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

Valérie Monbet, Pierre Ailliot

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

V. Monbet[1] and P. Ailliot[2]

[1]IRMAR, Universit de Rennes 1 & INRIA, Rennes, France
[2] LMBA, Universit de Bretagne Occidentale, Brest, France

January 4, 2016

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 (SCAD) penalization of the likelihood to shrink the parameters. The Expectation Maximization algorithm build for maximizing the penalized likelihood is described in details and tested on daily mean temperature time series.

Key words- Markov Switching Vector Autoregressive process, sparse, penalized likelihood, SCAD, dimension reduction.

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, the 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. The regime is often not observed directly and is thus introduced as a latent process in time series models in the spirit of Hidden Markov models. Similar phenomena are observed in economics time series with business cycles (see [Hamilton, 1989]). Markov Switching autoregressive (MSAR) models have been introduced as a generalization of autoregressive models and Hidden Markov Models. They are widely used for economic time series [Hamilton, 1989] and for meteorological time series (see for instance [Lu and Berliner, 1999] [Pinson and Madsen, 2012] and references therein). And they lead to flexible and interpretable models [Ailliot and Monbet, 2012]. MSAR models are also used in a multivariate context [Pinson and Madsen, 2012].

Markov Switching VAR models (MSVAR) 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. If $M=1$, $p=1$ and $d=12$, the number of parameters is equal to 446, 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 contains 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 statistic, consists in searching for parametric shapes for the autoregressive and covariance matrices. The most usual parametric shapes is 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 make sense to set zeros in the autoregressive matrices and in the precision matrices because they are linked with conditional independence.

Some well known reduction methods require no expertise. The most famous type of methods without any a priori is the information based-methods such as AIC (Akaike, 1974) and BIC (Schwarz, 1978). These criteria can be used for a best subset selection. Though such methods are generally only used to select the order of the model because it would be computationally too expensive to compute the criteria for all possible subsets of variables. In order to remove some coefficients from the model, some authors have proposed top-down or bottom-top strategies. Shortly said, the ideas is to build a research path defining nested models and to select the best models amoung them by AIC or BIC (see [Lütkepohl, 2005], [Krolzig and Hendry, 2001]). Such methods are more time efficient compared to exhaustive search but they are suboptimal. Granger causality and likelihood ratio tests provide an other way to select a model (see for example Anderson (1971) and Tiao and Box (1981)). But such tests generally eliminate the inappropriate models instead of selecting the best ones. Furthermore, it is delicate to find the right level of the test because several tests are usually performed with dependent statistics. More recently some authors have been interested in developing computationally efficient method for VAR selection based on shrinkage methods also known as penalized regression methods or regularization methods. These procedures have become popular since they are computationally feasible and perform variable selection and parameter estimation at the same time. The idea is to shrink to zero the irrelevant parameters. For the autoregressive part of VAR models as for high dimension regression models, the notion of sparsity is incorporated in the estimation procedure by penalization procedures like lasso [Hsu et al., 2008] and its variants [Medeiros et al., 2012] while for covariance matrices hard thresholding as graphical Lasso is used [Friedman et al., 2008] [Bickel and Levina, 2008]. 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 and Michailidis, 2013] 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 as Smoothly Clipped Absolute Deviation (SCAD) penalty.

MSVAR models are presented in Section 2.1 with their inference. The originality of the present

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. Here, it is more natural to focus on the precision matrices because the data exhibit strong colinearities. Both autoregressive matrices and precision matrices are penalized using SCAD penalties. It 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 build for the penalized likelihood is described in Section 3. The performance of the penalized estimators are 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) of daily mean temperatures in France. To the best of our knowledge, such models have never be implemented for temperature data and there exist no SWGEN for multisite temperature time series.

