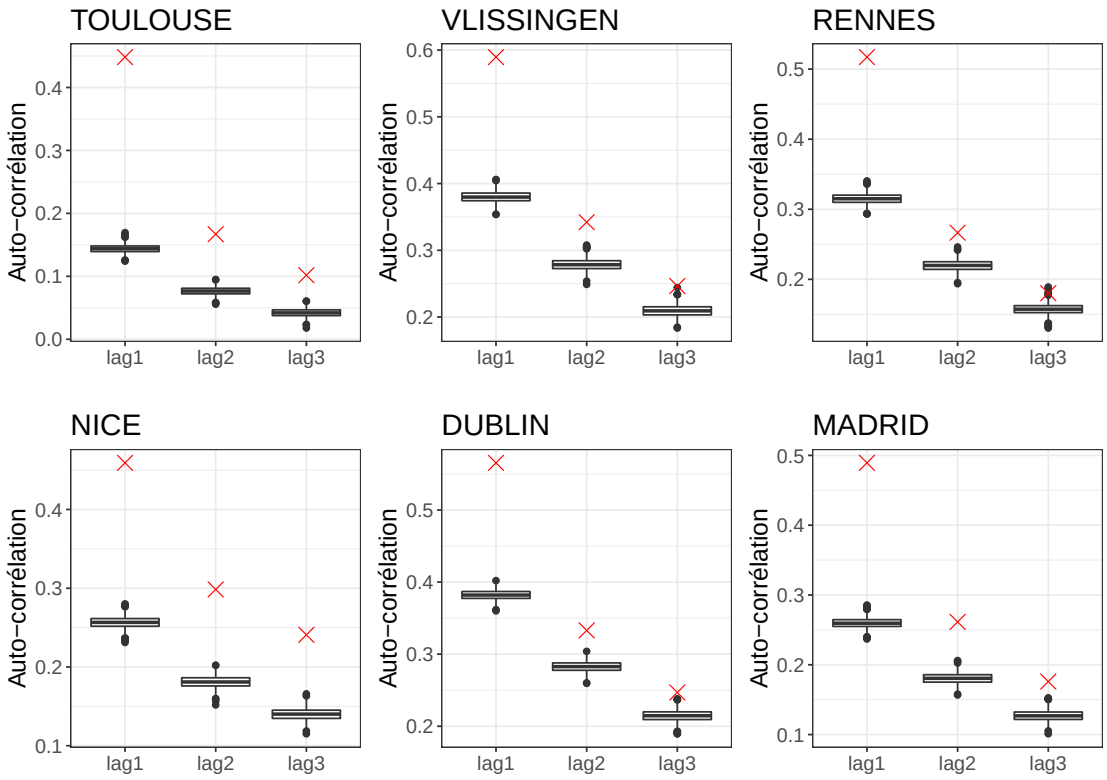

FIGURE 7.52 – Auto-corrélations de la vitesse du vent. En rouge : les valeurs calculées sur les observations. En noir : les boxplots des distributions obtenues par les simulations.

La Figure 7.52 représente les auto-corrélations de la vitesse du vent. On constate qu'elle sont largement sous-estimées par le modèle trivarié, contrairement au modèle univarié, qui fournissait des auto-corrélations qui n'étaient que légèrement sous-estimées.

Une conséquence de ce défaut d'auto-corrélation est la sous-estimation du VCX3 annuel (représentatif des tempêtes), comme le montre la Figure 7.53. Les extrêmes ponctuels (maxima annuels) sont en revanche correctement représentés (voir la Figure 7.54).

La sous-estimation de la variabilité interannuelle de la vitesse moyenne annuelle du vent, problème déjà identifié dans la Section 7.2 persiste dans ce modèle (Figure 7.55.

Après avoir examiné les résultats de validation variable par variable, nous allons allons examiner les distributions jointes des variables deux à deux, en commançant par la température et les précipitations.

Température/Précipitations

Sur la Figure 7.56, nous avons représenté, pour chacune des stations, les corrélations mensuelles entre la température et les précipitations. Le modèle parvient à reproduire correctement le comportement de ces corrélations. On constate qu'elles présentent une nette saisonnalité : elles sont plus élevées (et généralement positives) en hiver qu'en été, période pendant laquelle elles sont

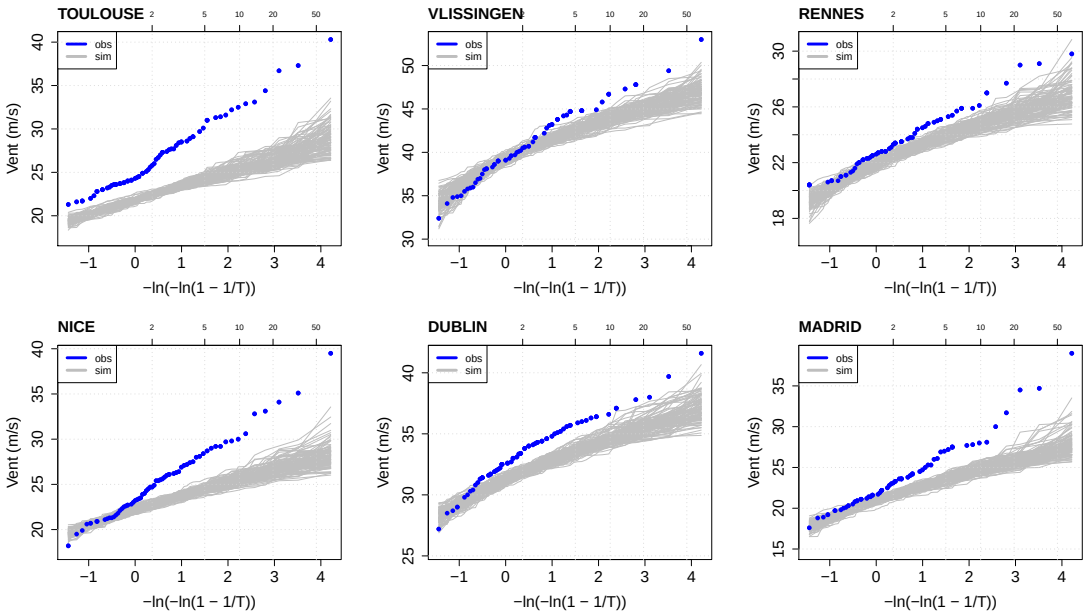

FIGURE 7.53 – VCX3 de la vitesse du vent.

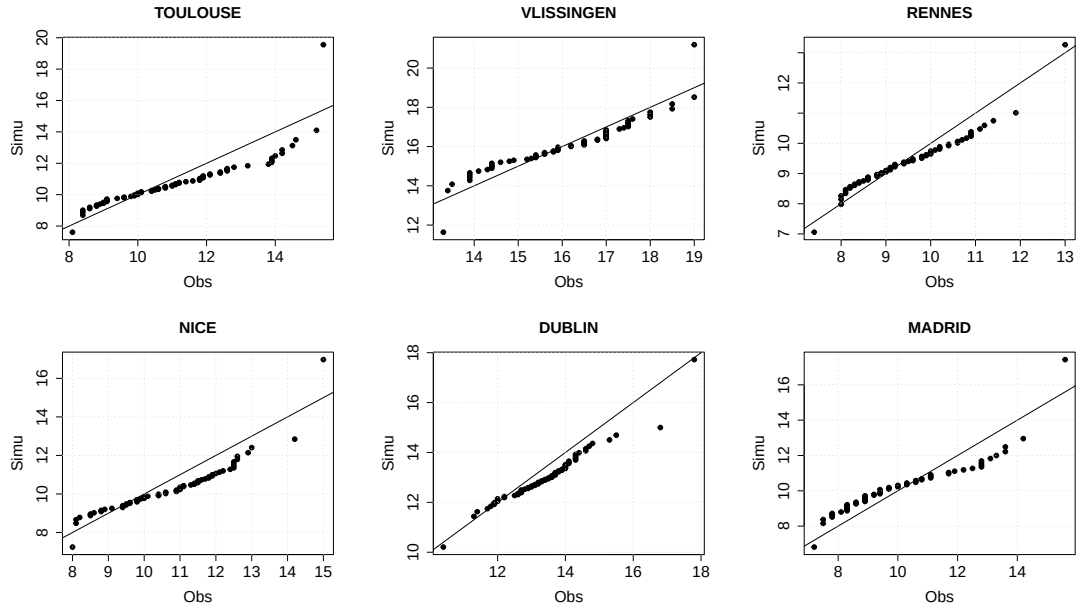

FIGURE 7.54 – QQ-plot des maxima annuels de la vitesse du vent.

