Monte Carlo smoothing for non-linear time series

By SIMON J. GODSILL, ARNAUD DOUCET
Signal Processing Group, University of Cambridge, Cambridge CB2 1PZ, U.K.
sjg@eng.cam.ac.uk,ad2@eng.cam.ac.uk

AND MIKE WEST
Institute of Statistics and Decision Sciences, Duke University, Durham N.C., U.S.A. 27708-0251
mw@isds.duke.edu

July 7, 2003

Abstract

We develop methods for performing smoothing computations in general state-space models. The methods rely on a particle representation of the filtering distributions, and their evolution through time using sequential importance sampling and resampling ideas. In particular, novel techniques are presented for generation of sample realizations of historical state sequences. This is carried out in a forward-filtering backward-smoothing procedure which can be viewed as the non-linear, non-Gaussian counterpart of standard Kalman filter-based simulation smoothers in the linear Gaussian case. Convergence in the mean-squared error sense of the smoothed trajectories is proved, showing the validity of our proposed method. The methods are tested in a substantial application for the processing of speech signals represented by a time-varying autoregression and parameterised in terms of time-varying partial correlation coefficients, comparing the results of our algorithm with those from a simple smoother based upon the filtered trajectories.

Key words: Bayesian inference; non-Gaussian time series; non-linear time series; particle filters; sequential Monte Carlo; state-space models

1 Introduction

In this paper we develop Monte Carlo methods for smoothing in general state-space models. To fix notation, consider the standard Markovian state-space model (West & Harrison, 1997)

$$
\begin{aligned}
x_{t+1} &\sim f(x_{t+1}|x_t) && \text{State evolution density} \\
y_{t+1} &\sim g(y_{t+1}|x_{t+1}) && \text{Observation density}
\end{aligned}
$$

where $\{x_t\}$ are unobserved states of the system and $\{y_t\}$ are observations made over some time interval $t \in \{1, 2, ..., T\}$. $f(.|.)$ and $g(.|.)$ are pre-specified state evolution and observation densities which may be non-Gaussian and involve non-linearity. It is assumed throughout that the distributions required can be represented by density functions, and that both $f(.|.)$ and $g(.|.)$ can be evaluated for any valid states and observations x_t and y_t. x_t and y_t may both in general be vectors. We assume that the process $\{x_t\}$ is Markov, generated according to the above state evolution, and that the observation process $\{y_t\}$ is independent conditional upon the state process $\{x_t\}$. Hence an expression for the joint distribution of states and observations can be obtained directly by the probability chain rule

$$
p\left(x_{1:t}, y_{1:t}\right) = f(x_1)\left(\prod_{i=2}^{t} f(x_i|x_{i-1})\right)\left(\prod_{i=1}^{t} g(y_i|x_i)\right)
$$

where $f(x_1)$ is the distribution of the initial state. Here $x_{1:t} = (x_1, ..., x_t)$ and $y_{1:t} = (y_1, ..., y_t)$ denote collections of observations and states from time 1 through t. In proving the validity of our proposed smoothing algorithm a more formal definition of the state space model will be required. This is presented in Appendix A.

A primary concern in many state-space inference problems is sequential estimation of the filtering distribution $p(x_t|y_{1:t})$. Updating of the filtering density can be achieved in principle using the standard filtering recursions

$$
\begin{aligned}
p(x_{t+1}|y_{1:t}) &= \int p(x_t|y_{1:t}) f(x_{t+1}|x_t) dx_t \\
p(x_{t+1}|y_{1:t+1}) &= \frac{g(y_{t+1}|x_{t+1}) p(x_{t+1}|y_{1:t})}{p(y_{t+1}|y_{1:t})}.
\end{aligned}
$$

Similarly, smoothing can be performed recursively backwards in time using the smoothing formula

$$
p(x_t|y_{1:T}) = \int p(x_{t+1}|y_{1:T}) \frac{p(x_t|y_{1:t}) f(x_{t+1}|x_t)}{p(x_{t+1}|y_{1:t})} dx_{t+1}.
$$

Inference in general state-space models has been revolutionized over the past decade by the introduction of cheap and massive computational resources, and the consequent development and widespread application of Monte Carlo methods. In batch-based scenarios Markov chain Monte Carlo (MCMC) methods have been widely used, and a variety of powerful tools has been developed and proven in application, see for example Carlin, Polson & Stoffer (1992), Carter & Kohn (1994), Shephard (1994), Shephard & Pitt (1997), De Jong (1997), Aguilar, Huerta, Prado & West (1999), Aguilar & West (1998) Pitt & Shephard (1999b) and Aguilar & West (2000). However, it is not always straightforward to construct an effective MCMC sampler in models with significant degrees of non-linearity and non-Gaussianity. Specifically, in these cases it can be hard to construct effective proposal distributions, either over collections of states simultaneously or even for single states conditional upon all others. The danger then is that the MCMC will be slowly mixing and may never converge to the target distribution within realistic time scales.

Alternative Monte Carlo strategies based upon sequential importance sampling, known generically as *particle filters*, have been rapidly emerging in areas such as target tracking for radar, communications, econometrics and computer vision (West, 1993; Gordon, Salmond & Smith, 1993; Kitagawa, 1996; Liu & Chen, 1998; Doucet, Godsill & Andrieu, 2000; Liu & West, 2001; Pitt & Shephard, 1999a; West & Harrison, 1997; Doucet, De Freitas & Gordon, 2001). These methods allow propagation of completely general target filtering distributions through time using combinations of importance sampling, resampling and local MCMC moves. The methods have been proven for many examples including highly non-linear models that are not easily implemented using standard MCMC.

In particle filtering methods the filtering density is approximated with an empirical distribution formed from point masses, or *particles*,

$$p(x_t|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta_{x_t^{(i)}}(x_t), \qquad \sum_{i=1}^{N} w_t^{(i)} = 1, \; w_t^{(i)} \geq 0 \tag{1}$$

where δ is the Dirac delta function and $w_t^{(i)}$ is a weight attached to particle $x_t^{(i)}$. Particles at time t can be updated efficiently to particles at time $t+1$ using importance sampling and resampling methods.

The theory of particle filtering is now quite well developed. For example, the empirical measure of equation (1) converges almost surely to the distribution associated to $p(x_t|y_{1:t})$ for all $t > 0$ as $N \rightarrow \infty$ under quite mild conditions on the state space model. Moreover, rates of convergence to zero have been established for expectations of mean-square error of functionals with respect to this filtering density. Hence particle filters are rigorously validated as a means for tracking the distribution of, and estimating the value of a hidden state over time. Some recent advances in convergence analysis can be found in Del Moral (1998), Crisan, Del Moral & Lyons (1999), Crisan & Lyons (1999), Crisan & Doucet (2000) and Crisan (2001).

While particle filtering theory and practice is now quite well established, smoothing aspects are less so. Existing approaches to smoothing with particle filters have been aimed at approximating the individual *marginal* smoothing densities $p(x_t|y_{1:T})$, either using the two-filter formula (Kitagawa, 1996) or forward filtering-backward smoothing (Doucet *et al.*, 2000; Hürzeler & Künsch, 2000). In many applications these marginal distributions are of limited interest, as investigations of historical states generally focus on *trajectories* and hence require consideration of collections of states together. If a single 'best' estimate for the smoothed trajectory is required, then Godsill, Doucet & West (2001) provide a sequential methodology for Maximum *a posteriori* sequence estimation based on dynamic programming and the Viterbi algorithm. However, a single best estimate is rarely appropriate in the Bayesian inference setting, especially when distributions are multimodal, and here we aim for random generation of state sequences.

The mew methods provide a completion of particle filtering methodology which allows random generation of entire historical trajectories drawn from the joint smoothing density $p(x_{1:t}|y_{1:t})$. The method relies firstly upon a forward filtering pass which generates and stores a particle-based approximation to the filtering density at each time step. Then a backwards 'simulation smoothing' pass is carried out in order to generate sampled realizations from the smoothing density. The method can be seen as the nonlinear/non-Gaussian analogue of the forward filtering/backwards sampling algorithms developed for linear Gaussian models and hidden Markov models (Carter & Kohn, 1994; Frühwirth-Schnatter, 1994; De Jong & Shephard, 1995). The proposed method is quite distinct from the MAP estimation procedure of **?** in which the forward particle filter is simply used to generate a grid of possible state values at each time point, with the Viterbi algorithm used to trace out the most probable state trajectory through that grid of state values.

The organization of the paper is as follows. In Section 2 the basic particle filtering and smoothing framework is described. Section 3 introduces the proposed simulation smoother algorithm for general state-space models. A proof of convergence is given for the new simulation smoother method in the appendices. The method is evaluated in Section 4 for a standard nonlinear model and in Section 5 with an extensive application to speech data analysis. Finally, some discussion is given in Section 6.

2 Filtering and Smoothing using Sequential Importance Sampling

In this section we review the standard procedure for filtering and smoothing using sequential importance sampling. In practice this is found to be highly effective in the filtering mode, but, as will be demonstrated in our simulations, it can give very poor results in the smoothing mode. Refer for example to Doucet *et al.* (2001) and Doucet *et al.* (2000) for a detailed overview of these standard methods. A more formal description of the particle filter is given in

Appendix A, including a statement the theorem required to prove the smoother of the next section.