As supplementary material to this paper, algorithms have been implemented in a 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]. They have been then 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 modelled a first order Markov chain $\{S_k\}_{k\in\mathbb{Z}}$ tacking its value 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 $\mathbf{Y}$. 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 $(\mathbf{Y}_{k-1}, \cdots, \mathbf{Y}_{k-p})$.

$$\mathbf{Y}_k = A_0^{(S_k)} + A_1^{(S_k)}\mathbf{Y}_{k-1} + \cdots + A_p^{(S_k)}\mathbf{Y}_{k-p} + \Sigma^{(S_k)}\boldsymbol{\epsilon}_k$$

  where the unknown parameters $A_0^{(s)}, \cdots, A_p^{(s)}$ are matrices of $\mathbb{R}^{d,d}$ and $\Sigma^{(s)}$ are positive matrix of $\mathbb{R}^{d,d}$. $\{\boldsymbol{\epsilon}_k\}_{k\in\mathbb{Z}}$ is a 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, Y_k will describe the temperature at $d = 12$ stations.

The various conditional independence assumptions are summarized by the directed graph below for $p = 1$.

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

Several well known models are included in the introduced MSVAR. Transition probabilities can be independent of the observation: $p(s_k|s_{k-1}, \mathbf{y}_{k-1}) = p(s_k|s_{k-1})$. If $p = 0$, one retrieves the classical Hidden Markov Model. If one has only one regime ($M = 1$), one retrieves the classical VAR model. If $p = 0$ and $M = 1$, one retrieves the Gaussian mixture model.

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 $p_{\boldsymbol{\theta}}(\mathbf{y}_1, \cdots, \mathbf{y}_n)$ 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 likelihood function $p_{\boldsymbol{\theta}}(\mathbf{y}_1, \cdots, \mathbf{y}_n)$ is therefore never directly used. In practice, the function which is maximized depend on the complete likelihood $p_{\boldsymbol{\theta}}(s_1, \cdots, s_n, \mathbf{y}_1 \cdots, \mathbf{y}_n|\mathbf{y}_{-p+1}, \cdots, \mathbf{y}_0)$.

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

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(s_1, \cdots, s_n, \mathbf{y}_1 \cdots, \mathbf{y}_n|\mathbf{y}_{-p+1}, \cdots, \mathbf{y}_0) &= \log p_{\boldsymbol{\theta}}(\mathbf{y}_1, \cdots, \mathbf{y}_n|s_1, \cdots, s_n, \mathbf{y}_{-p+1}, \cdots, \mathbf{y}_0) \\
&\quad + \log p_{\boldsymbol{\theta}}(s_1, \cdots, s_n|\mathbf{y}_{-p+1}, \cdots, \mathbf{y}_0)
\end{aligned}$$

The first term depends on the emission probabilities only while the second one depends on the transition probability. This particular form is exploited to propose efficient algorithms.

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

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\mathbf{y}_1, \cdots, \mathbf{y}_n|s_1, \cdots, s_n, \mathbf{y}_{-p+1}, \cdots, \mathbf{y}_0) &= \sum_{k=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{y}_k|\mathbf{y}_{k-1}, \cdots, \mathbf{y}_{k-p+1}, s_k) \\
&= \sum_{m=1}^{M} \sum_{\{k|s_k=m\}} \log p_{\boldsymbol{\theta}}(\mathbf{y}_k|\mathbf{y}_{k-1}, \cdots, \mathbf{y}_{k-p+1}, s_k = m)
\end{aligned}$$

2.3 Penalized likelihood

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

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

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

A good penalty shoud 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 to model complexity (sparsity property). And the function which associated the penalized estimate to the maximum likelihood estimate should be continuous in order to avoid prediction instability (continuity property).

We chose to work with the smoothly clipped absolute deviation penalty (SCAD). This penalty was introduced by Fan and Li [Fan and Li, 2001] and it verifies the three listed properties for many models. The SCAD penalty is usually defined by its derivative but it may easier to interpret if we give 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 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 with its derivatives 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 constraints are imposed to the λ_j in order that some λ_j are equal to each other.