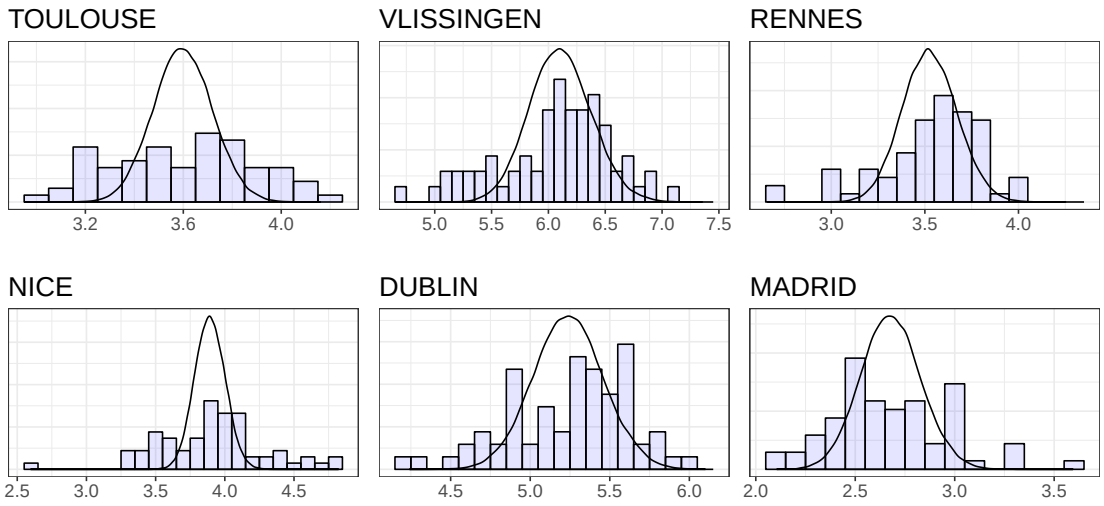

FIGURE 7.55 – Variabilité interannuelle de la vitesse moyenne annuelle du vent. L'histogramme représente les observations et la densité de probabilité a été obtenue à partir des simulations.

négatives. Cela peut s'expliquer simplement : lorsqu'il pleut en été, les températures sont en général plus faibles que les températures moyennes estivales, tandis que le phénomène inverse de produit en hiver : les périodes de froid intense sont souvent sèches, alors que les précipitations correspondent à des températures plus douces. On observe aussi sur certaines stations un second maximum local (de plus faible amplitude) de la corrélation. Celui-ci est situé entre juin et août et pourrait donc être induit par des précipitations orageuses liées à de fortes températures. Ceci est particulièrement visible sur les stations les plus méridionales (Toulouse, Nice, Madrid). Remarquons enfin qu'à Madrid et Nice, les corrélations sont négatives presque toute l'année.

La Figure 7.57 représente la probabilité d'occurrence de précipitations conditionnellement à la température. Ces graphes (et ceux qui suivent) ont été obtenus par la même méthode que dans le Chapitre 5. Précisons qu'ici nous ne tenons pas compte de la saisonnalité. Là aussi, le modèle reproduit bien le comportement observé. Pour la plupart des stations, la probabilité d'occurrence de pluie est maximale pour des températures *moyennes*. Les extrêmes de chaud ou de froid sont liés à une faible probabilité d'observer des précipitations. Notons cependant les exceptions de Nice et Madrid, stations pour lesquelles les précipitations sont plus probables lorsque les températures sont les plus faibles.

La Figure 7.58 présente une estimation de l'espérance de l'intensité des précipitations, conditionnellement au fait qu'il pleuve, et conditionnellement à la température. Cette statistique est bien reproduite par le modèle. Celui-ci permet donc, au moins du point de vue des statistiques simples que nous venons d'évoquer, un bon couplage entre la température et les précipitations. Nous allons également étudier le lien entre la température et la vitesse du vent.

Température/Vent

Les corrélations mensuelles entre la vitesse du vent et la température sont représentées sur la Figure 7.59. Même si les résultats ne sont pas parfaits, cette statistique est correctement reproduite par le modèle. Pour la plupart des stations, la forme de la saisonnalité de la corrélation

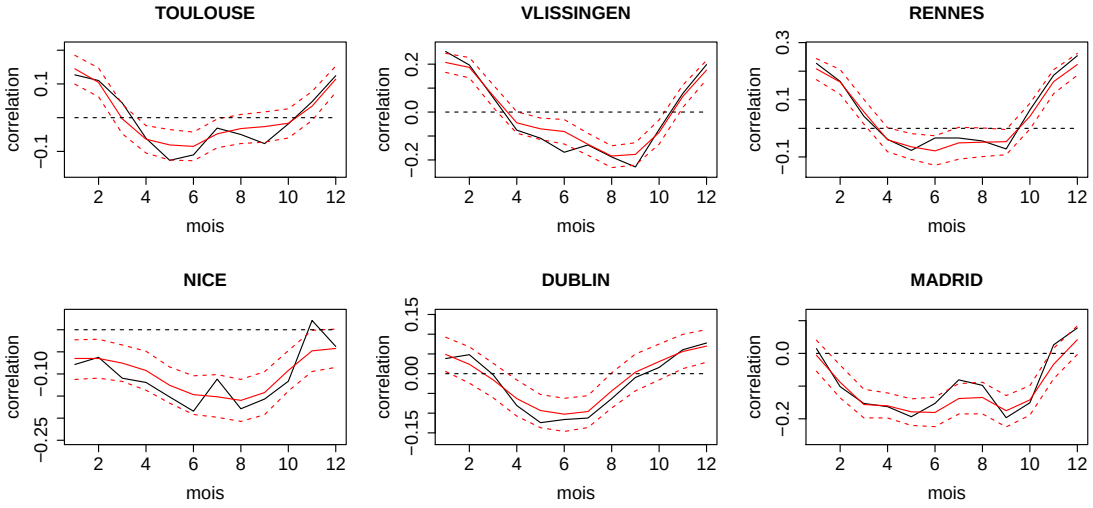

FIGURE 7.56 – Corrélations mensuelles entre la température et les précipitations. En noir, les observations et en rouge les simulations.

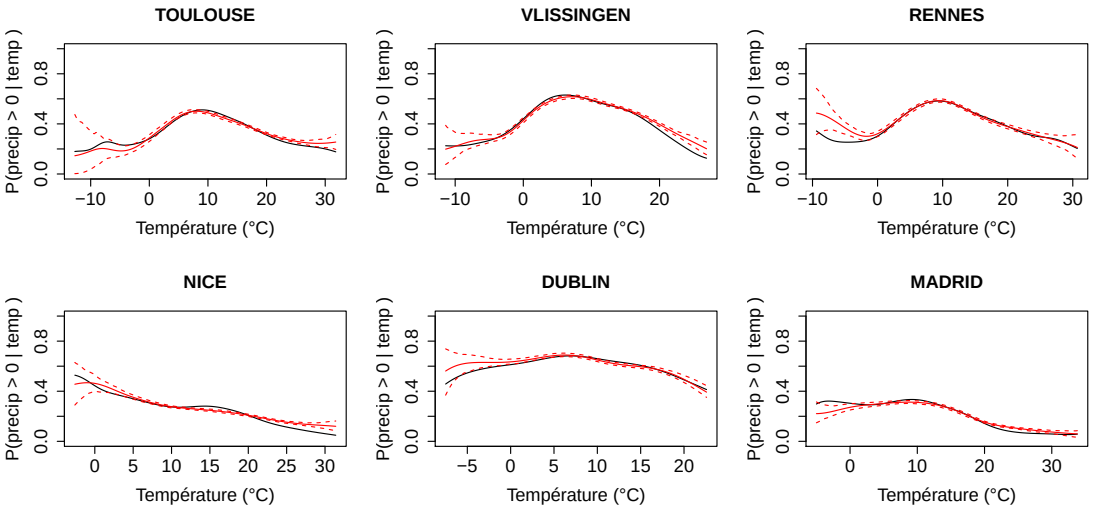

FIGURE 7.57 – Probabilité d'occurrence des précipitations conditionnellement à la température. En noir, les observations, en rouge les simulations.

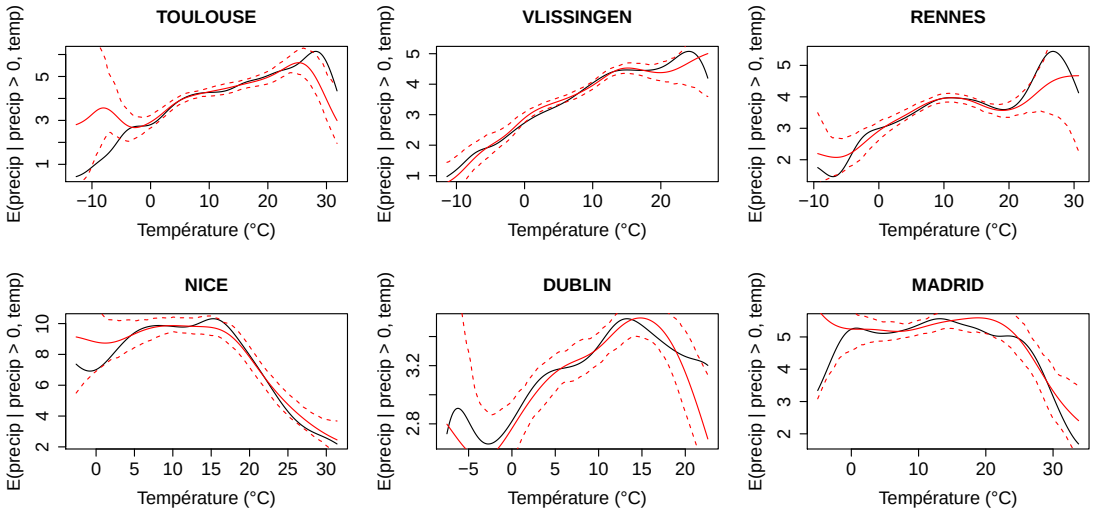

FIGURE 7.58 – Espérance des précipitations conditionnellement à la température pour les jours de pluie.

entre les températures et la vitesse du vent est semblable à celle observée entre la température et les précipitations. En hiver les corrélations sont positives : un vent fort est associé à des températures élevées et inversement en été : les corrélations sont négatives. Ceci est valable à l'exception notable de Nice, où la corrélation oscille autour de 0 toute l'année.

