Bull Math Biol
DOI 10.1007/s11538-010-9563-y

ORIGINAL ARTICLE

Decomposition of Neurological Multivariate Time Series by State Space Modelling

Andreas Galka · Kin Foon Kevin Wong · Tohru Ozaki · Hiltrud Muhle · Ulrich Stephani · Michael Siniatchkin

Received: 30 March 2010 / Accepted: 17 June 2010

Abstract Decomposition of multivariate time series data into independent source components forms an important part of preprocessing and analysis of time-resolved data in neuroscience. We briefly review the available tools for this purpose, such as Factor Analysis (FA) and Independent Component Analysis (ICA), then we show how linear state space modelling, a methodology from statistical time series analysis, can be employed for the same purpose. State space modelling, a generalization of classical ARMA modelling, is well suited for exploiting the dynamical information encoded in the temporal ordering of time series data, while this information remains inaccessible to FA and most ICA algorithms. As a result, much more detailed decompositions become possible, and both components with sharp power spectrum, such as alpha components, sinusoidal artifacts, or sleep spindles, and with broad power spectrum, such as FMRI scanner artifacts or epileptic spiking components, can be separated, even in the absence of prior information. In addition, three generalizations are discussed, the first relaxing the independence assumption, the second introducing non-stationarity of the covariance of the noise driving the dynamics, and the third

A. Galka (✉) · H. Muhle · U. Stephani · M. Siniatchkin
Department of Neuropediatrics, University of Kiel, 24098 Kiel, Germany
e-mail: a.galka@neurologie.uni-kiel.de

A. Galka
Institute of Experimental and Applied Physics, University of Kiel, 24098 Kiel, Germany

A. Galka
Information and Coding Theory Lab, Faculty of Engineering, University of Kiel, 24098 Kiel, Germany

K.F.K. Wong
Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

T. Ozaki
Tohoku University, 28 Kawauchi, Aoba-ku, Sendai 980-8576, Japan

Published online: 04 September 2010

allowing for non-Gaussianity of the data through a non-linear observation function. Three application examples are presented, one electrocardigram time series and two electroencephalogram (EEG) time series. The two EEG examples, both from epilepsy patients, demonstrate the separation and removal of various artifacts, including hum noise and FMRI scanner artifacts, and the identification of sleep spindles, epileptic foci, and spiking components. Decompositions obtained by two ICA algorithms are shown for comparison.

Keywords Time series analysis · Kalman filtering · Independent Component Analysis · Artifact removal · Electroencephalogram · EEG/FMRI fusion

1 Introduction

In contemporary neuroscientific research, multivariate time series are recorded in large quantities from modalities such as electroencephalography (EEG), magnetoencephalography (MEG), near-infrared spectroscopy (NIRS), and functional magnetic resonance imaging (FMRI). These time series typically will require further processing, in order to reduce their size and dimensionality, to remove artifacts and other unwanted components and to extract relevant information relating to characterization and classification of the data and the corresponding experiments or observed conditions. Recently, growing attention is devoted to algorithms for decomposition of given time series into sets of components representing physiologically meaningful entities of the underlying neurological systems and conditions (Choi et al. 2005; Delorme et al. 2007; James and Hesse 2005; Jung et al. 2001; Vigário et al. 2000).

The purpose of this paper is to provide a contribution to the development and dissemination of state space modelling as a powerful and flexible algorithm for the task of time series decomposition. We will present application examples for problems of analysis and filtering of time series typical of practical work in neurological data analysis, but the focus of the paper will remain methodological. In this Introduction, we begin with a brief review of the available tools for decomposing multivariate time series data into source components, as well as the tools for general dynamical modelling of time series.

The task of reconstructing a set of unknown source signals which have been mixed by an unknown linear function, is commonly known as *Blind Signal Separation* (*BSS*). Numerous algorithms for estimating these source signals, as well as the parameters of the mixing function (i.e., the *mixing matrix*), have been developed under the assumption of the existence of a set of source components that are mutually independent; this special case of BSS is known as *Independent Component Analysis* (*ICA*) (Cheung and Xu 2003; Choi et al. 2005; Cichocki and Amari 2002; Comon 1994; Hyvärinen et al. 2001; Jung and Kaiser 2003; Meinecke et al. 2002; Stögbauer et al. 2004).

Most algorithms for ICA assume two properties of the source components (i.e., the *independent components*): mutual independence and non-Gaussianity (more precisely, at most one Gaussian component is permitted). Furthermore, in most cases it is assumed that the number of source components is not larger than the dimension

of the data, and that measurement noise is negligible. Many algorithms for ICA are *instantaneous*, i.e., they do not take the temporal ordering of the data into account, as it is also the case with the classical statistical methods of Principal Component Analysis (PCA) and Factor Analysis (FA) (Basilevsky 1994). In contrast to these instantaneous algorithms, *dynamical* algorithms utilize also the information encoded by the temporal ordering of the data, i.e., recognize the time series aspect of the data.

While PCA, FA and ICA are applied to two-dimensional data arrays, where in most cases the dimensions represent space and time, also decompositions for three- and higher-dimensional data arrays have been explored, in particular Parallel Factor Analysis (ParaFac); here, in most cases, the additional dimension represents frequency (Miwakeichi et al. 2004). However, since usually frequency is not a separately measured quantity, these three-dimensional arrays have to be generated from the same two-dimensional original data arrays. As an alternative, subject-dependent variation has also been adopted as third dimension (Beckmann and Smith 2005). This paper, however, will not explore these methods further.

The analysis of multivariate time series has a long history, considerably predating the recent developments of BSS, ICA, and ParaFac algorithms. In particular, in the three decades from 1950 to 1980 the methodology of linear autoregressive (AR) modelling and its generalization, linear autoregressive moving-average (ARMA) modelling, was developed, both for univariate and multivariate time series (Brockwell and Davis 1987). These models aim at finding optimal predictors of the data at each point of time, based on previous data, therefore, they are dynamical by definition. Such predictors depend on the presence of temporal correlations in the data; these correlations can be quantified by the auto-covariance function, a function which in the general case of multivariate time series consists of a set of lagged covariance matrices. The parameters of AR/ARMA models may be estimated from the information contained in these matrices (Brockwell and Davis 1987).

Temporal correlations represent a possible approach to the problem of estimating independent components, thereby offering an alternative to non-Gaussianity, and several authors have proposed ICA algorithms based on simultaneous approximative diagonalization of instantaneous and lagged covariance matrices (Belouchrani et al. 1997; Molgedey and Schuster 1994; Ziehe and Müller 1998), while others are working directly with time-delay embedding vectors (Jung and Kaiser 2003; Stögbauer et al. 2004). It has also been suggested to model the source components by AR models (Barros and Cichocki 2001; Cheung and Xu 2003). Cheung and Xu (2003) have taken one step further by employing the concept of Generalized Autoregressive Conditional Heteroscedasticity (GARCH) modelling, originally developed in econometrics (Bollerslev 1986), for the purpose of allowing for time-dependent noise variance of the AR models of the source components. They then compute prediction errors of the sources, project them back into observation space, and finally decompose them by applying an instantaneous ICA algorithm. Thereby the decomposition task is split up into two steps which need to be iterated. Their work contributes a first link between ICA and state space modelling, and in this paper we intend to proceed further into this direction.

The concept of state space modelling itself has grown in parallel with the development of ARMA modelling, and it has been proved that for every ARMA model

there exists an equivalent state space model (Akaike 1974a). However, state space models are more general than ARMA models and offer additional freedom for time series modelling. A practical iterative algorithm for the estimation of dynamical states that are not accessible to direct observation, was provided by the introduction of the Kalman filter in 1960 (Kalman 1960), soon followed by algorithms for obtaining improved state estimates by retrospective smoothing (Rauch et al. 1965). A solution to the associated problem of estimating model parameters of state space models was found within the framework of maximum-likelihood estimation (Mehra 1971; Schweppe 1965), although at the price of considerable computational time expense.

In this paper, we will show that the problem of decomposing multivariate time series into independent components can be addressed directly by estimating appropriately designed state space models, within a rigorous framework for model comparison and without the need for reverting to instantaneous ICA algorithms or for assuming non-Gaussianity. Furthermore, we will show that the state space approach is flexible enough to allow for inclusion of problem-specific prior knowledge or for partly relaxing the independence assumption of ICA.

The modelling algorithm presented in this paper represents an extended and generalized version of an algorithm presented in earlier work (Wong et al. 2006). The main generalizations and variations which we introduce here, are as follows: the algorithm is generalized from univariate to multivariate time series; moving-average (MA) parameters are introduced, along with corresponding changes of the state space structure; an improved design of the application of GARCH modelling in state space is proposed; two further model generalization steps are proposed (interacting components and nonlinear observation function); and smoothing by the Rauch–Tung–Striebel smoother (Rauch et al. 1965) is introduced for the purpose of estimating components in state space. Furthermore, in the earlier paper the focus rested merely on the description of non-stationary phenomena through state space GARCH modelling, without reference to the BSS/ICA framework and the wider task of linear decomposition of multivariate time series. This extended focus is reflected in the choice of our application examples.

We remark that in recent years, several authors have proposed algorithms for estimating independent components which to some degree approach state space modelling, but so far usually without making use of Kalman filtering and smoothing methodology (Attias and Schreiner 1998; Dyrholm et al. 2007; Pearlmutter and Parra 1997). Most algorithms can be interpreted as special cases of the general state space model, such as the MA whitening models for source components in Attias and Schreiner (1998) and Dyrholm et al. (2007). Also, the distinction between *dynamical noise*, driving the dynamics, and *observational noise* is usually missing; more precisely, observational noise is neglected in most cases.

The structure of this paper is as follows. In Sect. 2, we will discuss the problem of fitting parametric models to time series data from a general perspective, mentioning input-output models and the important concept of identifiability; then the main classes of linear models will be presented, namely Factor Analysis, Independent Component Analysis and state space modelling. In Sect. 3 the estimation of states, input signals, and model parameters will be addressed in some detail, although for lack of space we will not be able to provide a thorough treatment; nevertheless we will try

to point out some key facts and connections with related fields, such as regularized least-squares estimation. In Sect. 4 we will introduce several generalizations of the basic linear state space model. Most technical details will be deferred to Appendix. In Sect. 5, a number of practical aspects relating to computational time consumption and limitations of data set size will be discussed. In Sect. 6, the performance of state space modelling will be illustrated by application to three examples of real-world time series decomposition problems, all taken from clinical practice; for comparison, two standard ICA algorithms will be applied to the same data. Section 7 concludes the paper with a discussion.

2 Models for Time Series Decomposition

2.1 Parametric Modelling and Identifiability

Let the data be denoted by $\mathbf{y}(t) = (y_1(t), \ldots, y_N(t))^\dagger, t = 1, \ldots, T$, where N denotes the dimension of the data (e.g., the number of EEG electrodes) and T the length of the time series, i.e., the number of time points at which the data was sampled. Any attempt to model such data set may be interpreted as the attempt to model the corresponding joint probability distribution

$$p(\mathbf{Y}|\boldsymbol{\vartheta}) = p\big(y_1(1), \ldots, y_N(1), \ldots, y_1(T), \ldots, y_N(T) \,\big|\, \boldsymbol{\vartheta}\big) =: L(\boldsymbol{\vartheta}; \mathbf{Y}), \qquad (1)$$

where $\mathbf{Y}$ represents the complete data set. Here, we assume that this probability distribution is described by a model which can be parameterized by a vector of model parameters $\boldsymbol{\vartheta}$; then (1) represents the likelihood of the data for given model parameters, denoted by $L(\boldsymbol{\vartheta}; \mathbf{Y})$. If there were no statistical dependencies between the variables $y_i(t)$, both with respect to i and t, the joint probability distribution would be given by the product of the marginal probability distributions $p(y_i(t))$; any deviations from independence give rise to corresponding conditional probability distributions. The model needs to be able to describe such dependencies, either explicitly or implicitly.

Many time series models possess an input-output structure, i.e., they describe a filter mapping one or several, usually unknown, input signals to an output signal, usually the data; typically these input signals may represent driving noise. The task of modelling the probability distribution of the data is then transferred to modelling the probability distribution of these input signals. Furthermore, the model may contain internal variables, sometimes known as *latent variables* or simply *states*, which are also unobserved. States will change with time and, therefore, do not belong to the set of model parameters; nevertheless they may be estimated from the data. Consequently, we have a twofold estimation problem: estimation of model parameters and estimation of states, which requires an estimate of model parameters.

An important issue relating to the estimation of model parameters in time series modelling is given by the concept of *identifiability*. It is interesting to note that at least two different definitions of this important concept exist. The earlier definition was provided in the context of system theory and control theory (Harvey et al. 2004;

Mehra 1974; Otter 1986). According to this definition, a model is regarded as identifiable, if the structure of the model permits the estimation of a *unique* set of parameters from given data.

More recently, an alternative definition was provided as part of the work on BSS and ICA (Tong et al. 1991). Here, it is assumed that there exist true values for the model parameters and the input signals; a set of these true values is defined to be identifiable with respect to a model class, if the estimates of model parameters and input signals reproduce the true values, or at least belong to the same equivalence class as the true values.

The most striking difference between these two definitions is the assumption of the existence of "true values" in the latter definition; it is equivalent to the assumption that the class of models which is employed for modelling the data, contains as a special case the "true model", i.e., a model equivalent to the process which actually produced the data. Whether this assumption of the "true model" being attainable is realistic or at least useful, is, to some extent, a philosophical question. In the system theory (or *system identification*) community the opposite position has become prevalent, namely that (almost) all modelling should be regarded merely as the search for the *best approximative* model within a chosen model class (Gevers 2006).

