Extracting Common pulse like signals from multivariate time series with a non linear Kalman Filter

Julien Gazeaux, D. Batista, C. Ammann, Philippe Naveau, C. Jégat, C. Cao

Extracting Common pulse like signals from multivariate time series with a non linear Kalman Filter

J. Gazeaux & D.Batista & C. Ammann & P.Naveau & C.Jégat & C.Gao

julien.gazeaux@latmos.ipsl.fr
philippe.naveau@lsce.ipsl.fr

Résumé : Afin de comprendre la nature et la cause de la variablité naturelle du climat, il est important d'expliquer les variations passées du climat par des forçages. L'objectif principal de ce travail est de présenter une procédure automatisée d'assimilation qui permette d'estimer l'amplitude de perturbations fortes mais temporellement très courtes, telles que des éruptions volcaniques, en utilisant des serie temporelles climatiques. La procedure d'extraction et de décomposition est exécutée sur des données réelles de sulfate issues de carottes glacières forées en différents sites du Groenland. Le sulfate propulsé par les éruptions volcaniques est transporté via la stratosphère jusque vers les pôles et déposé par sédimentation. Ainsi le sulfate extrait sur ces différents sites constitue un marqueur des plus importantes éruptions volcaniques qui se sont produites à travers le monde. De tels processus sont fortement non linéaires, que ce soit dans leurs dates mais aussi dans leurs amplitudes. Lorsqu'il ne sont pas correctement détéctés/estimés, ces évènements rares et extrèmes peuvent être à l'origine d'erreur d'estimation des différentes tendances de l'atmosphère. Ce travail peut être considéré à la fois comme un travail d'estimation d'évenements extrèmes et comme une première étapes d'estimation de tendance.

La méthode d'estimation s'applique sur des données multivariées ayant un signal caché inconnu commun. L'algorithme statistique mis en place est basé sur un model espace/état multivarié et sa résolution par un filtre de Kalman non linéaire. La non linéarités du filtre est résolue par le calcul d'une espérance et d'une vraiance doublement conditionnelle. Cette méthode fournit ainsi une estimation précise des dates et amplitudes des pics individuels à partir de différentes séries temporelles couvrant un intervalle de temps commun. En plus de l'amplitude des pics et de leurs effets à courts termes, cette méthode fournit également une mesure de significativité de chaques évènements extraits grâce au calcul d'une probabilité a posteriori. La souplesse, la robustesse et les limites de l'algorithme sont discutées par des tests de sensibilité de type Monte Carlo sur des jeux de données simulées.

Abstract : To understand the nature and cause of natural climate variability, it is important to possess an accurate estimate of past climate forcings. In this paper, we introduce an automatic procedure to estimate the magnitude of strong, but short-lived, volcanic signals in polar ice core series. Rather than treating individual records separately, our extraction algorithm jointly handles multiple time series to identify their common (but hidden) volcanic pulses. The statistical procedure is based on a multivariate multi-state space model. It provides an accurate estimator of the timing, peak magnitude and

duration of individual pulse-like events from a set of different series. It separates more effectively the real signals from spurious noise that can occur in any individual time series and at the same time provides a measure of confidence through the posterior probability for each pulse-like event. Using the joint signals, the algorithm is also more sensitive to identify smaller scale events while providing an objective estimate of the confidence associated with them. The flexibility and robustness of our approach, as well as its limitations, are discussed by applying our method to first simulated and then real world ice core time series.

1 Extraction Procedure

1.1 State Space Modeling

State space models have become a practical and powerful tool to model dynamic and complex systems. Closely related to the Kalman filter, it has been used in a wide range of disciplines: biology, economics, engineering, and statistics Guo(1999). The fundamental idea of the state space model is that the observed data is linearly dependent on latent variables of interest that vary in time. Mathematically, the observed data are governed by two equations, known as the *observational* and *system equations*. In our case, the observational equation expresses itself as a linear combination of three variables (common forcing, trends, and noise), while the system equations represent the temporal dynamics of the underlying hidden processes. For the reasons discussed above, we extend the univariate method to a multivariate setting. The statistical problem is to deduce the behavior of hidden variables of the pulse-like events from the observed data. The general implementation of multivariate extraction procedure is the main objective of this article. Before presenting some results on simulated data, we must introduce some notations and clarify our working hypotheses.

