TWO STATISTICAL METHODS FOR IMPROVING THE ANALYSIS OF LARGE CLIMATIC DATA SETS: GENERAL SKEWED KALMAN FILTERS AND DISTRIBUTIONS OF DISTRIBUTIONS

P. NAVEAU[(1)], M. VRAC[(2,3)], M.G. GENTON[(4)], A. CHÉDIN[(2)] AND E. DIDAY[(3)]
(1) Dept. of Applied Mathematics, University of Colorado, Boulder, USA, (2) Institut Pierre Simon Laplace, Ecole Polytechnique, France, (3) Université Paris IX Dauphine, France, (4) Dept. of Statistics, North Carolina State University, USA

Abstract. This research focuses on two original statistical methods for analyzing large data sets in the context of climate studies. First, we propose a new way to introduce skewness to state-space models without losing the computational advantages of the Kalman filter operations. The motivation stems from the popularity of state-space models and statistical data assimilation techniques in geophysics, specially for forecasting purposes in real time. The added skewness comes from the extension of the multivariate normal distribution to the *general multivariate skew-normal distribution*. A new specific state-space model for which the Kalman Filtering operations are carefully described is derived.

The second part of this work is dedicated to the extension of clustering methods into the *distributions of distributions* framework. This concept allows us to cluster distributions, instead of simple observations. To illustrate the applicability of such a method, we analyze the distributions of 16200 temperature and humidity vertical profiles. Different levels of dependencies between these distributions are modeled by copula's functions. The distributions of distributions are decomposed as mixtures and the algorithm to estimate the parameters of such mixtures is presented. Besides providing realistic climatic classes, this clustering method allows atmospheric scientists to explore large climate data sets into a more meaningful and global framework.

1. Introduction

In geophysical studies, the dimension of data sets from most oceanic, atmospheric numerical models and satellites is extremely large. There exists

a variety of recent techniques to deal with such an issue in the special context of climate studies. For example, Bayesian methods (e.g. Wikle et al. (2002)), data mining, imaging and statistical visualization procedures have provided interesting and innovative ways to analyze large climatic data sets. In addition to the computational problem, the distribution of climatic random vectors is often supposed to be Gaussian or a mixture of Gaussian distributions, although this assumption is not always satisfied for a wide range of atmospheric variables. For example, the distribution of daily precipitation amounts is by nature skewed. In this paper, we attend to address these two problems, large size and skewness, with two different approaches. Because the scope of these problems is very large, we will focus our attention on two specific statistical methods used in climate studies. In Section 2, we will present a simple way to incorporate skewness in data assimilation techniques without losing the computational advantages associated with the normal distribution. In Section 3, the concept of distributions of distributions (Diday et al. 1985, Vrac 2002, Vrac et al. 2001) will be used in order to improve classical clustering methods for large climatic data sets. This application is closely linked to the algorithm of inversion of the equation of radiative transfert (Chédin et al, 1985).

2. General Skewed Kalman Filters

Before presenting the details of our research on Kalman filters, we want to clarify some climatic terms to the statistician who may not be familiar with atmospheric sciences. In particular, we would like to recall the meaning of numerical models and data assimilation in the context of this work. For the former, a numerical computer model solves the governing physical, thermodynamics and micro-physical processes at different scales of interest and over a specific region (depending on the scientific problem under study). It provides deterministic outputs of different atmospheric variables (temperature, humidity, winds, etc) according to certain forcings (inputs). It is worthwhile to note that the evaluation of such computer simulations has generated an interdisciplinary effort between scientists and statisticians in recent years. The interested reader can look at Berk's work on the statistical assessment of such models. Data assimilation can be seen as a way of incorporating observations into a numerical model as it runs. From a statistical point of view, the objective of data assimilation is to use both sources of data, observations and model outputs, to provide a better statistical analysis, in particular to give better forecasts. In the context of numerical weather prediction, updates and forecasts has be performed routinely and in real time. This compounds with the large size of data sets and implies that very efficient but slow methods have to be disregarded. The

data-assimilation or update step is closely related to Kalman filter which is the best known filtering algorithm in the context of Gaussian distributions and linear system dynamics.

The overwhelming assumption of normality in the Kalman filter literature can be understood for many reasons. A major one is that the multivariate distribution is completely characterized by its first two moments. In addition, the stability of multivariate normal distribution under summation and conditioning offers tractability and simplicity. Therefore, the Kalman filter operations can be performed rapidly and efficiently whenever the normality assumption holds. However, this assumption is not satisfied for a large number of applications. For example, some distributions used in a state-space model can be skewed. In this work, we propose a novel extension of the Kalman filter by working with a larger class of distributions than the normal distribution. This class is called *general multivariate skew-normal distributions*. Besides introducing skewness to the normal distribution, it has the advantages of being closed under marginalization and conditioning. This class has been introduced by Domínguez-Molina et al. (2001) and is an extension of the multivariate skew-normal distribution first proposed by Azzalini and his coworkers (1996, 1999). These distributions are particular types of generalized skew-elliptical distributions recently introduced by Genton and Loperfido (2002), i.e. they are defined as the product of a multivariate elliptical density with a skewing function.