Suppose we have at time t weighted particles $\{x_{1:t}^{(i)}, w_t^{(i)}; i = 1, 2, ..., N\}$ drawn from the smoothing density $p(x_{1:t}|y_{1:t})$. We can consider this to be an empirical approximation for the density made up of point masses:

$$p(x_{1:t}|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta_{x_{1:t}^{(i)}}(x_{1:t}), \ \sum_{i=1}^{N} w_t^{(i)} = 1, \ w_t^{(i)} \geq 0 \tag{2}$$

In order to update the smoothing density from time t to time $t+1$, factorize it as follows:

$$p(x_{1:t+1}|y_{1:t+1}) = p(x_{1:t}|y_{1:t}) \times \frac{g(y_{t+1}|x_{t+1})f(x_{t+1}|x_t)}{p(y_{t+1}|y_{1:t})}$$

where the denominator is constant for a given dataset. In order to proceed from time t to $t+1$ one selects trajectories from the approximation (2). In the simplest case (the 'bootstrap' filter (Gordon *et al.*, 1993; Kitagawa, 1996)) N trajectories are drawn at random with replacement from $\{x_{1:t}^{(i)}; i = 1, 2, ..., N\}$ with probabilities $\{w_t^{(i)}; i = 1, 2, ..., N\}$. In more sophisticated schemes some part-deterministic variance reduction selection scheme is applied (Kitagawa, 1996; Liu & Chen, 1998; Doucet *et al.*, 2000; Carpenter, Clifford & Fearnhead, 1999) A new state is then generated randomly from an importance distribution $q(x_{t+1}|x_{1:t}, y_{t+1})$ and appended to the corresponding trajectory $x_{1:t}$. The importance weight is updated to

$$w_{t+1} \propto \frac{g(y_{t+1}|x_{t+1})f(x_{t+1}|x_t)}{q(x_{t+1}|x_{1:t}, y_{t+1})}.$$

Other selection schemes aim to improve performance at future time points by introducing a bias into the selection step. In these cases an additional term is included into the weight expression to allow for this bias. Examples of this include the most basic sequential imputations procedures of Liu & Chen (1995), where selection is only rarely carried out and weights are updated incrementally throughout the filtering pass, and the auxiliary particle filtering approach of Pitt & Shephard (1999a) in which a bias is intentionally introduced with the aim of boosting the number of particles in useful regions of the state space. For further discussion of these issues see Godsill & Clapp (2001).

The smoothing methods described in this section will be referred to as the standard trajectory-based smoothing method. Filtering is obtained as a simple corollary of the smoothing technique by simply discarding the past trajectories $x_{1:t}$ once the update has been made to $t+1$. It is clear that the selection (or 'resampling') procedure will lead to high levels of degeneracy in smoothed trajectories using this method. This is demonstrated in later simulations, and motivates the development of novel smoothing methods based upon backwards simulation in the next section.

3 Smoothing using backwards simulation.

The new method proposed here assumes that Bayesian filtering has already been performed on the entire dataset, leading to an approximate representation of $p(x_t|y_{1:t})$ for each time step $t \in \{1, ..., T\}$, consisting of weighted particles $\{x_t^{(i)}, w_t^{(i)};\ i = 1, 2, ..., N\}$. We note that the method is independent of the precise filtering algorithm and that any particle filtering scheme, whether deterministic or Monte Carlo, can be used. Recall that the primary goal here is to obtain sample realizations from the entire smoothing density in order to exemplify and generate insight into the structure of the smoothing distribution for collections of past states together. This can be based on the factorization

$$p(x_{1:T}|y_{1:T}) = p(x_T|y_{1:T}) \prod_{t=1}^{T-1} p(x_t|x_{t+1:T}, y_{1:T}) \tag{3}$$

where, using the Markovian assumptions of the model, we can write

$$\begin{aligned} p(x_t|x_{t+1:T}, y_{1:T}) &= p(x_t|x_{t+1}, y_{1:t}) \\ &= \frac{p(x_t|y_{1:t}) f(x_{t+1}|x_t)}{p(x_{t+1}|y_{1:t})} \\ &\propto p(x_t|y_{1:t}) f(x_{t+1}|x_t). \end{aligned} \tag{4}$$

Forward filtering generates a particulate approximation to $p(x_t|y_{1:t})$. Since we have from the above that $p(x_t|x_{t+1}, y_{1:T}) \propto p(x_t|y_{1:t}) f(x_{t+1}|x_t)$, we obtain immediately the following modified particle approximation:

$$p(x_t|x_{t+1}, y_{1:T}) \approx \sum_{i=1}^{N} w_{t|t+1}^{(i)} \delta_{x_t^{(i)}}(x_t)$$

with modified weights

$$w_{t|t+1}^{(i)} = \frac{w_t^{(i)} f(x_{t+1}|x_t^{(i)})}{\sum_{j=1}^{N} w_t^{(j)} f(x_{t+1}|x_t^{(j)})}. \tag{5}$$

This revised particulate distribution can now be used to generate states successively in the reverse-time direction, conditioning upon future states. Specifically, given a random sample $\widetilde{x}_{t+1:T}$ drawn approximately from $p(x_{t+1:T}|y_{1:T})$, take one step back in time and sample $\widetilde{x}_t$ from $p(x_t|\widetilde{x}_{t+1}, y_{1:T})$. The pair $(\widetilde{x}_t, \widetilde{x}_{t+1:T})$ is then approximately a random realization from $p(x_{t:T}|y_{1:T})$. Repeating this process sequentially back over time produces the following general "smoother-realization" algorithm:

Algorithm 1 ***- Sample realizations***

- Choose $\widetilde{x}_T = x_T^{(i)}$ with probability $w_T^{(i)}$.

- For $t = T-1$ to 1 :
  - Calculate $w_{t|t+1}^{(i)} \propto w_t^{(i)} f(\widetilde{x}_{t+1}|x_t^{(i)})$ for each $i = 1, ..., N$;
  - Choose $\widetilde{x}_t = x_t^{(i)}$ with probability $w_{t|t+1}^{(i)}$.
- $\widetilde{\mathbf{x}}_{1:T} = (\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_T)$ is an approximate realization from $p(x_{1:T}|y_{1:T})$.

Further independent realizations are obtained by repeating this procedure as many times as required. The computational complexity for each realization is $O(NT)$, in contrast with $O(N^2T)$ required for marginal smoothing procedures (Kitagawa, 1996; Doucet *et al.*, 2000; Hürzeler & Künsch, 2000), however it should be realised that the computations in our method are then repeated for each realization drawn.

In appendix A the convergence of the smoothed realizations is proved in terms of mean-squared error for state estimation as the number of particles tends to infinity.

4 Example 1: A nonlinear time series model

The new methods are first demonstrated for a standard nonlinear time series model (Kitagawa, 1996; West, 1993; Gordon *et al.*, 1993). This model has been used extensively for testing of numerical filtering techniques and here we use it to show the functionality and extended utility of the proposed smoother compared to the other available techniques (Kitagawa, 1996; Doucet *et al.*, 2000; Hürzeler & Künsch, 2000).

The state-space equations are as follows:

$$
\begin{aligned}
x_t &= \frac{x_{t-1}}{2} + 25\frac{x_{t-1}}{1+x_{t-1}^2} + 8\cos(1.2t) + v_t \\
y_t &= \frac{(x_t)^2}{20} + w_t
\end{aligned}
$$

where $v_t \sim \mathcal{N}(0, \sigma_v^2)$ and $w_t \sim \mathcal{N}(0, \sigma_w^2)$ and here $\sigma_v^2 = 10$ and $\sigma_w^2 = 1$ are considered fixed and known. The initial state distribution is $x_1 \sim \mathcal{N}(0, 10)$. The representation in terms of densities $f(x_t|x_{t-1})$ and $g(y_t|x_t)$ is straightforward.

A typical dataset simulated from this model is shown in Fig. 1. Filtering is performed using a standard bootstrap particle filter (Gordon *et al.*, 1993) with $N = 10,000$ particles. Filtering is applied to the same dataset as Fig. 1. There is clear evidence for strong non-Gaussianity and multimodality in the filtering distributions, see for example the density estimates obtained from the particle filter output at times $t = 64$ and $t = 3$ in Fig. 2.

Smoothing is carried out using the proposed smoother, drawing 10,000 realizations from the smoothing density. A small random selection of the smoothed trajectories drawn from $p(x_{1:100}|y_{1:100})$ is shown in Fig. 3. Now multimodality in the smoothing distribution can be seen, with separated clusters of trajectories visible in several parts of the time series. Histogram estimates of the individual smoothing distributions $p(x_t|y_{1:100})$ are shown in Fig. **??**.

The sampled realizations are useful in themselves, but they can also be used to study and visualize in more detail the characteristics of the multivariate smoothing distribution. Note that this is a capability well beyond that of (Kitagawa, 1996; Doucet *et al.*, 2000; Hürzeler & Künsch, 2000) which generate only the smoothed *marginals* $p(x_t|y_{1:T})$. Figs. 4-7 show visualizations of a selection of bivariate marginals estimated from $p(x_{1:100}|y_{1:100})$, using 2-dimensional scatter plots and kernel density estimates. The new smoothing method allows visualization of multimiodality and complex interactions between states in a way which is not possible with the existing (univariate) marginal smoothing techniques. Repeated independent runs of the particle filter/smoother identified essentially identical features in the smoothing distribution, which gives us some confidence in the accuracy of the results. Different types of interaction become important in the smoothing density as the parameters of the model are changed. For example, with $\sigma_v^2 = 1$ and $\sigma_w^2 = 9$, we expect the dynamics of the state to play a more important role than in the previous simulation. This is borne out by the computed smoothing densities from a new set of simulations; see for example Figs. 8 and 9, in which some diagonal structure is clearly present. Higher-dimensional smoothing marginals can be studied using 3-dimensional visualization tools such as those available in Matlab.

5 Example 2: Application to Speech signals represented by TVAR models

We now present a substantial application taken from the field of speech processing and analysis. Speech signals are inherently time-varying in nature, and any realistic representation should thus involve a model whose parameters evolve over time. One such model is the time-varying autoregression (TVAR) (Prado, West & Krystal, 1999; Kitagawa & Gersch, 1996) in which the coefficients of a standard autoregression are allowed to vary according to some probability law. These models are of very wide utility and importance in engineering, scientific and socio-economic applications, but are typically applied subject to severe restrictions on the models for time-variation in autoregressive coefficients for analytic reasons. More realistic models for patterns of variation over time in autoregressive parameters, and hence for the resulting "local" correlation and spectral structures, lead to intractable models and hence the need for simulation based approximations, especially in sequential analysis contexts.