De manière similaire à la Figure 7.58, nous avons représenté sur la Figure 7.60 l'espérance de la vitesse du vent conditionnellement à la température. Les résultats sont moins bons que pour le couple température/précipitations. En particulier, pour les stations de Vlissingen et Rennes, la vitesse moyenne du vent est sous-estimée par le modèle dans les queues de distribution de la température, et le même constat s'impose pour les températures les plus chaudes à Madrid.

Précipitations/Vent

Les corrélations mensuelles entre la vitesse du vent et les précipitations sont représentées sur la Figure 7.61. Ces corrélations présentent elles aussi une saisonnalité, dont la forme dépend de la station. Notons qu'elles sont toujours positives. La qualité du modèle de ce point de vue n'est pas entièrement satisfaisante.

En revanche, la probabilité d'occurrence de précipitations conditionnellement à la vitesse du vent est plutôt bien reproduite par le modèle, comme le montre la Figure 7.62.

Enfin, la Figure 7.63 présente les estimations de l'espérance de la quantité de précipitations, sachant qu'elle est strictement positive, et sachant la vitesse du vent.

7.3.3 Conclusion

En utilisant des simulations issues de notre modèle, nous l'avons évalué selon une série de critères concernant les trois variables à modéliser. On peut en conclure que la modélisation conjointe de la température et des précipitations est satisfaisante, au sens où selon les différents critères considérés, les simulations sont statistiquement proches des observations. Ainsi l'intégration du vent

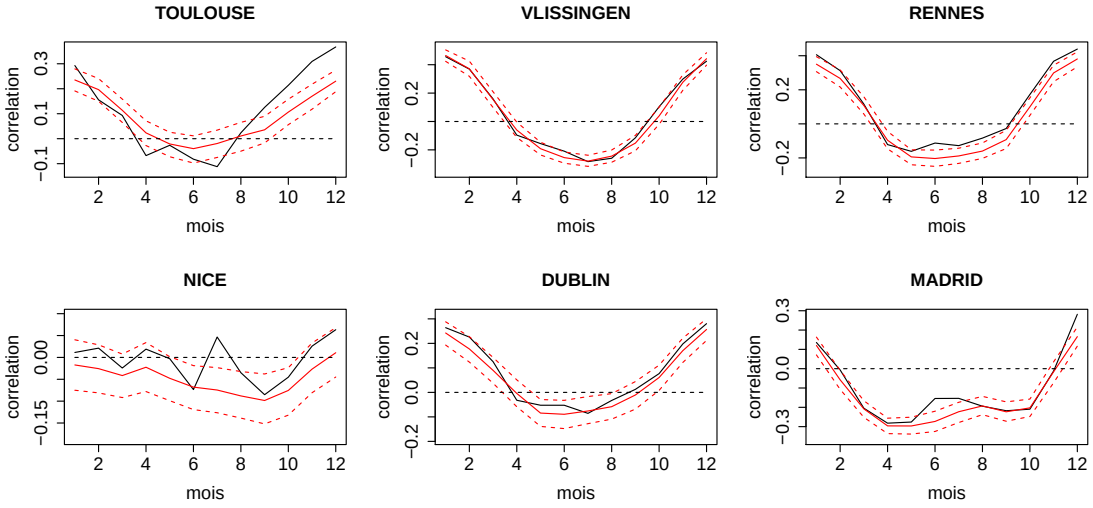

FIGURE 7.59 – Corrélations mensuelles entre la température et la vitesse du vent.

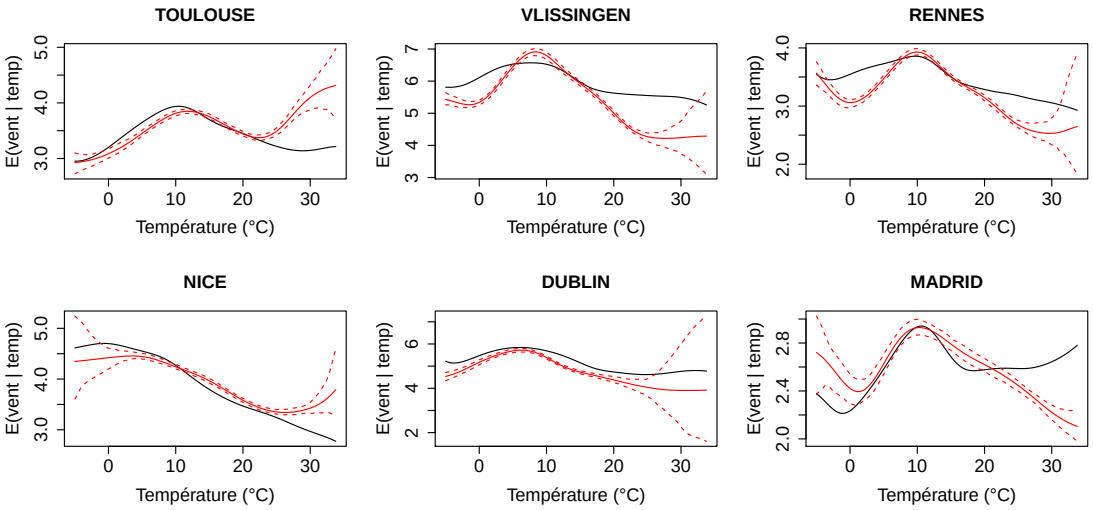

FIGURE 7.60 – Espérance de la vitesse du vent conditionnellement à la température.

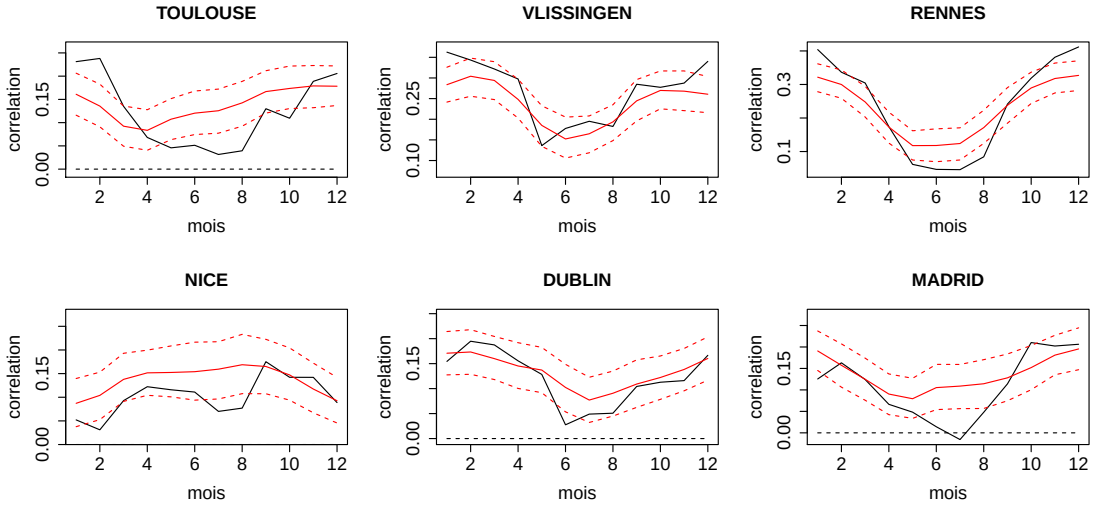

FIGURE 7.61 – Corrélations mensuelles entre les précipitations et la vitesse du vent

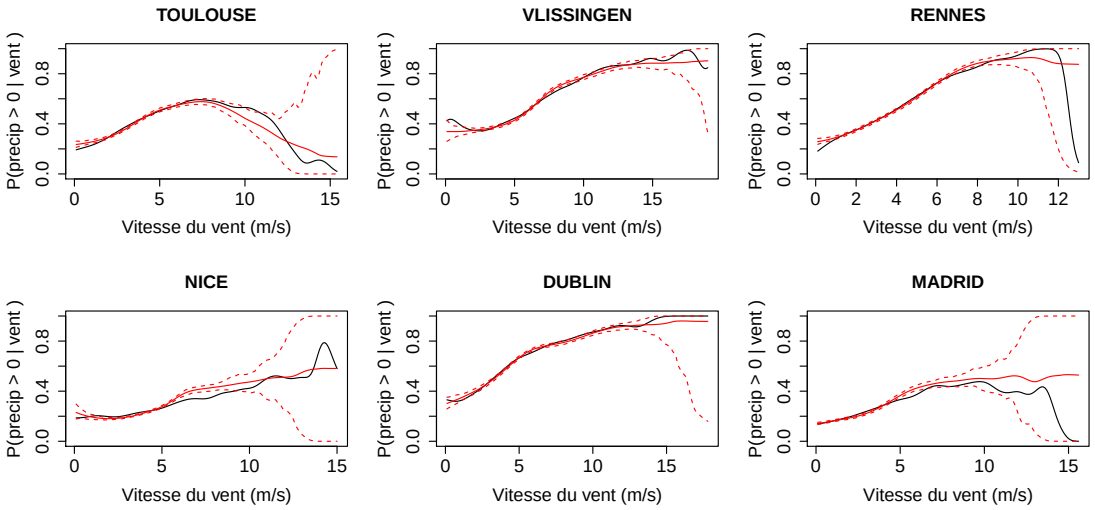

FIGURE 7.62 – Probabilité d'occurrence des précipitations conditionnellement à la vitesse du vent