2.2 Linear Instantaneous Models

A widely employed model is given by a linear *observation equation*

$$\mathbf{y}(t) = \mathsf{C}\mathbf{x}(t) + \boldsymbol{\epsilon}(t), \tag{2}$$

where $\mathbf{x}(t) = (x_1(t), \ldots, x_M(t))^\dagger, t = 1, \ldots, T$, denotes an input signal, given by a time series of unobserved M-dimensional vectors, C denotes a constant $(N \times M)$-dimensional parameter matrix, and $\boldsymbol{\epsilon}(t) = (\epsilon_1(t), \ldots, \epsilon_N(t))^\dagger, t = 1, \ldots, T$ denotes another input signal, given by a time series of unobserved N-dimensional vectors, which may represent observation noise or other un-modelled components.

We shall assume that $\mathbf{y}(t)$, $\mathbf{x}(t)$ and $\boldsymbol{\epsilon}(t)$ have zero mean. Let S_x and S_ϵ denote the covariance matrices of $\mathbf{x}(t)$ and $\boldsymbol{\epsilon}(t)$, respectively.

Both Independent Component Analysis (ICA) and Factor Analysis (FA) regard the components of $\mathbf{x}(t)$ as the source signals (or *independent components*, or *factors*) to be estimated from given data $\mathbf{y}(t)$, while C, S_x and S_ϵ represent the model parameters.

Without additional assumptions and constraints, (2) does not yet represent an identifiable model. Even after imposing suitable constraints, an indeterminacy with respect to linear transformations will remain: Let T denote a non-singular $(M \times M)$-dimensional matrix, then the transformation

$$\mathbf{x}(t) \to \mathsf{T}\mathbf{x}(t), \qquad \mathsf{C} \to \mathsf{C}\mathsf{T}^{-1} \tag{3}$$

will leave the data unchanged.

2.2.1 *Factor Analysis (FA)*

Standard FA is based on the following assumptions (Harman 1976):

- $M \leq N$, i.e., a reduction of dimensionality
- the probability distributions $p(x_i)$ of the factors $x_i(t)$ are Gaussian and do not change with time
- the probability distributions $p(\epsilon_i)$ of the un-modelled noise components $\epsilon_i(t)$ are Gaussian
- $\mathsf{S}_x = \mathsf{I}_M$, where I_M denotes the $(M \times M)$-dimensional identity matrix, corresponding to uncorrelated, standardized factors
- $\mathsf{S}_\epsilon = \mathrm{diag}(\sigma_{ii}^2)$, i.e., a diagonal matrix, corresponding to uncorrelated observation noise

If $(N - M)^2 = (N + M)$, the FA model becomes uniquely defined, i.e., identifiable, except for rotations in the $\mathbf{x}$-space, corresponding to the transformation matrix T in (3) being orthogonal; this indeterminacy is a consequence of the assumption of Gaussian distributions of factors $x_i(t)$ and can be removed through an additional constraint, such as the *Varimax* constraint (Harman 1976).

2.2.2 *Independent Component Analysis* (*ICA*)

Most ICA algorithms are based on the following assumptions (Hyvärinen et al. 2001):

- $M = N$
- the probability distributions $p(x_i)$ of the independent components x_i are *non-Gaussian* (except for at most one component)
- $p(x_i, x_j) = p(x_i)p(x_j)$ for all pairs $(i \neq j)$, corresponding to independence of components
- $\mathsf{S}_x = \mathsf{I}_M$, corresponding to uncorrelated, standardized components
- $\mathsf{S}_\epsilon = \mathsf{0}$, i.e., observation noise is negligible or absent

These assumptions are less specific than those of FA; for designing particular algorithms the shape of the non-Gaussian distributions $p(x_i)$ needs to be specified, explicitly or implicitly. If this is done, the model will typically become identifiable, in terms of being uniquely defined.

With respect to a "true" solution, models based on (2) cannot reconstruct the scaling and polarity of the true source components; this indeterminacy corresponds to the transformation matrix T in (3) being diagonal with non-zero real values on the diagonal. The standardization $\mathsf{S}_x = \mathsf{I}_M$ arbitrarily fixes the scale of all source components to unity. Another unavoidable indeterminacy exists with respect to the ordering of source components within the vector $\mathbf{x}(t)$, corresponding to the transformation matrix T being a permutation matrix. These indeterminacies divide the space of possible true solutions into equivalence classes; solutions within the same equivalence class cannot be distinguished by any ICA algorithm. Identifiability was defined by Tong et al. (1991) with respect to these equivalence classes.

We mention that ICA is sometimes also known as "non-Gaussian Factor Analysis" (Hyvärinen et al. 2001). Numerous ICA algorithms have been developed so far; most of them aim at attaining independence of components by minimizing the dependencies between the components, or, equivalently, by maximizing non-Gaussianity. In this paper, we select two ICA algorithms, "FastICA" (Hyvärinen 1999) and "MILCA"

(Stögbauer et al. 2004). FastICA employs an efficient fixed-point iteration for the purpose of maximizing non-Gaussianity. *Mutual Information Least-Dependent Component Analysis* (MILCA) is based on explicit minimization of a sophisticated estimator of *mutual information*; mutual information represents a widely employed measure for dependency between the components of bi- or multivariate vectors. In their basic implementations, both FastICA and MILCA are instantaneous algorithms, ignoring the temporal ordering of the data, although an extended version of MILCA which employs time-delay embedding vectors, has been proposed (Stögbauer et al. 2004). For both FastICA and MILCA the authors have made their implementations available for free download (http://www.cis.hut.fi/projects/ica/fastica/index.shtml and http://www.klab.caltech.edu/~kraskov/MILCA/).

2.3 State Space Modelling

We will now briefly summarize linear state space modelling, as an example for an intrinsically dynamical approach to time series modelling. Note that ARMA models, another well-known class of time series models, can be rewritten as state space models.

In linear state space modelling, (2) serves again as observation equation, but now the components of the vector $\mathbf{x}(t)$ are no longer regarded as input signals, but rather as latent variables or *states*, following an explicit first-order multivariate autoregressive (AR) dynamics:

$$\mathbf{x}(t) = \mathsf{A}\mathbf{x}(t-1) + \boldsymbol{\eta}(t); \tag{4}$$

consequently, $\mathbf{x}(t)$ is called the *state vector* of the model. $\boldsymbol{\eta}(t)$ denotes an input signal, given by a time series of unobserved M-dimensional vectors, representing *dynamical noise* (in contrast to the observational noise term $\boldsymbol{\epsilon}(t)$), with covariance matrix S_η. The $(M \times M)$-dimensional parameter matrix A represents the *state transition matrix* of the AR dynamics. The elements of the 4 matrices $\mathsf{A}, \mathsf{C}, \mathsf{S}_\eta, \mathsf{S}_\epsilon$ form the set of main model parameters of the linear state space model, to be summarized by the parameter vector $\boldsymbol{\vartheta}$. Equation (4) represents the *state dynamics equation* of the state space model.

In this paper, we propose to identify the state components $x_i(t)$ with the source components to be estimated, while we regard $\boldsymbol{\eta}(t)$ as a Gaussian white noise input, driving the AR dynamics. In this point, our approach differs fundamentally from the work of Zhang and Cichocki (2000) and Waheed and Salem (2005); these authors identify the unobserved input signal $\boldsymbol{\eta}(t)$ with the source components, while a driving noise input signal may be separately provided to the state dynamics equation (4). Their approach can be interpreted as employing a linear state space model as a model for a complicated observation process, corresponding to the case known as *convolutive mixing* in the BSS field, while the temporal correlation structure of the sources themselves is not explicitly modelled. In contrast, we choose to retain the *instantaneous mixing* observation equation, (2), while the state dynamics equation, (4) serves the purpose of explicitly modelling the temporal correlation structure of the source components.

In a linear state space model, the covariance matrix of the dynamical noise S_η describes instantaneous correlations between the components $x_i(t)$ of the state vector $\mathbf{x}(t)$, while the state transition matrix A describes temporal correlations between the $x_i(t)$; diagonal elements of A describe the autocorrelations of the $x_i(t)$, while off-diagonal elements describe cross-correlations. The correlation structure of the data $\mathbf{y}(t)$ then follows from multiplication with the observation matrix C.

2.4 Instantaneous Versus Convolutive Mixing

It is a characteristic feature of state space models that redundancies between the observation equation and the state dynamics equation easily arise; in particular, temporal correlation in the data may be described not only by the state dynamics equation, but also by the observation equation, resulting in a convolutive mixing situation (Dyrholm et al. 2007). In some situations, prior knowledge of the physics of the underlying observation process may suggest that convolutive mixing should be chosen, such as situations with multiple delayed signal propagation paths, e.g., in sonar signal processing.

But we remark that for any state space model with a convolutive mixing observation equation, an equivalent state space model with instantaneous mixing can be formulated, by choosing a different state space of higher dimension. Therefore, in the absence of prior knowledge in favor of convolutive mixing, we prefer to retain instantaneous mixing.

It is commonly believed that the EEG is generated by electromagnetic fields emitted by cortical current dipoles, while artifacts of technical origin arise from additional superimposed electromagnetic fields; since the propagation of these fields in the human head is very fast, there is no obvious motivation for including signal delays into the observation equation of the EEG. In sonar signal processing, the situation would be different, since propagation speeds are much slower, distances larger, and reflections provide multiple paths between sources and sensors.

2.5 Identifiability in State Space Modelling

In this paper, we adopt the viewpoint that in most cases of analyzing real-world data, we do not have a realistic chance of finding the "true model", but only the best approximative model within a chosen model class; consequently the definition of identifiability employed in the BSS field, as discussed above, cannot be applied. It remains identifiability with respect to the existence of a unique set of model parameters. The system identification community has provided a considerable amount of work on identifiability in general linear state space models (Harvey et al. 2004; Mehra 1974; Otter 1986).

Luckily, in this paper, we do not have to address the general case. Instead, we define a special class of state space models by the following assumptions:

- the dynamical noise $\boldsymbol{\eta}(t)$ is white Gaussian noise
- $\mathsf{S}_\eta = \mathsf{I}_M$, corresponding to uncorrelated, standardized dynamical noise
- the observation noise $\boldsymbol{\epsilon}(t)$ is white Gaussian noise
- $\mathsf{S}_\epsilon = \mathrm{diag}(\sigma_{ii}^2)$, corresponding to uncorrelated observation noise

- the state transition matrix is diagonal, $\mathsf{A} = \mathrm{diag}(a_{ii})$, where $a_{ii} \neq a_{jj}$ for all pairs $(i \neq j)$

These assumptions are sufficient to ensure identifiability. Note that we have replaced the standardization assumption $\mathsf{S}_x = \mathsf{I}_M$, which was applied to the factors in FA, or equivalently to the independent components in ICA, by a corresponding assumption for the dynamical noise $\boldsymbol{\eta}(t)$ (Attias and Schreiner 1998); in both cases, the assumption is necessary in order to fix the scaling in the space of unobserved source components or states. Through this step, S_η contains no free model parameters and can be removed from the vector of model parameters $\boldsymbol{\vartheta}$. We mention that in the case of univariate data, $N = 1$, as in Wong et al. (2006), we could alternatively choose the observation matrix C as a $(1 \times M)$-dimensional vector of ones, thereby fixing scaling of states; this choice becomes impossible for multivariate data, since different data channels require different scaling of components.

Since we have decided to regard the state components $x_i(t)$ as the source components to be estimated, we note that in this particular state space model each source component is modelled by an autoregressive model of first order, AR(1). A generalization of this approach will be introduced below.

The assumption of both S_η and A being diagonal matrices massively reduces the range of possible dynamical behavior, since it corresponds to the $x_i(t)$ possessing no instantaneous or temporal (i.e., delayed) cross-correlations; only autocorrelations remain possible. Since in this special state space model no path for interactions between the different components $x_i(t)$ exists, the $x_i(t)$ are mutually independent by construction; but the independence is not explicitly enforced through the model estimation procedure, as with most ICA algorithms, but implicitly built into the model structure. The advantage of this design lies in improved robustness with respect to coincidental mutual dependencies resulting from finite data set size; such correlations would influence model estimation procedures which try to minimize them, thereby leading to biased models.

3 Estimation of States and Model Parameters

3.1 Estimation of States by Least-Squares

As mentioned above, in any model with states as latent variables, a two-fold estimation problem needs to be addressed. Let us assume for a moment that estimates of model parameters are available and that we intend to compute estimates of states $\mathbf{x}(t)$ from given data $\mathbf{y}(t)$. For the case of $N = M$, non-singular C and negligible observation noise $\boldsymbol{\epsilon}(t)$, from (2) such estimate could simply be obtained by $\mathbf{x}(t) = \mathsf{C}^{-1}\mathbf{y}(t)$. If $\boldsymbol{\epsilon}(t)$ is not neglected and/or $N \neq M$, we can still obtain estimates by standard least-squares techniques, corresponding to computing the *Moore–Penrose pseudoinverse* of C.

We mention in particular the case $N < M$ which corresponds to an underdetermined system of equations; in the BSS field this case is known as the *overcomplete basis* case, where the number of sources is larger than the number of sensors. Pseudo-inverses can still be computed for this case by techniques such as Singular

Value Decomposition or Tikhonov regularization; this is sometimes known as *ridge regression*.

Solutions of an under-determined system of equations become uniquely defined only if additional constraints are applied, such as the standard *minimum-norm* constraint; this constraint then becomes part of the regularization term. We note that solutions obtained by such approach are identifiable in the sense of being uniquely defined; however, they are certainly not identifiable with respect to an underlying "true solution".

3.2 Estimation of States by Kalman Filtering and Smoothing

The methods for the estimation of states based on least-squares and inverses/pseudo-inverses of the observation matrix represent instantaneous methods, since they do not use the temporal ordering of the data; the solutions would not change under permutation of the data with respect to time. However, there exists a dynamical generalization of least-squares estimation: the famous Kalman filter (Grewal and Andrews 2001; Kalman 1960; Sorenson 1970). This filter is applied directly to the time series data in time domain and proceeds iteratively in forward direction through the time series, as will be briefly discussed now.

