Sparse vector Markov switching autoregressive models Application to multivariate time series of temperature

V. Monbet[a,*], P. Ailliot[b]

[a] *IRMAR, Université de Rennes 1 & INRIA, Rennes, France*

[b] *LMBA, Université de Bretagne Occidentale, Brest, France*

Abstract

Multivariate time series are of interest in many fields including economics and environment. The dynamical processes occurring in these domains often exhibit regimes so that it is common to describe them using Markov Switching vector autoregressive processes. However the estimation of such models is difficult even when the dimension is not so high because of the number of parameters involved. In this paper we propose to use a Smoothly Clipped Absolute Deviation penalization of the likelihood to shrink the parameters. The Expectation Maximization algorithm built for maximizing the penalized likelihood is described in details and tested on simulated data and real data of daily mean temperature.

Keywords: Markov Switching Vector Autoregressive process, sparsity, penalized likelihood, SCAD, EM algorithm, daily temperature, stochastic weather generator.

1. Introduction

Multivariate time series are of interest in many fields including economics and environment. The most popular tools for studying multivariate time series are the vector autoregressive (VAR) models because of their simple specification and the existence of efficient methods to fit these models. However, VAR models do not allow to describe time series mixing different dynamics. For instance, when meteorological variables are observed, the resulting time series exhibit an alternance of different temporal dynamics corresponding to weather regimes. Similar phenomena are observed in economics time series with business cycles (see (Hamilton, 1989)). In the sequel, we will consider daily mean temperature at 12 stations in France (see the map of Figure 1). A short sequence of the associated multivariate time series (one component for each station) is depicted on Figure 2. The regimes are highlighted by background white and gray boxes. Our objectif is to built a statistical model which allow to reproduce, the non linearities due to the regime switchings.

The regime is often not observed directly and is thus introduced as a latent process in time

*Corresponding author

Email addresses: valerie.monbet@univ-rennes1.fr (V. Monbet), pierre.ailliot@univ-brest.fr (P. Ailliot)

series models in the spirit of Hidden Markov models. Markov Switching autoregressive (MSAR) models have been introduced as a generalization of autoregressive models and Hidden Markov Models. They are widely used in many domains. For instance, (Lu and Berliner, 1999) propose a MSAR model to describe a runoff time series exhibiting pulsatile behavior. This model is a mixture of three autoregressive models which accommodate "rising", "falling" and "normal" states in the runoff process. (Bessac et al., 2016) and (Kazor and Hering, 2015) proposed Markov Switching VAR models to exhibit wind regimes from multivariate meteorological time series. These models are common for econometric time series too (Hamilton, 1989) (Krolzig, 2013). For instance, (Favero and Monacelli, 2005) and (Sims and Zha, 2006) have resorted to the MSVAR framework to detect shifts in the US monetary and fiscal policy.

Markov Switching VAR models (MSVAR) are simply VAR models with Makov switching parameters. They suffer of the same dimensionality problem as VAR models. For large (and even moderate) dimensions, the number of autoregressive coefficients in each regime can be prohibitively large. For example, a MSVAR model with M regimes, autoregressive processes of order p and dimension d involves $M(d + pd^2 + d(d-1)/2) + M(M-1)$ parameters. For a VAR model ($M = 1$) of order $p = 1$ in dimension $d = 12$, the number of parameters is equal to 222, which is huge and results in noisy estimates. When the variables are correlated, which is the standard situation in multivariate time series, over-learning is frequent. The estimated parameters contain spurious non-zero coefficients and are then difficult to interpret. The predictions associated to the model are usually unstable. Collinearity causes also ill-conditioning of the innovation covariance.

Several approaches have been proposed to overcome the dimensionality problem of VAR models. A first method, that is usual for environmental data and more generally in spatial statistics, consists in searching for parametric shapes for the autoregressive and covariance matrices. The most usual parametric shapes are such that the off-diagonal entries of the matrices decrease with the distance between the sites. This method has been used for MSVAR too (see e.g. (Hering et al., 2015), (Ailliot et al., 2006)). However, this approach often requires an important modelling effort. And, for some applications, it is really difficult to find convenient parametric models (Bessac et al., 2016). A typical difficulty, is that one needs to obtain a first (non parametric) fit of the model to be able to choose parametric shapes. In such situations, it is really helpful to have selection procedures which automatically set the spurious non zero coefficients to zero. Since one deals with colinearity, it makes sense to set zeros both in the autoregressive matrices and in the precision matrices because they are linked with conditional independence.

This can be performed using penalization procedures like lasso (Hsu et al., 2008) and its variants (Medeiros et al., 2012) which are known to be efficient to shrink to zero the useless parameters in the autoregressive part of VAR models as for high dimension regression models. For covariance matrices hard thresholding such as graphical lasso is usually used (Friedman et al., 2008) (Bickel and Levina, 2008). These procedures have become popular since they are computationally efficient and perform variable selection and parameter estimation at the same time. In most papers using penalties, the inference problem is formulated as a regression problem with independent innovations. It excludes all the models with the presence of cross-correlations among the error components. However, recently, Basu and Michailidis (Basu et al., 2015) show that consistent estimation is possible with L1-penalization for both least squares and log-likelihood based choices of loss functions for any stable VAR models. They consider the sparsity of the

covariance matrix too and discuss the choice of other penalties such as the Smoothly Clipped Absolute Deviation (SCAD) penalty.

The MSVAR models are presented in Section 2.1. The originality of the paper lies in the adaptation of a likelihood penalization method with hard thresholding for MSVAR models. The high dimensional parameters involved in MSVAR models are, for each regime, the autoregressive matrices and the covariance of the innovation or the precision matrices. It is more natural to focus on the precision matrices instead of covariance matrices when data exhibit with strong colinearities as it is the case for daily mean temperature. Both autoregressive matrices and precision matrices, the shrinkage is performed using SCAD penalties. That constitutes also an originality of the paper (see Section 2.3). In this section, asymptotic properties of the obtained estimators are quickly discussed too. The Expectation Maximization algorithm built for the penalized likelihood is described in Section 3. The performance of the penalized estimators is illustrated on simulated data in Section 4. And, in the last part of the paper (Section 5), sparse MSVAR are applied to build a multisite stochastic weather generator (SWGEN) for daily mean temperatures in France. To the best of our knowledge, such models have never been implemented for temperature data.

As supplementary material to this paper, algorithms have been implemented in a R package which can be freely download from CRAN R repository under the link
`https://cran.r-project.org/package=NHMSAR`.

2. Markov Switching Vector Autoregressive models

MSVAR models have been introduced for time series in economics as a generalization of autoregressive models and Hidden Markov Models (Hamilton, 1989) (Krolzig, 2013). Then they have been used for meteorological time series (see for instance (Lu and Berliner, 1999) (Pinson and Madsen, 2012) and references therein).

2.1. MSVAR model

A MSVAR model is defined as a discrete time stochastic process with two components $(S_k, \mathbf{Y}_k)$ with values in $\{1, \cdots, M\} \times \mathbb{R}^d$ and satisfying following conditions.