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 results 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 Pene, 2013] who have demonstrated that non homogeneous MSVAR models verify a property of ergodicity and that the estimator of maximum likelihood is consistent but no CLT. 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 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 $\boldsymbol{\theta}_\kappa$ of the maximum likelihood estimator $\boldsymbol{\theta}$ by cycling through the following steps.

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

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

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

where the term related to the initial distribution of the MC has been ommited for simplicicity reasons. The decomposition of function Q 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 $\boldsymbol{\theta}_{\text{tr}}$ requires numerical optimization when transitions probabilities are non homogeneous.

It is straighforward to verify that, when $p = 1$,

$$\max_{\boldsymbol{\theta}_{\text{em}}^{(m)}} \left\{ \sum_{k=2}^{n} p_{\boldsymbol{\theta}_\kappa}(s_k = m|\mathbf{y}_{(-p+1)}, \cdots, \mathbf{y}_n) \log p(\mathbf{y}_k|\mathbf{y}_{k-1}, \cdots, \mathbf{y}_{k-p+1}; \boldsymbol{\theta}_{\text{em}}^{(m)}) \right\}$$
$$= \max_{(A^{(m)}, \Omega^{(m)})} \left\{ \log\det \Omega^{(m)} - \text{trace}\left(S_{A^{(m)}} \Omega^{(m)} \right) \right\}$$

where $S_{A^{(m)}}$ is the empirical innovation covariance of regime m. In this estimator, each observation is weighted by its smoothing probability $p_{\boldsymbol{\theta}_\kappa}(s_k = m|\mathbf{y}_1 \cdots, \mathbf{y}_n)$ to be in regime m.

$$S_{A^{(m)}} = \frac{\sum_{k=2}^{n} p_{\boldsymbol{\theta}_\kappa}(s_k = m|\mathbf{y}_1 \cdots, \mathbf{y}_n) \left(\mathbf{y}_k - \left(A_0^{(m)} - A_1^{(m)} \mathbf{y}_{k-1} \right) \right) \left(\mathbf{y}_k - \left(A_0^{(m)} - A_1^{(m)} \mathbf{y}_{k-1} \right) \right)^T}{\sum_{k=2}^{n} p_{\boldsymbol{\theta}_\kappa}(s_k = m|\mathbf{y}_1 \cdots, \mathbf{y}_n)}$$

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

The smoothing probabilities $p_{\boldsymbol{\theta}_\kappa}(s_k = m|\mathbf{y}_1 \cdots, \mathbf{y}_n)$ and $p_{\boldsymbol{\theta}_\kappa}(s_k = m', s_{k-1} = m|\mathbf{y}_1 \cdots, \mathbf{y}_n)$ 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

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

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

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

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

We 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 the sequel (Section 5).

We proceed an iterative optimization of (4) in two stages. A direct consequence is that the dimension of the search space is reduced because we solve two problems in $d(d+1)/2$ 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

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

The penalized likelihood is non differentiable at the origine and non concave with respect to $\Omega^{(m)}$. To solve this problem, Zou and Li [Zou and Li, 2008] propose a 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

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

The maximization of the penalized likelihood can then be carried out by an iterative algorithm. One repeatedly solves

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

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 converge to a local maxima. Furthermore, at each iteration, problem (7) is a L^1 penalized problem which can be efficiently solve using the graphical Lasso algorithm proposed by Friedman et al. [Friedman et al., 2008]. This algorithm uses a block wise descent procedure and it is very fast.

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 obtain at the previous EM iteration.

We can note that when $\lambda'_m = 0$, (5) reduces to

$$\max_{\Omega^{(m)}} \left\{ \log\det \Omega^{(m)} - \text{trace}\left(S_{A^{(m)}} \Omega^{(m)} \right) \right\} \tag{8}$$

and this problem admits a closed form solution.