Here we consider a non-linear parameterization of the TVAR model in terms of partial correlations (PARCOR) coefficients (Friedlander, 1982). This is especially relevant in acoustical contexts such as speech processing since the PARCOR coefficients can be regarded as the parameters of a linear acoustical tube whose characteristics are time-varying. This acoustical tube model can be regarded as an approximation to the characteristics of the vocal tract (Proakis, Deller & Hansen, 1993). By allowing the width parameters of the acoustical tube model, and hence the instantaneous PARCOR coefficients, to vary over time we can allow for the physical changes which take place in the vocal tract shape as the speaker utters different sounds. The non-linear model implied by such a scheme is not readily estimated using standard optimal techniques such as MCMC or approximation methods such as the extended Kalman filter, owing to the strongly non-linear nature of the transformation between TVAR coefficients and instantaneous PARCOR coefficients, so we see sequential Monte Carlo filtering and smoothing as the method of choice in this case. A related, although not identical, TVAR model has been considered in a different application by Kitagawa & Gersch (1996).

5.1 Model specifications

A signal process $\{z_t\}$ is generated in the standard fashion from a Gaussian distribution centred at the linear prediction from the previous time step

$$f(z_t|z_{t-1:t-P}, a_t, \sigma_{e_t}) = \mathcal{N}\left(\sum_{i=1}^{P} a_{t,i} z_{t-i}, \sigma_{e_t}^2\right).$$

Here $a_t = (a_{t,1}, a_{t,2}, \ldots, a_{t,P})'$ is the time-varying AR(P) coefficient vector and $\sigma_{e_t}^2$ is the innovation variance at time t. Note that both the AR coefficient vector and the innovation variance are assumed time-varying, so we will specify evolution models for these too. As noted above, the changes in AR coefficients over time will model changes in the vocal tract shape as the speaker makes different utterances. Time variation in innovation variance will model changes in the strength of the excitation signal in the glottis. The speech signal $\{z_t\}$ is assumed to be partially observed in additive independent Gaussian background noise so that the observation process $\{y_t\}$ is generated according to

$$g(y_t|z_t, \sigma_v) = \mathcal{N}\left(z_t, \sigma_v^2\right)$$

where σ_v^2 is here assumed to be constant and known, corresponding to a fixed level of background noise in the environment which has been measured during a silent period. While some applications consider the environment to be noise-free, we argue that in any speech setting there will always be a certain degree of noise present which should be modelled in order to capture the true character of the unobserved signal $\{z_t\}$.

Furthermore, this framework allows us to perform noise reduction for noisy speech signals, a task which is frequently required for applications in mobile telephony and speech recognition, for example. Note that the extension of our methods to the case of time-varying and correlated observation noise is immediate.

For our simulations, a Gaussian first order autoregression is assumed for the log-standard deviation $\phi_{e_t} = \log(\sigma_{e_t})$, namely

$$f(\phi_{e_t}|\phi_{e_{t-1}}, \sigma^2_{\phi_e}) = \mathcal{N}\left(\alpha\phi_{e_{t-1}}, \sigma^2_{\phi_e}\right)$$

where α is a positive constant just less than unity and $\sigma^2_{\phi_e}$ is a fixed hyperparameter.

The model now requires specification of the time variation in TVAR coefficient vector a_t itself. This is the main interest in our application, as we wish to find a model which is physically meaningful for the application and easily interpretable. Possibly the simplest choice of all, and most common in previous work (Prado *et al.*, 1999; Kitagawa & Gersch, 1996), is a first order Gaussian autoregression directly on the coefficients a_t:

$$f(a_t|a_{t-1}, \sigma^2_a) = \mathcal{N}\left(\beta a_{t-1}, \sigma^2_a I\right)$$

More elaborate schemes of this sort are possible, such as a higher order autoregression involving AR coefficients from further in the past, or a non-identity covariance matrix, the latter being a key feature of such models for some authors and applications, as in the work of (and references in) Prado *et al.* (1999) and West & Harrison (1997). However, models of this form do not have a particularly strong physical interpretation for acoustical systems such as speech. Moreover, an unconstrained time-varying autoregression exhibits large ('unstable') oscillations inconsistent with real speech data. Improved stability can be achieved by ensuring that the instantaneous poles, i.e. the roots of the polynomial $(1 - \sum_{i=1}^{P} a_{t,i} L^{-i})$, lie strictly within the unit circle. This can be achieved by constraining the autoregression appropriately to have zero probability in the unstable region for the coefficients. Such a condition, however, is very expensive to simulate using rejection sampling methods. A possible means of achieving greater stability in the model would be to model the roots directly; see Huerta & West (1999a) and Huerta & West (1999b) for the time-invariant case. However, there are unresolved issues here of dealing with complex roots which evolve into real roots and *vice versa*; see Prado *et al.* (1999). These issues do not arise if one works in the reflection coefficient, or equivalently partial correlation (PARCOR) coefficient domain (Friedlander, 1982). Here the equivalent condition is that each reflection coefficient must simply be constrained to the interval $(-1, +1)$. The standard Levinson recursion is used to transform between a_t and the reflection coefficients ρ_t. Many models are possible for the time variation of ρ_t, including random walks based on beta distributions and inverse logit transformed normal, and all would be feasible in our sequential Monte Carlo framework. We have chosen a simple truncated normal autoregression for discussion here, namely:

Random PARCOR model

$$f_t(\rho_t|\rho_{t-1}, \sigma_a^2) \propto \begin{cases} \mathcal{N}(\beta\rho_{t-1}, \sigma_a^2 I), & \max\{|\rho_{t,i}|\} < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

where β is a coefficient just less than 1. All of the simulations in this paper have made use of this model. As we have already hinted, the TV-PARCOR framework is appealing not simply because of the ease of checking stability and evaluating the transition density but also because a reflection coefficient model has a strong physical interpretation in certain systems, notably speech and other acoustic sounds which are generated through tube-like mechanisms.

The state space model is now fully specified. The state vector is

$$x_t = (z_{t:t-P+1},\ \rho_t, \phi_{e_t})'.$$

The hyperparameters σ_a^2, α, β, $\sigma_{\phi_e}^2$ and σ_v^2 are assumed pre-specified and fixed in all the simulations. The initial state probability is truncated Gaussian over the 'stable' parameter region for the model.

5.2 Filtering and smoothing

The first step in analysing the data is to perform a complete forward sweep of a Monte Carlo filtering algorithm to produce weighted particles $\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N$ for $t = 1, 2, ...T$, drawn approximately according to $P(dx_t|y_{1:t})$. Since our smoothing method is quite general and not tied to any particular filtering method we do not consider all the possibilities in detail. We have experimented with various types of Monte Carlo filter, adapted to our specific TVAR model, using the standard sequential importance sampling (SIS) filter (Doucet *et al.*, 2000; Gordon *et al.*, 1993; Liu & Chen, 1998), the *auxiliary particle filter* (APF) (Pitt & Shephard, 1999a) and schemes which incorporate local MCMC moves to improve the filtering approximation (MacEachern, Clyde & Liu, 1999; Gilks & Berzuini, 2000). We observe empirically a moderate improvement in performance when the APF is used and for some challenging cases it is also worthwhile to incorporate MCMC moves. The importance function employed for the unobserved state is the state transition density, modified such that the current z_t is simulated exactly from its full conditional density which is Gaussian, i.e. we use:

$$q(x_t|x_{t-1}, y_t) = p(z_t|z_{t-1:t-P}, a_t, \sigma_{e_t}, y_t) f(\rho_t|\rho_{t-1}, \sigma_a^2) f(\phi_{e_t}|\phi_{e_{t-1}}, \sigma_{\phi_e}^2).$$

A full discussion of the relative merits of the various possible filtering schemes and importance functions applied to a related TVAR model can be found in Godsill & Clapp (2001). Following the filtering pass, smoothing is then carried out using the new method.

5.3 Results

A long section of noisy speech data is presented in Figure 10, see caption for details. The time-varying characteristics of the signal are clearly evident, with data points 1-6800 (approx.) representing the word 'rewarded' and 6800-10000 the word 'by'. A filtering and smoothing analysis is performed for the first 1000 data points. A white Gaussian noise signal with known standard deviation $\sigma_v = 0.02$ is added to the signal. The observed noisy data are shown in Figure 11 with known noise standard deviation $\sigma_v = 0.02$. The remaining fixed hyperparameters used were $\sigma_a = 0.01$, $\alpha = 0.99$, $\beta = 1$, $\sigma_{\phi_e} = 0.001$. These were determined empirically by trial and error and by past experience with similar audio datasets (Godsill & Rayner, 1998). The TVAR model order was $P = 4$; this was chosen in accordance with the findings of Vermaak, Andrieu, Doucet & Godsill (1999) and also gave the best results in subjective listening tests performed on the smoothed signal estimates.

The number of particles required will depend upon the dimensionality of the state vector, which can be quite high in a model such as this, and also on the degree of posterior uncertainty about those parameters; again, for this model the posterior distributions of parameters such as the TV-PARCOR coefficients can be quite diffuse, hence requiring a large number of particles for accurate representation. Filtering and smoothing were carried out using a wide variety of particle numbers, ranging from $N = 100$ up to $N = 20000$. Filtered estimates of quantities such as posterior mean parameter values were found to become quite robust from one realization of the filter to another provided the number of particles exceeded 1000.

The last 200 observed data points are shown in Figure 12. Filtered means and 5/95 percentiles for the TV-PARCOR coefficients are shown in Figure 13 for the last 200 data points, by which time it is assumed that the effects of the Gaussian initial state prior are negligible. The estimated trajectories are quite random in nature, reflecting the uncertainty about their value without the aid of retrospective smoothing. Similarly, Figure 14 shows the filtered mean estimate for the signal process. While some smoothing has occurred, it is fairly clear that the resulting estimate has followed the shape of the noisy signal too closely as a result of filtering without any retrospective sequence analysis.

By comparison, smoothing computations were then performed, using either our proposed backward sampling algorithm from Section 3 or the standard filtered trajectory approach outlined in Section 2. Figure 15 shows 10 realizations from the smoothing density using our proposed smoother. Figure 16 shows four separate realizations plotted on top of the observed noisy data and Figure 17 shows the mean of 10 realizations plotted with the observed data. These results show that good diversity is achieved by the backward sampling smoother and that (visually) much improved sequences are generated when compared with the filtered mean estimates of Figure 14.