- The first component is hidden and models a first order Markov chain $\{S_k\}_{k\in\mathbb{Z}}$ tacking its values in the set of states $\{1, \cdots, M\}$. The conditional distribution of S_k given $\{S_{k'}, \mathbf{Y}_{k'}\}_{k'<k}$ depends only of S_{k-1} and $\mathbf{Y}_{k-1}$. The transition probabilities are denoted $p(s_k|s_{k-1}, \mathbf{y}_{k-1}) = P(S_k = s_k|S_{k-1} = s_{k-1}, \mathbf{Y}_{k-1} = \mathbf{y}_{k-1})$.
  This process is often called regime. In meteorological applications, it usually describes the weather type (e.g. cyclonic, anticyclonic).
- The second component $\mathbf{Y}_k$ describes the evolution of the observed variables. The conditional distribution of $\mathbf{Y}_k$ given $\{\mathbf{Y}_{k'}\}_{k'<k}$ and $\{S_{k'}\}_{k'\leq k}$ only depends on S_k and $(\mathbf{Y}_{k-1}, \cdots, \mathbf{Y}_{k-p})$.

$$\mathbf{Y}_k = A_0^{(S_k)} + A_1^{(S_k)}\mathbf{Y}_{k-1} + \left(\Sigma^{(S_k)}\right)^{1/2}\boldsymbol{\epsilon}_k$$

where the unknown parameters $A_0^{(s)}$ are vectors of $\mathbb{R}^d$, $A_1^{(s)}$ are matrices of $\mathbb{R}^{d,d}$ and $\Sigma^{(s)}$ are positive symetric matrices of $\mathbb{R}^{d,d}$. $\{\boldsymbol{\epsilon}_k\}_{k\in\mathbb{Z}}$ is a multivariate sequence of independent and identically distributed Gaussian variables, with zero mean and unit variance, independent of the Markov chain $\{S_k\}$.
In the example below, $\mathbf{Y}_k$ will describe the daily mean temperature at $d = 12$ stations. It is straightforward to generalize this model and the mothodology described below for autoregressivemodels of higher order.

The various conditional independence assumptions are summarized by the directed graph below.

$$\begin{array}{lccccccccc}
\textbf{Hidden Regime} & \cdots & \rightarrow & S_{k-1} & \rightarrow & S_k & \rightarrow & S_{k+1} & \rightarrow & \cdots \\
 & & & \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & & \\
\textbf{Temperature} & \cdots & \rightarrow & \mathbf{Y}_{k-1} & \rightarrow & \mathbf{Y}_k & \rightarrow & \mathbf{Y}_{k+1} & \rightarrow & \cdots
\end{array}$$

It is generally assumed that the hidden process is an homogeneous Markov chain $p(s_k|s_{k-1}, \mathbf{y}_{k-1}) = p(s_k|s_{k-1})$. Here, we consider a slightly more general model where the Markov chain is non homogeneous. Its transition probabilities depend on the observation at the previous time step. We will see in Section 5 that it permits a better modelling of the particular data set considered in the work. Several well known models are included in the introduced MSVAR. The classical Hidden Markov Model corresponds to the case with autoregressive models of order 0. The Gaussian mixture model is obtained for autoregressive models of order 0 and $p(s_k|s_{k-1}, \mathbf{y}_{k-1}) = p(s_k)$. When there is only one regime ($M = 1$), MSVAR reduced to the classical VAR model. All these models can be fitted using the methodology introduced in this work.

2.2. Likelihood

For MSVAR models, given an observation sequence $\mathbf{y}_1 \cdots, \mathbf{y}_n$, the inference is performed by maximizing the likelihood $p_{\boldsymbol{\theta}}(\mathbf{y}_1, \cdots, \mathbf{y}_n|\mathbf{y}_0)$ where the subscript $\boldsymbol{\theta} \in \Theta$ denotes the dependence on the appropriate parameters. However, the unobserved Markov chain includes hidden variables which make intractable the analytical solution of the problem. The strategy used in this work, to maximise the likelihood function, is based on the EM algorithm. This algorithm takes advantage of the simplicity of the complete likelihood function $p_{\boldsymbol{\theta}}(s_1, \cdots, s_n, \mathbf{y}_1 \cdots, \mathbf{y}_n|\mathbf{y}_0)$.

It is straighforward to verify that the complete likelihood can be split in two terms as follows:

$$\log p_{\boldsymbol{\theta}}(s_1, \cdots, s_n, \mathbf{y}_1 \cdots, \mathbf{y}_n|\mathbf{y}_0) = \log p_{\boldsymbol{\theta}}(\mathbf{y}_1, \cdots, \mathbf{y}_n|s_1, \cdots, s_n, \mathbf{y}_0) + \log p_{\boldsymbol{\theta}}(s_1, \cdots, s_n|\mathbf{y}_0).$$

The first term depends only on the parameters of the VAR models while the second one depends on the transition probabilities. This particular form is exploited to propose efficient algorithms.

The first term can be further factorized using the conditionnal independence properties of the model.

$$\log p_{\boldsymbol{\theta}}(\mathbf{y}_1, \cdots, \mathbf{y}_n|s_1, \cdots, s_n, \mathbf{y}_0) = \sum_{k=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{y}_k|\mathbf{y}_{k-1}, s_k)$$

where $\sum_{k=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{y}_k|\mathbf{y}_{k-1})$ is the Gaussian likelihood associated with the autoregressive models of regime s_k. The second term is linked with the transition probabilities and it is standard.

2.3. Penalized likelihood

The introduction of a penality $\mathfrak{P}_\lambda(\boldsymbol{\theta})$ induces the maximization of a new function

$$\log p_\theta(\mathbf{y}_1, \cdots, \mathbf{y}_n|\mathbf{y}_0) - n\mathfrak{P}_\lambda(\boldsymbol{\theta}) \tag{1}$$

with $\boldsymbol{\theta} \in \Theta$ a compact subset of $\mathbb{R}^J$.

A good penalty should lead to an estimate which is nearly unbiased when the true parameter is large (unbiasedness property). The resulting estimate should automatically set the small estimated coefficients to zero to reduce the model complexity (sparsity property). And the function which associates the penalized estimate to the maximum likelihood estimate should be continuous in order to avoid prediction instability (continuity property).

Following Fan and Li (Fan and Li, 2001), the SCAD penalty is used. This penalty verifies the three properties listed above for many models and it has been demonstrated, on many examples, to perfom better than Lasso in finding the support of the parameters. Minimax Concave Penalty has similar performances. There is no single method that beats all other competitors in all situations and SCAD will the shown to perform well for the problem we are focusing on (see Sections 4 and 5).

The SCAD penalty is usually defined by its derivative but it may easier to interpret with a direct definition (2)

$$\mathfrak{p}_\lambda(|\theta|) = \begin{cases} \lambda|\theta| & \text{if } |\theta| \leq \lambda \\ -\frac{|\theta|^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)} & \text{if } \lambda \leq |\theta| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\theta| \geq a\lambda. \end{cases} \tag{2}$$

This corresponds to a quadratic spline function with knots at λ and $a\lambda$. The function is continuous and differentiable on $(-\infty, 0) \cup (0, \infty)$ but singular at 0 and its derivatives is equal to zero outside the range $[-a\lambda, a\lambda]$. SCAD penalty sets small coefficients to zero. A few other coefficients are shrunk towards zero. And the large coefficients are retained as they are. One can remark that when a tends to infinity the SCAD penalty tends to the Lasso penalty. Based on an argument of minimizing the Bayes risk, Fan and Li (2001) recommended the choice $a = 3.7$. We adopt the same choice in the sequel as most authors in the literature. When the parameter is of dimension $J > 1$, one defines the following penalty:

$$\mathfrak{P}_\lambda(\boldsymbol{\theta}) = \sum_{j=1}^{J} \mathfrak{p}_{\lambda_j}(|\theta_j|)$$

with J the dimension of $\boldsymbol{\theta}$. In practice, some equality constraints are imposed to the λ_j (see below).