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

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

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. One is based on an linear approximation of the penalty in the spirit of (6). An other way to deal with the autoregressive matrices 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. This algorithm can be plugged in the EM algorithm for MSVAR models since one just need to write a weighted version of a standard log-likelihood which takes into account of the smoothing probabilities $p(s_k|\mathbf{y_1}, \cdots, \mathbf{y_n})$. However one found, on simulated data, that the performances of this algorithm are not as good as the one based on a quadratic approximation of the penalized likelihood (see Section 4). Here, we follow the trick proposed in [Fan and Li, 2001] which consists in approximating the SCAD penalty by a quadratic function at each step of a Newton-Raphson algorithm.

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

This approximation is a concave minoration of the penalty and it is equal to the penalty at the current value $\boldsymbol{\theta}_\kappa$ of the parameter. As a consequence, if the new parameter increases this function it will increase the 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 anymore. We have checked on simulated data that the algorithm still converges to the expected solution (see Section 4).

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

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

where $(a_{ij})_\kappa$ are the entries of matrix A_κ. $(B)_{[n.z.]}$ (resp. $(B)_{[n.z.,n.z.]}$) is the vector (resp. matrix) of non zero coefficients of matrix B. G_κ and H_κ are respectively the gradient and the hessian

of the non penalized objective function trace $\left(S_A\left(\Sigma\right)^{-1}\right)$ computed for A_κ. $D_{1,\kappa}$ are matrices such that

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

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

In practice the penalized EM algorithm is initialized with the estimations obtained by the maximization of the non penalized likelihood. Furthermore, in order to initialize the algorithm for non penalized likelihood with realistic parameters and avoid convergence to non-interesting maxima, we use the nested nature of the MSVAR models as described in Section 4.

There exists 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 transition matrix of the Markov chain has values 0.76 and 0.84 on its diagonal. The chosen parameters (autoregressive matrices, precision matrices of the innovation) are shown on Figure 4 (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.

EM algorithm is well known to 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. One starts with the fit of an univariate model for each station. 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 check the consistence 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 zeros 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. Itjustifies our choice of Section 2.3 to prefer the quadratic approximation. 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 one is 2/3 of the time in the second regime; 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.

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. Our goal is to build a stochastic weather generator [Ailliot et al., 2015].

Energy consumption as well as crop yields depend on air temperature. For global management of energy or agricultural resources it is useful to have a good knowledge of simultaneous temperature variability in some chosen locations. Weather generators are useful for that. As far as we know, multisite temperature time series have seldom been studied and there exists only few statistical models for this variable.

We focus on 12 locations in France (see the map of Fig. 1). These locations have been chosen because they provide long time series and they are distributed over all the country. 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 volatility 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. The 55 years of available data leads to 55 independent realizations of the process in January. Examples of sequences are shown on Figure 2.

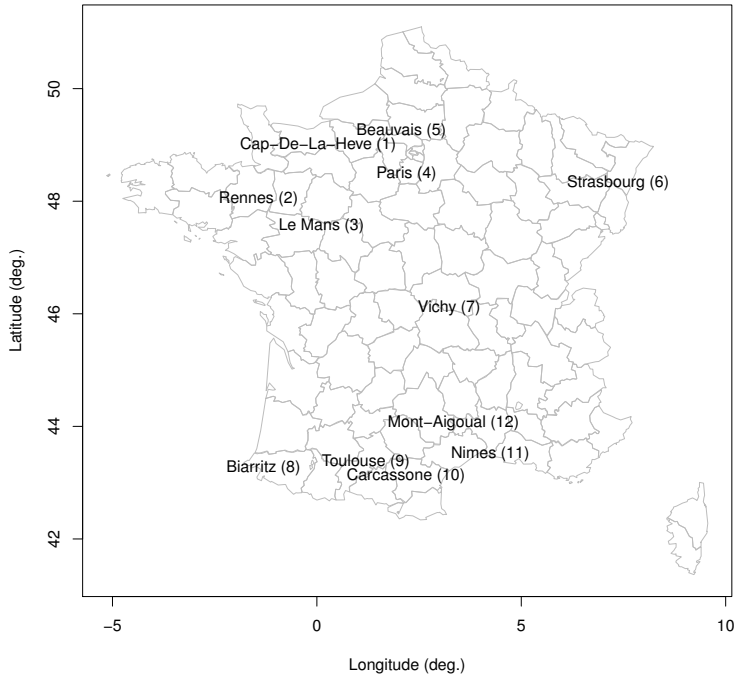

Figure 1: Considered stations. The names are centered on the location of the stations.