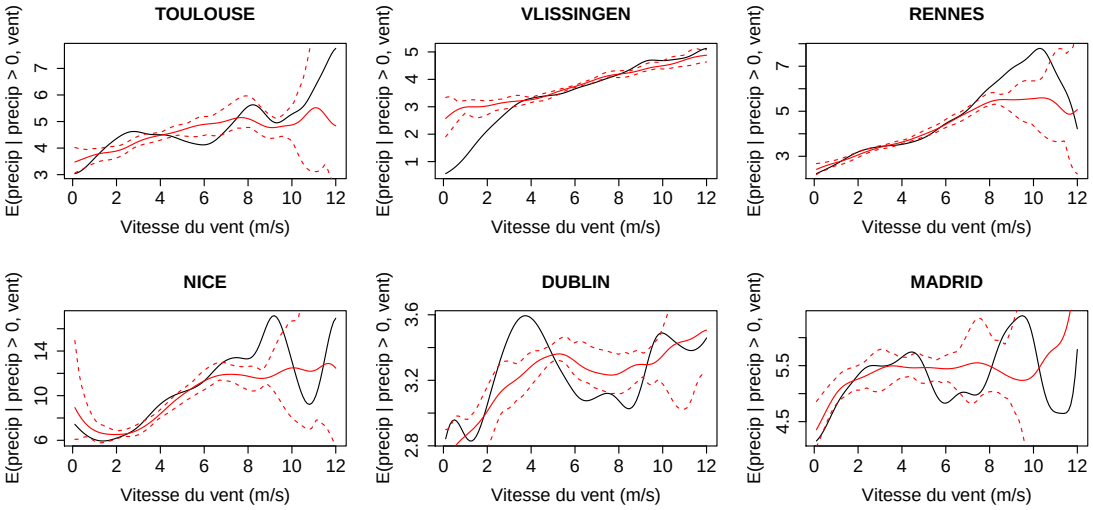

FIGURE 7.63 – Espérance des précipitations conditionnellement à la vitesse du vent pour les jours de pluie.

au modèle bivarié présenté dans le Chapitre 5 n'a pas particulièrement perturbé la modélisation de ces deux variables (rappelons que les trois variables ne sont pas générées indépendamment car elles sont liées par les états cachés). En revanche les résultats concernant la modélisation de la vitesse du vent dans le cadre du modèle trivarié sont plus mitigés. Nous avons constaté :

— Un défaut de saisonnalité dans les moments et maxima journaliers sur certaines stations.
— Un sévère défaut d'auto-corrélation qui se traduit en particulier par une sous-estimation du VCX3, caractéristique des périodes de vent fort.
— Une sous-estimation de la variabilité interannuelle de la vitesse moyenne annuelle du vent.

Notons que les deux premiers problèmes n'existaient pas dans les modèles univariés pour le vent présentés dans la Section 7.2, tandis que le troisième était déjà présent. Le problème du défaut de saisonnalité peut être résolu en introduisant de la saisonnalité dans les lois d'émission liées au vent, comme dans le Modèle 2 de la Section 7.2. Ainsi, dans l'Equation (7.2), $\nu_k^{(3)}$ est remplacée par $\nu_k^{(3)}(t)$, loi de Weibull dont le paramètre d'échelle est un polynôme trigonométrique. Les résultats sont représentés sur la Figure 7.64, à comparer avec ceux de la Figure 7.51. Cependant cela ne permet pas de résoudre les deux autres problèmes évoqués, qui persistent dans ce modèle alternatif. D'autre part, celui-ci a l'inconvénient d'ajouter encore des paramètres.

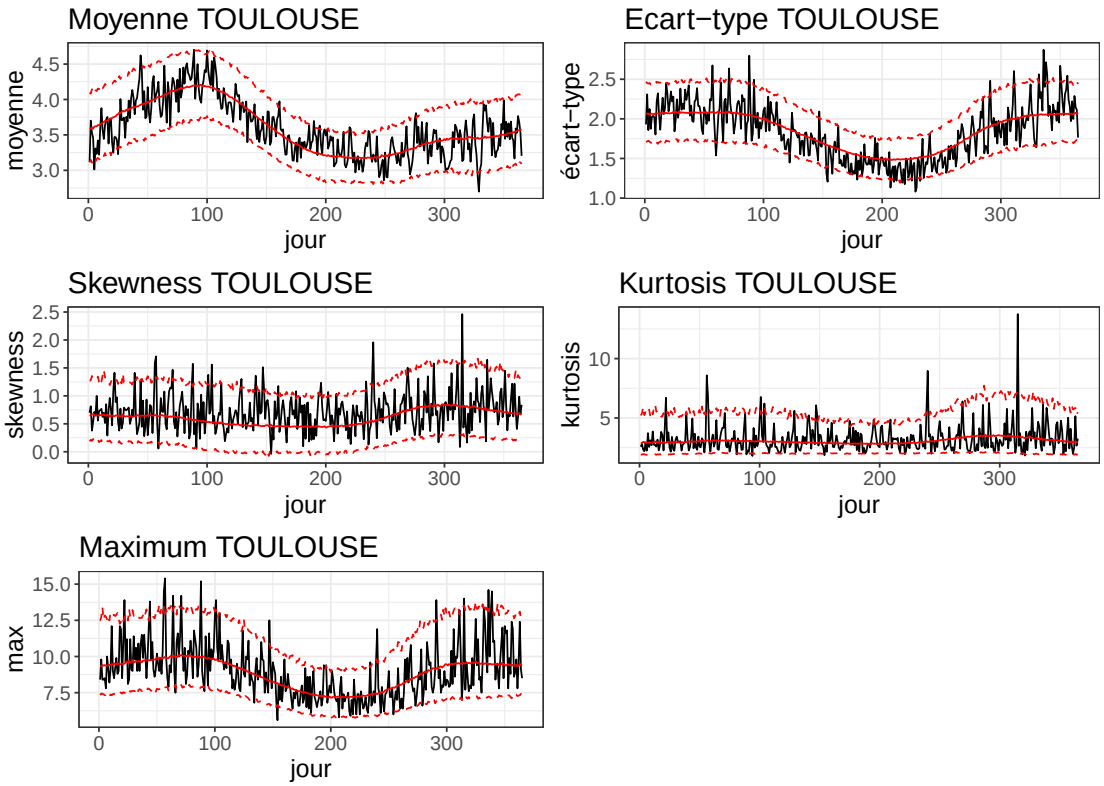

FIGURE 7.64 – Moments et maxima journaliers de la vitesse du vent, modèle trivarié avec paramètre d'échelle de Weibull périodique.

Chapitre 8

Conclusion

8.1 Résumé

Cette thèse portait sur la modélisation multivariée de variables météorologiques au pas journalier en utilisant des modèles de Markov cachés. Plus précisément, nous avions pour objectif la conception d'un générateur stochastique capable de simuler rapidement, conjointement et de façon réaliste des séries temporelles de variables météorologiques dont les propriétés statistiques se rapprochent de celles des séries réelles. Compte tenu de la non-stationnarité des variables climatiques, nous avons été amenés à nous intéresser à certaines propriétés théoriques de modèles de Markov cachés non-homogènes, ce qui a constitué la première partie de la thèse. Nous avons montré que sous des conditions faibles, l'estimateur du maximum de vraisemblance est fortement consistant dans un modèle de Markov caché dont les lois d'émission et les probabilités de transition sont des fonctions périodiques du temps. Nous avons aussi donné des conditions pour l'identifiabilité d'un tel modèle, généralisant ainsi des résultats concernant les modèles de Markov cachés homogènes. Nous nous sommes ensuite intéressés à un second cas de non-homogénéité : celui où les lois d'émission présentent des tendances polynomiales. Nous avons prouvé la consistence de l'estimateur du maximum de vraisemblance dans ce cadre.

Dans la seconde partie de la thèse, nous avons introduit un modèle de Markov caché pour la modélisation conjointe de la température et des précipitations au pas de temps journalier sur un seul site. Ce HMM non-homogène inclut une saisonnalité dans les probabilités de transition et des saisonnalités et tendances dans les lois d'émission, de façon à pouvoir représenter tous les comportements saisonniers des deux variables, ainsi que le changement climatique. Nous avons fait le choix de lois d'émission sous forme de mélanges, de façon à pouvoir approcher au mieux la distribution bivariée ciblée. Nous avons d'abord appliqué ce modèle à des données de température et précipitations issues de mesures réalisées sur des stations en Europe représentatives de climats variés. Après avoir estimé les paramètres du modèle par maximum de vraisemblance via l'algorithme EM, nous avons utilisé le modèle avec les paramètres estimés pour générer aléatoirement des séries synthétiques bivariées de température et de précipitations. Ces séries simulées ont été comparées aux séries observées selon un grand nombre de statistiques. Pour la plupart d'entre elles, le modèle donne de bons résultats. Il peut donc être utilisé pour générer des séries raisonnablement réalistes de température et précipitations. Ces variables sont en particulier essentielles pour la production hydroélectrique car elles influencent directement le débit des cours d'eau. Nous avons donc utilisé notre modèle en association avec un modèle hydrologique pour

générer des séries de débit et avons vérifié que celles-ci étaient statistiquement "proches" des débits observés. Nous nous sommes également servis de notre générateur stochastique pour estimer la probabilité d'observer des hivers simultanément froids et avec de faibles débits, cette probabilité étant difficile à estimer à partir des seules observations. Enfin, nous avons montré sur un exemple le lien qui peut être fait entre les états cachés du modèle et des variables climatiques à plus grande échelle. Dans le dernier chapitre de cette thèse, nous avons introduit un modèle de Markov caché non-homogène pour la modélisation de la vitesse du vent, puis nous avons intégré la vitesse du vent au modèle bivarié précipitations/température pour obtenir un modèle capable de simuler conjointement la température, les précipitations et la vitesse du vent.