Fan and Li (2001) obtained a weak consistency result for the SCAD penalized likelihood estimators for regression problems in finite dimension (see Theorem 1 of (Fan and Li, 2001)). More precisely, they proved that there exists a local minimizer of the SCAD penalized likelihood that tends to the true parameter at rate $n^{1/2}$ with probability tending to 1 when $n \to \infty$ and

$\lambda(n) \to 0$. This result holds under usual regularity conditions of the density function of the observations. The proof is mainly based on Taylor's expansions and on the Central Limit Theorem (CLT) for the gradient of the log-likelihood. This consistency result was extended for regression mixture models which is a particular case of MSVAR models (Khalili and Chen, 2007). It is still valid for homogeneous MSVAR models by the CLT obtained by Douc et al. (Douc et al., 2004), Theorem 2. As far as we know, no analogous CLT is available for non homogeneous MSVAR models yet. Indeed, the most recent theoritical results for non homogeneous MSVAR are due to Ailliot and Pène (Ailliot and Pène, 2015) who have demonstrated that non homogeneous MSVAR models verify a property of ergodicity and that the estimator of maximum likelihood is consistent. So, one can not directly generalize the asymptotic consistency result to non homogeneous MSVAR models.

Furthermore, Fan and Li (2001) also demonstrate a sparsity result which means that the SCAD estimate finds the zeros at the right places. Here again the proof is based on the same CLT and can then be easily adapted for homogeneous MSVAR models. It is not detailed here.

3. Inference and EM algorithm

3.1. EM algorithm for maximizing the likelihood function

The EM algorithm was initially introduced in (Baum et al., 1970) for HMMs and then generalized to models with latent variables in (Dempster et al., 1977). This recursive algorithm computes successive approximations $\boldsymbol{\theta}_\ell$ of the maximum likelihood estimator $\boldsymbol{\theta}^*$ by cycling through the following steps.

E-step Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_\ell) = E_{\boldsymbol{\theta}_\ell}(\log(p_{\boldsymbol{\theta}}(S_1, \cdots, S_n, \mathbf{Y}_1, \cdots, \mathbf{Y}_n))|\mathbf{y}_0, \cdots, \mathbf{y}_n)$ as a function of $\boldsymbol{\theta}$.
M-step Determine the updated parameter estimate $\boldsymbol{\theta}_{\ell+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_\ell)$.

In MSVAR models, $\boldsymbol{\theta}$ can be split in $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\rm tr}, \boldsymbol{\theta}_{\rm em}^{(m)}; m = 1, \cdots, M\}$ where $\boldsymbol{\theta}_{\rm tr}$ is the set of parameters which models the transition probabilities and $\boldsymbol{\theta}_{\rm em}^{(m)} = (A_0^{(m)}, A_1^{(m)}, \Sigma^{(m)})$ are the parameters of the autoregressive process of the regime m. It is easy to see that function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_\ell)$ can then be written as the sum of $M+1$ functions: one depending only on the transition parameters $\boldsymbol{\theta}_{\rm tr}$, and the others depending respectively on the parameters of each regime $\boldsymbol{\theta}_{\rm em}^{(m)}$, $m = 1, \cdots, M$. Indeed, the intermediate function Q has the following convenient decomposition

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_\ell) = \sum_{m,m'=1}^{M} \sum_{k=2}^{n} p_{\boldsymbol{\theta}_\ell}(s_k = m', s_{k-1} = m|\mathbf{y}_0, \cdots, \mathbf{y}_n) \log p_{\boldsymbol{\theta}_{\rm tr}}(s_k|s_{k-1}, \mathbf{y}_{k-1})$$
$$+ \sum_{m=1}^{M} \sum_{k=2}^{n} p_{\boldsymbol{\theta}_\ell}(s_k = m|\mathbf{y}_0, \cdots, \mathbf{y}_n) \log p_{\boldsymbol{\theta}_{\rm em}^{(m)}}(\mathbf{y}_k|\mathbf{y}_{k-1})$$

where the term related to the initial distribution of the Markov chain has been omitted for simplicity reasons. This decomposition permits to solve $M+1$ separate optimization problems on spaces with reduced dimension which is far more efficient than maximizing directly over all parameters. The estimation of $\boldsymbol{\theta}_{\rm tr}$ requires numerical optimization when transitions probabilities are non homogeneous but analytical expressions exist in the homogeneous case.

It is straighforward to verify that, for autoregressive models of order 1,

$$\begin{aligned}&\max_{\boldsymbol{\theta}_{\text{em}}^{(m)}} \left\{ \sum_{k=2}^{n} p_{\boldsymbol{\theta}_\ell}(s_k = m | \mathbf{y}_{(0)}, \cdots, \mathbf{y}_n) \log p_{\boldsymbol{\theta}_{\text{em}}^{(m)}}(\mathbf{y}_k | \mathbf{y}_{k-1}) \right\} \\ &= \max_{(A^{(m)}, \Omega^{(m)})} \left\{ \log\det \Omega^{(m)} - \text{trace}\left(S_{A^{(m)}} \Omega^{(m)} \right) \right\}\end{aligned}$$

where $S_{A^{(m)}}$ is the empirical innovation covariance of regime m.

$$S_{A^{(m)}} = \frac{\sum_{k=2}^{n} p_{\boldsymbol{\theta}_\ell}(s_k = m | \mathbf{y}_{(-p+1)}, \cdots, \mathbf{y}_n) \left(\mathbf{y}_k - \left(A_0^{(m)} - A_1^{(m)} \mathbf{y}_{k-1} \right) \right) \left(\mathbf{y}_k - \left(A_0^{(m)} - A_1^{(m)} \mathbf{y}_{k-1} \right) \right)^T}{\sum_{k=2}^{n} p_{\boldsymbol{\theta}_\ell}(s_k = m | \mathbf{y}_{(-p+1)}, \cdots, \mathbf{y}_n)}.$$

In this estimator, each observation is weighted by its smoothing probability $p_{\boldsymbol{\theta}_\ell}(s_k = m | \mathbf{y}_1 \cdots, \mathbf{y}_n)$ to be in regime m.

The smoothing probabilities $p_{\boldsymbol{\theta}_\ell}(s_k = m | \mathbf{y}_{(-p+1)}, \cdots, \mathbf{y}_n)$ and $p_{\boldsymbol{\theta}_\ell}(s_k = m', s_{k-1} = m | \mathbf{y}_{(-p+1)}, \cdots, \mathbf{y}_n)$ are computed using the so-called forward-backward recursions (see e.g. (Dempster et al., 1977) and references therein).

3.2. EM algorithm for maximizing the penalized likelihood function

Following (Green, 1990), for penalized likelihood estimators, the EM algorithm can be used substituting problem

$$\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} | \boldsymbol{\theta}_\ell) - \mathfrak{P}_\lambda(\boldsymbol{\theta}) \tag{3}$$

to $\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} | \boldsymbol{\theta}_\ell)$ in each M step.

For MSVAR models with Gaussian innovations in each regime, each sub-problem of the M step of the EM algorithm has the following formulation

$$\begin{aligned}\max_{\boldsymbol{\theta}_{em}^{(m)} = (A^{(m)}, \Omega^{(m)})} \Big\{ &\log\det \Omega^{(m)} - \text{trace}\left(S_{A^{(m)}} \Omega^{(m)} \right) \\ &- \sum_{i,j=1}^{d} \mathfrak{p}_{\lambda_m}(|a_{ij}^{(m)}|) - \sum_{i,j=1, i \neq j}^{d} \mathfrak{p}_{\lambda'_m}(|\omega_{ij}^{(m)}|) \Big\}. \end{aligned} \tag{4}$$

with $a_{ij}^{(m)}$ the entries of matrix $A_1^{(m)}$ and $\omega_{ij}^{(m)}$ the entries of $\Omega^{(m)} = \left(\Sigma^{(m)} \right)^{-1}$.

Note that we choose to penalize the off-diagonal coefficients of the precision matrix instead of the covariance matrix. It is generally convenient to focus on the precision matrix when there is colinearity between the variables because in such situation the precision matrix tends to be sparse. It is the case in the application considered in Section 5.