Given a state estimate $\mathbf{x}(t-1|t-1)$, where the notation $(t_1|t_2)$ denotes an estimate of the state at time t_1, obtained by using all data available at time t_2, the Kalman filter produces via (4) a state prediction

$$\mathbf{x}(t|t-1) = \mathsf{A}\mathbf{x}(t-1|t-1) \tag{5}$$

and via (2) a data prediction

$$\mathbf{y}(t|t-1) = \mathsf{C}\mathbf{x}(t|t-1), \tag{6}$$

from which a prediction error, also known as *innovation* (Kailath 1968),

$$\boldsymbol{\nu}(t) = \mathbf{y}(t) - \mathbf{y}(t|t-1) \tag{7}$$

can be computed; also a corresponding $(N \times N)$-dimensional prediction error covariance matrix S_ν is provided. From the innovation, a corrected state estimate (*filtered state estimate*) can be computed as

$$\mathbf{x}(t|t) = \mathbf{x}(t|t-1) + \mathsf{G}(t)\boldsymbol{\nu}(t), \tag{8}$$

where $\mathsf{G}(t)$ is known as the *Kalman gain matrix*. $\mathsf{G}(t)$ is computed from the condition of minimizing the expected squared state estimation error.

The corrected state estimate $\mathbf{x}(t|t)$ serves as starting point for the next iteration cycle, again consisting of a prediction and a correction step. The state estimate at the first time point $\mathbf{x}(1|1)$ cannot be estimated from any previous data; instead, it may be regarded as another set of model parameters, to be included in $\boldsymbol{\vartheta}$. Alternatively, initial state estimates can be set to zero; after a certain transient, the Kalman filter will nevertheless converge to a stable operation with reasonable state estimates.

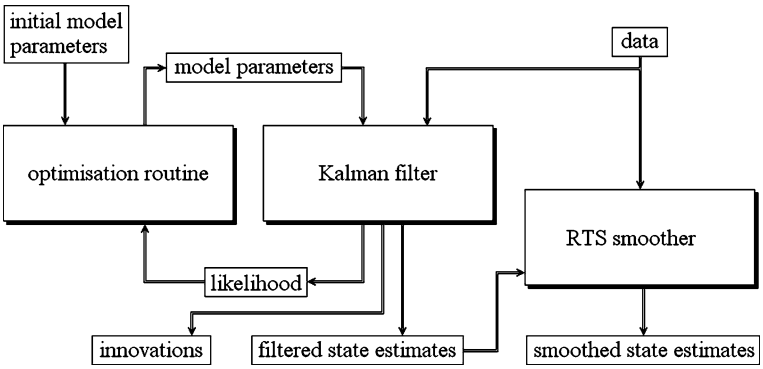

Fig. 1 Structure of the state space modelling algorithm. The core process is the transformation of the data into innovations (prediction errors) by the Kalman filter. In addition to the innovations, a corresponding value for the likelihood is provided. The optimization routine employs the Kalman filter for the task of finding model parameters such that the likelihood is maximized; this interaction is represented by a loop between both units. The RTS smoother performs an additional refinement of the filtered state estimates, as provided by the Kalman filter

When the Kalman filter iteration has reached the end of the time series, it is possible to iterate a second time through the time series, but this time in backward direction, in order to further improve state estimates by employing all available data, instead of just past data. These estimates are then written as $\mathbf{x}(t|T)$, where $t \leq T$ and T denotes the length of the time series. This step is known as *smoothing*; in this paper we employ a standard smoothing algorithm, known as the *Rauch–Tung–Striebel* (*RTS*) *smoother* (Rauch et al. 1965). In the earlier paper of Wong et al. (2006), no smoothing had been employed.

A graphical representation of the basic structure of state space modelling, including Kalman filtering and RTS smoothing, is given in Fig. 1. In this paper, we refrain from giving more detailed presentations of Kalman filtering and RTS smoothing and refer the reader to standard textbooks (Bar-Shalom and Fortmann 1988; Chui and Chen 1999; Grewal and Andrews 2001).

3.3 Observability and Underdetermined Models

The estimation of states within a given state space model will only be possible if the parameter matrices A and C possess a certain property, known as *observability* (Kailath 1980; Kalman et al. 1969).

Various tests for observability of state space models have been derived; a well-known test states that the pair (A, C) is observable, if and only if the observability matrix $\mathcal{O}$, being defined by

$$\mathcal{O} = \left(\mathsf{C}^\dagger, \mathsf{A}^\dagger\mathsf{C}^\dagger, \left(\mathsf{A}^\dagger\right)^2\mathsf{C}^\dagger, \ldots, \left(\mathsf{A}^\dagger\right)^{M-1}\mathsf{C}^\dagger\right)^\dagger, \tag{9}$$

has full rank: $\mathrm{rank}(\mathcal{O}) = M$ (Kailath 1980).

It is not surprising that a close relationship exists between observability and identifiability; see Mehra (1974) for a detailed discussion. However, again this remark applies only to identifiability in the sense of uniquely defined parameters,

not to "true models". As an important example, we consider again the underdetermined case $N < M$. Also for this case it is not hard to find state space models which are observable, and there exists a long tradition of developing and using such models in various fields, such as engineering and econometrics; in particular in econometrics it is not uncommon to have scalar data, $N = 1$, which are described by state spaces with state dimension $M > 5$ (Engle and Watson 1981; Pagan 1975). We repeat that such models do not claim to retrieve the underlying "true model"; they are to be understood merely as statistical models describing certain properties of the data, and as such they may be useful for certain tasks, such as prediction or control. Within this limited scope, the property of identifiability in the sense of uniquely defined model parameters is sufficient.

3.4 Estimation of Model Parameters

Now we remove the assumption that estimates of model parameters were already available. The estimation of the model parameters poses a considerably more difficult and time-consuming task than the estimation of states. Following the well established procedures of state space modelling (Åström 1980; Durbin and Koopman 2001; Mehra 1974; Otter 1986), we choose a maximum-likelihood approach.

In (1), we have defined the likelihood $L(\boldsymbol{\vartheta}; \mathbf{Y})$ of a given data set with respect to a model with parameters $\boldsymbol{\vartheta}$. We intend to find estimates for the parameters, such that $L(\boldsymbol{\vartheta}; \mathbf{Y})$, or equivalently $\log L(\boldsymbol{\vartheta}; \mathbf{Y})$, is maximized. For this purpose, we need to know the functional form of $\log L(\boldsymbol{\vartheta}; \mathbf{Y})$. This can be easily attained by employing the *innovation approach*, according to which for *causal* models the likelihood of a time series of prediction errors or *innovations*, as defined in (7), is approximately equal to the likelihood of the corresponding original data (Galka et al. 2006). The predictions are provided by the state space model, as presented above. If the model is able to remove most correlations from the data, both instantaneous and dynamical, the innovations will approximately be distributed as white Gaussian noise; for the case of a Markov process with continuous dynamics, this result has been proven rigorously (Protter 1990, Theorem 41).

Denoting the innovations by $\boldsymbol{\nu}(t)$, as in (7), the logarithmic likelihood of the data follows as

$$\log L(\boldsymbol{\vartheta}; \mathbf{Y}) = -\frac{1}{2}\left(T \log\left|\mathsf{S}_\nu(t)\right| + \sum_{t=1}^{T} \boldsymbol{\nu}^\dagger(t)\mathsf{S}_\nu^{-1}(t)\boldsymbol{\nu}(t) + TN\log(2\pi)\right), \qquad (10)$$

where $\mathsf{S}_\nu(t)$ denotes the covariance matrix of the innovations. This quantity needs to be maximized by suitable numerical optimization procedures, such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton algorithm (Dyrholm et al. 2007) or the Nelder–Mead simplex algorithm (Baldick 2006); we recommend employing both algorithms iteratively. In Fig. 1, the interplay between the Kalman filter and the optimization routine is graphically displayed.

The resulting model provides a decomposition of the data into components, as estimated by the Kalman filter and, possibly, the RTS smoother. Note that only the forward pass of the Kalman filter is required for computing the likelihood, while

the backward pass of the RTS smoother is not required; during the forward pass the predictions are performed which map the original time series to the innovations time series. Only after final estimates of parameters have been obtained, do we apply the RTS smoother, since it further improves state estimates.

4 Generalizations of the Basic State Space Model

4.1 Second- and Higher-Order Components

While the model, as discussed so far, suffices for modelling multivariate time series and decomposing them into independent components, better models can be obtained by allowing higher model orders for each source component; so far each component was modelled by a AR(1) component, but we may use AR(p) components with $p > 1$. This step also opens the possibility of including moving-average (MA) terms, which were absent from a similar model in the earlier paper of Wong et al. (2006).

The state space framework, as given by (4), seems to be limited to first order, i.e., modelling $\mathbf{x}(t)$ as a function of $\mathbf{x}(t-1)$ only, but it is common practice in time series analysis to define higher-order AR models within first-order state space by introducing additional state components, representing delayed copies or advanced predictions of the corresponding original state components.

In this paper, we need in particular the following case. Consider a second-order AR model with first-order MA term, i.e., an ARMA(2, 1) model:

$$x(t) = a_1 x(t-1) + a_2 x(t-2) + \eta(t) + b_1 \eta(t-1) \tag{11}$$

and rewrite it as

$$\begin{aligned} x(t) &= a_1 x(t-1) + \xi(t-1) + \eta(t), \\ \xi(t) &= a_2 x(t-1) + b_1 \eta(t), \end{aligned} \tag{12}$$

which is equivalent to

$$\begin{pmatrix} x(t) \\ \xi(t) \end{pmatrix} = \begin{pmatrix} a_1 & 1 \\ a_2 & 0 \end{pmatrix} \begin{pmatrix} x(t-1) \\ \xi(t-1) \end{pmatrix} + \begin{pmatrix} 1 \\ b_1 \end{pmatrix} \eta(t), \tag{13}$$

where $\xi(t)$ can be interpreted as $x(t+1|t)$, such that here we have a prediction of $x(t+1)$ as a new component of a state vector (Akaike and Nakagawa 1988). The state vector is defined as $\mathbf{x}(t) = (x(t), \xi(t))^\dagger$ and the corresponding observation equation as

$$x(t) = (1, 0)\mathbf{x}(t), \tag{14}$$

where the (1×2)-dimensional matrix $(1, 0)$ arises as observation matrix. Note that reformulation of ARMA models as state space models provides observation equations without observation noise terms. The specific form of the resulting transition matrix $\left(\begin{smallmatrix} a_1 & 1 \\ a_2 & 0 \end{smallmatrix}\right)$, known as *left companion form* or *observer canonical form*, is characteristic for advanced component models. This construction is easily extended to general AR or ARMA components of higher order $p \leq 2$.

We can then generalize the constraint on the transition matrix A, given in the previous section, by demanding that A be *block-diagonal*, with each block either being a scalar parameter (for first-order components), as before, or a $(p \times p)$-dimensional transition matrix in left companion form, for pth-order components. In this paper, we will employ only first order and second order components. When compared to first-order components, second-order components have the attractive property of providing a natural description for oscillations in the data.

By this generalization also, the constraint on the driving noise covariance matrix S_η needs to be modified: S_η will assume the same block-diagonal structure as A, with each block given by a corresponding $(p \times p)$-dimensional matrix resulting from the outer product $(1, b_1, \ldots, b_{p-1})^\dagger (1, b_1, \ldots, b_{p-1})$. Again, for the first-order case this reduces to the (1×1)-dimensional identity matrix, i.e., a scalar value of 1, as before. The vector of MA parameters $(b_1, \ldots, b_{p-1})$ becomes part of the parameter vector $\boldsymbol{\vartheta}$; these new parameters provide additional flexibility for improved fitting of the data. It is a characteristic advantage of observer canonical form models that MA parameters are conveniently accommodated in S_η. For a classical reference on the interplay of AR parameters and MA parameters within linear modelling of time series, we refer to Box and Jenkins (1970).

With the introduction of second-order components the definition of the basic state space model, as employed in this paper, is completed. In Appendix of this paper, a more detailed summary of the definition of the model and its parameter matrices can be found.

4.2 Further Remarks on Nonlinear Dynamics and Higher Model Orders

We remark that constructing a state space model from linear AR/ARMA components of first and second order only does not constrain the actual processes to be captured and decomposed from the data to be truly generated only by this class of models. If the dynamics of the system generating the data contains strong non-linearities, our linear state space modelling approach would aim at providing the best linear approximation to the non-linear dynamics. The qualitative properties of non-linear dynamics would not be preserved by this model, but we expect that for most applications short-term predictions would be only weakly inferior to predictions provided by non-linear predictors. Due to its two-stage structure, with predictor and corrector steps, the Kalman filter is well known to be very flexible with respect to misspecified or simplified models (Bar-Shalom and Fortmann 1988), while faithful reconstruction of global properties of the underlying dynamics is not required for meaningful state estimation. As an example, we have found through simulations that the time series decomposition approach, as proposed in this paper, is capable of decomposing mixtures of time series from nonlinear deterministic systems, such as the Lorenz, Rössler, and Mackey–Glass chaotic attractors, again by using only linear stochastic second-order models.

Full non-linear modelling remains a possibility, for example, by employing the extended Kalman filter, but for many application (especially in fields like neurology) results would differ only marginally from the linear results; only if it were possible to model accurately the particular type of non-linearity underlying the data, qualitatively different results could be expected.

On the other hand, we have also found that mixtures of time series with very rich and non-stationary dynamics, such as speech signals (corresponding to the classic "cocktail party problem"), cannot be properly decomposed by second-order models; in such situations 4th-order models were required for successful decomposition. The cocktail party problem differs from typical neuroscience applications in at least one important aspect: for the cocktail party problem the model assumption of a set of independent sources is known to be approximately correct, while in neuroscience usually no prior knowledge is available that would justify such specific assumption. Still, the assumption may be incorporated into the model, without the explicit claim that the resulting model would necessarily represent the true structure of the underlying dynamics; nevertheless, the model would probably remain a *statistically useful model* (in contrast to a *physically correct model*), with respect to particular processing and analysis tasks. It is due to this limited "absolute" meaning of the model structure and the set of extracted components that in the case of neurological time series we would usually not regard it as justified or necessary to employ model components with $p > 2$. Only in special situations, higher-order components might be justified and useful, such as in the case of pronounced oscillations displaying higher harmonics.