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 hold for the number of regimes: considering models with more than 2 regimes does not lead to significant improvements in simulation.

The regime is hidden and, when the transition kernel is homogeneous, it is not guided by any meteorological covariates. In this case, the switch 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 varying 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 described the transition probability functions of the Markov chain. 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.

The general MSVAR model includes many particular cases. Bayes information criteria (BIC) are computed to compare some them and they are reported on Table 2 with the number of non zero parameters. The BIC of MSVAR models is composed of a term, which is directly linked with the likelihood and which decreases when the likelihood increases, and a penalization term which increases with the number of parameters. The likelihood term has itself two components: one which is equivalent to a one step ahead prediction error and the other one which may be interpreted as a conditional entropy and which is linked to the quality of the classification. Table 2 shows that the shrinkage methods are efficient in reducing the number of parameters as it is expected from the sparsity property and the tests done on simulated data. Figure 4 shows which coefficients are shrunk to 0 in the autoregressive matrices and in the precision matrices. 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.

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 and number of 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. The BIC values reported for the saturated models (second column of Table 2) could lead us to select a model with only 1 regime which is a VAR model. But such a model will not be able to reproduce the non linearities observed in the data and its BIC value is just given for comparison.

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. This plot well illustrates the strong correlation between the time series. But it is difficult to make more interesting comments from this plot except that the first regime seems to be lightly less persistent than the second one. Indeed the diagonal of the transition matrix is equal to 0.76 and 0.84. It correspond to mean durations of about 4 and 6 days respectively. The bottom left panel shows the mean temperature in each station and each regime. The means are computed from the parameters of the MSVAR model. The mean temperature of Mont-Aigoual is particularly low in both regimes. These station is located at an altitude of about 2000 m and it is known to be very specific. One can see it on the time series too where the Mont-Aigoual temperature is the lowest (darkest) curve. Biarritz, Toulouse, Carcasonne and Nimes are located in the South of France and they benefit on sweeter 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 bottom panel illustrates the spatial dispersion of the temperature dynamic. One computes the frequency of k stations having the same sign of $y_t - y_{t-1}$. One sees that in the regime 1, most of the time, 11 stations have their temperature evolving similarly. That means that it is a large scale regime in which the variation of temperature is the same in the whole country. The autoregressive matrix of Regime 1 (see Figure 4) 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 auto-regressive matrix of the sparse model is more band diagonal. The empirical spatial correlation per regime (Fig. 3) and the precision matrices (Fig. 4) confirm that the first regime is associated to larger scale phenomena than Regime 2. To summary, regime 1 represents an oceanic large scale influence while regime 2 is characteristic of local influences.