We proceed an iterative optimization of (4) in two stages described hereafter. A direct consequence is that the dimension of the search space is reduced because we solve two problems, the first one in dimension $d(d+1)/2$ (number of free parameters in $\Omega^{(m)}$) and the second one in d^2

(number of free parameters in $A^{(m)}$), instead one problem in $d(3d+1)/2$. And it allows to use efficient algorithms for each stage. In the first stage, for fixed $A^{(m)}$, we optimize

$$\max_{\Omega^{(m)}} \left\{ \log\det\Omega^{(m)} - \text{trace}\left(S_{A^{(m)}}\Omega^{(m)}\right) - \sum_{i,j=1,i\neq j}^{d} \mathfrak{p}_{\lambda'_m}(|\omega_{ij}^{(m)}|) \right\}. \tag{5}$$

The penalized likelihood is non differentiable at the origin and non concave with respect to $\Omega^{(m)}$. To solve this problem, Zou and Li (Zou and Li, 2008) propose an unified algorithm based on a local linear approximation of the penalty function (6). Suppose that we are given an initial value Ω_0 that is close to the true value of $\Omega^{(m)}$, then one can approximate $\mathfrak{p}_\lambda(|\omega_{ij}^{(m)}|)$ by its tangent at Ω_0

$$\mathfrak{p}_{\lambda'}(|\omega_{ij}^{(m)}|) \approx \mathfrak{p}_{\lambda'}(|\omega_{0,ij}|) + \mathfrak{p}'_{\lambda'}(|\omega_{0,ij}|)(|\omega_{ij}^{(m)}| - |\omega_{0,ij}|). \tag{6}$$

Let us remark, when one subtitutes $\mathfrak{p}_\lambda(|\omega_{ij}^{(m)}|)$ by its approximation (6) in (5), the terms involving $\omega_{0,ij}$ act as constants and do not change the maximum. The maximization of the penalized likelihood can then be carried out by an iterative algorithm. One repeatedly solves

$$\Omega_\ell^{(m)} = \arg\max_{\Omega^{(m)}} \left\{ \log\det\Omega^{(m)} - \text{trace}\left(S_{A^{(m)}}\Omega^{(m)}\right) - \sum_{i,j=1,i\neq j}^{d} p'_{\lambda'_m}(|\omega_{\ell-1,ij}^{(m)}|)|\omega_{ij}^{(m)}| \right\} \tag{7}$$

where $\omega_{\ell-1,ij}^{(m)}$ are the entries of the estimation of $\Omega^{(m)}$ at iteration $\ell-1$. It has been demonstrated in (Zou and Li, 2008) that the algorithm converges to a local maxima. Furthermore, at each iteration, (7) is a L^1 penalized problem which can be efficiently solved using the graphical lasso algorithm proposed by Friedman et al. (Friedman et al., 2008). This algorithm uses a blockwise descent procedure and it is very efficient. It optimizes the target function with respect to a small block of parameters at a time, iteratively cycling through all blocks until convergence is reached.

At the first iteration of the EM algorithm, the initial value $\Omega_0^{(m)}$ is set to the unpenalized likelihood estimate. For the next iterations, $\Omega_0^{(m)}$ is set to the penalized estimation obtained at the previous EM iteration.

We can note that when $\lambda'_m = 0$, (5) reduces to $\max_{\Omega^{(m)}} \left\{\log\det\Omega^{(m)} - \text{trace}\left(S_{A^{(m)}}\Omega^{(m)}\right)\right\}$ and this problem admits a closed form solution.

In the second stage, for fixed $\Omega^{(m)}$, we optimize

$$\max_{A^{(m)}} \left\{ -\text{trace}\left(S_{A^{(m)}}\Omega^{(m)}\right) - \sum_{i,j=1}^{d} \mathfrak{p}_{\lambda_m}(|a_{ij}^{(m)}|) \right\}. \tag{8}$$

When $\lambda_m = 0$ this problem admits a closed form solution but otherwise a numerical optimization procedure needs to be used. This function is differentiable except at the origin but is non concave with respect to $A^{(m)}$. Different algorithms were proposed to approximate the solution

of such a problem.An attractive approach is to use the coordinate descent algorithm proposed by Breheny and Huang (Breheny and Huang, 2011). Coordinate descent algorithms optimize a target function with respect to a single parameter at a time, iteratively cycling through all parameters until convergence is reached. The main advantage of reducing the maximization problem to sub-problems of dimension one, is that one can calculate an explicit solution of each sub-problem. Breheny and Huang's algorithm is easy to be plugged in the EM algorithm for MSVAR models since one just needs to write a weighted version of the explicit solutions which takes into account of the smoothing probabilities $p(s_k|\mathbf{y}_{(-p+1)}, \cdots, \mathbf{y_n})$. However, simulation results (see Section 4) show that the performances of this algorithm are not as good as the one based on a quadratic approximation of the penalized likelihood. Here, we follow the trick proposed in (Fan and Li, 2001) which consists in approximating the SCAD penalty by a quadratic function.

$$\mathfrak{p}_\lambda(|\theta_j|) \approx \mathfrak{p}_\lambda(|\theta_j^{(0)}|) + \frac{\mathfrak{p}'_\lambda(|\theta_j^{(0)}|)}{|\theta_j^{(0)}|}\left(\theta_j^{(0)}(\theta_j - \theta_j^{(0)}) + \frac{1}{2}(\theta_j - \theta_j^{(0)})^2\right).$$

Substituting $\mathfrak{p}_\lambda$ by its approximation in (8) leads to a quadratic maximization problem which can be solved by a Newton-Raphson algorithm. At each step of the Newton-Raphson algorithm $\theta_j^{(0)}$ is set to the current value of the parameter. The quadratic approximation is a concave minoration of the penalty and it is equal to the penalty at the current value $\boldsymbol{\theta}_\ell$ of the parameter. As a consequence, if the new parameter increases the log likelihood penalized by the approximation, it will increase the original penalized log likelihood too.

Furthermore, at each step of the descent algorithm, the small valued a_{ij} coefficients are set equal to zero and the update is performed only for the non zero coefficients. The main drawback of the procedure is that, once a coefficient is set to zero it can not change any more. We have checked on simulated data that the algorithm has a good behaviour (see Section 4).

Let us denote $A_\ell = A_\ell^{(m)}$ the parameter at step ℓ of the Newton-Raphson algorithm and $\lambda = \lambda_m$. The update is given by

$$\begin{cases} (a_{ij})_{\ell+1} = 0 \text{ if } |(a_{ij})_\ell| < \lambda \\ (A_{\ell+1})_{[n.z.]} = (A_\ell)_{[n.z.]} + (H_\ell + D_{2,\ell})^{-1}_{[n.z.,n.z.]} (G_\ell + D_{1,\ell})_{[n.z.]} \end{cases} \tag{9}$$

where $(a_{ij})_\ell$ are the entries of matrix A_ℓ. $(B)_{[n.z.]}$ (resp. $(B)_{[n.z.,n.z.]}$) is the vector (resp. matrix) of non zero coefficients of matrix B. G_ℓ and H_ℓ are respectively the gradient and the hessian of the non penalized objective function trace $\left(S_A\,(\Sigma)^{-1}\right)$ computed for A_ℓ. $D_{1,\ell}$ are matrices such that

$$(D_{1,\ell})_{ij} = \sum_{i,j=1}^{d} \mathfrak{p}'_\lambda(|(a_{ij})_\ell|) \text{ for } i,j = 1, \cdots, d \tag{10}$$

and $D_{2,\ell}$ is a diagonal matrix of size $d^2 \times d^2$ with elements $(D_{1,\ell})_{ij} / |(a_{ij})_\ell|$ on its diagonal.

In practice the penalized EM algorithm is initialized with the estimations obtained by the maximization of the non penalized likelihood. The initialization of the non penalized maximization algorithm is described in Section 4.