Realizations from the smoothing density for the TV-PARCOR coefficients are given in Figure 18. A plausible degree of uncertainty is indicated by this result, and this may be compared with results obtained using the standard trajectory-based smoother. In order to highlight the differences between our proposed method and the standard method we smooth retrospectively back to $t = 600$. Once again, graphs of 10 realizations from the smoothing density are presented, this time comparing our proposed method and the standard method, see Figures 19 and 20. It is clear that the standard method degenerates to a single trajectory quite rapidly, while our proposed method achieves a plausible degree of variability between sampled realizations right back to $t = 600$. This is typical of the results we have observed over many different sections of the data. The proposed method is always observed to improve on the standard method when equal numbers of particles are used for the filtering pass. Of course, our method can also degrade if the number of filtered particles is too small to contain plausible smoothed trajectories. An example of this can be seen in Figure 21, where it is clear from inspection that the trajectories generated are too tightly clustered around a particular 'modal' trajectory, and hence a misleading inference about posterior uncertainty could be obtained. Nevertheless, the result is still visually improved compared with the standard method for the same number of particles, Figure 22.

Finally a filtering/smoothing pass was carried out on an entire sentence of speech, lasting some 2.5 seconds. Results can then be auditioned by listening to the simulated signal trajectories. An example of this can be found on the website `http://www-sigproc.eng.cam.ac.uk/~sjg/TV-PARCOR` where noisy speech and extracted speech can be compared. The results are satisfactory, eliminating some of the high-frequency frame-based artefacts ('musical noise') observed with standard speech enhancement algorithms, although the TVAR model causes some sibilance during unvoiced consonant sounds such as 'ss'.

6 Discussion

Recent years have seen a huge surge of interest in particle filtering, motivated by practical problems of sequential analysis in dynamic models in many areas of engineering, the natural sciences and socio-economics (Doucet *et al.*, 2001). Our work here is not specific to any one algorithm, and takes the established notion of sequentially updated particulate representations of posterior distributions in state space models as the starting point for smoothing. In speech processing as in other applications, it is often critically important to "look back over time" for several or many time steps in order to assess and evaluate how new data revises the view of the recent past. Hence smoothing algorithms are key, and our work here develops effective approaches that apply whatever filtering method is adopted.

We have developed and presented fairly simple and efficient methods for generation of sample realizations of joint smoothing densities in a general model context. Smoothing has not been stressed by earlier authors in the sequential simulation literature, and where it has been studied approaches have been limited to approximating the time-specific marginal smoothing distributions for individual states. We reiterate that this narrow focus is limiting and potentially misleading in many applications. Investigations of patterns of changes in historical states should focus on the joint trajectories of past states and hence necessarily involve consideration of joint smoothing densities, not simply the collection of marginals. Generating sample realizations is the most efficient, effective and intuitive approach to studying complicated, multivariate joint distributions, hence our focus on sampling algorithms for smoothing. This concept parallels that in forecasting, where studies of "sampled futures" is a key and critical exercise in any serious modelling-for-forecasting activity.

There are current research challenges in many aspects of the sequential simulation arena, including real needs for improved particle filtering algorithms, and reconciliation of the several variants of sequential importance sampling, resampling, and auxiliary particle methods. The current paper ignores issues of learning on fixed model parameters in addition to time-varying states, a broader problem also ignored by most other authors in the field, but critical in many applications such as the challenging multifactor models of Pitt & Shephard (1999b) and Aguilar & West (2000). In our current work with TVAR models we are developing analyses for both parameters and states using the *auxiliary particle plus* methods of Liu & West (2001). It should be noted that the smoothing methods developed and illustrated here apply directly in this context also, so providing a comprehensive smoothing algorithm.

Acknowledgement

This work was performed under partial support of grants from the Engineering and Physical Sciences Research Council and the National Science Foundation.

A Proof of convergence for backwards simulation smoother

We will specify a more formal measure-theoretic framework as in Crisan (2001) and Crisan & Doucet (2000) for proof of convergence. Let (Ω, F, P) be a probability space on which two vector real-valued stochastic processes $X = \{X_t,\ t \in \mathbb{N}^*\}$ and $Y = \{Y_t,\ t \in \mathbb{N}^*\}$ are defined; let n_x and n_y be the dimensions of the state space of X and Y, respectively. The process X is unobserved whereas the process Y is observed.

Let X be a Markov process with respect to the filtration $\mathcal{F}_t \triangleq \sigma(X_s, Y_s, s \in \{1, \dots, t\})$, having initial distribution $X_1 \sim f(dx_1)$ and transition kernel

$$f(dx_t|\, x_{t-1}) \triangleq P(X_t \in dx_t|\, X_{t-1} = x_{t-1})\,. \tag{7}$$

We assume that the observations are statistically independent given the signal and satisfy

$$g\left(dy_t \middle| x_t\right) \triangleq P\left(Y_t \in dy_t \middle| X_t = x_t\right). \tag{8}$$

For the sake of simplicity, we assume that $f\left(dx_t \middle| x_{t-1}\right)$ and $g\left(dy_t \middle| x_t\right)$ admit densities with respect to the Lebesgue measure; that is $f\left(dx_t \middle| x_{t-1}\right) = f\left(x_t \middle| x_{t-1}\right) dx_t$ and $g\left(dy_t \middle| x_t\right) = g\left(y_t \middle| x_t\right) dy_t$, corresponding to the densities $f()$ and $g()$ in the main text.

The following standard shorthands will be used. If μ is a measure, φ is a function and K is a Markov kernel,

$$(\mu, \varphi) \triangleq \int \varphi d\mu$$

$$\mu K\left(A\right) \triangleq \int \mu\left(dx\right) K\left(A \middle| x\right)$$

$$K\varphi\left(x\right) \triangleq \int K\left(dz \middle| x\right) \varphi\left(z\right).$$

We will also use the following notation:

$$\pi_{t|t-1}(dx_t) \triangleq P(dx_t|Y_{1:t-1} = y_{1:t-1})$$

and

$$\pi_{t|t}(dx_t) \triangleq P(dx_t|Y_{1:t} = y_{1:t}) \quad \text{(the‘filtering’measure)}$$

The general particle filtering method described in section 2 seeks successively to approximate these two measures using randomly propagated empirical measures.

More generally, the following shorthand is used for smoothing distributions:

$$\pi_{t_1:t_2|t_3}(dx_{t_1:t_2}) \triangleq P(dx_{t_1:t_2}|Y_{1:t_3=}y_{1:t_3})$$

The weighted approximation to the filtering measure is defined as

$$\pi_{t|t}^N\left(dx_t\right) \triangleq \sum_{i=1}^N w_t^{(i)} \delta_{x_t^{(i)}}\left(dx_t\right). \tag{9}$$

while the unweighted measure following resampling (selection) is:

$$\pi_{t|t}^{\prime\,N}\left(dx_t\right) \triangleq \frac{1}{N} \sum_{i=1}^N \delta_{x_t^{\prime(i)}}\left(dx_t\right). \tag{10}$$

Similarly, N smoothed realizations generated using Algorithm 1 can be used to form an unweighted approximation to the joint smoothing density:

$$\pi_{t:T|1:T}^{N}\left(dx_{t:T}\right) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta_{x_{t:T}^{(i)}}\left(dx_{t:T}\right). \tag{11}$$

We will prove in Theorem 2 the convergence of this approximating measure to the true smoothing measure $\pi_{t:T|1:T}$ for any $t \in \{1, 2, ..., T\}$.

Define a (joint) measure $\rho_t\left(dx_{t-1:t}\right) \triangleq \pi_{t-1|t-1}\left(dx_{t-1}\right) q\left(dx_t \middle| y_t, x_{t-1}\right)$, chosen such that it is absolutely continuous with respect to $\pi_{t-1:t|t-1}\left(dx_{t-1:t}\right)$ and such that h_t is the following (strictly positive) Radon Nykodym derivative:

$$h_t\left(y_t, x_{t-1}, x_t\right) \propto \frac{\pi_{t-1:t|t-1}\left(dx_{t-1:t}\right)}{\rho_t\left(dx_{t-1:t}\right)}.$$

We now state sufficient conditions for convergence of the particle filter. Let $B\left(\mathbb{R}^n\right)$ be the space of bounded, Borel measurable functions on $\mathbb{R}^n$ and denote for any $\varphi \in B\left(\mathbb{R}^n\right)$

$$\|\varphi\| \triangleq \sup_{x\in\mathbb{R}^n} \varphi\left(x\right).$$

Assume from now on that the observation process is fixed to a given observation record $Y_{1:T} = y_{1:T}$. All subsequent convergence results will be presented on this basis.

Consider the following assumption:

Assumption 1 $\pi_{t-1:t|t}$ *is absolutely continuous with respect to* ρ_t. *Moreover,* $g\left(y_t \middle| x_t\right) h_t\left(y_t, x_{t-1}, x_t\right)$ *is positive and bounded in argument* $\left(x_{t-1}, x_t\right) \in \left(\mathbb{R}^{n_x}\right)^2$.

The following theorem for the particle filter is then a direct consequence of Crisan (2001) and Crisan & Doucet (2000).

Theorem 1 *(Crisan & Doucet (2000)) Under Assumption 1, for all* $t > 0$, *there exists* $c_{t|t}$ *independent of* N *such that for any* $\phi \in B\left(\mathbb{R}^{n_x}\right)$

$$E\left[\left(\left(\pi_{t|t}^{N}, \phi\right) - \left(\pi_{t|t}, \phi\right)\right)^2\right] \leq c_{t|t}\frac{\|\phi\|^2}{N}.$$

where the expectation is over all realizations of the random particle method.

Now define

$$\overline{f}\left(x\right) \triangleq \left\|f\left(x\middle|\cdot\right)\right\|$$

and consider the following assumption.