2.1. THE GENERAL MULTIVARIATE SKEW-NORMAL DISTRIBUTION

The general multivariate skew-normal distribution is a family of distributions including the normal one, but with extra parameters to regulate skewness. It allows for a continuous variation from normality to non-normality, which is useful in many situations, see e.g. Azzalini and Capitanio (1999) who emphasized statistical applications for the skew-normal distribution. An n-dimensional random vector X is said to have a general multivariate skew-normal distribution (Domínguez-Molina et al., (2001)), denoted by $GMSN_{n,m}(\mu, \Sigma, D, \nu, \Delta)$, if it has a density function of the form:

$$\frac{1}{\Phi_m(D\mu; \nu, \Delta + D\Sigma D^T)} \phi_n(x; \mu, \Sigma)\Phi_m(Dx; \nu, \Delta), \qquad x \in \mathbb{R}^n, \qquad (1)$$

where $\mu \in \mathbb{R}^n$, $\nu \in \mathbb{R}^m$, $\Sigma \in \mathbb{R}^{n\times n}$ and $\Delta \in \mathbb{R}^{m\times m}$ are both covariance matrices, $D \in \mathbb{R}^{m\times n}$, $\phi_n(x; \mu, \Sigma)$ and $\Phi_n(x; \mu, \Sigma)$ are the n-dimensional normal pdf and cdf with mean μ and covariance matrix Σ. When $D = 0$, the density (1) reduces to the multivariate normal one, whereas it reduces to Azzalini and Capitanio's (1999) density when $m = 1$ and $\nu = D\mu$. The matrix parameter D is referred to as a "shape parameter". The moment

generating function $M(t)$ for a GMSN distribution is given by:

$$M(t) = \frac{\Phi_m(D(\mu + \Sigma t); \nu, \Delta + D\Sigma D^T)}{\Phi_m(D\mu; \nu, \Delta + D\Sigma D^T)} \exp\{\mu^T t + \frac{1}{2}(t^T \Sigma t)\}, \qquad t \in \mathbb{R}^n. \tag{2}$$

The simulation of random vectors from the GMSN distribution is rather simple. Indeed, Domínguez-Molina et al. (2001) showed that if $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ are two random vectors with joint distribution given by:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_{n+m}\left(\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \Sigma & -\Sigma D^T \\ -D\Sigma & \Delta + D\Sigma D^T \end{pmatrix}\right), \tag{3}$$

then the conditional distribution of X given $Y \leq D\mu$ is a general multivariate skew-normal distribution $GMSN_{n,m}(\mu, \Sigma, D, \nu, \Delta)$.

The three basic tools when implementing the Kalman filter are the closure under linear transformation, under summation and conditioning. In section 2.3, we will present how the general skew-normal distribution behaves under such constraints.

2.2. THE STATE-SPACE MODEL AND THE KALMAN FILTER

The *State Space Model* has been widely studied (e.g. Shepard (1994), Shumway and Stoffer (1991), Harrison and Stevens (1971, 1976)). This model has become a powerful tool for modeling and forecasting dynamical systems and it has been used in a wide range of disciplines such as biology, economics, engineerings and statistics (Guo et al. (1999), Kitagawa and Gersch (1984). The basic idea of the state-space model is that the d-dimensional vector of observation Y_t at time t is generated by two equations, the *observational* and the *system* equations. The first equation describes how the observations vary in function of the unobserved state vector X_t of length h: $Y_t = F_t X_t + \epsilon_t$, where ϵ_t represent an added noise and F_t is a $d \times h$ matrix of scalars. The essential difference between the state-space model and the conventional linear model is that the state vector X_t is not assumed to be constant but may change in time. The temporal dynamical structure is incorporated via the system equation: $X_t = G_t X_{t-1} + \eta_t$, where η_t represents an added noise and G_t is an $h \times h$ matrix of scalars. There exists a long literature about the estimation of the parameters for such models. In particular, the Kalman filter provides an optimal way to estimate the model parameters if the assumption of gaussianity holds. Following the definition by Meinhold and Singpurwalla (1983). the term "Kalman filter" used in this work refers to a recursive procedure for inference about the state vector. To simplify the exposition, we assume that the observation errors ϵ_t are independent of the state errors η_t and that the sampling is equally spaced, $t = 1, ..., n$.

The results shown in this paper could be easily extended without such constraints. But, the loss of clarity in the notations would make this work more difficult to read without bringing any new important concepts.

2.3. KALMAN FILTERING AND GENERAL SKEW-NORMAL DISTRIBUTIONS

From Equation (2), it is straightforward to see that the sum of two independent general multivariate skew-normal distributions is not necessary a general multivariate skew-normal distribution. In order to obtain the closure under summation needed for the Kalman Filtering, we extend the linear state-space model to a wider state-space model for which the stability under summation is better preserved. In order to pursue this goal, we need the following lemma. Its proof can be found in Domínguez-Molina et al. (2001).

Lemma 1 Suppose $Y = GMSN_{n,m}(\mu, \Sigma, D, \nu, \Delta)$ and A is a $r \times n$ matrix. Then, we have $X = AY \sim \mathrm{GMSN}_{r,m}(A\mu, A\Sigma A^T, DA^{\leftarrow}, \nu, \Delta)$ where $A^{\leftarrow}$ is the left inverse of A and $A^{\leftarrow} = A^{-1}$ when A is an $n \times n$ nonsingular matrix. If Y is partitioned into two components, Y_1 and Y_2, of dimensions h and $n - h$ respectively and with a corresponding partition for μ, Σ, D, and ν. Then the conditional distribution of Y_2 given $Y_1 = y_1$ is:

$$GMSN_{n-h,m}(\mu_2+\Sigma_{21}\Sigma_{11}^{-1}(y_1-\mu_1), \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}, D_2, \nu-D_1 y_1, \Delta). \quad (4)$$

The converse is also true, i.e. if (4) is the conditional distribution of Y_2 given $Y_1 = y_1$ and $Y_1 \sim GMSN_{h,m}(\mu_1, \Sigma_{11}, D_1, \nu_1, \Delta)$, then the joint distribution of Y_1 and Y_2 is $GMSN_{n,m}(\mu, \Sigma, D, \nu, \Delta)$.