8.2 Avantages et limitations des modèles

Outre leur capacité à générer des séries de variables météorologiques, les différents modèles que nous avons présentés dans cette thèse présentent plusieurs avantages :

- Ils sont facilement compréhensibles. Même si la forme des lois d'émission peut être complexe, la dynamique d'un modèle de Markov cachée est simple.
- Ils sont interprétables. Comme nous nous sommes attachés à le montrer, l'examen des paramètres estimés permet de donner directement une interprétation aux différents états cachés en termes météorologiques. Cette interprétation peut être complétée par l'estimation de la séquence d'états la plus probable. On peut en effet, grâce à l'algorithme de Viterbi, associer à chaque pas de temps un état, dans une logique de classification non supervisée. On peut alors interpréter les groupes obtenus via des variables exogènes (par exemple, des champs de géopotentiel). Nous avons illustré ce processus sur un exemple dans le Chapitre 6.
- Contrairement aux techniques de rééchantillonnage, ils permettent de générer des valeurs ou suites de valeurs qui ne figurent pas dans les données.
- Ils sont suffisamment flexibles pour s'adapter à différents climats.
- Le processus de simulation est facile à implémenter en quelques lignes de code (R ou Python par exemple) et est peu coûteux numériquement : sur un ordinateur standard, quelques minutes suffisent pour générer 1000 trajectoires de plusieurs décennies au pas journalier.

Néanmoins, certaines limitations sont à signaler :

- Ils possèdent des hyper-paramètres qu'il peut être délicat de choisir :
  - le nombre d'états cachés, que nous avons noté K,
  - les paramètres M et M_1 de complexité des mélanges qui constituent les lois d'émission,
  - le degré d des polynômes trigonométriques qui modélisent les différentes saisonnalités.
- Ils possèdent un grand nombre de paramètres (jusqu'à plusieurs centaines), l'estimation nécessite donc un volume de données suffisant.
- A cause du grand nombre de paramètres et de la convergence lente de l'algorithme EM, l'estimation des paramètres peut être longue pour les modèles les plus complexes (de quelques heures à deux jours au maximum). D'autre part, il est nécessaire de lancer plusieurs fois l'algorithme puisque celui-ci ne converge que vers un maximum local de la fonction objectif.
- Ils ne permettent pas de capturer toutes les caractéristiques des processus que l'on veut modéliser. En particulier, la dépendance temporelle est souvent insuffisante : nous l'avons constaté sur les canicules et vagues de froid, les périodes sèches ou pluvieuses, et sur la vitesse du vent.

8.3 Perspectives

Dans cette section, nous proposons des pistes d'amélioration pour les modèles étudiés dans cette thèse, en lien avec les limitations que nous venons d'évoquer, ainsi que de nouveaux axes de recherche pour généraliser nos résultats.

Aspects théoriques D'un point de vue théorique, nous avons montré la consistance forte de l'EMV dans deux cas particuliers de HMM non-homogènes. Une prochaine étape pourrait être de montrer un théorème central limite. Rappelons que la normalité asymptotique de l'EMV a été obtenue par Bickel et al. (1998) pour les HMM stationnaires. D'autre part, nous avons montré séparément des résultats sur les HMM avec une saisonnalité d'une part, et sur les HMM avec des tendances d'autre part. Ces résultats pourraient être unifiés de façon à prendre en compte simultanément ces deux types de non-stationnarité.

HMM autorégressif Nous avons vu que l'un des principaux défauts de la modélisation que nous avons proposée est une dépendance temporelle insuffisante. Plusieurs pistes peuvent être envisagées pour répondre à ce problème. La première d'entre elles est l'ajout d'un terme autorégressif. Le modèle ne serait alors plus un HMM mais un "HMM autorégressif", aussi appelé MS-AR (pour *Auto-Regressive Markov Switching*) ou modèle autorégressif à changement de régime Markovien. Dans cette généralisation des HMM proposée par Hamilton (1989), la loi de l'observation Y_t conditionnellement à l'état X_t dépend aussi des s observations précédentes $Y_{t-1}, \dots, Y_{t-s}$. Pour $s = 2$, les relations de dépendance entre les variables peuvent être représentées par le graphe de la Figure 8.1.

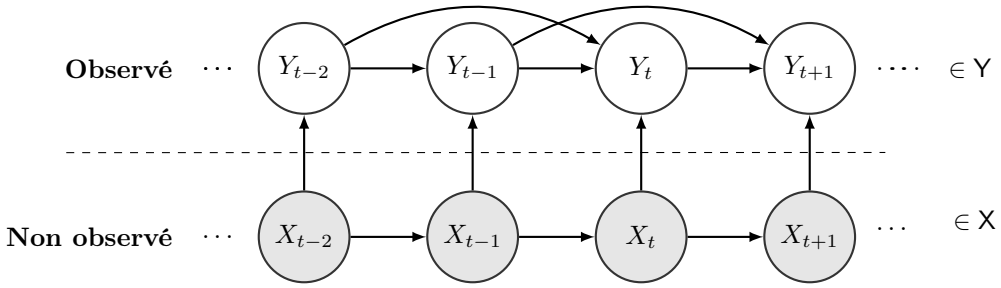

FIGURE 8.1 – Dynamique d'un modèle autorégressif à régime Markovien

Exemple. Soit $(X_t)_{t\geq 1}$ une chaîne de Markov d'ordre 1 à valeurs dans $\{1, \dots, K\}$ et $(\varepsilon_t)_{t\geq 1}$ une suite de variables aléatoires i.i.d., de même loi normale standard et indépendante de $(X_t)_{t\geq 1}$. Soit $\lambda_1, \dots, \lambda_K \in (-1, 1)$ et $\sigma_1, \dots, \sigma_K > 0$. On considère le processus $(Y_t)_{t\geq 0}$ défini par $Y_0 = 0$ et pour tout $t \geq 1$, $Y_t = \lambda_{X_t} Y_{t-1} + \sigma_{X_t} \varepsilon_t$. Alors $(X_t, Y_t)_{t\geq 1}$ est un modèle MS-AR avec $s = 1$.

La consistance forte et la normalité asymptotique de l'EMV pour de tels modèle ont été démontrées par Douc et al. (2004). Un modèle de ce type a été utilisé par Ailliot and Monbet (2012) pour la modélisation de séries temporelles de vent. Une autre approche pour introduire de la dépendance est de faire dépendre la probabilité de transition dans l'état suivant de l'observation dans l'état actuel, comme représenté par la Figure 8.2. Ainsi la suite des états cachés forme une chaîne de Markov non homogène dont les transitions dépendent des observations.

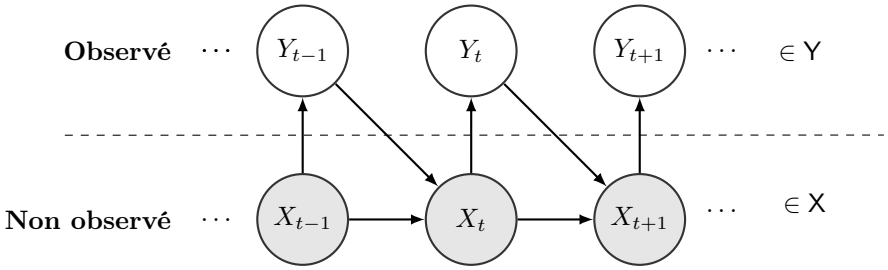

FIGURE 8.2 – Dynamique d'un modèle à régime Markovien non homogène

Prenons l'exemple de la modélisation des températures avec un état chaud et un état froid. On peut supposer que si l'état à l'instant t est l'état chaud et que la température à l'instant t est très élevée, la probabilité de rester dans l'état chaud à l'instant $t+1$ sera elle aussi élevée, car nous sommes entrés dans une période de canicule. L'observation Y_{t+1} sera donc plus probablement générée par l'état chaud. Le modèle que nous venons de décrire permet de produire ce type de comportement. Les deux approches précédentes peuvent être adoptées simultanément. On obtient alors un modèle appelé NHMS-AR. Un tel modèle a été utilisé pour la modélisation du vent dans Ailliot et al. (2015b) et la consistance de l'EMV a été montrée dans Ailliot and Pene (2015) (voir aussi Pouzo et al. (2016)).