Assumption 2 *For any $t \in (1, \ldots, T)$, one has*

$$\left(\pi_{t|T}, \left(\frac{\overline{f}}{\pi_{t|t-1}}\right)^2\right) < \infty.$$

We can now state the main theorem concerning convergence of the proposed simulation smoothing algorithm:

Theorem 2 *Under Assumptions 1 and 2, for all $t \in (1, \ldots, T)$, there exists $c_{t|T}$ independent of N such that for any $\phi \in B\left(\mathbb{R}^{n_x T}\right)$*

$$E\left[\left(\left(\pi^N_{1:T|T}, \phi\right) - \left(\pi_{1:T|T}, \phi\right)\right)^2\right] \leq c_{1|T} \frac{\|\phi\|^2}{N}$$

where $c_{1|T}$ can be computed using the backward recursion

$$c_{t|T} = \left(\left(c_{t+1|T}\right)^{1/2} + 2\left(c_{t|t}\right)^{1/2}\left(\pi_{t+1|T}, \left(\frac{\overline{f}}{\pi_{t+1|t}}\right)^2\right)^{1/2}\right)^2. \qquad (12)$$

and $c_{t|t}$ is given by Theorem 1.

Proof. We prove here that for any $\phi_t \in B\left(\mathbb{R}^{n_x(T-t+1)}\right)$, there exists $c_{t|T}$ independent of N such that

$$E\left[\left(\left(\pi^N_{t:T|T}, \phi_t\right) - \left(\pi_{t:T|T}, \phi_t\right)\right)^2\right] \leq c_{t|T} \frac{\|\phi_t\|^2}{N}$$

The proof proceeds by induction. Theorem 1 ensures that the result is true for $t = T$. Now, rewrite $\left(\pi_{t:T|T}, \phi_t\right)$ as follows:

$$\left(\pi_{t:T|T}, \phi_t\right) = \left(\pi_{t+1:T|T}, \frac{\left(\pi_{t|t}, \phi_t f\right)}{\pi_{t|t} f}\right).$$

Then decompose the error term $\left(\pi^N_{t:T|T}, \phi_t\right) - \left(\pi_{t:T|T}, \phi_t\right)$ as

$$\begin{aligned}
&\left(\pi^N_{t:T|T}, \phi_t\right) - \left(\pi_{t:T|T}, \phi_t\right) \qquad (13)\\
&= \left(\left(\pi^N_{t+1:T|T} - \pi_{t+1:T|T}\right), \frac{\left(\pi^N_{t|t}, \phi_t f\right)}{\pi^N_{t|t} f}\right)\\
&\quad + \left(\pi_{t+1:T|T}, \frac{\left(\pi^N_{t|t}, \phi_t f\right)\left(\pi_{t|t} - \pi^N_{t|t}, f\right)}{\left(\pi^N_{t|t}, f\right)\pi_{t+1|t}}\right)\\
&\quad + \left(\pi_{t+1:T|T}, \frac{\left(\pi^N_{t|t} - \pi_{t|t}, \phi_t f\right)}{\pi_{t+1|t}}\right).
\end{aligned}$$

Then one has by Minkowski's inequality

$$
\begin{aligned}
&E\left[\left(\pi^N_{t:T|T},\phi_t\right)-\left(\pi_{t:T|T},\phi_t\right)\right]^{1/2}\\
&\qquad\le E\left[\left(\left(\pi^N_{t+1:T|T}-\pi_{t+1:T|T}\right),\frac{\left(\pi^N_{t|t},\phi_t f\right)}{\pi^N_{t|t}f}\right)^2\right]^{1/2}\\
&\qquad\quad+E\left[\left(\pi_{t+1:T|T},\frac{\left(\pi^N_{t|t},\phi_t f\right)\left(\pi_{t|t}-\pi^N_{t|t},f\right)}{\left(\pi^N_{t|t},f\right)\pi_{t+1|t}}\right)^2\right]^{1/2}\\
&\qquad\quad+E\left[\left(\pi_{t+1:T|T},\frac{\left(\pi^N_{t|t}-\pi_{t|t},\phi_t f\right)}{\pi_{t+1|t}}\right)^2\right]^{1/2}. \qquad (14)
\end{aligned}
$$

Consider now the three terms on the right hand side.
First term. For any $x_{t+1:T}$, one has

$$
\left\|\frac{\left(\pi^N_{t|t},\phi_t f\right)}{\pi^N_{t|t}f}\right\|\le\|\phi_t\|,
$$

thus we obtain using the induction hypothesis

$$
E\left[\left(\left(\pi^N_{t+1:T|T}-\pi_{t+1:T|T}\right),\frac{\left(\pi^N_{t|t},\phi_t f\right)}{\pi^N_{t|t}f}\right)^2\right]\le c_{t+1|T}\frac{\|\phi_t\|^2}{N}. \qquad (15)
$$

Second term. For any $x_{t+1:T}$, one has

$$
\left|\frac{\left(\pi^N_{t|t},\phi_t f\right)\left(\pi_{t|t}-\pi^N_{t|t},\phi_t f\right)}{\left(\pi^N_{t|t},f\right)\pi_{t+1|t}}\right|\le\frac{\left|\left(\pi_{t|t}-\pi^N_{t|t},f\right)\right|}{\pi_{t+1|t}}\|\phi_t\|,
$$

thus

$$
\begin{aligned}
&E\left[\left(\pi_{t+1:T|T},\frac{\left(\pi^N_{t|t},\phi_t f\right)\left(\pi_{t|t}-\pi^N_{t|t},f\right)}{\left(\pi^N_{t|t},f\right)\pi_{t+1|t}}\right)^2\right]\\
&\qquad\le E\left[\left(\pi_{t+1:T|T},\frac{\left|\left(\pi_{t|t}-\pi^N_{t|t},f\right)\right|}{\pi_{t+1|t}}\right)^2\right]\|\phi_t\|^2\\
&\qquad=\left(\pi_{t+1:T|T},\pi^{-2}_{t+1|t}E\left(\left(\pi_{t|t}-\pi^N_{t|t},f\right)^2\right)\right)\|\phi_t\|^2\ \text{(Jensen's inequality)}
\end{aligned}
$$

By Theorem 1 and Assumption 2,

$$E\left(\left(\pi_{t|t}-\pi^N_{t|t},f\right)^2\right)\le c_{t|t}\frac{\overline{f}^2}{N}$$

and hence

$$E\left[\left(\pi_{t+1:T|T},\frac{\left(\pi^N_{t|t},\phi_t f\right)\left(\pi_{t|t}-\pi^N_{t|t},f\right)}{\left(\pi^N_{t|t},f\right)\pi_{t+1|t}}\right)^2\right]\le\frac{c_{t|t}}{N}\left(\pi_{t+1|T},\frac{\overline{f}^2}{\pi^2_{t+1|t}}\right)\|\phi_t\|^2. \tag{16}$$

Third term. One has

$$\begin{aligned}E&\left[\left(\pi_{t+1:T|T},\frac{\left(\pi^N_{t|t}-\pi_{t|t},\phi_t f\right)}{\pi_{t+1|t}}\right)^2\right]\\&\le E\left[\left(\pi_{t+1:T|T},\frac{\left(\pi^N_{t|t}-\pi_{t|t},\phi_t f\right)^2}{\pi^2_{t+1|t}}\right)\right]\quad\text{(Jensen)}\\&=\left(\pi_{t+1:T|T},\pi^{-2}_{t+1|t}E\left(\left(\pi^N_{t|t}-\pi_{t|t},\phi_t f\right)^2\right)\right)\end{aligned}$$

and, by Theorem 1 and Assumption 2, one obtains

$$E\left(\left(\pi^N_{t|t}-\pi_{t|t},\phi_t f\right)^2\right)\le\frac{c_{t|t}}{N}\overline{f}^2\|\phi_t\|^2,$$

so that

$$E\left[\left(\pi_{t+1:T|T},\frac{\left(\pi^N_{t|t}-\pi_{t|t},\phi_t f\right)}{\pi_{t+1|t}}\right)^2\right]\le\frac{c_{t|t}}{N}\left(\pi_{t+1|T},\frac{\overline{f}^2}{\pi^2_{t+1|t}}\right)\|\phi_t\|. \tag{17}$$

The result follows by combining Eq. (13), (15), (16) and (17). ■

References

Aguilar, O., Huerta, G., Prado, R. & West, M. (1999). Bayesian inference on latent structure in time series (with discussion). In J. O. Berger, J. M. Bernardo, A. P. Dawid & A. F. M. Smith (eds.), *Bayesian Statistics 6* 3–26. Oxford: University Press.

Aguilar, O. & West, M. (1998). Analysis of hospital quality monitors using hierarchical time series models. In C. Gatsonis, R. E. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli & M. West (eds.), *Case Studies in Bayesian Statistics in Science and Technology* vol. 4 287–302. New York: Springer-Verlag.

AGUILAR, O. & WEST, M. (2000). Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *J. Bus. Econ. Statist.* **18**.

CARLIN, B. P., POLSON, N. G. & STOFFER, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Am. Statist. Assoc.* **87**(418) 493–500.

CARPENTER, J., CLIFFORD, P. & FEARNHEAD, P. (1999). Building robust simulation-based filters for evolving datasets. *Tech. Rep. University of Oxford* .

CARTER, C. K. & KOHN, R. (1994). On Gibbs Sampling for state space models. *Biometrika* **81**(3) 541–553.

CRISAN, D. (2001). Parficle filters - a theoretical perspective. In A. Doucet, J. F. G. De Freitas & N. J. Gordon (eds.), *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

CRISAN, D., DEL MORAL, P. & LYONS, T. (1999). Discrete filtering using branching and interacting particle systems. *Markov proc. & rel. flds.* **5**(3) 293–318.

CRISAN, D. & DOUCET, A. (2000). Convergence of sequential Monte Carlo methods. Tech. Rep. CUED/F-INFENG/TR.381 Cambridge University Engineering Department.

CRISAN, D. & LYONS, T. (1999). A particle approximation of the solution of the Kushner-Stratonovich equation. *Prob. theor. & rel. flds.* **5**(3) 293–318.