The proof is the same as for the multivariate Gaussian distribution.

2.4. EXTENSION OF THE LINEAR STATE-SPACE MODEL

Our strategy to derive a model with a more flexible skewness is to directly incorporate a skewness term, say S_t, into the observation equation

$$\begin{aligned} Y_t &= F_t X_t + \epsilon_t \\ &= P_t U_t + Q_t S_t + \epsilon_t, \text{ with } F_t = (P_t, Q_t) \text{ and } X_t = (U_t^T, S_t^T)^T, \end{aligned} \quad (5)$$

where the random vector U_t of length k and the $d \times k$ matrix of scalar P_t represent the linear part of the observation equation. In comparison, the random vector S_t of length l and the $d \times l$ matrix of scalar Q_t correspond to the additional skewness. The most difficult task in this construction is to propose a simple dynamical structure of the skewness vector S_t and the "linear" vector U_t while keeping the independence between these two

vectors (the last condition is not theoretically necessary but it is useful when interpreting the parameters). To reach this goal, we suppose that the bi-variate random vector $(U_t^T, V_t^T)^T$ is generated from a linear system:

$$\begin{cases} U_t &= K_t U_{t-1} + \eta_t^* \\ V_t &= -L_t V_{t-1} + \eta_t^+ \end{cases} \tag{6}$$

where the Gaussian noise $\eta_t^* \sim N(\mu_\eta^*, \Sigma_\eta^*)$ is independent of $\eta_t^+ \sim N(\mu_\eta^+, \Sigma_\eta^+)$ and where K_t , respectively L_t represents a $k \times k$ matrix of scalars, respectively a $l \times l$ matrix of scalars. The multivariate normal distribution of the vector $(U_t^T, V_t^T)^T$ is denoted by

$$\begin{pmatrix} U_t \\ V_t \end{pmatrix} \sim N_{k+l}\left(\begin{pmatrix} \psi_t^* \\ \psi_t^+ \end{pmatrix}, \begin{pmatrix} \Omega_t^* & 0 \\ 0 & \Omega_t^+ \end{pmatrix} \right). \tag{7}$$

The parameters of such vectors can be sequentially derived from any initial vector $(U_0^T, V_0^T)^T$ with a normal distribution. From (3), we define the skewness part S_t of the state vector $X_t = (U_t^T, S_t^T)^T$ as the following conditional variable $S_t = [V_{t-1} | V_t \leq L_t \psi_{t-1}^+]$. It follows a general multivariate skew-normal distribution $S_t \sim GMSN_{l,l}(\psi_{t-1}^+, \Omega_{t-1}^+, L_t, \psi_t^+, \Sigma_\nu^+)$. Consequently the state vector has also a general multivariate skew-normal distribution

$$X_t = \begin{pmatrix} U_t \\ S_t \end{pmatrix} \sim GMSN_{k+l,k+l}(\psi_t, \Omega_t, D_t, \nu_t, \Delta_t), \text{ with } \psi_t = \begin{pmatrix} \psi_t^* \\ \psi_{t-1}^+ \end{pmatrix}, \tag{8}$$

$$\Omega_t = \begin{pmatrix} \Omega_t^* & 0 \\ 0 & \Omega_{t-1}^+ \end{pmatrix}, D_t = \begin{pmatrix} 0 & 0 \\ 0 & L_t \end{pmatrix}, \nu_t = \begin{pmatrix} 0 \\ \psi_t^+ \end{pmatrix}, \text{ and } \Delta_t = \begin{pmatrix} I & 0 \\ 0 & \Sigma_\nu^+ \end{pmatrix}.$$

The price for this gain in skewness flexibility is that this state vector does not have anymore a linear structure like the one defined by the system equation. If $P_t = 0$ or $L_t = 0$ then the classical state-space model is obtained.

Proposition 1 Suppose that the initial vector $(U_0^T, V_0^T)^T$ of the linear system defined by (6) follows the normal distribution defined by

$$\begin{pmatrix} U_0 \\ V_0 \end{pmatrix} \sim N_{k+l}\left(\begin{pmatrix} \psi_0^* \\ \psi_0^+ \end{pmatrix}, \begin{pmatrix} \Omega_0^* & 0 \\ 0 & \Omega_0^+ \end{pmatrix} \right). \tag{9}$$

Then both the state vector $X_t = (U_t^T, S_t^T)^T$ and the observation vector Y_t follow general multivariate skew-normal distributions, $X_t \sim GMSN_{h,m}(\psi_t, \Omega_t, D_t, \nu_t, \Delta_t)$ and $Y_t \sim GMSN_{d,m}(\mu_t, \Gamma_t, E_t, \nu_t, \Delta_t)$ for $t \geq 1$. The parameters of these distributions satisfy

$$\psi_t^* = K_t \psi_{t-1}^* + \mu_\eta^*, \;\; \psi_t^+ = -L_t \psi_{t-1}^+ + \mu_\eta^+ \text{ and } \mu_t = F_t \psi_t + \mu_\epsilon,$$

and

$$\Omega_t^* = K_t \Omega_{t-1}^* K_t^T + \Sigma_\eta^*, \;\; \Omega_t^+ = L_t \Omega_{t-1}^+ L_t^T + \Sigma_\eta^+ \text{ and } \Gamma_t = F_t \Omega_t F_t^T + \Sigma_\epsilon$$

$$E_t = D_t F_t^{\leftarrow}, D_t = D_{t-1} G_t^{\leftarrow} \text{ and } \nu_t = (0^T, \psi_t^{+T})^T.$$

The proofs of our propositions about the Skewed Kalman filter can be found in Naveau and Genton (2002).