4.3 Further Generalizations of the Model

The state space framework readily permits further generalizations which may be useful and in many cases even essential for modelling time series in neuroscience. Here, we will discuss three such generalizations; again, technical details are deferred to Appendix.

4.3.1 Interacting Components

For certain groups of state components the independence constraint may be relaxed by allowing those off-diagonal elements of the transition matrix A corresponding to these components to be non-zero. Typically, first a model with independent components is estimated, then a generalized model with interacting components is estimated, starting from the independent model as initial solution. The decision about which components are allowed to interact, needs to be contributed by the data analyst, who interprets the components from the independent model by subjective assessment, possibly using prior knowledge.

We could remove this subjective element by repeatedly refitting the model, each time adding just one additional component-component interaction, while all others remain zero, such that for each pair of components (and even each direction of interaction) a value of the likelihood would be generated; then only those interactions which yield high improvements of likelihood, would be retained in the final model. In this paper, we have not employed this alternative, since there were no ambiguities in the decision on which components to include in the set of interacting components; see Sect. 6.3 of this paper.

This generalization is particularly convenient for components in the data that are too complicated to be reliably captured by a single component of the state space model, such as residual MRI scanner noise in EEG, as will be demonstrated in Sect. 6.3.

4.3.2 Non-stationary Covariance of Driving Noise

The elements of the driving noise covariance matrix S_η may change with time by following themselves from a specific AR/ARMA dynamics; in econometrics this is known as *stochastic volatility* modelling (Aït-Sahalia and Kimmel 2007). While full stochastic volatility modelling is a challenging task, a computationally efficient deterministic variant, known as the *Generalized Autoregressive Conditional Heteroscedasticity* (GARCH) model, is available, as also employed by Cheung and Xu (2003) in the context of ICA. A GARCH model does not contain a separate noise term (i.e., a stochastic term), therefore we call it deterministic; instead of a separate noise term, the innovations are used.

Recently, GARCH modelling has been extended to the state space framework (Galka et al. 2004, 2010; Wong et al. 2006). In this paper, we essentially follow the algorithm given by Wong et al. (2006), with the exception of formulating it directly for the standard deviations of the driving noise, instead of the logarithm of the variance (see Appendix for details).

The state space GARCH dynamics typically contains AR and MA terms, such as shown by (11), giving rise to further model parameters to be included in $\boldsymbol{\vartheta}$. We remark that it is advisable to give the freedom of GARCH dynamics only to one or very few components within a state space model. Experience has shown that if most or all components use GARCH dynamics, the components of the decomposition tend to become featureless.

The detailed implementation of the state space GARCH algorithm employed in this paper is given in Appendix.

4.3.3 Nonlinear Observation

If the distribution of the data deviates from a Gaussian distribution, nonlinear modelling is required. The easiest way to introduce a non-linearity is given by replacing (2) by

$$\mathbf{y}(t) = g\big(\mathbf{x}(t)\big) + \boldsymbol{\epsilon}(t), \tag{15}$$

where $g(.)$ is a monotone non-linear function (Ozaki and Iino 2001). In this paper, we use

$$g(x) = \frac{1}{\gamma}\sinh(\gamma x) \quad \text{where } \gamma = \begin{cases} \gamma_1 & \text{if } x > 0, \\ \gamma_2 & \text{if } x < 0. \end{cases} \tag{16}$$

For $\gamma = 0$, the linear Gaussian case is employed. The two new model parameters γ_1, γ_2 need to be included in $\boldsymbol{\vartheta}$. Furthermore, an appropriate correction term has to be added to the logarithmic likelihood, corresponding to the standard transformation of probability density functions with respect to coordinate transformations; details are given in Appendix.

5 Practical Considerations

5.1 Data Set Size and Computational Time Consumption

Fitting a state space model, possibly with the generalizations presented in the previous section, to multivariate time series data may easily become a computationally challenging task, due to the need to perform nonlinear numerical optimization. Large models are characterized by parameter vectors $\boldsymbol{\vartheta}$ with more than 100 dimensions, such that, depending on the efficiency of the implementation and the clock rate of the computer, full optimization of these parameters may consume dozens of hours. Furthermore, state space models should not be fitted in a pure batch mode, but rather a certain amount of interaction by an experienced data analyst is recommended. In this respect, many of the instantaneous ICA algorithms offer better performance, since they usually provide a result within a few minutes of computation, or less. However, some more advanced methods, like MILCA (Stögbauer et al. 2004), or the convolutive mixing method of Dyrholm et al. (2007), are also more demanding in terms of computational time consumption.

Typical data set sizes that can be analyzed with our state space methodology are up to $N = 10$ channels and a few 10^4 of time points; for data sets with considerably larger number of channels, the computational time consumption becomes prohibitively large, at least if optimizations are to be performed until convergence of the estimates, as we have done in the application examples reported in this paper.

For data sets with more than $N = 10$ channels, the analysis may be limited to a subset of most relevant channels; this is the solution we have chosen in the two EEG examples reported in this paper. As an extreme, but not uncommon case, even just a single channel may be chosen and modelled; and in many cases it will still be possible to extract several meaningful components, since the state space model will typically be observable (see Sect. 3.3). Employing more channels provides additional information, but also raises the question whether common source components exist that can be identified in all channels.

If data sets with dimension $N \gg 10$ need to be modelled, there is the option of dividing the channels of the data set into several subsets each of which having dimension $\tilde{N} \leq 10$. Then only correlations within the channels of each subset will be modelled, but not across subsets. Alternatively, the full data set may be analyzed by a fast instantaneous method, such as PCA or some fast ICA method; and then only a subset of principal components, or independent components, is retained for the state space modelling step.

If data sets with length $T \gg 10^4$ need to be modelled, there is the option of fitting the model on a subset of reduced length $\tilde{T} \leq 10^4$; then the Kalman filter forward pass, and also the smoother backward pass, can be applied to the complete data set, without refitting parameters. For data sets with good stationarity properties, this approach can be expected to provide good results; for the case of strong non-stationarity, the state space GARCH generalization may be helpful, although it will slow down the Kalman filter, since it prevents it from attaining its stable state.

5.2 Initial Values for the Parameters

Next we discuss the issue of choosing initial values for the non-linear numerical optimization. In principle, the parameter vector $\boldsymbol{\vartheta}$ could be initialized with random numbers, subject to certain constraints, such as stability of AR(1) and AR(2) components, Nyquist limits for frequencies, etc. But we have found that it is more efficient to choose the initial values for the main sets of parameters according to the linear properties of the data. For this purpose, we recommend to make use of the close relationship between multivariate AR(MA) models and state space models (Akaike 1974a).

Fitting a multivariate AR model can be done very fast, and the resulting model can easily be transformed into a state space model without observational noise; this transformation can be performed analogous to (11)–(14). Then this state space model is again transformed as follows: The eigenvalues of the transition matrix of the original state space model are computed, and each real eigenvalue gives rise to a AR(1) component within the transformed state space model, while each pair of complex conjugated eigenvalues gives rise to a ARMA(2, 1) component; see Neumaier and Schneider (2001) for a detailed discussion. The model order of the multivariate AR model does not play an important role, as long as a "moderate" value is chosen; typical values would be 15 or 20.

At first, the resulting set of AR(1) and ARMA(2, 1) components will usually be too large, so the most important components need to be chosen, while the other components are discarded; this can be done by observing the change of the likelihood, if one component is omitted from the full model. The final number of components to retain, and hence the state dimension M, remains a somewhat subjectively chosen parameter. A well-justified approach to choosing the number of components would be based on information criteria, such as the Akaike Information Criterion (AIC) (Akaike 1974a) or the Bayesian Information Criterion (BIC) (Schwarz 1978), in order to compare the improvement of likelihood due to inclusion of each component with the increase of the number of parameters. But this approach would require full optimization of many initial candidate models, resulting in huge computational time expense. We therefore recommend to initially include a larger number of components, and later, after some optimization, to remove those components which contribute little or nothing to improving the likelihood.

By this preparatory analysis, we ensure that the initial transition matrix $\mathsf{A}^{(\mathrm{init})}$ already reflects the main temporal correlation information in the data; the other system matrices, namely the observation matrix $\mathsf{C}^{(\mathrm{init})}$ and the dynamical noise covariance matrix $\mathsf{S}_\eta^{(\mathrm{init})}$ are also initialized through this procedure. At first, $\mathsf{S}_\eta^{(\mathrm{init})}$ will typically be a full matrix, but it needs to be diagonal, or rather block-diagonal (in case of components of order $p > 1$), if independent components are desired. There exists no linear transformation that would diagonalize both $\mathsf{A}^{(\mathrm{init})}$ and $\mathsf{S}_\eta^{(\mathrm{init})}$ simultaneously. For this reason, we suggest to replace all off-diagonal elements of $\mathsf{S}_\eta^{(\mathrm{init})}$ by zeros, or rather replace all elements outside the blocks on the diagonal by zeros (in case of components of order $p > 1$). By appropriate rescaling the remaining diagonal values can be set to ones; for $p > 1$ only the upper left element within each block on the diagonal will be scaled to one, while the remaining elements will assume other values.

The step of replacing most elements of $\mathsf{S}_\eta^{(\mathrm{init})}$ by zeros will severely deteriorate the likelihood of the resulting initial state space model; it is then left to the subsequent non-linear numerical optimization step to improve the likelihood again, within the chosen structure of the state space model.

If the same data is fitted repeatedly by using different initial models, the finally resulting models may differ with respect of many of their parameters; however, for methods of comparable quality, in terms of the likelihood, the components within the state vector describing prominent features within the data will closely resemble each other. Examples will be shown in Sect. 6. Models may also contain "weak" components which would not be reproduced in a different model fit; such components can be removed from the model with little or no effect to the likelihood.

5.3 Practical Model Design

The model resulting from the initialization and optimization approach proposed in this paper so far, will be composed of a set of AR(1) components, and a set of ARMA(2, 1) components, such that all components are independent of each other. Higher-order components, representing ARMA(p, $p-1$) processes with $p > 2$, could be created by merging subsets of these elementary components; we have already discussed this point briefly in Sect. 4.2. After merging, the complete model needs to be improved by numerical optimization again.

The close relation ship between the AR model order p and the corresponding MA model order $p-1$ is a consequence of the state space representation of general ARMA(p, q) models (Akaike 1974b). The total number of AR(1) components and ARMA(2, 1) components to be employed in a particular model will, more or less, remain a subjective decision. As mentioned above, AIC or BIC could be employed for estimating an optimal number of components; but in many applications, the analyst can preferably use his experience and prior knowledge to decide which components are required in the model. Large initial models could be gradually pruned, or smaller models could be augmented (e.g., by extracting additional components directly from the innovations), until the innovations appear sufficiently white, and all expected components are well accommodated within the set of components. In the case of prior knowledge or expectations being completely unavailable, whiteness of the innovations becomes the main criterion.

Also, the decision of which component within a state space model should be given the additional freedom of non-stationary covariance through state space GARCH is usually based on subjective decision based on visual inspection of the components of a stationary model; examples will be presented in Sect. 6. We regard the partly subjective approach to model design described in this section as justified by our intention to create statistically useful models, as opposed to physically correct models.

6 Applications to Neurological Data

6.1 General Procedure

We will now present three neurological time series as examples for time series decomposition problems. Each time series will be decomposed by state space mod-

elling, as presented in this paper, and additionally by FastICA and by MILCA. For FastICA, the stabilized deflation algorithm and the "pow3" non-linearity are chosen. For MILCA, all parameters of the algorithm are kept at their default values, as predefined in the implementation provided by the authors (number of nearest neighbors: 12; MI estimator: rectangular; number of harmonics for curve fit: 1; number of fine tuning angles: 128). For FastICA and MILCA, the number of independent components is equal to the dimension of the data, while for state space modelling it can be freely chosen, e.g., by removing weak components from a model with a large number of components; in our first example we choose this number again as the data dimension, while in the other two examples we choose numbers considerably larger than the data dimension. In this paper, the comparison of the three decomposition algorithms is performed simply by visual comparison and interpretation of the components. Furthermore, in the third example, we employ power spectra for the purpose of demonstrating the performance of artifact rejection.

6.2 Fetal ECG Time Series

6.2.1 Description of Data Set

We begin by studying the decomposition of a time series representing the electrocardiogram (ECG) of a pregnant woman. The data was sampled from $N = 8$ electrodes placed at thorax and abdomen, arbitrarily labelled $A, B, \ldots, H$, at a sampling rate of 500 Hz; the length of the time series is $T = 2500$ points. The data is shown in Fig. 2 (top left panel); the sharp spikes corresponding to the mother's heartbeat are clearly visible, while the spikes of the fetus' heartbeat, having much smaller amplitude, but higher frequency, are much weaker and can only be seen in few channels. Furthermore, in channel D a very slow oscillation appears, probably due to breathing. This data set has been studied previously in the context of BSS by other authors (Meinecke et al. 2002; Stögbauer et al. 2004); it is suitable as a benchmark data set since it represents a case where we can expect the basic assumption of a linear instantaneous mixture of independent sources to be approximately correct.

6.2.2 Results of Analysis

We analyze the ECG data by FastICA, MILCA and state space modelling; the resulting components are shown in Fig. 2. The state space model is arbitrarily defined to consist of 4 first-order (AR(1)) and 4 second-order (ARMA(2, 1)) components, such that the total number equals the dimension of the data; other choices for the relative numbers of first and second order components would, after sufficient optimization, yield similar results, but it is advisable to have at least some second-order components in a model, since they provide natural models for oscillations.