DE JONG, P. (1997). The scan sampler for time series models. *Biometrika* **84**(4) 929–937.

DE JONG, P. & SHEPHARD, N. (1995). The simulation smoother for time series models. *Biometrika* **82**(2) 339–350.

DEL MORAL, P. (1998). Measure valued processes and interacting particle systems. *Ann. Appl. Prob.* **8**(2) 438–495.

DOUCET, A., DE FREITAS, J. F. G. & GORDON, N. J. (eds.) (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

DOUCET, A., GODSILL, S. J. & ANDRIEU, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing* **10** 197–208.

FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *J. Time Ser. Anal.* **15** 183–202.

Friedlander, B. (1982). Lattice filters for adaptive processing. *Proc. IEEE* **70**(8) 829–867.

Gilks, W. R. & Berzuini, C. (2000). Following a moving target - Monte Carlo inference for dynamic Bayesian models. *J. R. Statist. Soc. B* **63** 1–20.

Godsill, S. J. & Clapp, T. C. (2001). Improvement strategies for Monte Carlo particle filters. In A. Doucet, J. F. G. De Freitas & N. J. Gordon (eds.), *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Godsill, S. J., Doucet, A. & West, M. (2001). Maximum *a posteriori* sequence estimation using Monte Carlo particle filters. *Ann. Inst. Stat. Math.* **53**(1) 82–96.

Godsill, S. J. & Rayner, P. J. W. (1998). *Digital Audio Restoration: A Statistical Model-Based Approach*. Berlin: Springer, ISBN 3 540 76222 1.

Gordon, N. J., Salmond, D. J. & Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F* **140**(2) 107–113.

Huerta, G. & West, M. (1999a). Bayesian inference on periodicities and component spectral structure in time series. *J. Time Ser. Anal.* **20**(4) 401–416.

Huerta, G. & West, M. (1999b). Priors and component structures in autoregressive time series models. *J. R. Statist. Soc. B* **61** 881–899.

Hürzeler, M. & Künsch, H. R. (2000). Monte Carlo approximations for general state space models. *J. Comp. Graph. Statist.* **7**(2) 175–193.

Kitagawa, G. (1996). Monte Carlo filter and smoother for nonlinear non-Gaussian state space models. *J. Comp. Graph. Statist.* **5** 1–25.

Kitagawa, G. & Gersch, W. (1996). *Smoothness Priors Analysis of Time Series, Lecture Notes in Statistics #116*. New York: Springer-Verlag.

Liu, J. & West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In A. Doucet, J. F. G. De Freitas & N. J. Gordon (eds.), *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Liu, J. S. & Chen, R. (1995). Blind deconvolution via sequential imputations. *J. Am. Statist. Assoc.* **90**(430) 567–576.

Liu, J. S. & Chen, R. (1998). Sequential Monte Carlo methods for dynamical systems. *J. Am. Statist. Assoc.* **93** 1032–44.

MacEachern, S. N., Clyde, M. & Liu, J. S. (1999). Sequential importance sampling for nonparametric Bayes models: the next generation. *Can. J. Stats.* **27** 251–267.

Pitt, M. K. & Shephard, N. (1999a). Filtering via simulation: auxiliary particle filters. *J. Am. Statist. Assoc.* **94** 590–9.

Pitt, M. K. & Shephard, N. (1999b). Time varying covariances: A factor stochastic volatility approach (with discussion). In J. O. Berger, J. M. Bernardo, A. P. Dawid & A. F. M. Smith (eds.), *Bayesian Statistics 6* 547–570. Oxford: University Press.

Prado, R., West, M. & Krystal, A. D. (1999). Evaluation and comparison of eeg traces: Latent structure in non-stationary time series. *J. Am. Statist. Assoc.* **94** 1083–1095.

Proakis, J., Deller, J. & Hansen, J. (1993). *Discrete-Time Processing of Speech Signals*. New York: Macmillan.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika* **81**(1) 115–131.

Shephard, N. & Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84** 653–67.

Vermaak, J., Andrieu, C., Doucet, A. & Godsill, S. J. (1999). On-line Bayesian modelling and enhancement of speech signals. Tech. Rep. CUED/F-INFENG/TR.361 Cambridge University Engineering Department Cambridge, England.

West, M. (1993). Mixture models, Monte Carlo, Bayesian updating and dynamic models. In J. H. Newton (ed.), *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface* 325–333 Fairfax Station, Virginia. Interface Foundation of North America.

West, M. & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag 2nd edn.

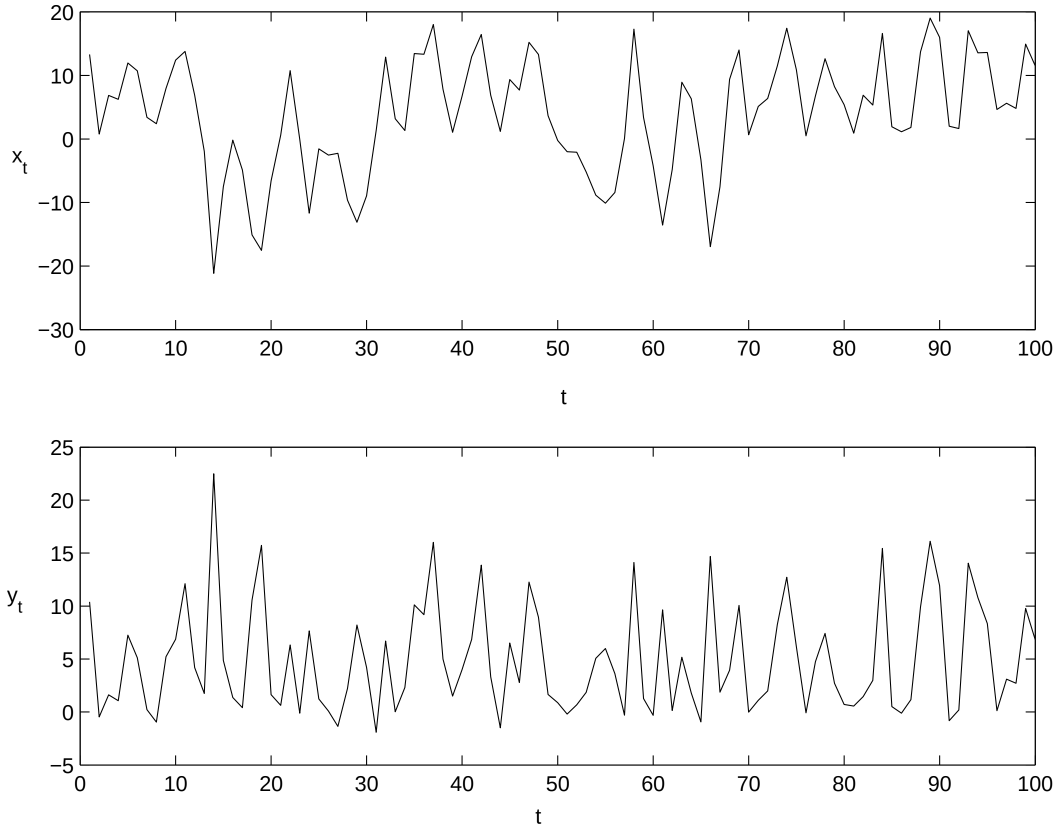

Figure 1: Simulated data from nonlinear time series model. Top: hidden state sequence x_t; bottom: observed data sequence y_t.

.

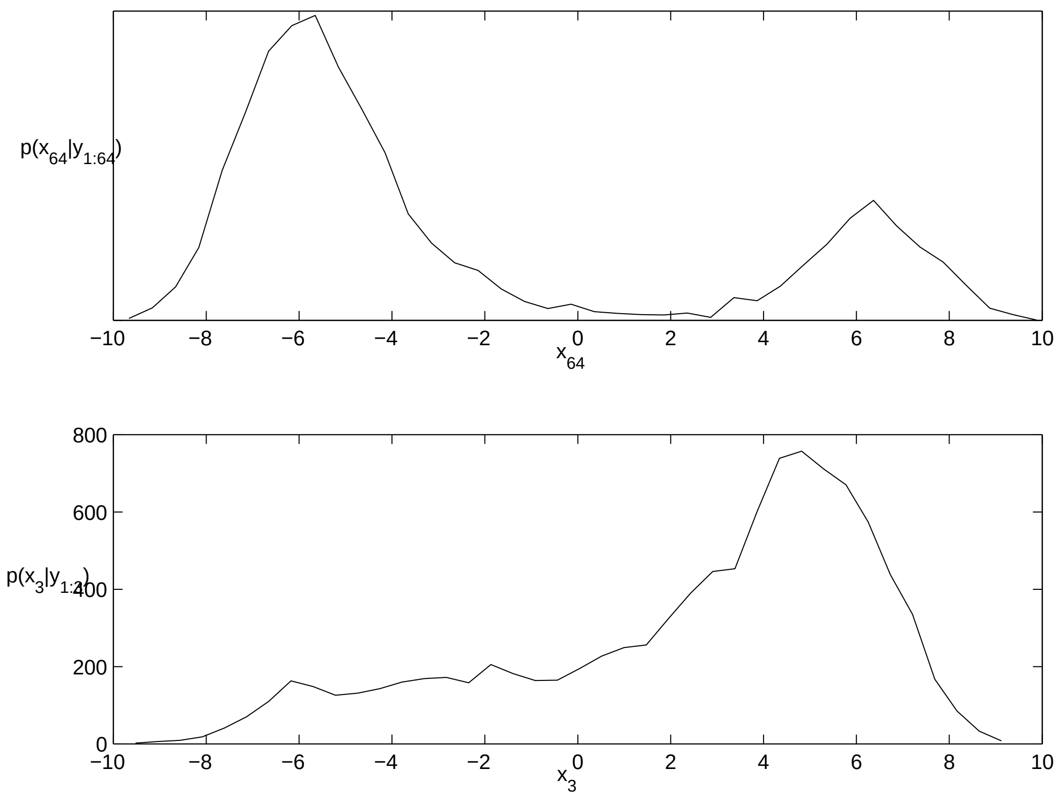

Figure 2: Filtering density estimates from the particle filter output.