2.5. SEQUENTIAL ESTIMATION PROCEDURE: KALMAN FILTERING

To extend the Kalman filter to general skewed normal distributions, we follow the work of Meinfold and Singpurwalla (1983) who derived a Bayesian formulation to derive the different steps of the Kalman filtering. The key notion is that given the data $\mathbf{Y}_t = (Y_1, ..., Y_t)$, inference about the state vector values can be carried out through a direct application of Bayes' theorem. In the Kalman literature, the conditional distribution of $(X_{t-1}|\mathbf{Y}_{t-1})$ is usually assumed to follow a Gaussian distribution at time $t-1$. In our case, this assumption at time $t-1$ is expressed in function of the general multivariate skew-normal distribution:

$$(X_{t-1}|\mathbf{Y}_{t-1}) = GMSN_{n,m}(\hat{\psi}_{t-1}, \hat{\Omega}_{t-1}, \hat{D}_{t-1}, \hat{\nu}_{t-1}, \hat{\Delta}_{t-1}), \qquad (10)$$

where $\hat{.}$ represents the location, scale, shape, and skewness parameters of $(X_{t-1}|\mathbf{Y}_{t-1})$. Then, we look forward in time t, but in two stages: prior to observing Y_t, and after observing Y_t. To implement these two steps, Lemma 1 is used to determine the conditional distribution of a general multivariate skew-normal distribution.

Proposition 2 Suppose that the initial vector $(U_0^T, V_0^T)^T$ follows the normal distribution defined by (9), that the posterior distribution of X_t follows (10) at time $t-1$ and that we have for U_t and V_t introduced in (5)

$$\left(\begin{array}{c} U_{t-1} \\ V_{t-1} \end{array} \middle| \mathbf{Y}_{t-1} \right) \sim N_{k+l} \left(\left(\begin{array}{c} \hat{\psi}_{t-1}^* \\ \hat{\psi}_{t-1}^+ \end{array} \right), \left(\begin{array}{cc} \hat{\Omega}_{t-1}^* & \hat{\Omega}_{t-1}^{*+} \\ \hat{\Omega}_{t-1}^{*+} & \hat{\Omega}_{t-1}^+, \end{array} \right) \right), \qquad (11)$$

where $\hat{.}$ represents the posterior mean and covariance. We define the following quantities: $R_t^+ = L_t \hat{\Omega}_{t-1}^+ L_t + \Sigma_\eta^+$, $R_t^* = K_t \hat{\Omega}_{t-1}^* K_t + \Sigma_\eta^*$,

$$\Sigma_t = Q_t R_t^* Q_t^T + P_t R_t^+ P_t^T + \Sigma_\epsilon, \text{ and } \tilde{\Omega}_t = L_t(\hat{\Omega}_{t-1}^+ + C_t P_t^T \Sigma_t^{-1} P_t C_t) L_t^T$$

and $e_t = Y_t - Q_t[K_t \hat{\psi}_{t-1}^* + \mu_\eta^*] - P_t[E(S_t|\mathbf{Y}_{t-1})] - \mu_\epsilon$, where $E(S_t|\mathbf{Y}_{t-1})$ is the conditional expectation of S_t given $\mathbf{Y}_{t-1}$ and C_t is the conditional

covariance $C_t = cov(V_{t-1}, S_t|\mathbf{Y}_{t-1})$. The parameters of the posterior distributions are computed through the next cycle by the following sequential procedure:

$$(\, X_t \,|\mathbf{Y}_t) \sim GMSN_{k+l,k+l}(\hat{\psi}_t, \hat{\Omega}_t, \hat{D}_t, \hat{\nu}_t, \hat{\Delta}_t), \text{ with } \hat{\psi}_t = \begin{pmatrix} \hat{\psi}_t^* \\ \hat{\psi}_{t-1}^+ \end{pmatrix},$$

and where

$$\hat{\Omega}_t = \begin{pmatrix} \hat{\Omega}_t^* & 0 \\ 0 & \tilde{\Omega}_t \end{pmatrix}, \hat{D}_t = \begin{pmatrix} 0 & 0 \\ 0 & L_t \end{pmatrix}, \hat{\nu}_t = \begin{pmatrix} 0 \\ \hat{\psi}_t^+ \end{pmatrix}, \text{ and } \hat{\Delta}_t = \begin{pmatrix} I & 0 \\ 0 & \hat{\Sigma}_\nu^+ \end{pmatrix}$$

and with

$$\begin{pmatrix} \hat{\psi}_t^* \\ \hat{\psi}_t^+ \end{pmatrix} = \begin{pmatrix} K_t\hat{\psi}_{t-1}^* + \mu_\eta^* + R_t^* Q_t^T \Sigma_t^{-1} e_t \\ -L_t\hat{\psi}_{t-1}^+ + \mu_\eta^+ - L_t C_t P_t^T \Sigma_t^{-1} e_t \end{pmatrix},$$

and

$$\begin{pmatrix} \hat{\Omega}_t^* & \hat{\Omega}_t^{*+} \\ \hat{\Omega}_t^{*+} & \hat{\Omega}_t^+ \end{pmatrix} = \begin{pmatrix} R_t^* - R_t^* Q_t^T \Sigma_t^{-1} Q_t R_t^* & -K_t\hat{\Omega}_{t-1}^{*+} L_t^T + R_t^* Q_t^T \Sigma_t^{-1} P_t C_t L_t^T \\ -L_t\hat{\Omega}_{t-1}^{*+} K_t^T + L_t C_t P_t^T \Sigma_t^{-1} Q_t R_t^* & R_t^+ - L_t C_t P_t^T \Sigma_t^{-1} P_t C_t L_t^T \end{pmatrix}.$$

Although the notations are a little more complex, the Kalman filtering steps for the skewed extended state-space model does not present any particular computational difficulties.