The bottom left panel of Fig. 2 shows state estimates obtained for the state space model by the RTS smoother. At the top of the panel first-order components are shown, ordered according to the size of the corresponding AR parameter (increasing from top to bottom), followed by second-order components, ordered according to the size of the corresponding main frequency (also increasing from top to bottom). From the

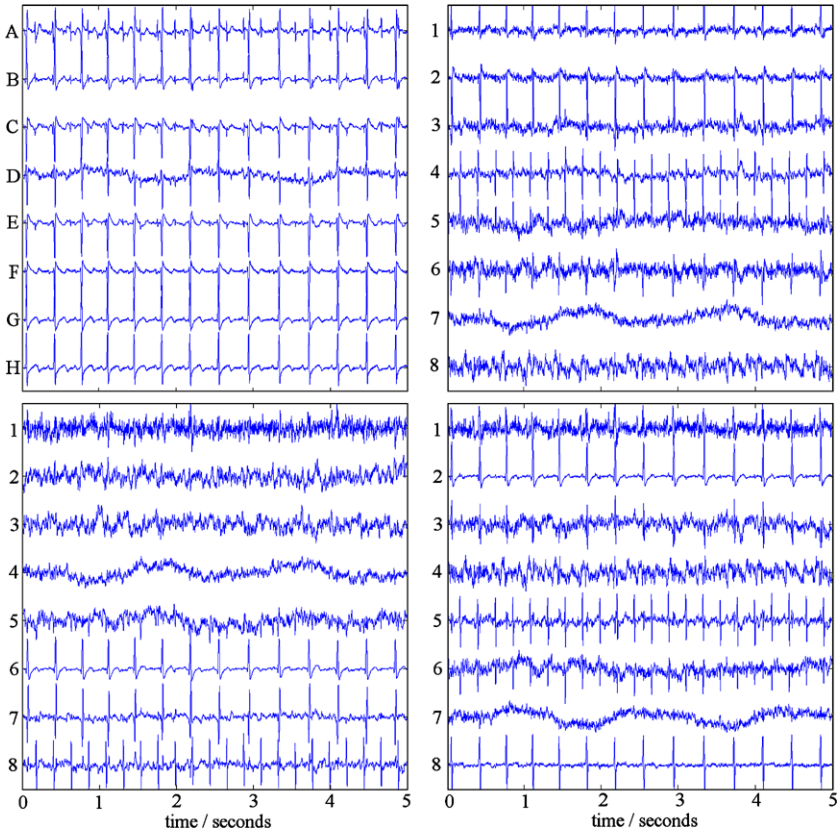

Fig. 2 Electrocardiogram data (*top left panel*); components estimated by FastICA (*top right panel*), by MILCA (*bottom right panel*) and by state space modelling, using the RTS smoother (*bottom left panel*). In the state space modelling decomposition, components 1–4 represent AR(1) components, while components 5–8 represent ARMA(2, 1) components

figure, it can be seen that the state space model collects the mother's heartbeat in two second-order components (components 6 and 7) and the fetus' heartbeat in another second-order component (c. 8), while the remaining 5 components contain breathing (c. 4 and 5) and background noise.

When comparing with the results of FastICA and MILCA, also shown in Fig. 2 (right panels), it can be seen that these algorithms also succeed in finding two components for the mother and one for the fetus. However, for both algorithms parts of the mother and fetus signals reappear in some of the remaining 5 components, i.e., the separation of the physiologically meaningful signals from the background noise has not been completely successful.

6.3 Epileptic Spiking EEG Time Series

6.3.1 Description of Data Set

The second example is given by a clinical EEG time series recorded from an awake 9-year old male patient suffering from Rolandic epilepsy. The EEG equipment was a

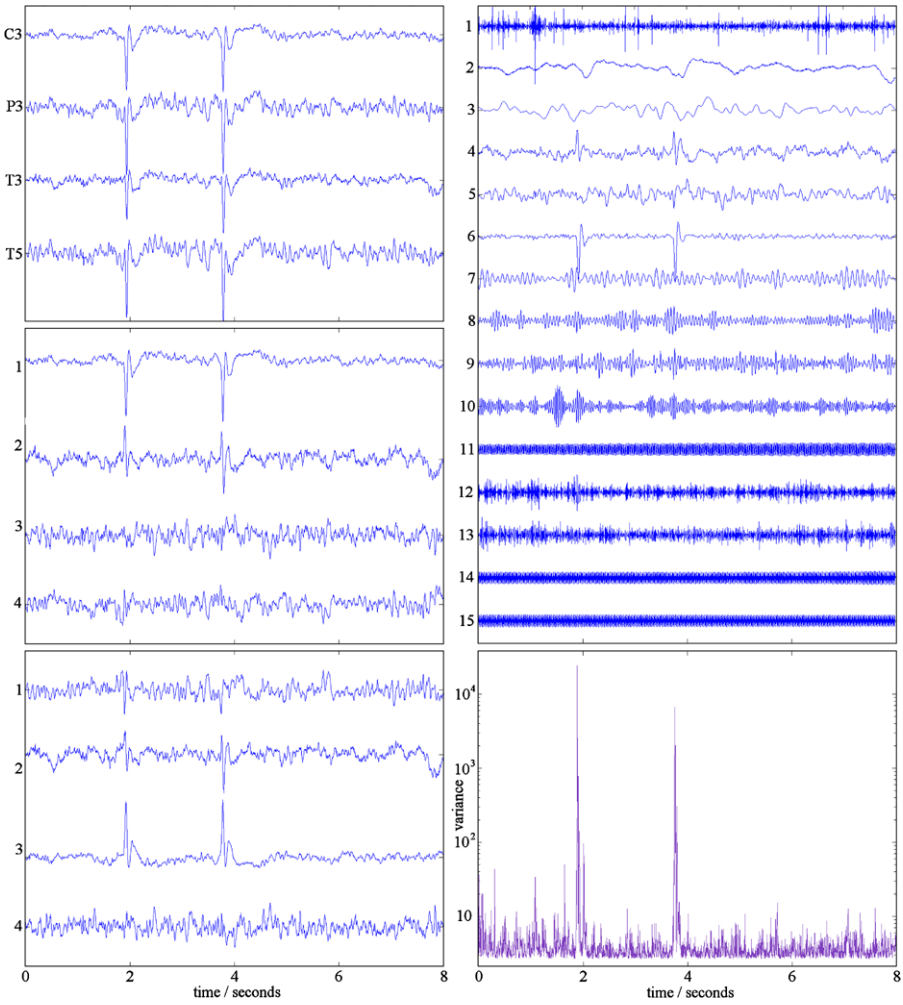

Fig. 3 Electroencephalogram data (*top left panel*); components estimated by FastICA (*middle left panel*), by MILCA (*bottom left panel*) and by state space modelling, using the RTS smoother (*top/middle right panel*). In the state space modelling decomposition, components 1–2 represent AR(1) components, while components 3–15 represent ARMA(2, 1) components. The state variance of the 6th component of the state space model was allowed to change with time (state space GARCH); the time course of this variance is shown *in the bottom right panel* (note the logarithmic scale of the *vertical axis*)

Nihon–Kohden Neurofax device for routine EEG recordings. The electrical reference was the average of F3 and F4, and the sampling rate was 256 Hz. The data is shown in Fig. 3 (top left panel); for this analysis a subset of 4 EEG channels (C3, P3, T3, T5) was selected from the complete set of 20 channels. The length of the chosen data set was 8.0 seconds. In the figure, it can be seen that within the chosen time interval for the chosen electrodes two pronounced epileptic spike-wave events have occurred. The power spectrum of this data (not shown) reveals that the data is weakly contaminated by hum noise from the electric power supply, as it is not uncommon in

clinical data sets. While the main hum frequency is at 50 Hz, two higher harmonics are also present, at 100 Hz and 106 Hz (reflected back from 150 Hz due to aliasing).

6.3.2 Results of Analysis

Again we analyze the EEG data by FastICA, MILCA and state space modelling; the resulting components are shown in Fig. 3. Based on a preparatory analysis by multivariate AR modelling, as described above, the state space model is defined to consist of 2 first-order (AR(1)) and 13 second-order (ARMA(2, 1)) components, thus we are using $M = 15$ components for describing $N = 4$ data channels. The possibility to choose $M > N$ is a characteristic strong point of state space methods. As a technical detail, we note that the actual state space dimension of this analysis was not 15, but 28, since each second-order component contributes two state space dimensions.

State estimates obtained by the RTS smoother are shown in the upper right panel of Fig. 3; first-order and second-order components are ordered again in the same way as described above for the ECG data set. Components 11, 14, and 15 represent the three sharp spectral lines at 50 Hz, 100 Hz, and 106 Hz, corresponding to the hum noise and its harmonics. These frequencies are already detected by the preparatory multivariate AR analysis, such that no prior information needs to be provided for the analysis. However, it is advisable to fix the damping coefficient (i.e., the modulus of the complex root, see Appendix) of these three components to unity during optimization of parameters, since it is known that these sinusoidal components are undamped.

The sixth component represents the two spike-wave events in the data. In order to help this component to capture the highly non-stationary behavior of these events, it was defined as a component with non-stationary covariance, i.e., with state space GARCH; this was not done for any other component of this analysis. In the lower right panel of Fig. 3, the time-dependent variance of this component is displayed. It can be seen that at the onset of each spike the variance jumps from its resting value of about 2 to values about 10^4, thereby illustrating that the additional freedom of state space GARCH dynamics is actually employed for improved modelling of this data. Thanks to this step, the sudden rise of the EEG does not lead to large data prediction errors (i.e., innovations), but is rapidly captured by the model itself. During model fitting, the model is rewarded for reducing innovations by improved values of the likelihood. We remark that also this component has been identified without being provided with any prior information, such as the timings of the spikes.

The decision which component should become a state space GARCH component was based on visual inspection of the best stationary model; already in the corresponding decomposition there was one component showing strongest correlation with the spike-wave events, and consequently this component was given the additional freedom of state space GARCH. Since the likelihood framework has the power to improve predictions and, implicitly through the model structure, mutual independence of components, the additional freedom in the model led to more spike power being collected into this component, and less into other components.

However, the spike-wave events are not yet fully captured by the sixth component, since certain parts of these events are clearly visible in the second and fourth

components of the decomposition. Further improvement of the model design may be required, in order to fully extract these events from other components and background activity. As an alternative, we may also consider the possibility, that in this data set the dynamics of the spike-wave component *intrinsically* possesses mutual dependencies with other components, such that a full decomposition into independent components would be infeasible.

In the analysis of this data set, we have also applied the model generalization of employing a non-linear observation function, in order to reduce deviations from Gaussianity which may result from the high amplitudes of the sharp spikes.

Among the remaining components we note a clear alpha component (c. 7), with a frequency of 9.63 Hz, and two beta components (c. 8 and 9), with frequencies 17.57 Hz and 17.68 Hz; these components appear completely unaffected by the spike-wave events. Components 1, 12, and 13 represent mainly high-frequency noise.

In comparison, the FastICA and MILCA decompositions, shown in the middle and lower left panels of Fig. 3, fail to identify the hum noise and alpha components, while roughly succeeding to allocate most of the spike-wave events within one component. However, also in these decompositions parts of these events appear in one or two other components. Clearly the limitation of $M = N$ renders it impossible for these algorithms to provide a decomposition as detailed as it is possible with state space modelling. Furthermore, it is well known that most ICA algorithms perform poorly in the identification of periodic or narrow-band sources, like hum noise. However, since hum noise artifacts can easily be removed by other methods, this does not represent a serious weakness for practical work.

6.4 EEG Time Series Recorded Inside MRI Scanner

6.4.1 Description of Data Set

As the third example, we study another EEG data set, recorded from an 8-year old male patient also suffering from Rolandic epilepsy, with centro-temporal spikes. This data set was recorded while the patient was sleeping inside a magnetic resonance imaging scanner (Philips Achieva 3T); EEG and FMRI were recorded simultaneously, using MRI-compatible EEG equipment. FMRI recording parameters were: 8 channel SENSE head coil, TR = 2250 ms, TE = 45 ms, 30 slices, 64×64 matrix, slice thickness = 3.5 mm, FOV = 200 mm, flip angle = 900, 540 scans in 20 min.

The EEG was recorded from 30 scalp sites using a reference located between Fz and Cz. Sintered Ag/AgCl ring electrodes were attached using the BrainCap (Falk-Minow Services, Herrsching-Breitbrunn, Germany), which is part of the MR-compatible EEG recording system BrainAmp-MR (Brain Products Co., Munich, Germany). Electrode impedance was kept below 7 kΩ. Data were transmitted from the amplifier (5 kHz sampling rate, 250 Hz lowpass filter, 10 s time constant) via an optic fibre cable to the computer located outside the scanner room.

This EEG data set represents a particular challenge for time series decomposition, since it contains several distinct physiologically relevant components, as well as artifacts, resulting from the FMRI acquisition process. Removal of these artifacts has

recently become a field of intense research (Allen et al. 2000; Negishi et al. 2004; Niazy et al. 2005).

The main scanner artifact results from the electromagnetic gradient fields which are applied consecutively for each slice of voxels. Only after a complete scan of all slices has been performed, there is a short interval without artifact, before the next scan starts. The gradient fields produce huge inductive responses in all EEG channels which completely obscure the actual EEG signal. Luckily, this artifact is to a certain degree reproducible from scan to scan, therefore, most of its power can be removed by subtracting a template of the artifact, which is generated by averaging (Allen et al. 2000). However, experience has shown that high-frequency components (>30 Hz) of this artifact cannot be removed well by this approach, such that at higher frequencies residual artifacts remain; therefore EEG data sets recorded during FMRI acquisition are usually low-pass filtered, and analysis of higher-frequency phenomena in the EEG becomes impossible. We mention that in the recording of this data set, scanner and EEG systems were not synchronized, whereby the problems with high-frequency artifacts will be aggravated.