Parcimonie Comme nous l'avons signalé, les modèles que nous proposons comportent un grand nombre de paramètres, en particulier le modèle trivarié. Leur estimation est donc délicate car d'une part la fonction de vraisemblance compte de nombreux maxima locaux, et d'autre part, elle nécessite de longues séries d'observations. Une façon de réduire le nombre de paramètres est de jouer sur les hyper-paramètres. Nous avons fait varier le nombre K d'états mais nous n'avons pas exploré toutes les possibilités pour les autres paramètres que sont M, M_1 (qui déterminent la complexité des lois d'émission) et d (qui détermine la complexité des saisonnalités). Plusieurs solutions peuvent être envisagées pour rendre le modèle plus parcimonieux :

- Pour les modèles bivarié (précipitations, température) et trivarié (précipitations, température, vent), les lois d'émission ont été choisies comme des mélanges de $M = 4$ composantes. Il n'est peut-être pas nécessaire d'avoir une telle complexité dans chaque état. On pourrait reformuler le modèle en remplaçant M par un M_k dépendant de l'état, avec M_k éventuellement inférieur à 4. Cela réduirait le nombre de paramètres mais obligerait à choisir les M_k, ce qui augmenterait le nombre d'hyper-paramètres à fixer a priori.
- Pour tenir compte des différentes saisonnalités, le modèle inclut des polynômes trigonométriques (pour les probabilités de transition, la saisonnalité de la température et celle de l'intensité des précipitations). Nous avons choisi des polynômes de degré $d = 2$ pour tous ces polynômes et dans chaque état. Il est cependant possible que choisir $d = 1$ pour certains d'entre eux (ou dans certains états) soit suffisant.
- La matrice de transition périodique contient $K(K-1)(2d+1)$ paramètres, soit par exemple 210 paramètres si $K = 7$ et $d = 2$. Or on constate en pratique qu'une telle complexité n'est pas nécessaire car de nombreuses transitions restent nulles tout au long de l'année (voir par exemple la Figure 5.13). On peut donc réduire drastiquement le nombre de paramètres en constraignant certains coefficients de la matrice à être nuls (par exemple $Q_{ij}(t) = 0$ pour tout t dès lors que $|i-j| > m$ pour un entier $m \in \{1, \ldots, K-2\}$: matrice bande).

Lois d'émission pour la vitesse du vent Classiquement, nous avons utilisé la loi de Weibull dans notre modélisation de la vitesse du vent (Section 7.2). D'autres choix peuvent être fait,

notamment la loi de Rayleigh-Rice (Drobinski et al., 2015).

Tendance dans les précipitations Nous avons vu dans le Chapitre 5 que le processus des précipitations exhibe des tendances, au moins sur certaines stations. Le phénomène est cependant plus complexe que pour les températures. En effet, une tendance peut exister dans la fréquence ou dans l'intensité des précipitations (éventuellement les deux) et le comportement peut être différent en hiver et en été. Nous n'avons pas inclus de tendance pour les précipitations dans nos modèles car celles-ci restent modestes la plupart du temps, et ne concernent pas toutes les stations. Néanmoins, sur certaines stations, il pourrait être nécessaire d'en tenir compte pour améliorer les résultats.

Modélisation au pas horaire ou tri-horaire Cette thèse porte sur la modélisation de variables météorologiques au pas de temps journalier. Cependant pour certaines applications, notamment dans le domaine de la production électrique, il est préférable d'observer les variables au pas de temps tri-horaire (une observation toutes les trois heures) ou horaire. Notre modélisation peut être transposée à ce cas en ajustant un modèle différent pour chaque pas de temps (donc 24 modèles au pas horaire).

Rayonnement Le rayonnement solaire est une variable importante pour un producteur d'électricité comme EDF car il conditionne la production photovoltaïque. On s'intéresse au rayonnement solaire moyen sur chaque journée (plus précisément le SSRD : *Surface Solar Radiation Downwards*). Même si nous n'avons pas intégré cette variable dans notre modèle, nous proposons ici quelques idées pour le faire.

Le rayonnement peut se décomposer sous la forme $R_t = C(t)\tau_t$. Le terme $C(t)$ est une fonction déterministe et périodique appelée *courbe de ciel clair*. Il représente le rayonnement maximal possible à l'instant t, c'est-à-dire le rayonnement que l'on observe dans le cas où il n'y a pas de couverture nuageuse. C'est donc une borne supérieure physique qui dépend essentiellement de la latitude. La renormalisation par la courbe de ciel clair $\tau_t \in (0, 1]$ est la partie stochastique du rayonnement : c'est elle que nous souhaitons modéliser. La Figure 8.3 montre que même après renormalisation par la courbe de ciel, il subsiste une saisonnalité : le processus $(\tau_t)_t$ n'est pas stationnaire. Cela s'explique par le fait que le ciel est plus souvent couvert en hiver qu'en été. Notons que 0 est une borne inférieure pour τ_t mais elle n'est jamais atteinte : cela signifierait un rayonnement nul toute la journée, ce qui n'arrive pas aux moyennes latitudes. On peut aussi décomposer le rayonnement en $R_t = C(t) - SRD_t$ où $SRD_t \geq 0$ (*Solar Radiation Difference*) est le rayonnement absorbé par la couche nuageuse, soit ce qu'il manque pour atteindre le ciel clair.

Il semble difficile d'intégrer le rayonnement à notre modèle trivarié de la même façon que nous avons intégré le vent à notre modèle bivarié (voir le Chapitre 7). En effet, nous ne disposons pas de mesures in-situ de rayonnement sur une durée aussi longue que celle des autres variables (soit environ 60 années). D'autres sources de données existent pour le rayonnement solaire :

- Les données satellite HelioClim 3. Ce sont des données grillées à une résolution spatiale de 3 à 8km (selon la latitude) au pas horaire ou journalier, depuis 2004. La longueur de l'historique est donc bien inférieure à celle dont on dispose pour les autres variables. D'autre part, même en choisissant le point de grille le plus proche de la localisation des mesures in-situ des autres variables, on peut perdre la cohérence entre les variables car le rayonnement est une variable très locale (il dépend des nuages).
- Les réanalyses ERA5 (Dee et al., 2011). Elles présentent l'avantage de fournir un historique plus long (depuis 1979) au pas horaire sur toutes les variables qui nous intéressent.

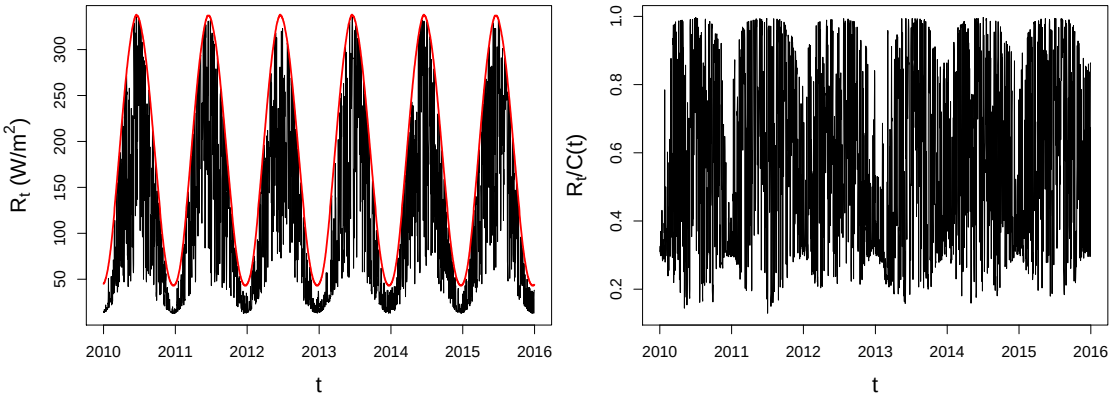

FIGURE 8.3 – A gauche : rayonnement au pas journalier, en rouge la courbe de ciel clair. A droite : $R_t/C(t)$.

Cependant, la résolution est plus faible (30km) et ce ne sont pas des observations. Elles peuvent néanmoins servir comme "preuve de concept" pour un modèle.

Ces deux sources de données incluent aussi le rayonnement ciel clair.

Une première solution pour obtenir un générateur de temps incluant le rayonnement est de l'intégrer à notre HMM trivarié et d'utiliser les réanalyses sur les 40 années disponibles pour ajuster un modèle, avec les réserves que nous avons émises concernant les réanalyses. Une autre possibilité est de simuler le rayonnement conditionnellement aux autres variables. Ainsi la procédure de simulation pour quatre variables sur une station serait la suivante :

1. Estimation du modèle HMM trivarié (précipitations, température, vent) décrit dans le Chapitre 7 en utilisant les données ECA&D.
2. Estimation de la loi du rayonnement conditionnellement aux autres variables en utilisant les données HelioClim (au point de grille le plus proche de la station).
3. Simulation des trois premières variables via le HMM trivarié.
4. Simulation du rayonnement conditionnellement aux simulations précédentes, en utilisant la loi estimées à l'étape 2.