3. Distributions of distribution with application to climatology

3.1. MOTIVATIONS AND DATA

The data set under study comes from the European Center for Meteorological Forecasting (ECMWF). The temporal resolution is of 6 hour (0 a.m., 6 a.m., 12 a.m., 6 a.m.) and the data covers the period from December 1998 to December 1999. For each latitude and each longitude, the values of different atmospheric variables (pressure values, temperature, specific humidity, winds, etc) are available at 50 different vertical levels. These levels are not equally spaced and vary from one location to another. This implies that we can not choose a specific altitude (or pressure level) and simply apply classical methods at different chosen altitudes. Despite this difficulty, the atmospheric scientist would like to summarize the information contained in this multi-variate 3D grid into a 2D map, i.e. on the surface of the Earth. Being able to recognize different climatic behaviors is of particular interest. An accurate partition of these vertical profiles is essential

to interpret satellite observations into atmospheric variables (inversion of equations of radiative transfer, Chédin et al, 1985). From a statistical point of view, we rephrase this scientific question as a clustering problem, classifying multi-variate vertical profile distributions into clusters with similar physic properties inside a cluster and distinct physic characteristics between clusters. Consequently, a fundamental difference with classical clustering algorithms is that a classification method has be directly applied to distributions (vertical profiles) instead of observations. As an application, 16200 multi-variate vertical profile distributions have to be decomposed as a mixture of $K = 7$ classes. This number was chosen by atmospheric scientists and each class should correspond to a specific climatic situation. The distributions will either be of temperatures, humidities, or both. To illustrate the clustering procedure, we will focus on a particular date (the 15th of December 1998 at midnight). Before showing the results of this analysis, we need to establish a basic statistical framework.

3.2. DEFINING DISTRIBUTIONS OF DISTRIBUTIONS

Suppose that the vector $\mathbf{F} = (F_1, ..., F_n)$ represents the temperature vertical profile distributions over the entire globe, respectively $\mathbf{H} = (H_1, ..., H_n)$ for the humidity. To work with such sets of distributions, the concept of *distributions of distributions* developed by E. Diday (2001) is needed. The details of the clustering methodology of distribution of distributions can be found in the work by M. Vrac (2002, 2001).

Let t be a real. A *distribution function of distributions* is defined by

$$D_t(x) = \mathbb{P}(\{F \in \Omega_F \text{ such that } F(t) \leq x\}), \text{ for any real } t,$$

where Ω_F is the set of all possible temperature distributions. From a more practical point of view, $D_t(x)$ could be estimated by

$$\hat{D}_t(x) = \frac{1}{n} \sum_{i=1}^{n} I[\hat{F}_i(t) \leq x], \text{ with } \hat{F}_i(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} I[X_{i,j} \leq x],$$

where $I[A]$ represents the indicator function, equal to 1 is if A true and 0 otherwise, and $\hat{F}_i(t)$ denotes the empirical distribution of the ith profile that has n_i observations. Although this estimation strategy has the advantage of being simple, the clustering algorithm converges slowly due to the step-functions. Instead, we use the "Parzen estimation method" to model the vertical profile distributions

$$\hat{f}_i(x) = \frac{1}{n_i h_i} \sum_{j=1}^{n_i} K\left(\frac{x - X_{i,j}}{h_i}\right), \text{ and } \hat{F}_i(t) = \int_{-\infty}^{t} \hat{f}_i(x)dx,$$

where K is a kernel function and h_i the window width (Silverman 1986). Because the density $d_t(x) = D'_t(x)$ takes its values on $[0, 1]$, we choose to model it by a Beta density

$$d_{t,\gamma_t}(x) = \frac{\Gamma(\rho_t + \nu_t)}{\Gamma(\rho_t)\Gamma(\nu_t)} x^{\rho_t - 1}(1 - x)^{\nu_t - 1}, \text{ with } \gamma_t = (\rho_t, \nu_t) > 0. \qquad (12)$$

Hence, $\hat{D}_{t,\hat{\gamma}_t}(x) = \int_0^x \hat{d}_{t,\hat{\gamma}_t}(u)du$ with $\hat{\gamma}_t$ estimated from the sample $\{\hat{F}_i(t)\}_{i=1,\ldots,n}$.

For the practitioner, studying the relationship between two given temperatures, say t_1 and t_2, is of primary interest. To investigate such a link, the definition of D_t with t real is extended to the bi-vector $\mathbf{t} = (t_1, t_2)$ by setting

$$D_{\mathbf{t}}(x_1, x_2) = \mathbb{P}(\{F \in \mathbf{\Omega_F} \text{ such that } F(t_1) \leq x_1 \text{ and } F(t_2) \leq x_2\}).$$

The extension to higher dimensions does not present any major difficulty, but to reduce the notational complexity we restrict our exposition to the bi-variate case for the remainder of this paper.