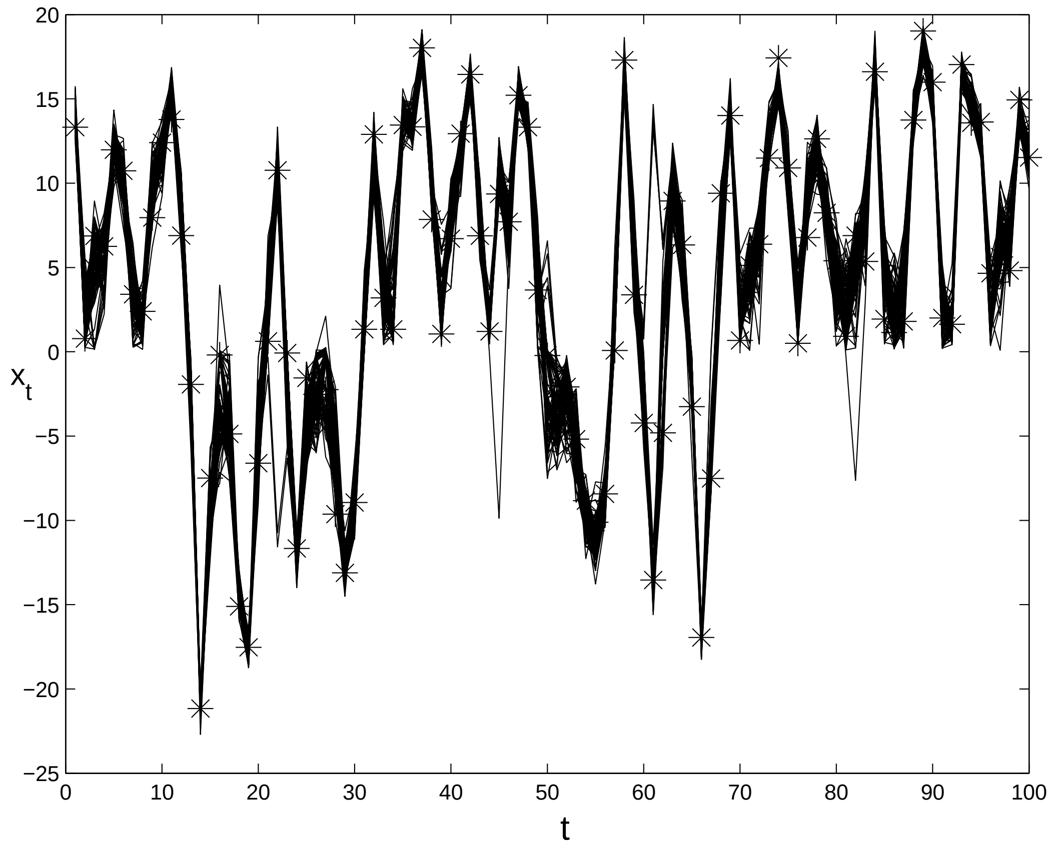

Figure 3: Smoothing trajectories drawn from $p(x_{1:100}|y_{1:100})$. True simulated states shown as '*'.

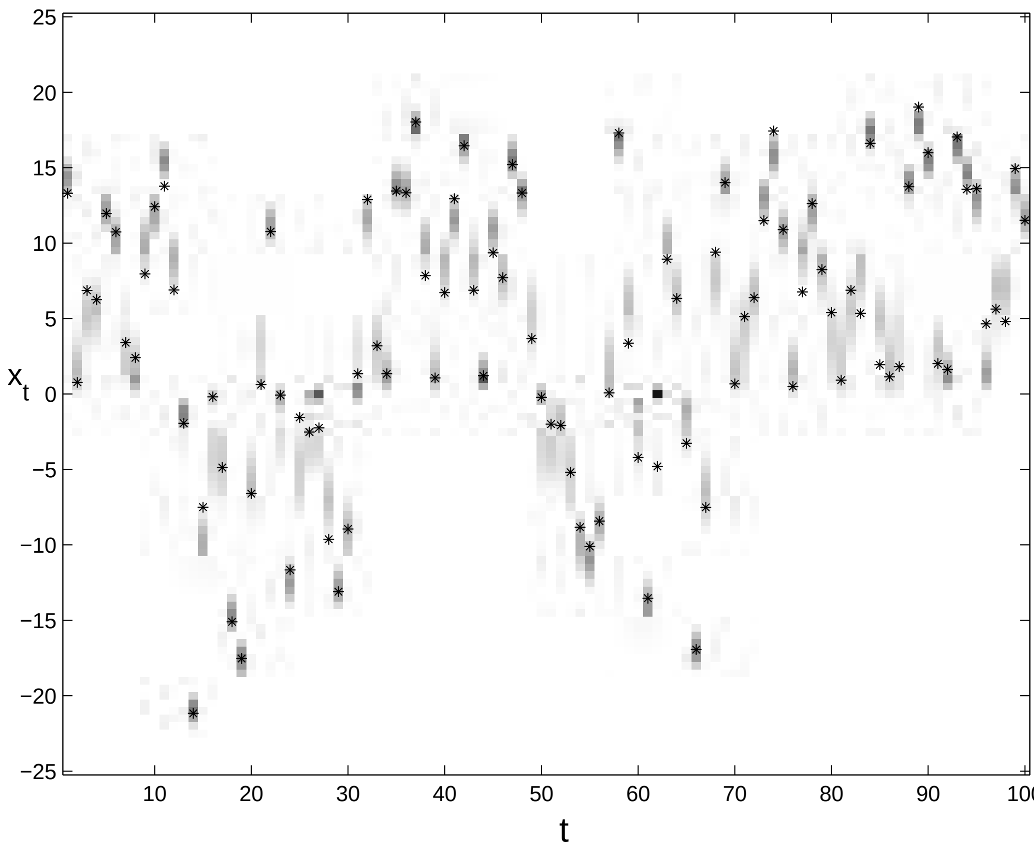

Figure 4: Histogram estimates of smoothing densities, $p(x_t|y_{1:100})$, shown as gray scale intensities in vertical direction. True simulated states shown as '*'.

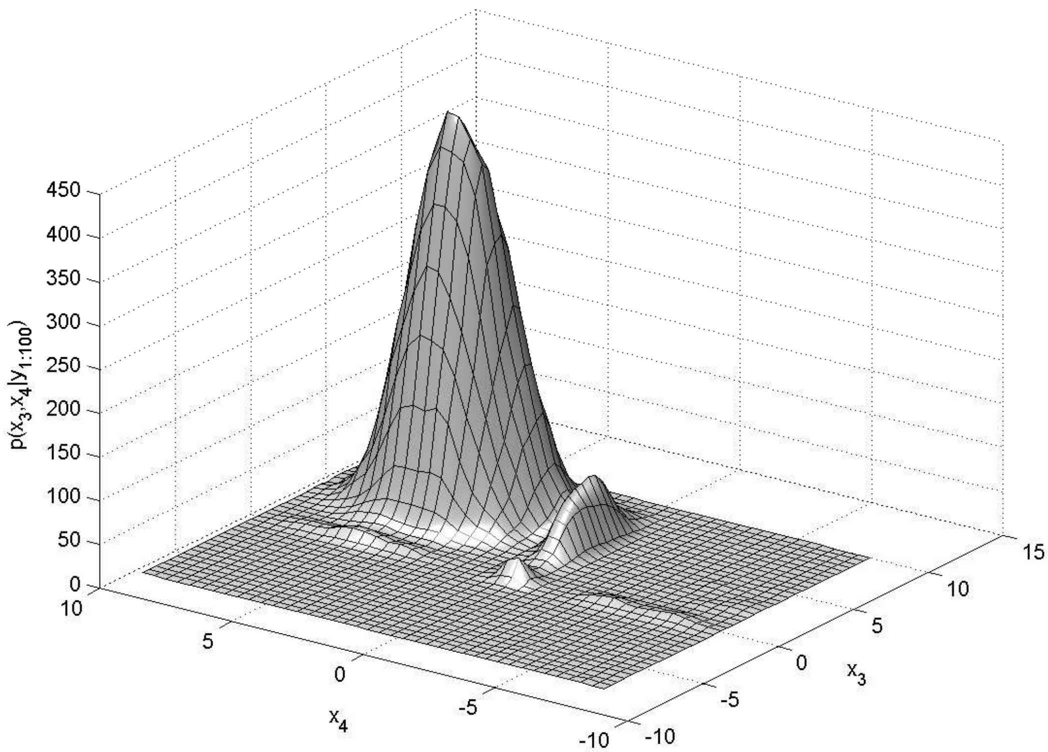

Figure 5: Kernel density estimate for $p(x_{3:4}|y_{1:100})$

.