Nevertheless, we have chosen to perform this template-based prefiltering, using BrainVision Analyser software (Brain Products Co., Munich, Germany). The markers for each volume were set based on gradient criterion of 1000 μV/ms. For correction, a sliding average calculation for 10 consecutive volumes was used. The data was then sub-sampled to 250 Hz; no additional low-pass filtering was applied.

The data, resulting from this preprocessing, is shown in Fig. 4 (top left panel); for this analysis a subset of 6 EEG channels (Fp1, Fp2, F4, C4, F8, T8) was selected from the complete EEG. The length of the chosen data set was 27.012 seconds. In the figure, a series of epileptic spikes in channels C4 and T8 can be seen, corresponding to two separate epileptic foci. Furthermore, we see spikes related to the heartbeat, especially in channels Fp1 and Fp2, and sleep-related patterns, such as three spindles (at around 3, 13, and 21 seconds) and a K-complex (at 21 seconds). Although no low-pass filtering has been applied, the residual scanner artifacts are not visible to the eye.

However, a closer inspection reveals that at the onset of each new fMRI scan very sharp spikes occur that are not removed by the template filtering step. From a statistical perspective, these can be interpreted as *outliers*, i.e., values affected by artifactual noise of such amplitude that they contain practically no useful information; we have therefore decided to regard the data at 14 such time points within this data set as *missing data* and to replace them by values interpolated from their neighbors. This procedure is not entirely satisfying, and in future work improved methods should be developed; we will briefly return to this point in the discussion.

6.4.2 Results of Analysis

Again we analyze the EEG data by FastICA, MILCA and state space modelling; the resulting components are shown in Fig. 4. Based on a preparatory analysis by multivariate AR modelling, the state space model is defined to consist of 4 first-order and 28 second-order components; this large number of components simplifies the separation of physiological and artifactual components. The actual state dimension

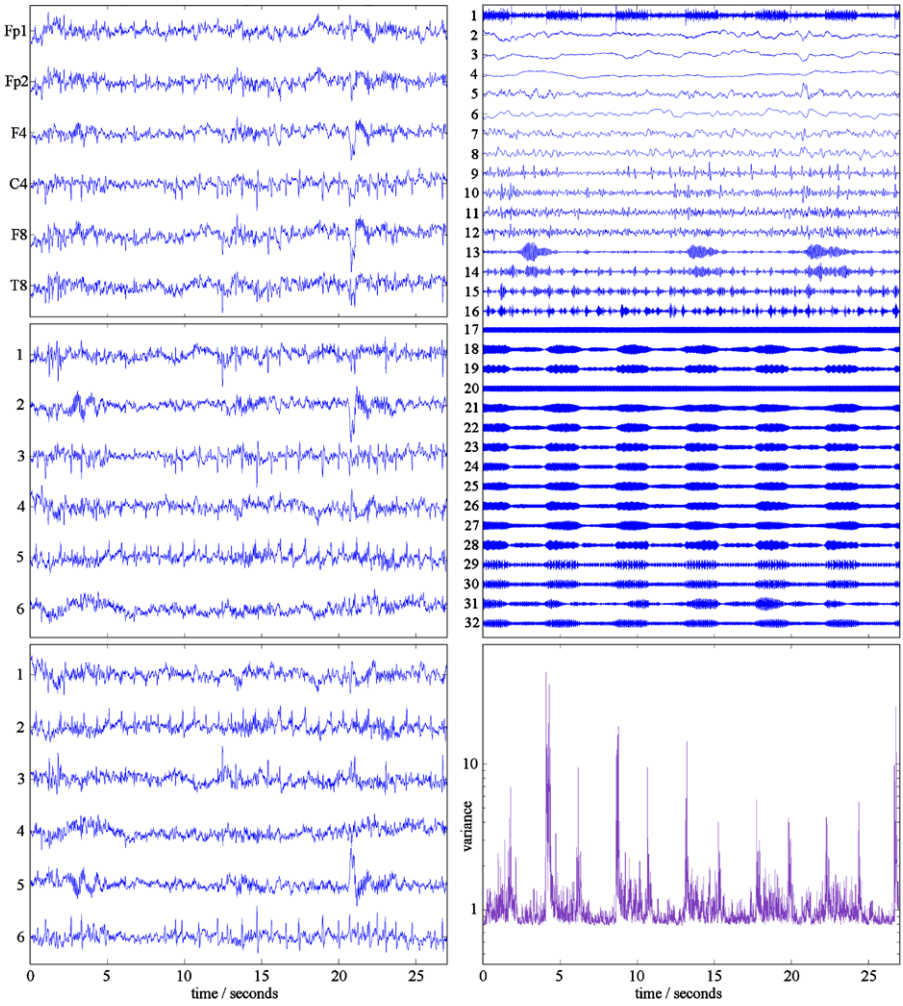

Fig. 4 Electroencephalogram data, recorded during FMRI scanning (*top left panel*); components estimated by FastICA (*middle left panel*), by MILCA (*bottom left panel*) and by state space modelling, using the RTS smoother (*top/middle right panel*). In the state space modelling decomposition, components 1–4 represent AR(1) components, while components 5–32 represent ARMA(2, 1) components. The state variance of the 26th component of the state space model was allowed to change with time (state space GARCH); the time course of this variance is shown *in the bottom right panel* (note the logarithmic scale of the *vertical axis*)

is 60. Here, we again have made use of the freedom of choosing the state space dimension M higher than the data dimension N. State estimates obtained by the RTS smoother are shown in the upper right panel of Fig. 4; first-order and second-order components are ordered again in the same way as described above for the ECG data set.

In the figure, it can be seen that 14 components at higher frequency display a characteristic regular periodic structure with a periodicity of about 4 seconds. It is obvious

that these components represent the residual scanner artifact, with each 4 seconds interval representing two complete head scans. Apparently every second scan produces a weaker artifact, a phenomenon for which the manufacturer of the scanner could not yet offer an explanation. We also note that the topmost component in the panel, a first-order component, displays the same periodic structure. In fact, the AR parameter of this component is negative, whence consecutive values of this component have opposite signs. For this reason, this component could also be regarded as a high-frequency component, corresponding to half the sampling frequency, i.e., the highest possible frequency.

The scanner artifact is known to possess an extraordinarily complicated (albeit approximately reproducible) structure, comprising various periodical components, as well as sudden spikes; for this reason it gives rise to a fairly large number of components in this decomposition. In order to obtain a better description of this artifact, we have employed the model generalization of interacting components, i.e., we have allowed the 15 scanner artifact components to interact via mutual pairs of autoregressive parameters within the state transition matrix A. It would seem inappropriate to describe one complicated artifact by a set of 15 mutually independent components. By this generalization, these 15 components—14 second-order, 1 first-order—are effectively linked into a single component of much higher order.

This step adds no less than 210 additional model parameters to the parameter vector $\boldsymbol{\vartheta}$; computing estimates for these parameters via numerical maximization of likelihood requires considerable additional computational effort. As expected, the likelihood improves indeed, compared to the case without interactions, but it remains to be shown that this improvement is worth the price of 210 new parameters. It is precisely for such purposes that information criteria like AIC or BIC were created. For the sake of brevity, we refrain from going into more detail on this point, but we mention that this introduction of additional parameters passes the test of both AIC and BIC.

The choice of the components to be selected for inclusion into this set of interacting components was done by visual inspection of the components of a decomposition without interacting components; due to the periodic structure of the scanner artifact there were no ambiguous cases, and all components could uniquely be classified as artifactual of non-artifactual. In fact, a plot of the components of the decomposition without interacting components would look quite similar to the plot in the upper right panel of Fig. 4, therefore, we omit it here.

Furthermore, we define one of the scanner artifact components to be a component with state space GARCH; the 26th component is chosen for this purpose, since it is found to contribute strongly to improving the likelihood. Starting from the best non-GARCH model, not only the GARCH parameters, but all model parameters are refitted. No other component is chosen to become a GARCH component. This step adds only three parameters to $\boldsymbol{\vartheta}$, an investment which again is easily justified by comparing the values of AIC or BIC. In the lower right panel of Fig. 4, the time-dependent variance of the GARCH component is displayed. It can be seen that the variance increases and decreases in the same rhythm of the scans that is also visible in the scanner artifact components themselves, with maximum values above 10, while minimum values are below 1. Again, this result confirms that the additional freedom

of GARCH dynamics of the state variance is actually employed for improved modelling of this data, even in a case where in the data the artifact is invisible to the eye.

Also in this analysis, we have used the generalization of employing a nonlinear observation function, in order to reduce possible deviations from Gaussianity. Again, the two additional parameters of this generalization need to justify their existence by improved AIC or BIC values.

Two of the components, numbered 17 and 20, do not show the periodic structure of the scanner artifacts, but instead stationary sine waves (in the figure appearing as dark bands, due to insufficient graphical resolution), similar to the hum noise of the previous example, but with frequencies 48.1676 Hz and 67.7493 Hz; these components certainly represent also artifacts of technical origin.

Components 14, 15 and 16 seem to represent mainly the heartbeat; while many EEGs recorded inside the scanner display pronounced heartbeat artifacts (known as *ballistocardiogram*), this particular data set is only weakly affected, and predominantly high-frequency components are identified.

The remaining 12 components can be regarded as neurophysiologically meaningful. In particular, we see the two epileptic foci (c. 9 and 10), the three sleep spindles (c. 13 and 14; c. 14 also contains heartbeat) and the K-complex (c. 3 and 5).

If we compare these results with the FastICA and MILCA decompositions, shown in the middle and lower left panels of Fig. 4, we see that also these algorithms succeed in identifying some of the components, such as the foci (1 and 3 for FastICA, 3 and 6 for MILCA) and the heartbeat artifact (5 for FastICA, 2 for MILCA). However, the sleep spindles remain spread over several components, and the residual scanner artifacts, as well as the sinusoidal artifacts, remain completely undetected. Like in the previous example, this failure is partly due to the weakness of ICA to identify periodic sources, and partly due to the small number of available components. The result would improve somewhat if all 30 channels of the EEG data set were analyzed by FastICA or MILCA.

6.4.3 Artifact Removal

The state space model decomposition of Fig. 4 can be employed for various purposes; particular physiologically meaningful components within the data, such as epileptic spikes or sleep spindles, may be selected for further analysis, or they may be employed as regressors for subsequent FMRI analysis. Since the model contains an estimate of the observation matrix, any chosen component may be transformed back to "data space", while suppressing all other components, such that the spatial distribution may be studied.

If we regard artifact suppression as the main goal of the decomposition, we would typically remove all obvious artifact components and transform the remaining components back into data space. The filtered data set can then be further analyzed by other methods; in particular, higher-frequency properties may be investigated, since they have not been removed by low-pass filtering.

We demonstrate the latter possibility briefly by comparing the power spectra of the unfiltered data set (as shown in the upper left panel of Fig. 4) with the power spectra

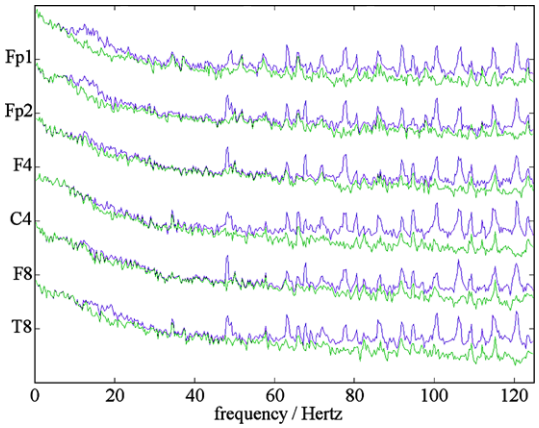

Fig. 5 Power spectrum of the EEG data analyzed in Fig. 4 before removal of residual scanner artifacts (*blue curves*) and after artifact removal (*green curves*); residual artifacts were identified by state space modelling and removed as described in the text. The *vertical axis* represents the logarithm of spectral power

of the filtered data set, where the filtering consists of removing components 1 and 14–32. The resulting spectra are shown (in logarithmic scale) in Fig. 5. The spectra of the unfiltered data (in blue) show a pattern of various peaks at higher frequencies, but it remains unclear which peaks are certainly artifactual and which may belong to physiologically relevant components. After filtering, we obtain the spectra shown in green; most of the peaks have been removed, especially the removal of the sinusoidal components with frequencies 48.1676 Hz and 67.7493 Hz is clearly visible. Some peaks, however, have remained, although with considerably reduced power (note the logarithmic scale of the figure); it is difficult to decide whether they still represent residual artifacts or physiologically meaningful components. We remark that in some channels also at lower frequencies (10–30 Hz) a considerable amount of power has been removed by this filtering. Altogether, it can be seen from Fig. 5 that the artifact removal step considerably improves the quality of the EEG at higher frequencies.

A similar artifact removal step could also be applied to the previous example of the epileptic spiking EEG time series; in that case, the artifact components to be removed from the data represented power supply hum noise and its higher harmonics (components 11, 14, and 15). Also, the epileptic spike component may be employed for various purposes, such as automatic spike detection and counting.

Finally, we remark that since the estimates of the state components are obtained by a smoother and residual data prediction errors (innovations) have been separated, it can be expected that the smoothed components will also lead to substantially improved signal-to-noise ratio, when transformed back into data space.

7 Discussion and Conclusion

In this paper, we have proposed to employ state space modelling as a method for decomposing multivariate time series data into source components, and we have demonstrated the application of this algorithm to three neurological multivariate time series. The model may be pre-specified such that mutually independent components result, but this constraint can easily be generalized to interacting components. State space modelling represents a very well developed concept for time series modelling, which

also includes autoregressive moving-average (ARMA) modelling as a special case; also its application to the task of decomposing time series into independent components forms only a special case. The broader framework behind fitting state space models to time series data is given by the venerable science of *system identification* (Ljung 1999).