There exist several methods to select the penalization constants. Here we decided to simply use BIC. BIC is computed for a grid of penalization constants and the M-uplet of penalization constants corresponding to the smallest BIC is selected. The method has been validated on simulated data. Its main drawback is the discretization of the research space. A proper optimization algorithm run for this problem would be very time expensive.

4. Some simulation results

Asymptotic properties of the proposed estimators have been discussed in Section 3. A simulation study is performed in order to validate the proposed estimators on samples of finite length. The model considered for simulation is inspired from the real data of Section 5. It is a sparse MSVAR of order 1 with 2 regimes, the observations are in $\mathbb{R}^{12}$. The Markov chain is homogeneous. Its transition matrix has values 0.76 and 0.84 on its diagonal. The chosen parameters (autoregressive matrices, precision matrices of the innovation) are shown on Figure 3 (columns 2 and 4). The autoregressive matrices are sparse and the precision matrices have zero coefficients organized by blocks. In order to have similar conditions as the ones of Section 5, we generate 55 independent replications of time series of length 31. Parameters are successively estimated for 100 different samples.

A well known limitation of the EM algorithm is that it may converge to local maxima so that it needs a careful initialization. In a MSVAR framework, the initialization step can take advantage of the nested nature of the models. The following heuristic strategy, which seems to give good results for the data set considered in this paper, has been used hereafter. One starts with the fit of an univariate model for each component of the multivariate process. Each univariate model is itself initialized as follows: a classification is performed by kmeans on the one time step difference of the observations and an autoregressive model is fitted in each class. In the second step, a multivariate model with diagonal A and Σ is fitted, initialized with the univariate models. This diagonal model is used to initialize a third model which is a saturated multivariate model in which all the coefficients of A and Σ matrices are free. The next step consists in maximizing the SCAD penalized likelihood to obtain the model referred to as sparse model. Two algorithms introduced in the previous section are compared. The first one is based on a quadratic approximation of the penalized likelihood coupled with a Newton-Raphson algorithm (see Section 3). The second one is an adpatation for MSVAR models of the component wise descent algorithm of Breheny and Huang (Breheny and Huang, 2011).

In order to quantify the performances of the algorithms, one computes two types of criteria. The first one is the Frobenius norm between the true parameters and the estimations. This index can be interpreted as a root mean square error. The second one computes the mean percentage of "true" zeros and "false" zeros over 100 samples. It allows to evaluate if the estimators set to zero the right coefficients of the matrices. In practice, one computes, for each simulated sample, the number of zero parameters the estimators have found and the number of non zero parameters which have been set to zero. Then, one computes the mean of these numbers over the 100 samples. The results are reported in Table 1.

One observes that the sparse estimator based on a quadratic approximation of the penalty leads to smaller error than the saturated estimator. One also remarks that the estimations of the sparse model with quadratic approximation are the best, except for $A^{(1)}$, the autoregressive

matrix of the first regime. The algorithm based on a quadratic approximation allows to better detect the zeros than the component wise descent algorithm. And the estimation error are also globally better. It suggest to prefer the quadratic approximation for applications with real data. The parameters of the second regime are better estimated than the ones of the first regime: more zero coefficients are found and the estimation error is lower. The transition matrix is such that the Markov chain spends 2/3 of the time in the second regime that exhibit more time variability than regime 1. This can explain the better estimations in this regime.

Model (algorithm)	Indices	$A^{(1)}$	$A^{(2)}$	$\Omega^{(1)}$	$\Omega^{(2)}$
Saturated	Error	**0.95**	0.66	1.07	0.27
		[0.68,1.32]	*[0.53,0.83]*	*[0.73,1.56]*	*[0.2, 0.37]*
Sparse (Q)	Error	0.97	**0.42**	**1.00**	**0.23**
		[0.51,1.59]	*[0.29,0.60]*	*[0.65,1.38]*	*[0.15,0.33]*
	% "True" zeros	**97**	**99**	**85**	85
		[89,100]	*[96,100]*	*[70,98]*	*[78, 95]*
	% "False" zeros	**27**	**25**	**19**	**09**
		[11,45]	*[11,34]*	*[00,42]*	*[00, 25]*
Sparse (CW)	Error	1.84	0.82	2.38	0.32
		[1.60, 2.04]	*[0.69,1.05]*	*[1.69,3.17]*	*[0.20,0.60]*
	% "True" zeros	85	87	81	**86**
		[81, 89]	*[83, 92]*	*[72,91]*	*[78, 92]*
	% "False" zeros	55	26	29	12
		[49,64]	*[78,92]*	*[13,47]*	*[0, 25]*

Table 1: Performances of saturated and sparse estimators. Q holds for quadratic approximation and CW for the component wise descent algorithm. The bold values correspond to the algorithm which gives the best results.

5. Daily temperature time series

The models and methods described in the previous section are now tested and discussed on daily mean temperature time series.

Energy consumption as well as crop yields depend on air temperature. For global management of energy or agricultural resources, a good knowledge of simultaneous temperature variability at some chosen locations is needed. Weather generators are useful for that. Modelling of daily temperature time series has been studied for application in weather derivatives markets (Šaltytė Benth and Benth, 2012) (Cao et al., 2015) and for stochastic weather generators (Wilks and Wilby, 1999). The temperature is typically modelled as a sum of a mean process which describes the local climate dynamic and a residual process which captures small variability (see (Kleiber et al., 2013)). The mean process is itself the sum of a deterministic season and an autoregressive process. The residual process is a spatial Gaussian process. For weather derivatives markets application, its standard deviation is written as the product of a deterministic season and a GARCH process (Campbell and Diebold, 2011) while in case of the weather generators, full spatial models are used and the modelling effort mainly focuses on the covariance of the residual process (Kleiber, 2016). In the context of stochastic weather generators, it has been

found that the introduction of regimes corresponding to typical weather patterns may help to better capture the non linearities of the meteorological variables (see (Ailliot et al., 2015a) and references therein). The air temperature can be modelled as a regime switching VAR model with observed regimes specified by the observed precipitations. It leads to a VAR model for dry days and an other VAR model for wet days (Richardson, 1981). The MSVAR model is more general since the regimes are not observed (see (Bessac et al., 2016) for a comparison of models with observed and latent regimes for wind data).

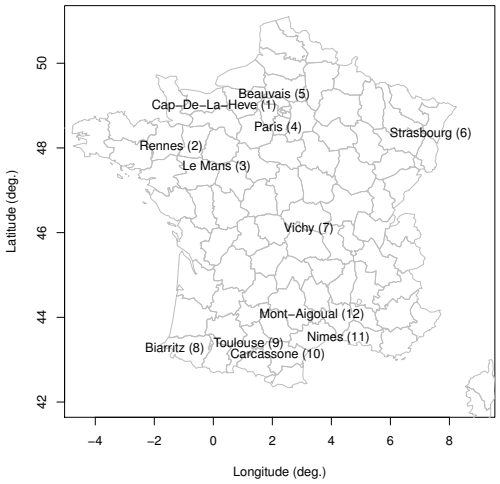

Figure 1: Considered stations. The names are centered on the location of the stations.

We focus on 12 locations in France (see map of Figure 1). These locations have been chosen because they provide long time series and they are distributed over all the country. Some tests have been performed with more stations and other datasets and the results were similar. But full spatial model may be needed for dense station network. Data are extracted from the European Climate Assessment & Dataset. They can be freely downloaded and used for scientific purposes at the URL: `http://eca.knmi.nl/dailydata/index.php`. We kept the years of data with only isolated missing data so that one has 55 multivariate sequences. Isolated missing data have been imputed by linear interpolation of the data at the same location and the nearest dates. The time series are not stationary. There is a weak increasing trend in the data which is neglected here. The temperature time series present a strong seasonal component with colder temperatures and higher temporal variability in winter than in summer. There are several approaches to treat such non stationarity (Ailliot and Monbet, 2012). Here, we decided to block the data month by month and we focus on January month. Finally we have 55 independent time series of length 31 and dimension 12. Examples of sequences are shown on Figure 2.