L'idée de simuler le rayonnement conditionnellement à l'occurrence des précipitations remonte au modèle WGEN (Richardson, 1981). Dans ce modèle, on génère d'abord les précipitations, puis les "résidus" (c'est-à-dire les variables désaisonnalisées) de la température maximale, minimale et du rayonnement sont générés en utilisant un modèle autorégressif multivarié (VAR(1)). Enfin, ces trois variables sont reconstruites en multipliant par une variance et en ajoutant une moyenne, lesquelles dépendent de l'occurrence de pluie. Des extensions de ce modèle telles que Wilks (1999) et Parlange and Katz (2000) utilisent la même idée. Dans Larsen and Pense (1981), les auteurs simulent le rayonnement conditionnellement à l'occurrence de pluie en utilisant une autre approche. Les jours de pluie, le SRD est simulé en utilisant une loi beta, tandis qu'une loi gamma est utilisée pour les jours secs. Les paramètres de ces lois sont calculés séparément mois par mois pour tenir compte de la saisonnalité. Nous avons testé ce modèle sur une station, avec les données HelioClim pour le rayonnement et les données ECA&D pour les précipitations. Nous avons examiné les distributions mensuelles du rayonnement pour les jours pluvieux, non pluvieux,

et les avons comparées à celle issues des simulations dans le modèle de Larsen and Pense (1981). Nous avons obtenu un ajustement correct pour les distributions mensuelles, avec néanmoins une surestimation du rayonnement pour les jours pluvieux en hiver. D'autre part, comme la loi du rayonnement simulé ne dépend que du mois et de l'occurrence de pluie, la dépendance temporelle du rayonnement est sous-estimée par les simulations.

Par conséquent, nous suggérons de retenir l'idée d'une simulation conditionnelle du rayonnement, mais d'utiliser toute l'information disponible et non seulement l'occurrence de précipitations. Cela inclut cette dernière, mais aussi l'intensité des précipitations, la température, le vent, ainsi que toutes ces variables au pas de temps précédent, voire suivant.

Bibliographie

Pierre Ailliot and Valérie Monbet. Markov-switching autoregressive models for wind time series. *Environmental Modelling & Software*, 30 :92–101, 2012.

Pierre Ailliot and Françoise Pene. Consistency of the maximum likelihood estimate for non-homogeneous Markov–switching models. *ESAIM : Probability and Statistics*, 19 :268–292, 2015.

Pierre Ailliot, Craig Thompson, and Peter Thomson. Space–time modelling of precipitation by using a hidden Markov model and censored gaussian distributions. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 58(3) :405–426, 2009.

Pierre Ailliot, Denis Allard, Valérie Monbet, and Philippe Naveau. Stochastic weather generators : an overview of weather type models. *Journal de la Société Française de Statistique*, 156 (1) :101–113, 2015a.

Pierre Ailliot, Julie Bessac, Valérie Monbet, and Françoise Pene. Non-homogeneous hidden markov-switching models for wind time series. *Journal of Statistical Planning and Inference*, 160 :75–88, 2015b.

Grigory Alexandrovich, Hajo Holzmann, and Anna Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2) :423–434, 2016.

Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pages 3099–3132, 2009.

Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory*, pages 33–1, 2012.

Andrew R Barron. The strong ergodic theorem for densities : generalized Shannon-McMillan-Breiman theorem. *The annals of Probability*, 13(4) :1292–1303, 1985.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The annals of mathematical statistics*, 37(6) :1554–1563, 1966.

Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1) :164–171, 1970.

Enrica Bellone, James P Hughes, and Peter Guttorp. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate research*, 15(1) :1–12, 2000.

Fred Espen Benth and Jūratė Šaltytė Benth. Weather derivatives and stochastic modelling of temperature. *International Journal of Stochastic Analysis*, 2011, 2011.

Julie Bessac, Pierre Ailliot, Julien Cattiaux, and Valérie Monbet. Comparison of hidden and observed regime-switching autoregressive models for (u, v)-components of wind fields in the northeastern Atlantic. *Advances in Statistical Climatology, Meteorology and Oceanography*, 2 (1) :1–16, 2016.

Peter J Bickel, Ya'acov Ritov, and Tobias Ryden. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics*, pages 1614–1635, 1998.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7) :719–725, 2000.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3) :561–575, 2003.

J Boé and L Terray. A weather-type approach to analyzing winter precipitation in France : Twentieth-century trends and the role of anthropogenic forcing. *Journal of Climate*, 21(13) : 3118–3133, 2008.

Nadine Brisson, Christian Gary, Eric Justes, R Roche, Bruno Mary, Dominique Ripoche, Daniel Zimmer, Jorge Sierra, Patrick Bertuzzi, P Burger, et al. An overview of the crop model STICS. *European Journal of agronomy*, 18(3-4) :309–332, 2003.

M Broniatowski, G Celeux, and J Diebolt. Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. *Data analysis and informatics*, 3 :359–373, 1983.

Sean D Campbell and Francis X Diebold. Weather forecasting for weather derivatives. *Journal of the American Statistical Association*, 100(469) :6–16, 2005.

Olivier Cappé, Eric Moulines, and Tobias Rydén. Inference in hidden Markov models. In *Proceedings of EUSFLAT Conference*, pages 14–16, 2009.

Jose A Carta, Penelope Ramirez, and Sergio Velazquez. A review of wind speed probability distributions used in wind energy analysis : Case studies in the canary islands. *Renewable and sustainable energy reviews*, 13(5) :933–955, 2009.

Gilles Celeux and Jean Diebolt. L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. *Revue de statistique appliquée*, 34(2) :35–52, 1986.

Gilles Celeux and Jean-Baptiste Durand. Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4) :541–564, 2008.

Gilles Celeux, Didier Chauveau, and Jean Diebolt. *On stochastic versions of the EM algorithm.* PhD thesis, INRIA, 1995.

Didier Dacunha-Castelle, Thi Thu Huong Hoang, and Sylvie Parey. Modeling of air temperatures : preprocessing and trends, reduced stationary process, extremes, simulation. *Journal de la Société Française de Statistique*, 156(1) :138–168, 2015.

Yohann De Castro, Élisabeth Gassiat, and Claire Lacour. Minimax adaptive estimation of nonparametric hidden Markov models. *Journal of Machine Learning Research*, 17 :1–43, 2016.

Yohann De Castro, Élisabeth Gassiat, and Sylvain Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Transactions on Information Theory*, PP(99) :1–1, 2017. ISSN 0018-9448. doi : 10.1109/TIT.2017.2696959.

Dick P Dee, SM Uppala, AJ Simmons, Paul Berrisford, P Poli, S Kobayashi, U Andrae, MA Balmaseda, G Balsamo, d P Bauer, et al. The era-interim reanalysis : Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656) :553–597, 2011.

Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, pages 94–128, 1999.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

Manuel Diehn, Axel Munk, and Daniel Rudolf. Maximum likelihood estimation in hidden Markov models with inhomogeneous noise. *arXiv preprint arXiv :1804.04034*, 2018.

Randal Douc, Catherine Matias, et al. Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7(3) :381–420, 2001.

Randal Douc, Eric Moulines, Tobias Rydén, et al. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics*, 32(5) :2254–2304, 2004.

Randal Douc, Eric Moulines, Jimmy Olsson, Ramon Van Handel, et al. Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 39(1) :474–513, 2011.

Randal Douc, Eric Moulines, et al. Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *The Annals of Statistics*, 40(5) :2697–2732, 2012.

Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear time series : theory, methods and applications with R examples*. CRC Press, 2014.

Philippe Drobinski, Corentin Coulais, and Bénédicte Jourdier. Surface wind-speed statistics modelling : alternatives to the weibull distribution and performance evaluation. *Boundary-Layer Meteorology*, 157(1) :97–123, 2015.

C Flecher, P Naveau, D Allard, and N Brisson. A stochastic daily weather generator for skewed data. *Water Resources Research*, 46(7), 2010.

G David Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3) :268–278, 1973.

KR Gabriel and J Neumann. A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society*, 88(375) :90–95, 1962.

Mark Gales and Steve Young. The application of hidden Markov models in speech recognition. *Foundations and trends in signal processing*, 1(3) :195–304, 2008.

F Garavaglia, J Gailhard, E Paquet, M Lang, R Garçon, and P Bernardara. Introducing a rainfall compound distribution model based on weather patterns sub-sampling. *Hydrology and Earth System Sciences Discussions*, 14 :p–951, 2010.

R Garçon. Prévision opérationnelle des apports de la Durance à Serre-Ponçon à l'aide du modèle MORDOR. Bilan de l'année 1994-1995. *La Houille Blanche*, (5) :71–76, 1996.

Elisabeth Gassiat, Alice Cleynen, and Stéphane Robin. Finite state space non parametric hidden Markov models are in general identifiable. *Statistics and Computing*, 26(1–2) :61–71, 2016a.

Elisabeth Gassiat, Judith Rousseau, et al. Nonparametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1) :193–212, 2016b.

Peter W Glynn and Dirk Ormoneit. Hoeffding's inequality for uniformly ergodic markov chains. *Statistics & probability letters*, 56(2) :143–146, 2002.

James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica : Journal of the Econometric Society*, pages 357–384, 1989.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5) :1460–1480, 2012.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1) : 193–218, 1985.

James P Hughes and Peter Guttorp. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water resources research*, 30(5) :1535–1546, 1994.

James P Hughes, Peter Guttorp, and Stephen P Charles. A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 48(1) :15–30, 1999.

Richard W Katz. Precipitation as a chain-dependent process. *Journal of Applied Meteorology*, 16(7) :671–676, 1977.