We may regard Factor Analysis and Principal Component Analysis as "ancestors" of most data decomposition algorithms (Harman 1976), and it should not be surprising that Factor Analysis has been generalized into various directions, one of them being "non-Gaussian Factor Analysis", better known as Independent Component Analysis (Hyvärinen et al. 2001). If it is intended to take the temporal aspect of the data into account, state space modelling is a natural choice, and the first attempts to merge state space modelling and Factor Analysis have already been proposed around 1985 in the econometrics community, under the denomination of "Dynamical Factor Analysis" (Molenaar 1985), but so far this method has not been widely employed. Independently, a number of authors have developed algorithms for Blind Signal Separation (BSS) based on simultaneous approximative diagonalization of instantaneous and lagged covariance matrices (Molgedey and Schuster 1994; Belouchrani et al. 1997; Ziehe and Müller 1998). While these algorithms do indeed employ some of the information contained in the temporal ordering of the data, they are still not equivalent to state space modelling, for several reasons:

- Lagged covariance matrices are ignorant of the direction of time, while dynamical time series models are based on prediction, and prediction is always directed from past and/or present values to future values. If a multivariate time series was inverted with respect to the direction of time, ICA models would remain invariant, but ARMA and state space models would change; only in the special case of univariate linear models we would have invariance.
- It is well known that ARMA modelling can achieve better predictions with smaller number of parameters, as compared to pure AR modelling, such that it corresponds well to the concept of "parsimonious modelling". In order to achieve the same prediction performance as an ARMA model, a pure AR model would possibly need a large model order, corresponding to a large number of lagged covariance matrices; but simultaneous approximative diagonalization of such a set of many covariance matrices would be impractical.
- The absence of observation noise terms in the ICA models corresponds to these models lacking a probabilistic data model, which represents a severe constraint. The same situation is given with Factor Analysis (representing a proper probabilistic data model) and Principal Component Analysis (lacking a proper probabilistic data model). From this, we see that a generalization of Factor Analysis that would truly deserve the name "non-Gaussian Factor Analysis", necessarily needs to retain the observation noise term of Factor Analysis; this point has recently been emphasised by Beckmann and Smith (2004), who also discuss the consequences of omitting the noise term.

A further advantage of state space modelling is given by the possibility of choosing the state space dimension independent of the data dimension. As long as the (rather mild) condition of observability is fulfilled, the state vectors, as functions of time,

can be uniquely estimated by Kalman filtering, provided a sufficient amount of data is analyzed. The Kalman filter provides state estimates while iterating in forward direction through the data; improved state estimates can be obtained by additionally running the RTS smoother backward through the data.

For given model parameters, estimation of states by the Kalman filter and the RTS smoother is a straightforward procedure. The estimation of suitable model parameters itself is done by non-linear numerical optimization (maximization) of the log-likelihood (or preferably, minimization of information criteria like AIC or BIC). We remark that both the model parameters resulting from this procedure and the state estimates resulting from application of Kalman filtering and RTS smoothing do not claim to reproduce the correct values with respect to an underlying "true model"; i.e., we make no claim for identifiability in the sense of Tong et al. (1991). The models which we fit to given data, are to be regarded merely as *statistically useful models*, i.e., best approximative models within a chosen model class, but this class will certainly not include the "true model". Nobody could reasonably define a model class that would possibly contain the "true model" of human brain.

Fitting large models (corresponding to high-dimensional state spaces) to large data sets by full non-linear numerical optimization is an expensive task in terms of computational time consumption; currently we recommend to analyze at most 10 channels and a few 10^4 time points of data. Further research should be devoted to the development of more efficient approaches to this high-dimensional optimization problem. In this respect, many currently available ICA algorithms offer an advantage, since they can be applied to higher-dimensional data sets, while still consuming only moderate computational time.

The neurological time series analyzed in this paper were chosen for the purpose of demonstrating the potential of state space modelling for typical filtering problems arising in contemporary neuroscience and in practical analysis of clinical time series. The first time series, the fetal electrocardiogram, has already been employed by other authors as a benchmark data set for BSS/ICA (Meinecke et al. 2002; Stögbauer et al. 2004). Here, we have presented a state space model for this data set which provides a superior decomposition, as compared with the competing ICA algorithms, in terms of residual mixing between mother, fetus, and remaining background noise components.

It is interesting to note that if we evaluate quantitatively the residual mutual information of the sets of components, as estimated by MILCA and by state space modelling (RTS smoother), we find for the MILCA components a value 20% smaller than the value for the state space modelling components, i.e., the MILCA decomposition is clearly superior with respect to the strict independence constraint of ICA. Nevertheless, the state space modelling components form a physiologically more plausible decomposition. We interpret this result as illustrative of the difference between imposing an independence constraint either on the level of the particular components, or on the level of the underlying dynamical model: as already mentioned, part of the dependencies within the data set will be due to finite-length effects, and artificially reducing these will result in distorted components, even if, in terms of the value of a sample estimate of mutual information, the constraint has been successfully imposed. Implementing the independence constraint on the level of the dynamical model offers a convenient alternative, since the model will be less susceptible to finite-length effects or other sources of coincidental correlations.

Unlike the fetal electrocardiogram, the other two time series, taken from clinical EEG recordings of patients suffering from epilepsy, have not been discussed in published work so far. They represent much more challenging decomposition tasks, in terms of the number and complexity of the components to be separated and the validity of the assumption of independence. For this reason, application of standard ICA algorithms provides decompositions of only limited usefulness, as has been demonstrated. On the other hand, by state space modelling a number of components could be extracted which correspond well to physiologically meaningful signals, or to known artifacts. In order to achieve these results, three additional generalizations were incorporated into the model: interacting components, non-stationary covariance of driving noise (by state space GARCH modelling) and nonlinear observation functions. The resulting generalized state space models provide superior descriptions of the data, as can be proved by comparing the improvement of likelihood with the increase of the number of model parameters.

By introducing these generalizations, it has been demonstrated that the methodology of classical state space modelling can be easily generalized into various directions. If such generalizations remain within the field of parametric modelling, they can be conveniently accommodated into the standard maximum-likelihood (or, preferably, minimum-AIC/BIC) framework. For the filtering and decomposition tasks discussed, especially the state space GARCH method forms an essential part of our algorithm. In this paper we have, for the sake of brevity, refrained from discussing the theoretical background of state space GARCH modelling in detail. In this context, we would like to mention that putting more effort into the description of the noise driving the state dynamics, as it is done with the state space GARCH method, forms an alternative to designing improved predictive model terms. Any state space model can be decomposed into a "deterministic" predictive term and a stochastic term; this fact is related to the famous Doob–Meyer decomposition theorem of stochastic calculus (Kallenberg 2002).

The usefulness of a general-purpose filter which can separate components with overlapping power spectra and remove artifacts without resorting to low-pass or band-block filtering (such as *notch filtering* for hum noise removal), will be obvious, and for the two EEG data sets presented in this paper several examples for such filtering were demonstrated. Especially the EEG data set recorded during FMRI acquisition represents a challenging example of a filtering task, which attracts considerable attention in current research. We have presented a model which succeeded in identifying a set of residual scanner artifact components, such that these can be easily removed from the data, without sacrificing physiologically meaningful higher-frequency components within the data. In this paper, we have not yet conducted a detailed quantitative study of the performance of this filter, in comparison to other filters which have been proposed for the same task; this point remains as a subject for future work.

Acknowledgements This work was supported by the by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through project SFB 855 "Biomagnetic Sensing" and by the Japanese Society for the Promotion of Science (JSPS) through fellowship ID No. P 03059 and grants KIBAN B No. 173000922301 and WAKATE B No. 197002710002.

The authors are grateful to A. Hyvärinen and to H. Stögbauer and co-workers for making their ICA software packages available for download on the Internet. The ECG data set was made available to the public as part of the DAISY database (http://www.esat.kuleuven.ac.be/sista/daisy).

Appendix

In this Appendix, we summarize in detail the linear state space model, as used in this paper. We begin with the basic, fully linear model with independent components, then we discuss its generalizations.

A.1 Basic Model

For given time series data $\mathbf{y}(t) = (y_1(t), \ldots, y_N(t))^\dagger, t = 1, \ldots, T$, where N denotes the dimension of each data vector and T the length of the time series, the observation equation is given by

$$\mathbf{y}(t) = \mathsf{C}\mathbf{x}(t) + \boldsymbol{\epsilon}(t),$$

where $\mathbf{x}(t) = (x_1(t), \ldots, x_M(t))^\dagger$ denotes the state vector, M the state space dimension, C the $(N \times M)$-dimensional observation matrix and $\boldsymbol{\epsilon}(t)$ a vector of observation noise; the covariance matrix of $\boldsymbol{\epsilon}(t)$ is given by

$$\mathsf{S}_\epsilon = \mathrm{diag}\big(\sigma_{ii}^2\big), \quad i = 1, \ldots, N.$$

Let the dynamical model for the state vector $\mathbf{x}(t)$ be defined by n_1 first-order (AR(1)) components and n_2 second-order (ARMA(2, 1)) components; the generalization to components of higher order is straightforward, but not needed in this paper. Then we have $M = n_1 + 2n_2$. In the observation matrix C the columns numbered $n_1 + 2$, $n_1 + 4, \ldots, M$ contain only zeros, all other elements may assume non-zero values and form part of the set of model parameters.

The dynamical equation is given by

$$\mathbf{x}(t) = \mathsf{A}\mathbf{x}(t-1) + \boldsymbol{\eta}(t),$$

where the $(M \times M)$-dimensional matrix A denotes the transition matrix and $\boldsymbol{\eta}(t)$ denotes a vector of dynamical noise; the covariance matrix of $\boldsymbol{\eta}(t)$ is given by

$$\mathsf{S}_\eta = \begin{pmatrix}
1 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\
0 & 1 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & 1 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & 1 & b_1 & 0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & b_1 & b_1^2 & 0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & 0 & 0 & 1 & b_2 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & 0 & 0 & b_2 & b_2^2 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 1 & b_{n_2} \\
0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & b_{n_2} & b_{n_2}^2
\end{pmatrix},$$

where the identity matrix in the upper left corner of S_η has dimension $(n_1 \times n_1)$. The parameters $b_1, \ldots, b_{n_2}$ represent the moving average (MA) parameters of the model.

The transition matrix A is given by

$$\mathsf{A} = \begin{pmatrix} a^{(1)} & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & a^{(2)} & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & a^{(n_1)} & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & a_1^{(1)} & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & a_2^{(1)} & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & a_1^{(2)} & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & a_2^{(2)} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & a_1^{(n_2)} & 1 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & a_2^{(n_2)} & 0 \end{pmatrix},$$

where $a^{(1)}, \dots, a^{(n_1)}$ is the set of dynamical parameters of AR(1) components, and $(a_1^{(1)}, a_2^{(1)}), \dots, (a_1^{(n_2)}, a_2^{(n_2)})$ is the set of pairs of dynamical parameters of ARMA(2, 1) components, as given in (11). For practical optimization, it is advisable to replace each pair of these parameters by the pair of phase φ (i.e., frequency) and modulus r of the corresponding complex root:

$$a_1 = 2r\cos\varphi, \qquad a_2 = -r^2.$$

The frequency φ can be kept constant for components of known frequency, and the modulus r should be constrained to the interval [0, 1]. For sinusoidal components, the modulus should be exactly 1.0 or very close to this value; such constraints facilitate the optimization procedure.

A.2 Interacting Components

If two components (labelled k and l) are selected for interaction, the corresponding pair of off-diagonal elements of the transition matrix A is allowed to deviate from zero. For the example of two ARMA(2, 1) components, this would give us

$$\mathsf{A} = \begin{pmatrix} \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ \dots & a_1^{(k)} & 1 & \dots & a^{(l,k)} & 0 & \dots \\ \dots & a_2^{(k)} & 0 & \dots & 0 & 0 & \dots \\ \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ \dots & a^{(k,l)} & 0 & \dots & a_1^{(l)} & 1 & \dots \\ \dots & 0 & 0 & \dots & a_2^{(l)} & 0 & \dots \\ \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \end{pmatrix},$$

this is, for ARMA(2, 1) components the interaction term refers only to the first of the two state dimensions belonging to the component within the state vector. In principle, it would also be possible to introduce non-zero interaction terms for the remaining three matrix elements, involving the second state components, but we do not use this variant here. The case of interacting AR(1) components is straightforward.

A.3 Non-stationary Covariance of Driving Noise

If component k is chosen to be a component with non-stationary variance/covariance, a dynamical equation is set up for its standard deviation (known as a state space GARCH model):

$$\sigma^{(k)}(t) = \sigma_0^{(k)} + \sum_{\tau=1}^{p} \alpha_\tau^{(k)} \sigma^{(k)}(t-\tau) + \sum_{\tau=1}^{q} \beta_\tau^{(k)} \hat{\upsilon}^{(k)}(t-\tau),$$

where $\hat{\upsilon}^{(k)}(t)$ represents the noise driving the covariance dynamics. This model corresponds to an ARMA(p, q) model. In this paper we have chosen the simplest case $p = q = 1$. Then there are new model parameters $(\sigma_0^{(k)}, \alpha_1^{(k)}, \beta_1^{(k)})$. Note that the basic state space model (without non-stationary variance/covariance) corresponds to $(\sigma_0^{(k)}, \alpha_1^{(k)}, \beta_1^{(k)}) = (1, 0, 0)$, and these values need to be chosen for those components which remain without non-stationary variance/covariance. It is then possible to formally define the state space GARCH model as a vector model for all state space dimensions in parallel, which simplifies the implementation.

In the earlier papers of Galka et al. (2004) and Wong et al. (2006), state space GARCH models were introduced in which the logarithm of the variance, $2 \log \sigma^{(k)}(t)$, was used as dynamical variable, but in this paper we have decided to formulate the model directly in the standard deviations $\sigma^{(k)}(t)$. Strictly speaking, these variables then should not be called "standard deviations", since they may become negative.