MSVAR models of order 1 have been fitted to the data. Models with higher order have been tried but no significant improvement was found for the corresponding weather generators (results not shown). A similar remark holds true for the number of regimes: we only consider models with 2 regimes. A partition into only 2 regimes may seem too simple. But keep in mind that our

goal is to build a weather generator and we found that a finer partition did not help to better reproduce the statistical features of the data. Furthermore the number of parameters involved in the models grows quickly with the number of regimes.

The regime is hidden and, when the transition kernel is homogeneous, it is not guided by any meteorological covariates. In this case, the switches between the different regimes may be not coherent with the observed phenomena. It usually gives more (physical) meaning to consider non homogeneous transition probabilities for the Markov chain (Hughes et al., 1999). Several approaches are possible. Here we let the transition probabilities vary with the observation at a chosen location. The idea is that the probability to switch from a hot to a cold regime is higher when the temperature at the reference site is low than when the temperature is high. Several choices are possible to describe the transition probability functions of the Markov chain (Ailliot et al., 2015b). We have chosen a logistic shape

$$p\left(s_k|s_{k-1}, y_{k-1}(j^*)\right) = \frac{\exp\left(\beta_0^{(s_k)} + \beta_1^{(s_k)} y_{k-1}(j^*)\right)}{1 + \exp\left(\beta_0^{(s_k)} + \beta_1^{(s_k)} y_{k-1}(j^*)\right)},\ s_k \in \{1, 2\} \tag{11}$$

where $y_{k-1}(j^*)$ is the temperature observed at one given location and $(\beta_0^{(s)}, \beta_1^{(s)}) \in \mathbb{R}^2$ for all $s \in \{1, 2\}$. In the sequel, the reference station j^* is Cap de la Hève which is located North West. Similar results were obtained with other stations. We could also use other covariates such as the spatial mean or large scale forcing instead of a reference station. However we have done various tests and the results were not significantly different.

The general MSVAR model includes many particular cases. Bayes information criteria (BIC) are computed to compare some of them and are reported on Table 2 with the number of non zero parameters. Table 2 shows that the shrinkage methods are efficient in reducing the number of parameters. It also shows that sparse models have smaller BIC than the saturated models. It is mainly due to the lower number of parameters: any sparse model is a constrained version of its saturated counterpart and its log-likelihood is therefore slightly larger. Figure 3 shows which coefficients are shrunk to 0 in the autoregressive matrices and in the precision matrices.

M	Diagonal	Saturated	Sparse	Saturated, N.H.	Sparse, N.H.
1	90761 (36)	73958 (234)	72984 (100)	-	-
2	88483 (74)	74149 (470)	72119 (195)	71778 (474)	69771 (178)

Table 2: BIC for MSVAR models with 1 and 2 regimes together with number of non zero parameters in brackets. "N.H." referred to as non homogeneous models.

The non homogeneous models (N.H.) have lower BIC than the homogeneous models because the finer modelling of the transition probabilities leads to an improvement of the prediction of the switches.

Hereafter we focus on the sparse non homogeneous model which is selected by BIC. Left panel of Figure 2 shows the temperature time series of the 12 stations for the January month of 2011 as well as the succession of the 2 regimes materialized by white (regime 1) and gray (regime 2) boxes. Regime s_k at time k is defined as the one that maximizing the smoothing probability

$$s_k = \arg \max_{s_k \in \{1, \cdots, M\}} P\left(S_k = s_k | \mathbf{y}_1 \cdots, \mathbf{y}_n\right). \tag{12}$$

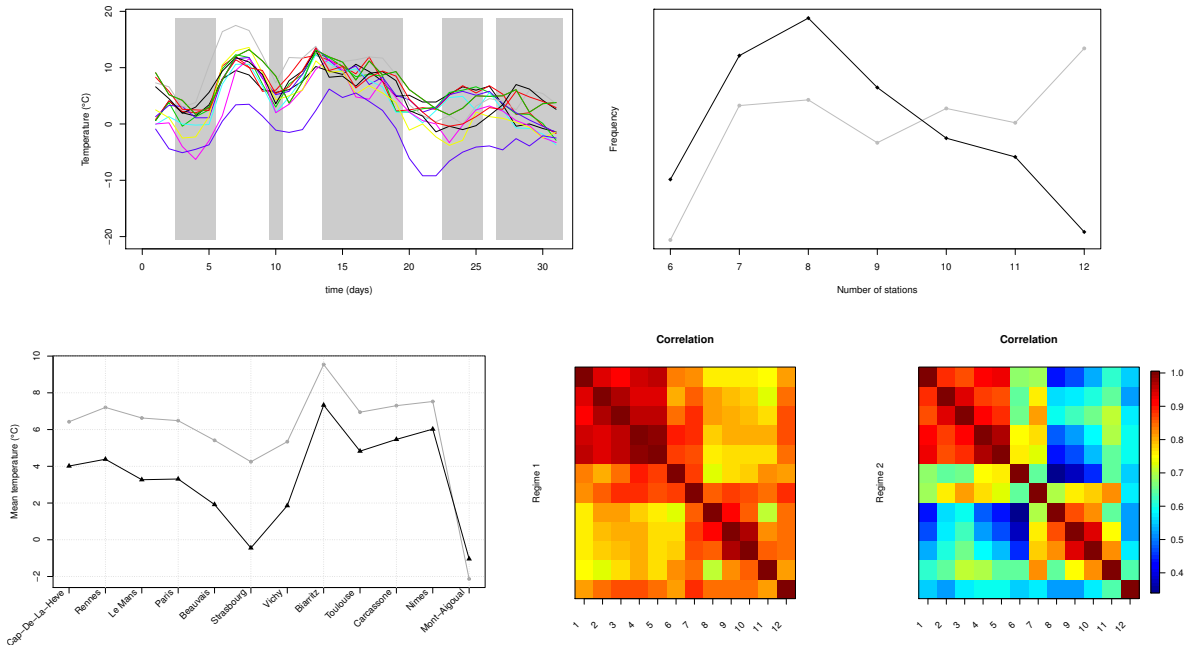

Figure 2: Left panel : Time series for January 2007 with regimes highlighted by white and gray boxes (top left panel). Regimes are infered by local decoding (see Eq. (12)). Number of stations having the same sign for $y_t - y_{t-1}$ (top right panel). Empirical mean temperature in each station for both regimes (bottom left panel). Empirical spatial correlation for each regime of the non homogeneous sparse MSVAR model (bottom right panel). The regimes are the ones of the non homogeneous sparse MSVAR. The darker color correspond to the 2nd regime.

This plot well illustrates the strong correlation between the 12 components of the multivariate time series. The first regime seems to be slightly less persistent than the second one. The bottom left panel shows the mean temperature at each station and in each regime. The mean temperature of Mont-Aigoual is particularly low in both regimes. This station is located at an altitude of about 2000 m and it is known to be very specific. Biarritz, Toulouse, Carcasonne and Nimes are located in the South of France and they benefit on milder temperatures. Strasbourg and Vichy are in the East of France and are submitted to a more continental climate. The first regime is clearly associated with higher temperatures than the second one. The right plot of top panel illustrates the spatial dispersion of the temperature dynamic. The empirical distribution of $N_k = \max(I_k, D_k)$ is computed, I_k (resp. D_k) is the number of stations where the temperature has increased (resp. decreased) between time $k-1$ and time k. In the regime 1, most of the time, more temperature time series evolve similarly than in regime 2. That means that it is a large scale regime in which the variation of temperature is similar over the whole country. The autoregressive matrix of Regime 1 (see Figure 3) shows that the temperature of all stations is mainly driven by the temperature at Cap de la Hève, Rennes and Biarritz (stations/columns 1, 2 and 8) which are located on the west part of France, close to the ocean. In Regime 2, the stations have less influence to each others and the autoregressive matrix of the sparse model is more band diagonal. The empirical spatial correlation per regime (Fig. 2) and the precision matrices (Fig. 3) confirm that the first regime is associated to larger scale phenomena than Regime 2. In summary, regime 1 represents an oceanic large scale influence while regime 2

corresponds to more local conditions.