1.2 Our model

Suppose we observe J time series over the same time length, say T, and with the same temporal resolution. Each time series is denoted $y_j(t)$. We also assume that each of these time series is affected by a similar pulse-like forcing, say $x(t)$, that is unobserved and has to be estimated. This forcing corresponds to abrupt events and therefore is nonlinear and non-Gaussian. Our first equation explains how the three elements of our statistical model (trends, cycles, pulse-like events and noises) interact

$$y_j(t) = \beta_j\, x(t) + f_j(t) + \epsilon_j(t) \quad \begin{array}{l} \text{for } j = 1, \ldots, J \\ \text{and } t = 1, \ldots, T. \end{array} \tag{1}$$

The hidden, but common, pulse-like signal is represented by the random variable $x(t)$. The scalar β_j can be viewed as a scaling factor that reflects the impact of $x(t)$ on the

j-th time series. The second component $f_j(t)$ corresponds to a smooth trend. The last term, $\epsilon_j(t)$, is simply a background iid Gaussian random noise process centered about zero with standard deviation σ_j. The different noises in Equation (1) are assumed to be independent.

The two main differences of our hidden signal $x(t)$ with classical regression models comes from its pulse-like nature and its short term memory. To obey this constraint, we simply construct $x(t)$ as an autoregressive model defined by

$$x(t) = \alpha\, x(t-1) + v(t), \text{ for } t = 1, \ldots, T \tag{2}$$

where $|\alpha| < 1$ is an unknown constant and $v(t)$ corresponds to an iid random sequence and we set $x(0) = 0$. To create a pulse like effect, we impose that the iid random sequence $v(t)$ follows a mixture of a normal random variables

$$v(t) = \begin{cases} N(\mu_v, \sigma_v^2) & \text{, if } \quad I_t = 1, \\ 0 & \text{, if } \quad I_t = 0. \end{cases}, \tag{3}$$

where $N(\mu_v, \sigma_v^2)$ represents a Gaussian variable with mean μ_v and standard deviation σ_v. In Equation (3), I_t is a sequence of iid Bernoulli random variables, whose parameter $\pi = \Pr[I_t = 1]$ denotes the probability of observing a pulse-like event. The random variable $v(t)$ corresponds to the strength associated with a rare event. In contrast, $v(t)$ is set to zero if I_t equals to zero. Equation (2) allows for a short lived temporal effect of such a forcing. Despite its low number of parameters (π, α, μ_v, β_j, σ_j), the additive model defined by Equation (1) with this hidden dynamical structure (2) and its pulse-like nature defined by (3) offers enough flexibility to mimic pulse-like events behaviors at an annual scale.

The trends f_j in Equation (1) are modeled by cubic smoothing splines represented in state space form Wahba(1978), Wecker(1983). This representation allows the trends $f_j(t)$ to be expressed as functions of its first derivatives,

$$\mathbf{F_j}(t) = B\mathbf{F_j}(t-1) + \mathbf{E}_{f_j}(t),$$

where $\mathbf{F_j}(t) = (f_j(t), f_j^{(1)}(t))$, $B[i,k] = 1/(k-i)!$ for $k \geq i$ or zero otherwise. The two-dimensional vector $\mathbf{E}_{f_j}(t)$ represents a zero mean Gaussian vector with covariance elements $\lambda_j \sigma_j^2 / [(i+k-1)(i-1)!(k-1)!]$ where λ_j denotes the smoothing parameter. With these notations, it is possible to combine $x(t)$, $f_j(t)$, and their associated noises, and thus to rewrite equations (1-3) in matrix form. With the state vector $X_t = (v(t), x(t), \mathbf{F_1}(t), \ldots, \mathbf{F_J}(t))^T$ We can define