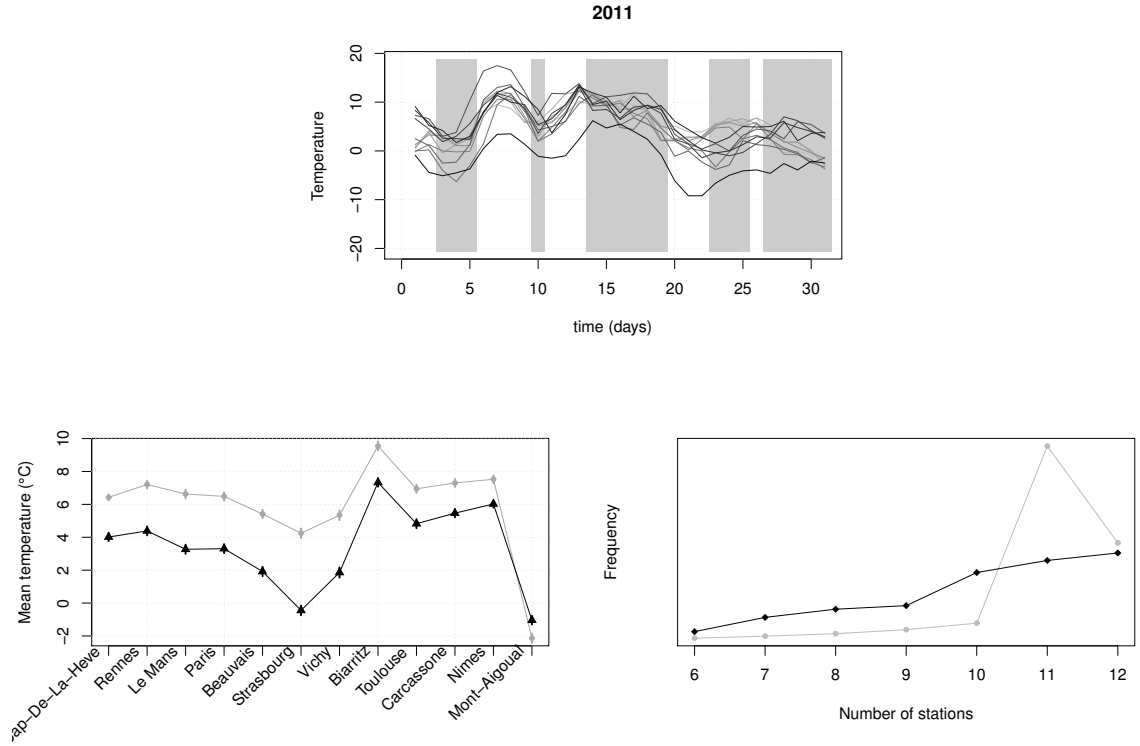

Figure 2: Left panel : Time series for January 2007 with regimes highlighted by white and gray boxes (top panel). Mean temperature in each station for both regimes (bottom left panel). Mean are computed from the parameters of the fitted MSVAR. Number of stations having the same sign for $y_t - y_{t-1}$ (bottom right panel). The darker color correspond to the 2nd regime.

This partition into only 2 regimes may seem simple. But keep in mind that our main 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 parameter involved in the models grows quickly with the number of regimes.

Let us now look at some performances of the 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.

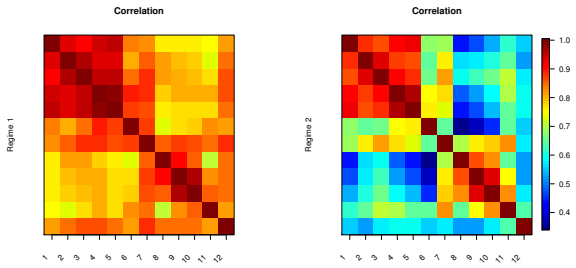

Figure 3: Empirical spatial correlation for each regime.