3.3. MIXTURE OF DISTRIBUTION OF DISTRIBUTIONS AND COPULAS

Our goal is to cluster the different vertical profile distributions into $K = 7$ classes. To perform this task, we assume that the distribution $D_{\mathbf{t}}$ can be expressed as a mixture of distributions

$$D_{\mathbf{t}}(x_1, x_2) = \sum_{k=1}^{K} \pi_k D_{\mathbf{t},k}(x_1, x_2)$$

where $\sum \pi = 1, 0 < \pi_k < 1$ and $D_{\mathbf{t},k}$ represents a bi-variate distribution. We express the relationship between the distribution $D_{\mathbf{t},k}$ and its two marginals by directly applying Sklar's theorem (Sklar 1959, Nelsen 1998). This gives

$$D_{\mathbf{t}}(x_1, x_2) = \sum_{k=1}^{K} \pi_k C_{\mathbf{t},k}\left(D_{t_1,k}(x_1), D_{t_2,k}(x_2)\right),$$

where $C_{\mathbf{t},k}$ is a copula function. There exists a variety of parametric forms to model this copula. In our applications, we use Frank's copula (Nelsen 1998)

$$C_{\mathbf{t},k}(u, v) = \frac{1}{\log \beta_{\mathbf{t},k}} \log\left(1 + \frac{(\beta^u_{\mathbf{t},k} - 1)(\beta^v_{\mathbf{t},k} - 1)}{\beta_{\mathbf{t},k} - 1}\right), \text{ with } u, v \in [0, 1],$$

where the positive parameter $\beta_{\mathbf{t},k} i \neq 1$ is a indicator of dependence, $C_{\mathbf{t},k}(u, v) \sim uv$ for $\beta_{\mathbf{t},k} \uparrow 1$, $C_{\mathbf{t},k}(u, v) \sim \min(u, v)$ for $\beta_{\mathbf{t},k} \downarrow 0$ and $C_{\mathbf{t},k}(u, v) \sim \max(u +$

$v-1,0)$ for $\beta_{\mathbf{t},k} \uparrow \infty$. The first case, respectively the second case, corresponds to the independence, respectively to the total dependence.

3.4. PARAMETERS ESTIMATION AND CLUSTERING ALGORITHM

The next step is to sequentially cluster the $n = 16200$ vertical profile distributions and to estimate all parameters from the previous sections. The chosen method is an extension to distributions of the "Nuées Dynamiques" method (Diday et al., 1974). Given a partition $\mathbf{\Pi} = \{\Pi_1, ..., \Pi_K\}$ (the first one is randomly generated), the clustering algorithm constitutes of 3 main steps: (1) estimation of the mixture proportions $\{\pi_k\}$, (2) estimation of other mixture parameters, $(\gamma_{t_1,k}, \gamma_{t_2,k})$ for the Beta laws and $\{\beta_{\mathbf{t},k}\}$ for the copula's parameter, (3) re-allocation of all individuals ω_i into K new classes with $i = 1, ..., n$. This 3 step procedure is repeated until the desired convergence is reached. The first step is undertaken by setting π_k as the number of elements in the kth class divided by the total number of individuals. Other alternatives can be used (Celeux and Govaert, 1993). The second step is realized by maximizing the *classifier log-likelihood*

$$l(\mathbf{\Pi}, \theta) = \sum_{k=1}^{K} \sum_{\omega_i \in \Pi_k} \log \left[d_{k,\mathbf{t}}(x_1^{(i)}, x_2^{(i)}; \theta_k) \right], \text{ with } \theta = \{\beta_{\mathbf{t},k}, \gamma_{t_1,k}, \gamma_{t_2,k}\}_{k=1,...,K},$$

where $\omega_i = \{i : \hat{F}_i(t_1) \leq x_1, \hat{F}_i(t_2) \leq x_2\}$ and $d_{k,\mathbf{t}}(x_1, x_2; \theta_k)$ is the density derived from $D_{k,\mathbf{t}}(x_1, x_2; \theta_k)$. The last step is implemented by defining the new classes as $\Pi_k = \{\omega : \pi_k d_{k,\mathbf{t}}(\omega; \theta_k) \geq \max\{\pi_l d_{l,\mathbf{t}}(\omega; \theta_l) : l = 1, ..., K\}\}$

3.5. APPLICATION TO THE TEMPERATURE PROFILES

Figure 1 shows a classification of the 16200 vertical temperature profiles into 7 clusters. This result was obtained after applying the clustering procedure for two iterations. Although not spatial dependence was introduced in the model, the spatial coherence obtained from the clustering procedure is a positive indicator of the quality of the algorithm. From a scientific perspective, the clusters provides realistic classes. Cluster 4 can be identified as a "tropical class". Two "polar" clusters can be linked to the winter season at the South pole (cluster 1) and to the summer season at the North pole (cluster 7). Cluster 3 makes the transition between moderate and tropical zones, cluster 6 between polar and moderate zones. The high reliefs are clearly identified (Himalaya, Andes).