Richard W Katz. Use of conditional stochastic models to generate climate change scenarios. *Climatic Change*, 32(3) :237–255, 1996.

Richard W Katz and Marc B Parlange. Overdispersion phenomenon in stochastic modeling of precipitation. *Journal of Climate*, 11(4) :591–601, 1998.

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598) :671–680, 1983.

Sergey Kirshner. *Modeling of multivariate time series using hidden Markov models*. PhD thesis, University of California, Irvine, 2005.

Willem Kruijer, Judith Rousseau, Aad Van Der Vaart, et al. Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4 :1225–1257, 2010.

Joseph B Kruskal. Three-way arrays : rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2) : 95–138, 1977.

Anna Kuentz, T Mathevet, J Gailhard, and B Hingray. Building long-term and high spatio-temporal resolution precipitation and air temperature reanalyses by mixing local observations and global atmospheric reanalyses : the ANATEM model. *Hydrology and Earth System Sciences*, 19(6) :2717–2736, 2015.

Martin F Lambert, Julian P Whiting, and Andrew V Metcalfe. A non-parametric hidden Markov model for climate state identification. *Hydrology and Earth System Sciences Discussions*, 7 (5) :652–667, 2003.

Greg A Larsen and Roberta B Pense. *Stochastic simulation of daily climate data*. Research Division, Statistical Reporting Service, Department of Agriculture, 1981.

François Le Gland and Laurent Mevel. Asymptotic behaviour of the MLE in hidden Markov models. In *Proceedings of the 4th European Control Conference, Bruxelles 1997*, 1997.

Joseph Cheuk Yi Lee, Michael Jason Fields, and Julie K Lundquist. Assessing variability of wind speed : comparison and validation of 27 methodologies. *Wind Energy Science (Online)*, 3(NREL/JA-5000-72768), 2018.

Luc Lehéricy. Consistent order estimation for nonparametric hidden Markov models. *Bernoulli*, 25(1) :464–498, 2019.

Jan Lennartsson, Anastassia Baxevani, and Deliang Chen. Modelling precipitation in Sweden using multiple step Markov chains and a composite model. *Journal of Hydrology*, 363(1) : 42–59, 2008.

Brian G Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1) :127–143, 1992.

Rogemar S Mamon and Robert J Elliott. *Hidden Markov models in finance*, volume 4. Springer, 2007.

Xiao-Li Meng and Donald B Rubin. Maximum likelihood estimation via the ECM algorithm : A general framework. *Biometrika*, 80(2) :267–278, 1993.

Laurent Mevel and Lorenzo Finesso. Asymptotical statistics of misspecified hidden Markov models. *IEEE Transactions on Automatic Control*, 49(7) :1123–1132, 2004.

Mohammed Mraoua. Temperature stochastic modeling and weather derivatives pricing : empirical study with Moroccan data. *Afrika Statistika*, 2(1), 2007.

EV Newnham. The persistence of wet and dry weather. *Quarterly Journal of the Royal Meteorological Society*, 42(179) :153–162, 1916.

Marc B Parlange and Richard W Katz. An extended version of the Richardson model for simulating daily weather variables. *Journal of Applied Meteorology*, 39(5) :610–622, 2000.

Toby A Patterson, Alison Parton, Roland Langrock, Paul G Blackwell, Len Thomas, and Ruth King. Statistical modelling of individual animal movement : an overview of key methods and a discussion of practical challenges. *Advances in Statistical Analysis*, 101(4) :399–438, 2017.

Nadav Peleg, Simone Fatichi, Athanasios Paschalis, Peter Molnar, and Paolo Burlando. An advanced stochastic weather generator for simulating 2-d high-resolution climate variables. *Journal of Advances in Modeling Earth Systems*, 2017.

Jennifer Pohle, Roland Langrock, Floris M van Beest, and Niels Martin Schmidt. Selecting the number of states in hidden Markov models : pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3) :270–293, 2017.

Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428) :1303–1313, 1994.

Demian Pouzo, Zacharias Psaradakis, and Martin Sola. Maximum likelihood estimation in possibly misspecified dynamic models with time-inhomogeneous markov regimes. *arXiv preprint arXiv :1612.04932*, 2016.

Lawrence Rabiner and B Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1) :4–16, 1986.

P Racsko, L Szeidl, and M Semenov. A serial approach to local stochastic weather models. *Ecological modelling*, 57(1-2) :27–41, 1991.

Balaji Rajagopalan and Upmanu Lall. A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water resources research*, 35(10) :3089–3101, 1999.

Clarence W Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1) :182–190, 1981.

Andrew W Robertson, Sergey Kirshner, and Padhraic Smyth. Hidden Markov models for modeling daily rainfall occurrence over Brazil. *Information and Computer Science, University of California*, 2003.

Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2) : 461–464, 1978.

SG Smith. The seasonal variation of wind speed in the United Kingdom. *Weather*, 38(4) :98–103, 1983.

R Srikanthan and TA McMahon. Stochastic generation of annual, monthly and daily climate data : A review. *Hydrology and Earth System Sciences Discussions*, 5(4) :653–670, 2001.

Sidney Teweles Jr and Hermann B Wobus. Verification of prognostic charts. *Bulletin of the American Meteorological Society*, pages 455–463, 1954.

Augustin Touron. Consistency of the maximum likelihood estimator in seasonal hidden Markov models. *Statistics and Computing*, pages 1–21, 2018.

EJM Van den Besselaar, AMG Klein Tank, and TA Buishand. Trends in European precipitation extremes over 1951–2010. *International Journal of Climatology*, 33(12) :2682–2689, 2013.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2) :260–269, 1967.

Mathieu Vrac, Michael Stein, and Katharine Hayhoe. Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Climate Research*, 34(3) :169–184, 2007.

Greg CG Wei and Martin A Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411) :699–704, 1990.

Robert L Wilby and TML Wigley. Downscaling general circulation model output : a review of methods and limitations. *Progress in physical geography*, 21(4) :530–548, 1997.

Daniel S Wilks. Use of stochastic weather generators for precipitation downscaling. *Wiley Interdisciplinary Reviews : Climate Change*, 1(6) :898–907, 2010.

Daniel S Wilks and Robert L Wilby. The weather generation game : a review of stochastic weather models. *Progress in physical geography*, 23(3) :329–357, 1999.

DS Wilks. Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1) :178–191, 1998.

DS Wilks. Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Agricultural and Forest Meteorology*, 96(1-3) : 85–101, 1999.

Larry L Wilson, Dennis P Lettenmaier, and Eric Skyllingstad. A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation. *Journal of Geophysical Research : Atmospheres*, 97(D3) :2791–2809, 1992.

CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.

Liao Yaoming, Zhang Qiang, and Chen Deliang. Stochastic modeling of daily precipitation in China. *Journal of Geographical Sciences*, 14(4) :417–426, 2004.

Byung-Jun Yoon. Hidden Markov models and their applications in biological sequence analysis. *Current genomics*, 10(6) :402–415, 2009.

Walter Zucchini and Peter Guttorp. A hidden Markov model for space-time precipitation. *Water Resources Research*, 27(8) :1917–1923, 1991.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Modélisation multivariée de variables météorologiques

Mots Clefs : Mathématiques, Statistiques, Modélisation stochastique, Modèles de Markov caché, Climat, Météorologie

Résumé : La production d'énergie renouvelable et la consommation d'électricité dépendent largement des conditions météorologiques : température, précipitations, vent, rayonnement solaire... Ainsi, pour réaliser des études d'impact sur l'équilibre offre-demande, on peut utiliser un générateur de temps, c'est-à-dire un modèle permettant de simuler rapidement de longues séries de variables météorologiques réalistes, au pas de temps journalier. L'une des approches possibles pour atteindre cet objectif utilise les modèles de Markov caché : l'évolution des variables à modéliser est supposée dépendre d'une variable latente que l'on peut interpréter comme un type de temps. En adoptant cette approche, nous proposons dans cette thèse un modèle permettant de simuler simultanément la température, la vitesse du vent et les précipitations, en tenant compte des non-stationnarités qui caractérisent les variables météorologiques. D'autre part, nous nous intéressons à certaines propriétés théoriques des modèles de Markov caché cyclo-stationnaires : nous donnons des conditions simples pour assurer leur identifiabilité et la consistance forte de l'estimateur du maximum de vraisemblance. On montre aussi cette propriété de l'EMV pour des modèles de Markov caché incluant des tendances de long terme sous forme polynomiale.

Title : Multivariate modelling of weather variables

Keys words : Mathematics, Statistics, Stochastic modelling, Hidden Markov models, Climate, Meteorology

Abstract : Renewable energy production and electricity consumption both depend heavily on weather : temperature, precipitations, wind, solar radiation... Thus, making impact studies on the supply/demand equilibrium may require a weather generator, that is a model capable of quickly simulating long, realistic times series of weather variables, at the daily time step. To this aim, one of the possible approaches is using hidden Markov models : we assume that the evolution of the weather variables are governed by a latent variable that can be interpreted as a weather type. Using this approach, we propose a model able to simulate simultaneously temperature, wind speed and precipitations, accounting for the specific non-stationarities of weather variables. Besides, we study some theoretical properties of cyclo-stationary hidden Markov models : we provide simple conditions of identifiability and we show the strong consistency of the maximum likelihood estimator. We also show this property of the MLE for hidden Markov models including long-term polynomial trends.