Let us now look at some performances of the model as a stochastic weather generator. The stochastic weather generator is expected to reproduce the marginal distribution, the dependence structure of order two and more complex statistics such as the intensity of up-crossings. In practice, one simulates 100 times 55 sequences of 1 month from the non homogeneous sparse MSVAR(1) model with two regimes and compare their statistics to the ones of the observations.

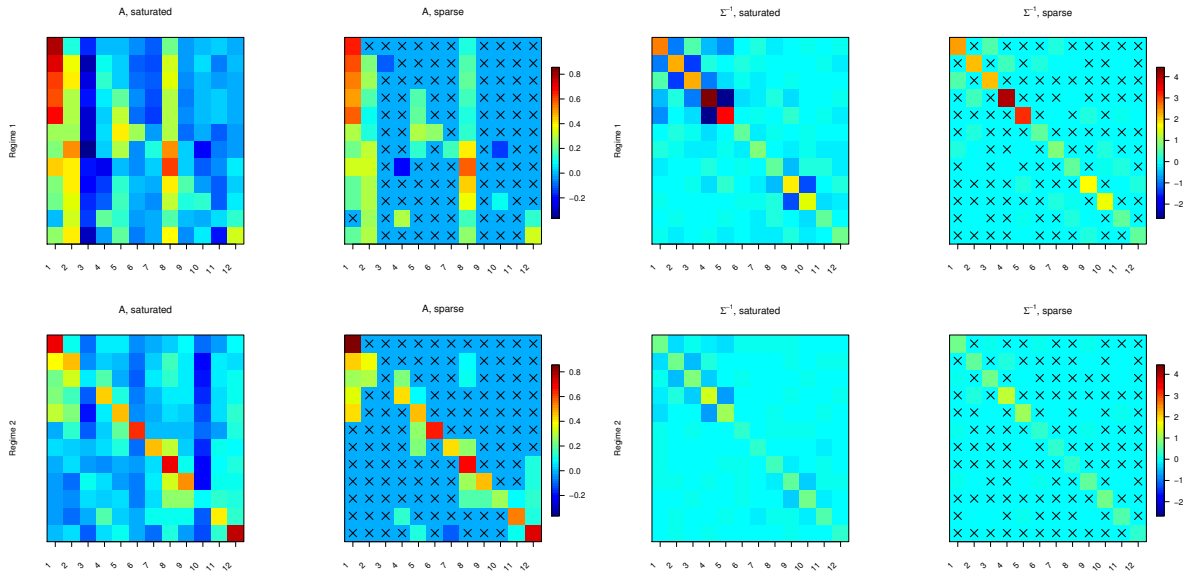

Figure 3: Autoregressive matrices (left panel) and precision matrices (right panel). For each regime the matrices of the saturated and sparse non homogeneous models are depicted. The crossed squares correspond to 0 coefficients. Results for January. The order of the lines of the matrices corresponds to the numbers given on the map of Fig. 1 so that Mont-Aigoual corresponds to the last line for example.

The quantile-quantile plots of Figure 4 compare how several sparse models reproduce the marginal distribution of the data, namely the VAR model (M=1), the homogeneous MSVAR model (M=2) and the non homogeneous MSVAR model (M=2). MSVAR models does better than the VAR model and the non homogeneous MSVAR model is the best. However none of the model captures correctly the upper tail which is lighter than the Gaussian one. An extreme value analysis shows that the upper tail of the temperature distribution is bounded. The same observation was made for instance in (Ferrez et al., 2011) (Dacunha-Castelle et al., 2015). Although such a result has to be interpreted with care because the rate of convergence to the true domain of attraction can be slow (Gomes, 1984), we think that it still explains why a MSVAR model with Gaussian innovations may not be able to reproduce it. More generally, we found that it is often difficult for MSAR models to well reproduce the marginal distribution. It may be due to the estimation procedure which is based on the maximization of the likelihood. It leads to optimizing a criteria which mainly deals with the one step-ahead forecast error. A solution would be to apply a posteriori marginal transformations and match the distribution of the data.

The plot of up-crossing intensity in Cap de la Hève (see Fig. 4) is representative of what is

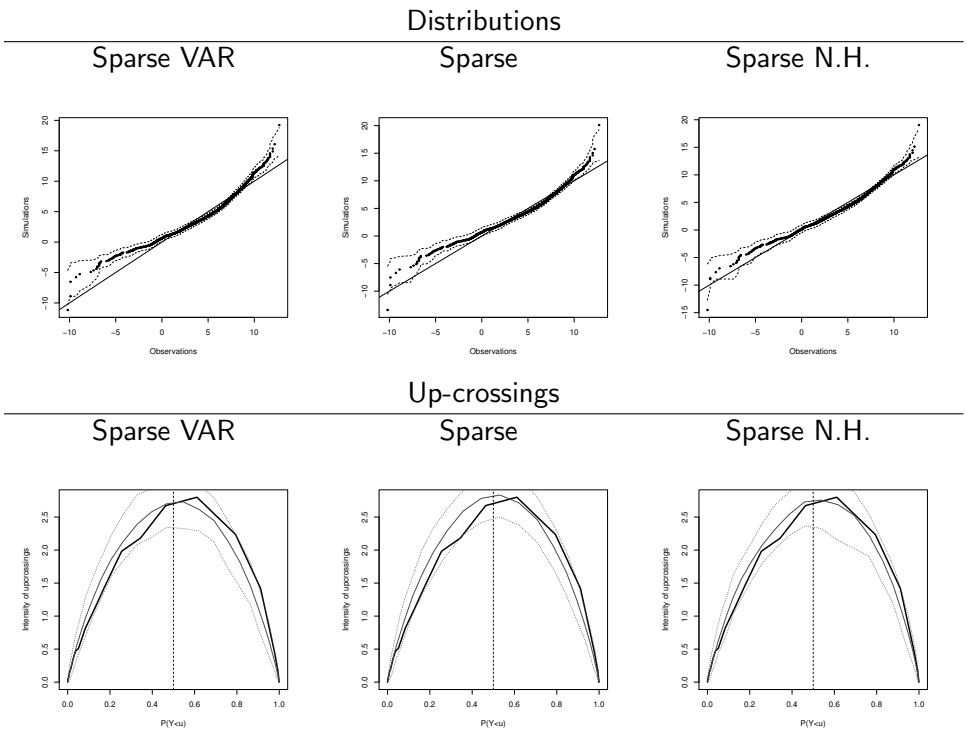

Figure 4: Statistics for January at Cap de la Hève. The bold line corresponds to the data and the thin one to the simulations. The dotted lines materialize a 95% empirical confidence interval. Sparse and Sparse N.H. hold for MSVAR models. For the up-crossings plots, the vertical lines materialize the median.

observed at the other stations. The bold curve corresponds to the data and it is not symmetric with respects to the median. It means that the temperature time series cross the high levels more often than the low ones. For a Gaussian process the intensity of up-crossings is symmetric as it can be seen for the VAR model. The non homogeneous MSVAR model better catches this non linearity than the homogeneous MSVAR model especially for the temperatures which are under the median.