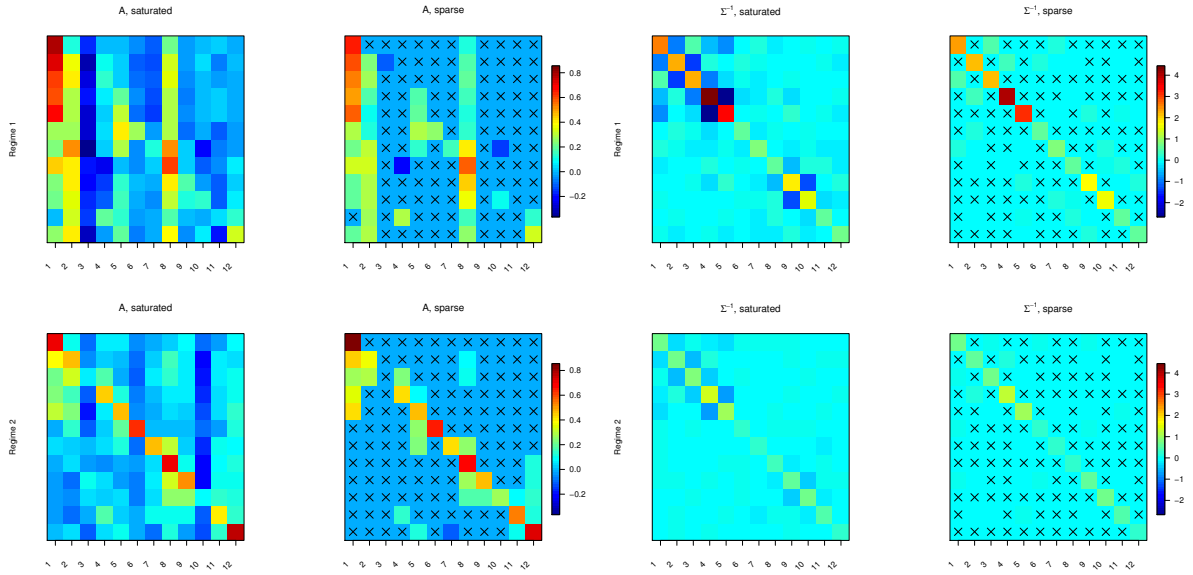

Figure 4: Autoregressive matrices (left panel) and precision matrices (right panel). For each regime the matrices of the saturated and sparse models are depicted. The crossed squares correspond to 0 coefficients. Results for January. The order of the lines of the matrices corresponds to the numbers given on the map of Fig. 1 so that Mont-Aigoual corresponds to the last line for example.

The quantile-quantile plots of Figure 5 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). One observes that MSVAR models do better than the VAR model and the non homogeneous MSVAR model the best. An extreme value analysis shows that the upper tail of the temperature distribution is bounded. It explains why a MSVAR model with Gaussian innovations may not be able to reproduce it. More generally, it well known that it is difficult for MSAR models to well fit the marginal distribution. Indeed, the maximization of the likelihood leads to a cost criteria which mainly deals with the second order dependence structure. The marginal distribution is then only a derivative.

The plot of up-crossing intensity in Cap de la Hève is representative of what is observed for 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. It is a well known non linear behaviour. Indeed for a Gaussian process the intensity of up-crossings is symmetric as it can be seen for the VAR model. One observes that the non

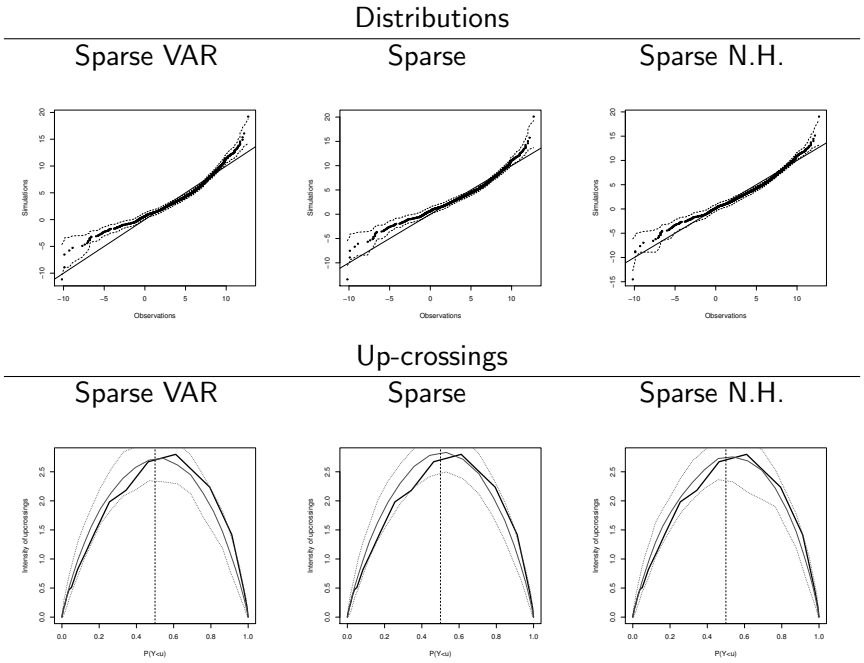

Figure 5: Statistics for January at Cap de la Hève. The bold line corresponds to the data and the thin one to the simulations. The dotted lines materialize a 95% empirical confidence interval. Sparse and Sparse N.H. hold for MSVAR models. For the up-crossings plots, the vertical line materializes the median.

homogeneous MSVAR model better catches this non linearity than the homogeneous MSVAR model especially for the temperatures which are under the median.