In a standard GARCH model, the driving noise term $\hat{\upsilon}^{(k)}(t)$ would be given by the prediction errors of the data, i.e., the *innovations* $\boldsymbol{\nu}(t)$. In the state space case, the appropriate state prediction errors are not directly available and need to be estimated, using the innovations $\boldsymbol{\nu}(t)$. The estimator which we use here is the same as derived by Wong et al. (2006); for the complete state vector it is given by

$$\hat{\boldsymbol{\upsilon}}(t) = \mathsf{S}_\eta(t) - \mathsf{S}_\eta(t)\mathsf{C}^\dagger \mathsf{S}_\nu^{-1}(t)\mathsf{C}\mathsf{S}_\eta(t) + \mathsf{G}(t)\boldsymbol{\nu}(t)\boldsymbol{\nu}^\dagger(t)\mathsf{G}^\dagger(t),$$

where $\mathsf{S}_\nu(t)$ and $\mathsf{G}(t)$ denote the innovation covariance matrix and the Kalman gain matrix, respectively, both provided by the Kalman filter. From this equation, $\hat{\boldsymbol{\upsilon}}(t)$ is a square matrix; in order to obtain the noise estimates for the individual components, we pick out the diagonal values from this matrix. While this uniquely defines the noise terms for AR(1) components, for each ARMA(2, 1) component there are two diagonal elements; here, we have chosen to simply average over these two elements, but other choices would be possible. The resulting average is denoted by $\hat{\upsilon}^{(k)}(t)$ for the kth component.

Finally, for each component with non-stationary variance/covariance the time-dependent variances are fed into the covariance matrix of $\boldsymbol{\eta}(t)$, which for an

ARMA(2, 1) component corresponds to inserting a (2×2)-dimensional submatrix into the diagonal of S_η according to

$$\mathsf{S}_\eta(t) = \begin{pmatrix} \ddots & \vdots & \vdots & \ddots \\ \dots & (\sigma^{(k)}(t))^2 & b_k(\sigma^{(k)}(t))^2 & \dots \\ \dots & b_k(\sigma^{(k)}(t))^2 & b_k^2(\sigma^{(k)}(t))^2 & \dots \\ \ddots & \vdots & \vdots & \ddots \end{pmatrix},$$

where for simplicity b_k denotes the MA parameter of component k, although in the set of all ARMA(2, 1) components this parameter would probably carry a different label. The case of AR(1) components is straightforward.

A.4 Nonlinear Observation

From (15) and (16), it can be seen that this generalization corresponds to

$$\frac{1}{\gamma}\operatorname{arsinh}\big(\gamma \mathbf{y}(t)\big) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t).$$

Here, we ignore the small distortion of the distribution of $\boldsymbol{\epsilon}(t)$, or preferably, we change the assumptions of our model, such that after the transformation of the noise term its distribution would be a multivariate Gaussian. Thus, before any of the main modelling steps are applied, the data is transformed by the inverse hyperbolic sine function. Since the likelihood should refer to the original, untransformed data, we need to apply a correction. According to elementary probability theory (Gnedenko 1969), for a transformation $z = f(y)$ the probability density functions are related according to

$$p_y(y) = p_z(z)\left|\frac{\partial f(y)}{\partial y}\right| = p_z\big(f(y)\big)\left|\frac{\partial f(y)}{\partial y}\right|,$$

where in case of multivariate variables the absolute value becomes a determinant (Jacobi determinant).

In our case, this determinant is simply the product of the absolute values of the derivatives of the individual elements of the vector $\mathbf{y}(t)$. The derivative of the inverse hyperbolic sine function is given by

$$\frac{1}{\gamma}\frac{\partial}{\partial y}\operatorname{arsinh}(\gamma y) = \frac{1}{\sqrt{\gamma^2 y^2 + 1}}$$

and the value of this function needs to be computed for all values of the data set. For each time point, the log-likelihood, given by (10), is then corrected by adding the logarithm of the corresponding Jacobi determinant.

References

Aït-Sahalia, Y., & Kimmel, R. (2007). Maximum likelihood estimation of stochastic volatility models. *J. Financ. Econ.*, *83*, 413–452.

Akaike, H. (1974a). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. *Ann. Inst. Stat. Math.*, *26*, 363–387.
Akaike, H. (1974b). A new look at the statistical model identification. *IEEE Trans. Autom. Control*, *19*, 716–723.
Akaike, H., & Nakagawa, T. (1988). *Statistical analysis and control of dynamic systems*. Dordrecht: Kluwer Academic.
Allen, P. J., Josephs, O., & Turner, R. (2000). A method for removing imaging artifact from continuous EEG recorded during functional MRI. *NeuroImage*, *12*, 230–239.
Åström, K. J. (1980). Maximum likelihood and prediction error methods. *Automatica*, *16*, 551–574.
Attias, H., & Schreiner, C. E. (1998). Blind source separation and deconvolution: the dynamic component analysis algorithm. *Neural Comput.*, *10*, 1373–1424.
Baldick, R. (Ed.) (2006). *Applied optimization: formulation and algorithms for engineering systems*. Cambridge: Cambridge University Press.
Bar-Shalom, Y., & Fortmann, T. (1988). *Tracking and data association*. San Diego: Academic Press.
Barros, A. K., & Cichocki, A. (2001). Extraction of specific signals with temporal structure. *Neural Comput.*, *13*, 1995–2000.
Basilevsky, A. (1994). *Statistical factor analysis and related methods: theory and applications*. New York: Wiley-Interscience.
Beckmann, C., & Smith, S. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging*, *23*, 137–152.
Beckmann, C., & Smith, S. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage*, *25*, 294–311.
Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., & Moulines, E. (1997). A blind source separation technique using second order statistics. *IEEE Trans. Signal Process.*, *45*, 434–444.
Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econom.*, *31*, 307–327.
Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis, forecasting and control*. San Francisco: Holden-Day.
Brockwell, P. J., & Davis, R. A. (1987). *Time series: theory and methods*. Berlin: Springer.
Cheung, Y. M., & Xu, L. (2003). Dual multivariate auto-regressive modeling in state space for temporal signal separation. *IEEE Trans. Syst. Man Cybern.*, *33*, 386–398.
Choi, S., Cichocki, A., Park, H., & Lee, S. (2005). Blind source separation and independent component analysis: a review. *Neural Inf. Process. Lett. Rev.*, *6*, 1–57.
Chui, C. K., & Chen, G. (1999). *Springer series in information sciences: Vol. 17. Kalman filtering: with real-time applications* (3rd ed.). Berlin: Springer.
Cichocki, A., & Amari, S. (2002). *Adaptive blind signal and image processing*. Chichester: Wiley.
Comon, P. (1994). Independent component analysis, a new concept? *Signal Process.*, *36*, 287–314.
Delorme, A., Sejnowski, T., & Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*, *34*, 1443–1449.
Durbin, J., & Koopman, S. J. (2001). *Time series analysis by state space methods*. Oxford: Oxford University Press.
Dyrholm, M., Makeig, S., & Hansen, L. K. (2007). Model selection for convolutive ICA with an application to spatiotemporal analysis of EEG. *Neural Comput.*, *19*, 934–955.
Engle, R. F., & Watson, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. *J. Am. Stat. Assoc.*, *76*, 774–781.
Galka, A., Yamashita, O., & Ozaki, T. (2004). GARCH modelling of covariance in dynamical estimation of inverse solutions. *Phys. Lett. A*, *333*, 261–268.
Galka, A., Ozaki, T., Bosch-Bayard, J., & Yamashita, O. (2006). Whitening as a tool for estimating mutual information in spatiotemporal data sets. *J. Stat. Phys.*, *124*, 1275–1315.
Galka, A., Wong, K., & Ozaki, T. (2010). Generalized state space models for modeling non-stationary EEG time series. In A. Steyn-Ross & M. Steyn-Ross (Eds.), *Springer series in computational neuroscience*. *Modeling phase transitions in the brain* (pp. 27–52). Berlin: Springer.
Gevers, M. (2006). A personal view of the development of system identification. *IEEE Control Syst. Mag.*, *26*, 93–105.
Gnedenko, B. V. (1969). *The theory of probability*. Moscow: Mir Publishers.
Grewal, M. S., & Andrews, A. P. (2001). *Kalman filtering: theory and practice using MATLAB*. New York: Wiley-Interscience.
Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.
Harvey, A., Koopman, S. J., & Shephard, N. (Eds.) (2004). *State space and unobserved component models*. Cambridge: Cambridge University Press.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, *10*, 626–634.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

James, C., & Hesse, C. (2005). Independent component analysis for biomedical signals. *Physiol. Meas.*, *26*, R15–R39.

Jung, A., & Kaiser, A. (2003). Considering temporal structures in independent component analysis. In: *Proc. 4th int. symp. ICA BSS, ICA 2003* (pp. 95–100). Nara, Japan, Apr. 2003.

Jung, T.-P., Makeig, S., McKeown, M., Bell, A., Lee, T.-W., & Sejnowski, T. (2001). Imaging brain dynamics using independent component analysis. *IEEE Proc.*, *88*, 1107–1122.

Kailath, T. (1968). An innovations approach to least-squares estimation—Part I: linear filtering in additive white noise. *IEEE Trans. Autom. Control*, *13*, 646–655.

Kailath, T. (1980). *Information and system sciences series. Linear systems*. Englewood Cliffs: Prentice-Hall.

Kallenberg, O. (2002). *Foundations of modern probability*. Berlin: Springer.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.*, *82*, 35–45.

Kalman, R. E., Falb, P. L., & Arbib, M. A. (1969). *International series in pure and applied mathematics. Topics in mathematical system theory*. New York: McGraw-Hill.

Ljung, L. (1999). *System identification: theory for the user* (2nd ed.). Englewood Cliffs: Prentice-Hall.

Mehra, R. K. (1971). Identification of stochastic linear systems using Kalman filter representation. *AIAA J.*, *9*, 28–31.

Mehra, R. K. (1974). Identification in control and econometrics: similarities and differences. *Ann. Econ. Soc. Meas.*, *3*, 21–47.

Meinecke, F., Ziehe, A., Kawanabe, M., & Müller, K.-R. (2002). A resampling approach to estimate the stability of one- or multidimensional independent components. *IEEE Trans. Biomed. Eng.*, *49*, 1514–1525.

Miwakeichi, F., Martínez-Montes, E., Valdés-Sosa, P., Nishiyama, N., Mizuhara, H., & Yamaguchi, Y. (2004). Decomposing EEG data into space-time-frequency components using parallel factor analysis. *NeuroImage*, *22*, 1035–1045.

Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*, 181–202.

Molgedey, L., & Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, *72*, 3634–3637.

Negishi, M., Abildgaard, M., Nixon, T., & Constable, R. (2004). Removal of time-varying gradient artifacts from EEG data acquired during continuous fMRI. *Clin. Neurophysiol.*, *115*, 2181–2192.

Neumaier, A., & Schneider, T. (2001). Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Trans. Math. Softw.*, *27*, 27–57.

Niazy, R., Beckmann, C., Iannetti, D., Brady, J., & Smith, S. (2005). Removal of FMRI environment artifacts from EEG data using optimal basis sets. *NeuroImage*, *28*, 720–737.

Otter, P. (1986). Dynamic structural systems under indirect observation: identifiability and estimation aspects from a system theoretic perspective. *Psychometrika*, *51*, 415–428.

Ozaki, T., & Iino, M. (2001). An innovation approach to non-Gaussian time series analysis. *J. Appl. Probab.*, *38*, 78–92.

Pagan, A. R. (1975). A note on the extraction of components from time series. *Econometrica*, *43*, 163–168.

Pearlmutter, B. A., & Parra, L. C. (1997). Maximum likelihood blind source separation: a context-sensitive generalization of ICA. In M. C. Mozer, M. I. Jordan & T. Petsche (Eds.), *Advances in neural information processing systems* (Vol. 9, pp. 613–619). Cambridge: MIT Press.

Protter, P. (1990). *Stochastic integration and differential equations*. Berlin: Springer.

Rauch, H. E., Tung, G., & Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA J.*, *3*, 1445–1450.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, *6*, 461–464.

Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Trans. Inf. Theory*, *11*, 61–70.

Sorenson, H. W. (1970). Least-squares estimation: from Gauss to Kalman. *IEEE Spectr.*, *7*, 63–68.

Stögbauer, H., Kraskov, A., Astakhov, S. A., & Grassberger, P. (2004). Least-dependent-component analysis based on mutual information. *Phys. Rev. E*, *70*, 066123.

Tong, L., Liu, R., Soon, V. C., & Huang, Y. (1991). Indeterminacy and identifiability of blind separation. *IEEE Trans. Circuits Syst.*, *38*, 499–509.

Vigário, R., Sarela, J., Jousmiki, V., Hamalainen, M., & Oja, E. (2000). Independent component approach to the analysis of EEG and MEG recordings. *IEEE Trans. Biomed. Eng.*, *47*, 589–593.

Waheed, K., & Salem, F. M. (2005). Linear state space feedforward and feedback structures for blind source recovery in dynamic environments. *Neural Process. Lett.*, *22*, 325–344.

Wong, K. F. K., Galka, A., Yamashita, O., & Ozaki, T. (2006). Modelling non-stationary variance in EEG time series by state space GARCH model. *Comput. Biol. Med.*, *36*, 1327–1335.

Zhang, L., & Cichocki, A. (2000). Blind deconvolution of dynamical systems: a state space approach. *J. Signal Process.*, *4*, 111–130.

Ziehe, A., & Müller, K.-R. (1998). TDSEP—an efficient algorithm for blind separation using time structure. In L. Niklasson, M. Bodén & T. Ziemke (Eds.), *Proc. 8th int. conf. artificial neural networks, ICANN'98* (pp. 675–680). Berlin: Springer.