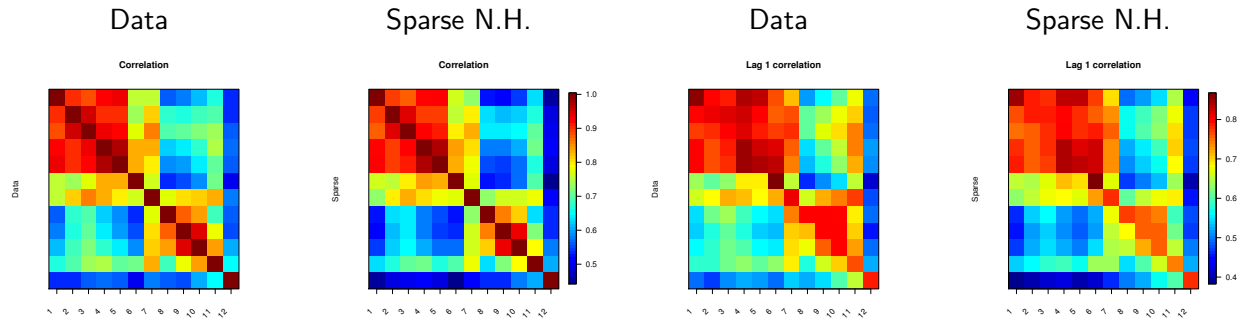

Figure 5: Statistics for January. Space time correlations at lag 0 (left panel) and lag 1 (right panel). Results for the non homogeneous sparse MSVAR model.

Figure 5 shows the lag 0 and lag 1 spatial correlation matrices for the data and the non ho-

mogeneous MSVAR model. The model well captures the space time dependence structure. Iit slightly underestimates the large scale correlation (see for instance columns 9 and 10 of the first line).

As a short summary, we found that the fitted non homogeneous sparse MSVAR model is quite efficient to simulate multisite January temperature time series. In particular, it captures some non linearities induced by the regime switchings. The main drawback of the model is its low ability to reproduce the tails of the marginal distribution. A more global validation has been performed using other datasets. The conclusions were similar.

6. Concluding remarks

Weather type models provide a flexible and interpretable family of models for meteorological time series. Considering temperature time series, we have shown that non homogeneous sparse MSVAR models allow to well reproduce non linearities existing in the data such as non separable covariances or non symmetric up-crossings and that non homogeneous sparse MSVAR models are more convenient than homogeneous models for simulating multisite time series of temperature in France.

The proposed shrinkage significantly decreases the number of parameters in the models and helps for interpretation. One of the most sensible points in the proposed inference algorithm concerns the search of the penalization constants. It is computationally expensive and far from exhaustive. However, the values of the penalization constants mainly impact the number of parameters in the models and not the simulation performances which is the final goal of the considered application. An other way for decreasing the number of parameters would be to find parametric models for the auto-regressive matrices and the covariance matrices. This is proposed in full spatial modelling (Verdin et al., 2015). Such models allow to consider a much larger number of stations and may be more convenient for many applications. But, it may be difficult to find convenient parametric shapes because of the complex space time structure of the data. And the inference would be difficult if latent regimes are introduced in the model.

References

Ailliot, P., Allard, D., Monbet, V., and Naveau, P. (2015a). Stochastic weather generators: an overview of weather type models. *Journal de la Société Française de Statistique*, 156(1):101–113.

Ailliot, P., Bessac, J., Monbet, V., and Pene, F. (2015b). Non-homogeneous hidden markov-switching models for wind time series. *Journal of Statistical Planning and Inference*, 160:75–88.

Ailliot, P. and Monbet, V. (2012). Markov-switching autoregressive models for wind time series. *Environmental Modelling & Software*, 30:92–101.

Ailliot, P., Monbet, V., and Prevosto, M. (2006). An autoregressive model with time-varying coefficients for wind fields. *Environmetrics*, 17(2):107–117.

Ailliot, P. and Pène, F. (2015). Consistency of the maximum likelihood estimate for non-homogeneous Markov-switching models. *ESAIM: PS*, 19:268–292.

Basu, S., Michailidis, G., et al. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, pages 164–171.

Bessac, J., Ailliot, P., Cattiaux, J., and Monbet, V. (2016). Comparison of hidden and observed regime-switching autoregressive models for (u, v)-components of wind fields in the northeast atlantic. *Adv. Stat. Clim. Meteorol. Oceanogr.*, 2(1).

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232.

Campbell, S. D. and Diebold, F. X. (2011). Weather forecasting for weather derivatives. *Journal of the American Statistical Association*.

Cao, X., Okhrin, O., Odening, M., and Ritter, M. (2015). Modelling spatio-temporal variability of temperature. *Computational Statistics*, 30(3):745–766.

Dacunha-Castelle, D., Hoang, T. T. H., and Parey, S. (2015). Modeling of air temperatures: preprocessing and trends, reduced stationary process, extremes, simulation. *Journal de la Société Française de Statistique*, 156(1):138–168.

Dempster, A. P., Laird, N. M., Rubin, D. B., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1):1–38.

Douc, R., Moulines, E., Ryden, T., et al. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics*, 32(5):2254–2304.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Favero, C. A. and Monacelli, T. (2005). Fiscal policy rules and regime (in) stability: evidence from the us.

Ferrez, J., Davison, A., and Rebetez, M. (2011). Extreme temperature analysis under forest cover compared to an open field. *Agricultural and Forest Meteorology*, 151(7):992–1001.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gomes, M. I. (1984). Penultimate limiting forms in extreme value theory. *Annals of the Institute of Statistical Mathematics*, 36(1):71–85.

Green, P. J. (1990). On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 443–452.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384.

Hering, A. S., Kazor, K., and Kleiber, W. (2015). A Markov-switching vector autoregressive stochastic wind generator for multiple spatial and temporal scales. *Resources*, 4(1):70–92.

Hsu, N.-J., Hung, H.-L., and Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis*, 52(7):3645–3657.

Hughes, J. P., Guttorp, P., and Charles, S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(1):15–30.

Kazor, K. and Hering, A. S. (2015). Assessing the performance of model-based clustering methods in multivariate time series with application to identifying regional wind regimes. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(2):192–217.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479).

Kleiber, W. (2016). High resolution simulation of nonstationary gaussian random fields. *Computational Statistics & Data Analysis*.

Kleiber, W., Katz, R. W., Rajagopalan, B., et al. (2013). Daily minimum and maximum temperature simulation over complex terrain. *The Annals of Applied Statistics*, 7(1):588–612.

Krolzig, H.-M. (2013). *Markov-switching vector autoregressions: Modelling, statistical inference, and application to business cycle analysis*, volume 454. Springer Science & Business Media.

Lu, Z.-Q. and Berliner, L. M. (1999). Markov switching time series models with application to a daily runoff series. *Water Resources Research*, 35(2):523–534.

Medeiros, M. C., Mendes, E., et al. (2012). Estimating high-dimensional time series models. *CREATES Research Paper*, 37.

Pinson, P. and Madsen, H. (2012). Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models. *Journal of Forecasting*, 31(4):281–313.

Richardson, C. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1):182–190.

Šaltytė Benth, J. and Benth, F. E. (2012). A critical view on temperature modelling for application in weather derivatives markets. *Energy Economics*, 34(2):592–602.

Sims, C. A. and Zha, T. (2006). Were there regime switches in us monetary policy? *The American Economic Review*, pages 54–81.

Verdin, A., Rajagopalan, B., Kleiber, W., and Katz, R. W. (2015). Coupled stochastic weather generation using spatial and generalized linear models. *Stochastic Environmental Research and Risk Assessment*, 29(2):347–356.

Wilks, D. and Wilby, R. (1999). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, 23(3):329–357.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509.