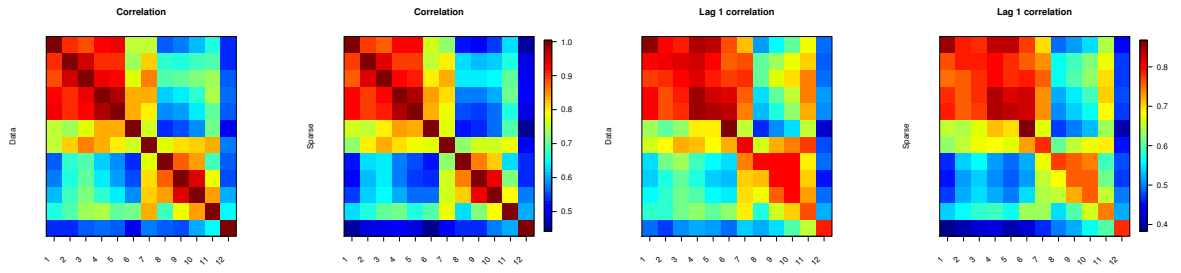

Figure 6: Statistics for January. Space time correlations at lag 0 (left panel) and lag 1 (right panel). Results for the non homogeneous MSVAR model.

Figure 6 shows the lag 0 and lag 1 spatial correlation matrices for the data and the non homogeneous MSVAR model. The model well catch the space time dependence structure. One observes that it slightly under estimate the large scale correlation (see for instance columns 9 and 10 of the first line).

As a short resume, one could say that the fitted non homogeneous sparse MSVAR model is quite efficient to simulate multisite January temperature time series. In particular, it captures the non linearities existing in the data. The main drawback of the model is its low ability to reproduce the marginal distribution tails. A more global validation has been performed including data from other months. The conclusions was similar as the ones of January.

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 allows to well reproduce non linearities existing in the data such as non separable covariances or non symmetric up-crossing. We have exhibit 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 decrease the number of parameters in the models and may help for interpretation. The simulation performances of the sparse models are equivalent to the one of the full models. One of the most sensible point 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. An other way for decreasing the number of parameters would be to find parametric models for the auto-regressive matrices and the covariance matrices. It would be more drastic. But, it is difficult to find convenient parametric shape because of the complex space time structure of the data (see for instance the autoregressive matrices).

References

[Ailliot et al., 2015] Ailliot, P., Allard, D., Monbet, V., and Naveau, P. (2015). Stochastic weather generators : an overview of weather type models. *Journal de la SFdS*.

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

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

[Ailliot and Pene, 2013] Ailliot, P. and Pene, F. (2013). Consistency of the maximum likelihood estimate for non-homogeneous markov-switching models. *arXiv preprint arXiv:1306.2116*.

[Basu and Michailidis, 2013] Basu, S. and Michailidis, G. (2013). Estimation in high-dimensional vector autoregressive models. *arXiv preprint arXiv:1311.4175*.

[Baum et al., 1970] 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 et al., 2016] 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.

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

[Breheny and Huang, 2011] 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.

[Dempster et al., 1977] 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 et al., 2004] 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 and Li, 2001] 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.

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

[Green, 1990] 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, 1989] 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 et al., 2015] 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 et al., 2008] 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 et al., 1999] 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.

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

[Krolzig and Hendry, 2001] Krolzig, H.-M. and Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, 25(6):831–866.

[Lu and Berliner, 1999] 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.

[Lütkepohl, 2005] Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

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

[Pinson and Madsen, 2012] 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.

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