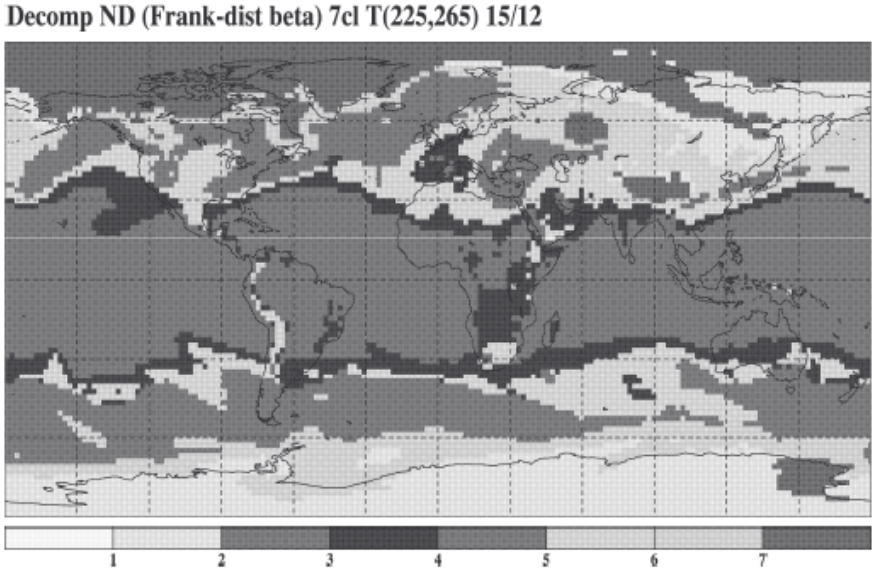

Figure 1. Clustering of the 16200 temperature vertical profiles into 7 clusters

3.6. EXTENSION TO MULTI-DIMENSIONAL DISTRIBUTIONS

In the previous sections, we exclusively focused on the temperature profiles but extending the procedure to multi-dimensional atmospheric vectors, e.g. the bi-variate vector of the temperature and humidity profiles, will greatly incrase the range of applications of this work. The coupling method is based on the following mixture decomposition

$$D_{(r)}(\mathbf{x}^{(r)}) = \sum_{k=1}^{K} \pi_{r,k} C_{r,\mathbf{t}^{(r)},k} \left(D_{r,t_{1,r},k}(x_1), D_{r,t_{2,r},k}(x_2) \right), \text{ with } \mathbf{x}^{(r)} = (x_1^{(r)}, x_2^{(r)}),$$

where the integer r represents either the temperature ($r = 1$) or the humidity ($r = 2$). Then this couple of distributions can be linked by Sklar's theorem. There exists a copula function C such that

$$D(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = C\left(D_{(1)}(\mathbf{x}^{(1)}), D_{(2)}(\mathbf{x}^{(2)}) \right).$$

Although the notations become more complex, the same overall principles of the algorithm described in Section 3.4 can be applied. A main difference is that, in addition of setting two temperature levels $(t_1^{(1)}, t_2^{(1)})$, we also need to fix two humidity levels $(t_1^{(2)}, t_2^{(2)})$. Figure 2 represents the output of such a coupling procedure. Cluster 7, respectively cluster 1, corresponds to the winter season at the North pole, respectively the summer season at

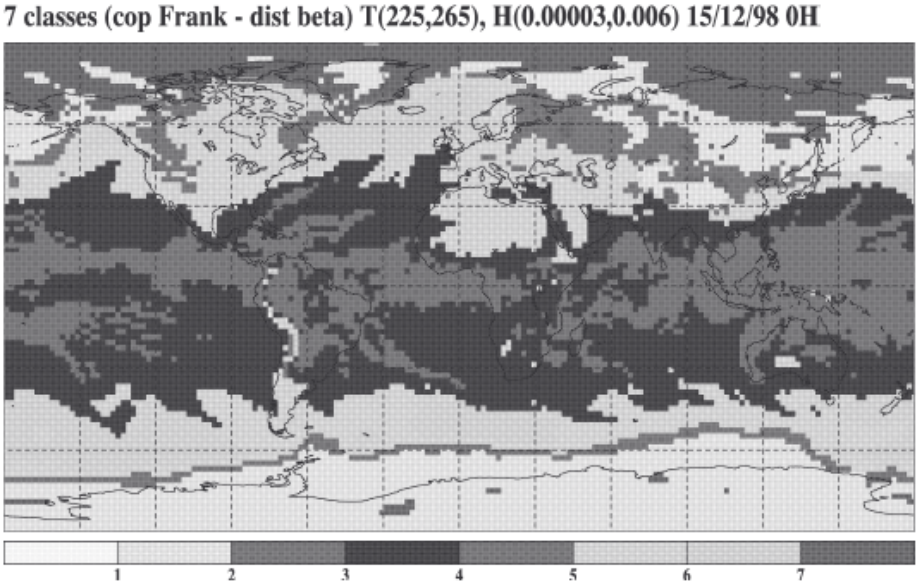

Figure 2. Clustering in 7 classes by coupling the temperature and the humidity

the South pole. This two regions were already identified in the temperature clustering, but additional variations are generated from humidity in Figure 2. Two tropical classes are identified, very humid (cluster 4) and humid (clusters 3). Cluster 4 is in better agreement with existing humid zones than the ones obtained before. The other clusters represent transition regions from tropical classes (hot and humid) to polar classes (dry and cold).

4. Conclusions

In this first part of this work, we showed that extending the normal distribution to the general multivariate skew-normal distribution for state-space models did neither reduce the flexibility nor the traceability of the operations associated with Kalman filtering. To the contrary, the introduction of a few skewness parameters provides a simple source of asymmetry needed for many applications. Further research is currently conducted to illustrate the capabilities of such extended state-space models for real case studies.