$$Y_t = HX_t + E_t, \tag{4}$$

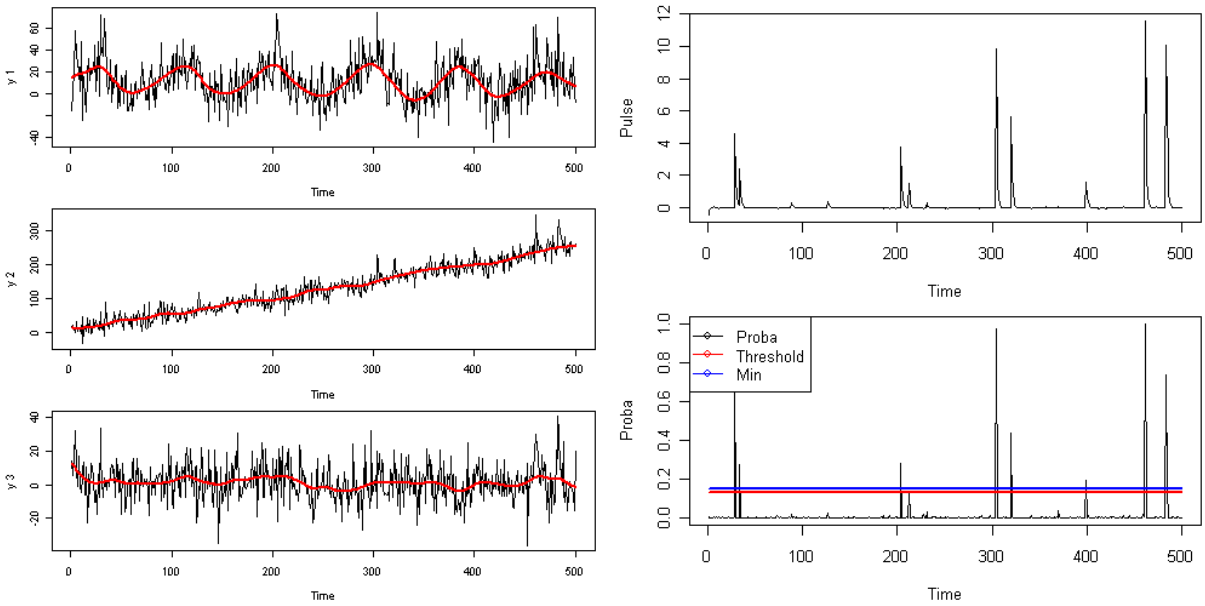
where the temporal dynamics is then described by another matrix equation

$$X_t = \Phi X_{t-1} + E_t^* \tag{5}$$

The matrices H and Φ and the random vectors E_t and E_t^* have explicit (but complex) forms.

A rich literature Guo(1999), Shepard(1994) exists to estimate parameters of the state space models. Such techniques are closely related to statistical data assimilation schemes. In Gaussian state space models, the Kalman filter provides an optimal recursive estimate of $x(t)$ from observations $Y_t = (y_1(t), \ldots, y_J(t))$. Unfortunately, the nature of the pulse-like events (the mixture of distribution) implies that the overall assumption of normality is not satisfied (see Equation 3). To solve this problem, we drew inspiration from original work of Guo(1999) who offered a variation of the Kalman filter. The principal idea is to approximate the distribution of the mixture of normals by a normal distribution whose first two moments are identical to that of the mixture. The details of this technique within the univariate framework can also be found in Naveau(2003). When the last evaluations of this modified Kalman filter are found, then a sequential backward algorithm is applied Guo(1999).

2 Extraction procedure on simulated data



In the three left figures are shown the time series, in black the input known signal. The procedure split the different part of the signal respectively to the model presented above. The red lines in the left figures present the extracted trend which is different from a serie to an other. The pulse like signal and its associated probabilities are presented on the right figure. The panel above show the amplitude of each pulse, the panel below show the associated probability of significance.

Bibliographie

[1] Batista, Gazeaux, Ammann, Naveau, Jégat and Gao. (2009) Extracting Common pulse like signals from multivariate ice cores time series, Submitted to JGR
[2] Guo, Wang and M. Brown, (1999) A signal extraction approach to modeling hormone time series with pulses and a changing baseline, J. of Am. Stat. Ass..
[3] Shepard (1994), Partially non-gaussian state-space models, Biometrika
[4] Wahba (1978), Improper priors, spline smoothing and the problem of guarding against model errors in regression, J. of the Royal Stat. Soc.
[5] Wecker, and Ansley (1983) The signal extraction approach to nonlinear regression and spline smoothing, J. of the Amer Stat. Ass.