By introducing a higher abstraction level in clustering methods, the concept of distributions of distributions and copulas extends the applicability of current procedures (Diday et al. 2001, Vrac 2002). In addition, it allows to model different dependence levels for probabilistic data, internal dependencies inside a distribution of distributions (see Section 3.5) and external ones, for example between the humidity and temperature verti-

cal profile distributions. Besides providing realistic climatic classifications, these results emphasize the strong potential of this clustering method at helping the understanding of other atmospheric variables and their inter-relationships. Other algorithms have been generalized in the same way with copulas : the theoretically extensions of the algorithms EM, SEM, SAEM, and CEM was derived by Vrac (2002). Comparisons between these extended methods and "classical" algorithms of classification indicate that the procedures based on the concept of distributions of distributions perform better in the context of climatic studies (Vrac 2002). It is worthwhile to note that the proposed method can also be applied to classical numerical observations and functional data. Finally, multi-variate versions of the algorithm exist and are based on multidimensional generalized Archimedian copulas (Vrac 2002). This extension to multi-variate cases constitutes a strong axis of current research.

References

J. Anderson, An ensemble adjustment Kalman filter for data assimilation, *Monthly Weather Review*, **129**, 2884-2903, (2001).

Azzalini, A., Dalla Valle A., The multivariate skew-normal distribution, *Biometrika*, **83**, 715-726 (1996).

Azzalini, A., Capitanio, A., Statistical applications of the multivariate skew normal distribution, *J. R. Statist. Soc. B*, **61**, 579-602 (1999).

T. Bengtsson, D. Nychka, C. Snyder A frame work for data assimilation and forecasting in high-dimensional non-linear dynamical systems. Submitted to *J. R. Statist. Soc. B*, (2002).

H.H. Bock and E. Diday, Analysis of symbolic data. Exploratory methos for extracting statistical information from complex data, publisher Springer-Verlag, Heidelberg, (2000).

R. Chen and J. S. Liu, Mixture Kalman filters, *J. R. Statist. Soc. B*, **62**, 493-508, (2000).

Cressie, N. and C.K. Wikle, (2002). Space-time Kalman filter. *Entry in Encyclopedia of Environmetrics*, **4**, eds. A.H. El-Shaarawi and W.W. Piegorsch. Wiley, New York, pp.2045-2049.

E. Diday, A generalisation of the mixture decomposition problem in the symbolic data analysis framework, *Cahiers du CEREMADE*, **0112**, (2001).

G. Celeux and G. Govaert, Comparison of the mixture and the classification maximum likelihood in cluster analysis, *Journal of statist. computer,* **47,** 127-146, (1993).

A. Chédin, N. Scott, C. Wahiche, P. Moulinier, The improved initialization inversion method : a high resolution physical method for temperature retrievals from satellites of the TIROS-N series, *J. Clim. Appl. Meteor.,* **24,** 128-143, (1985).

E. Diday, A generalisation of the mixture decomposition problem in the symbolic data analysis framework, *Cahiers du CEREMADE,* **0112,** (2001).

Domínguez-Molina, González-Farías, G. & Gupta, A. K., A general multivariate skew-normal distribution. Submitted (2001).

GENTON, M. G. & LOPERFIDO, N., Generalized skew-elliptical distributions and their quadratic forms. Submitted (2002).

GUO W. & WANG Y. & BROWN M., A signal Extraction Approach to Modeling Hormones Time series with Pulses and a Changing Baseline. *J. Amer. Stat. Assoc.*, **vol. 94 No. 447**, 746-756 (1999).

HARRISON, P. J. & STEVENS, C. F., A Bayesian approach to short-term forecasting. *Operational Res. Quart.* **22**, 341-362 (1971).

HARRISON, P. J. & STEVENS, C. F., Bayesian forecasting. *J. Roy. Statist. Soc. Ser. B* **38**, 205-247 (1976).

KITAGAWA & GERSCH, A Smoothness Priors State-Space Modeling of Times Series With Trend and Seasonality. *J. Amer. Statist. Assoc.* **79**, 378-389 (1984).

MEINHOLD, R. J. & SINGPURWALLA, N. D., Understanding the Kalman filter. *The American Statistician* **37**, 123-127 (1983).

P. NAVEAU AND M. GENTON , The Multivariate General Skewed Kalman Filter. *Submitted to the J. of Multi-Variate Analysis*, (2002)

R. B. NELSEN, An introduction to Copulas, publisher Springer Verlag, Lectures Notes in Statistics, (1998).

SHEPARD N., Partially Non-Gaussian State-space Models. *Biometrika*, **81**, 115-131 (1994).

SHUMWAY, R. H. & STOFFER, D. S., Dynamic linear models with switching. *J. Amer. Statist. Assoc.* **86**, 763-769 (1991).

B.W. SILVERMAN, Density Estimation for Statistics and Data Analysis, publisher Chapman and Hall, London, (1986).

M. VRAC, E. DIDAY, A. CHÉDIN, P. NAVEAU, Mélange de distributions de distributions, *SFC'2001 8èmes Rencontres de la Société Francophone de Classification*, Université des Antilles et de Guyane, Guadeloupe, (2001).

M. VRAC, Analyse et modélisation de données probabilistes par Décomposition de Mélange de Copules et Application à une base de données climatologiques, *Thèse de doctorat*, Université Paris IX Dauphine, (2002).

WIKLE, C.K. AND N. CRESSIE, (1999) A dimension reduced approach to space-time Kalman filtering. *Biometrika* , 86 , 815-829.

WIKLE, C.K., MILLIFF, R.F., NYCHKA, D., AND L.M. BERLINER, (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *JASA* , 96 , 382-397.