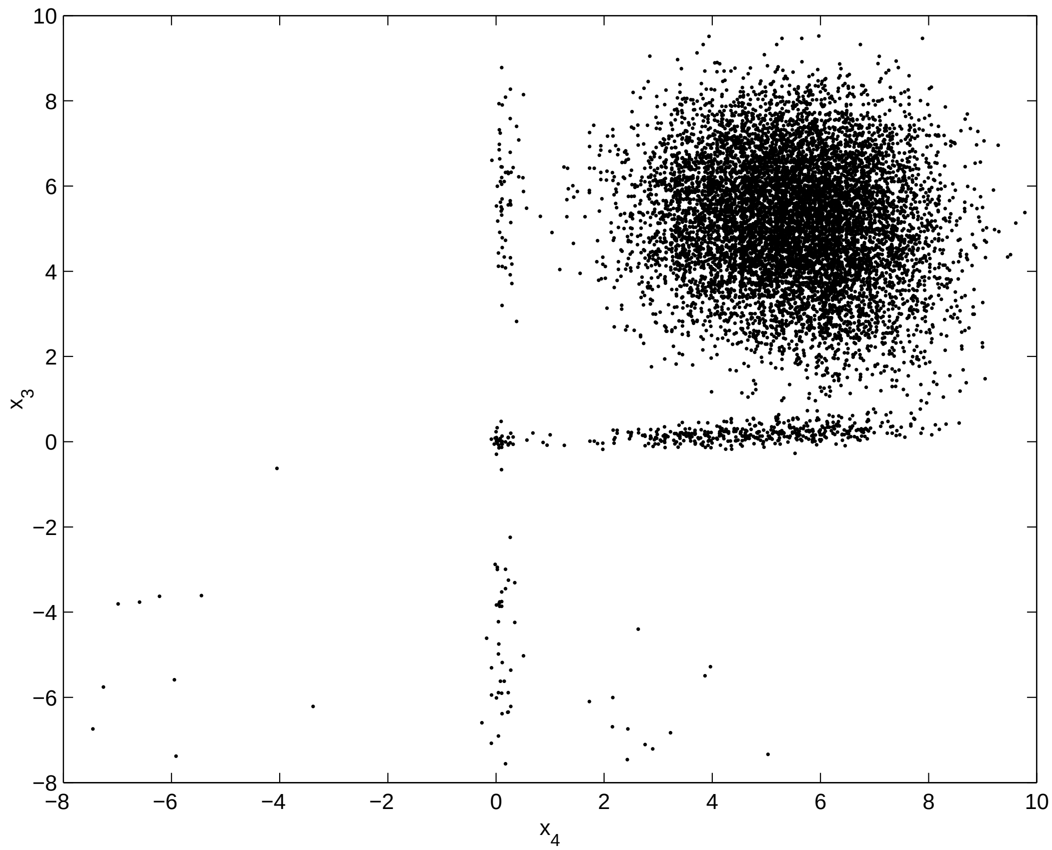

Figure 6: Scatter plot of points drawn from $p(x_{3:4}|y_{1:100})$.

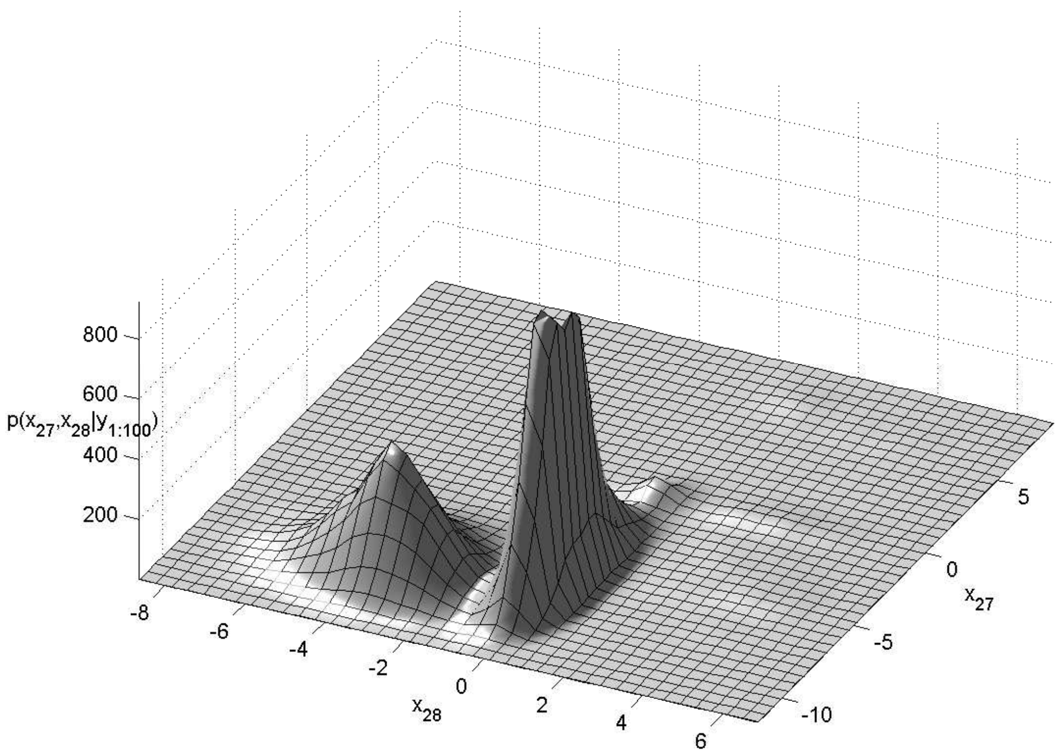

Figure 7: Kernel density estimate for $p(x_{27:28}|y_{1:100})$.

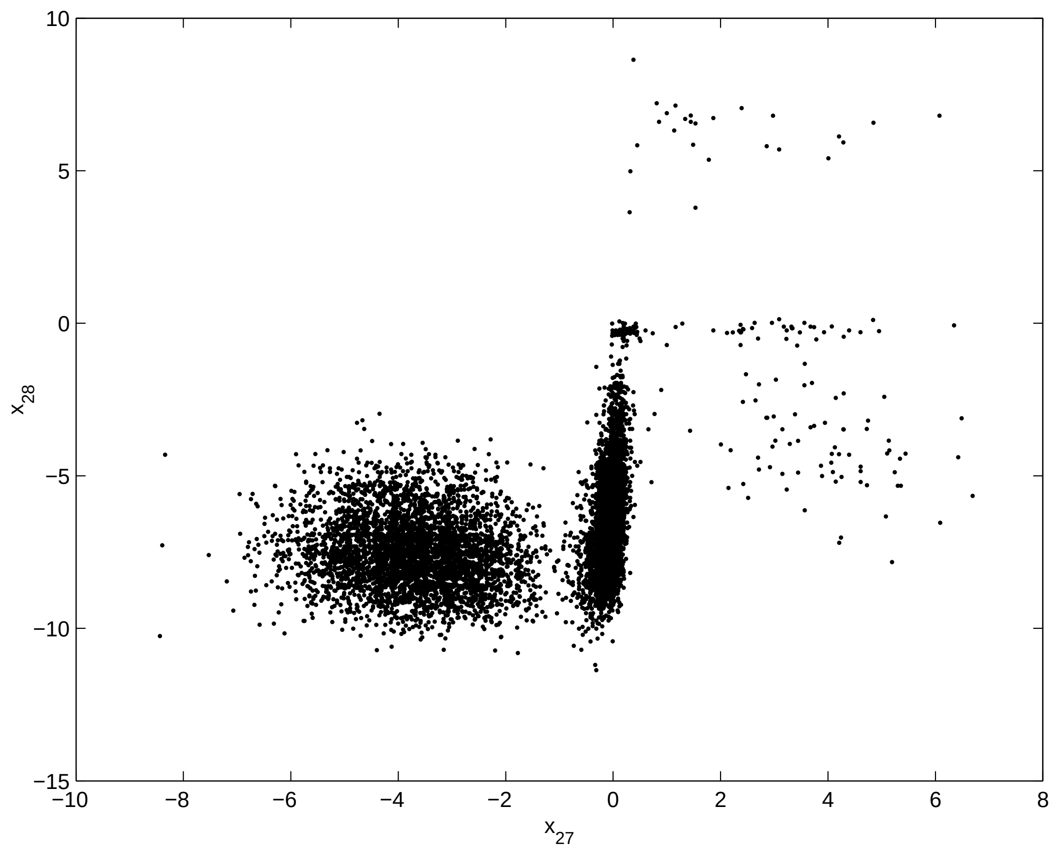

Figure 8: Scatter plot of points drawn from $p(x_{27:28}|y_{1:100})$.

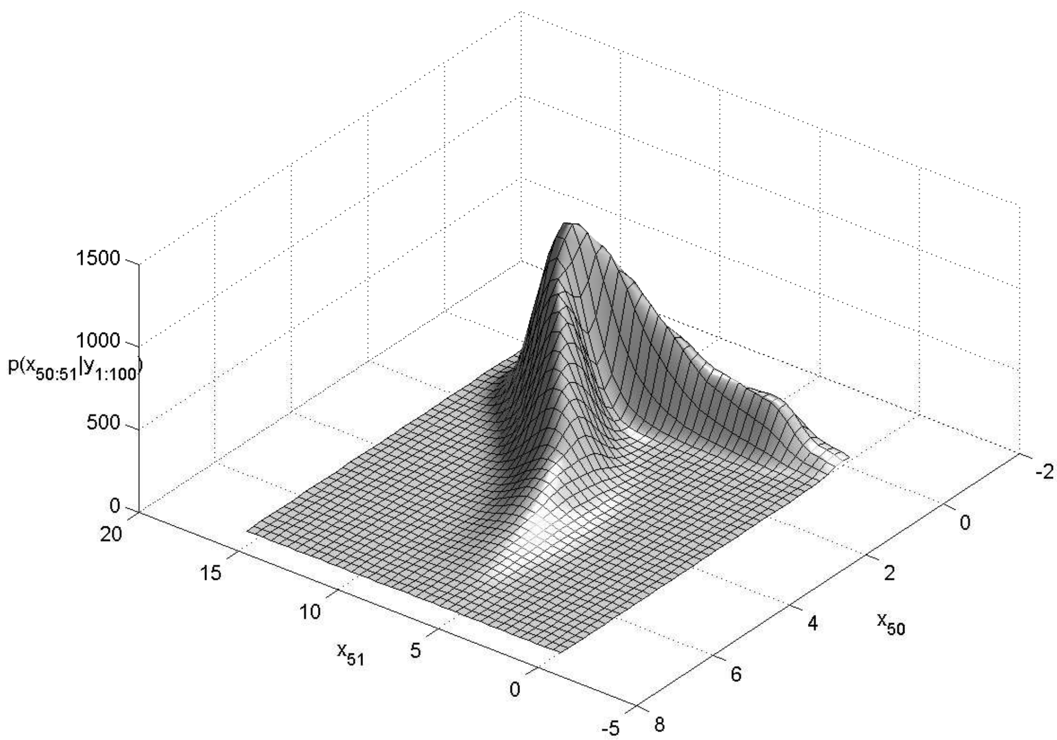

Figure 9: Kernel density estimate for $p(x_{50:51}|y_{1:100})$, using $\sigma_v^2 = 1$ and $\sigma_w^2 = 9$.

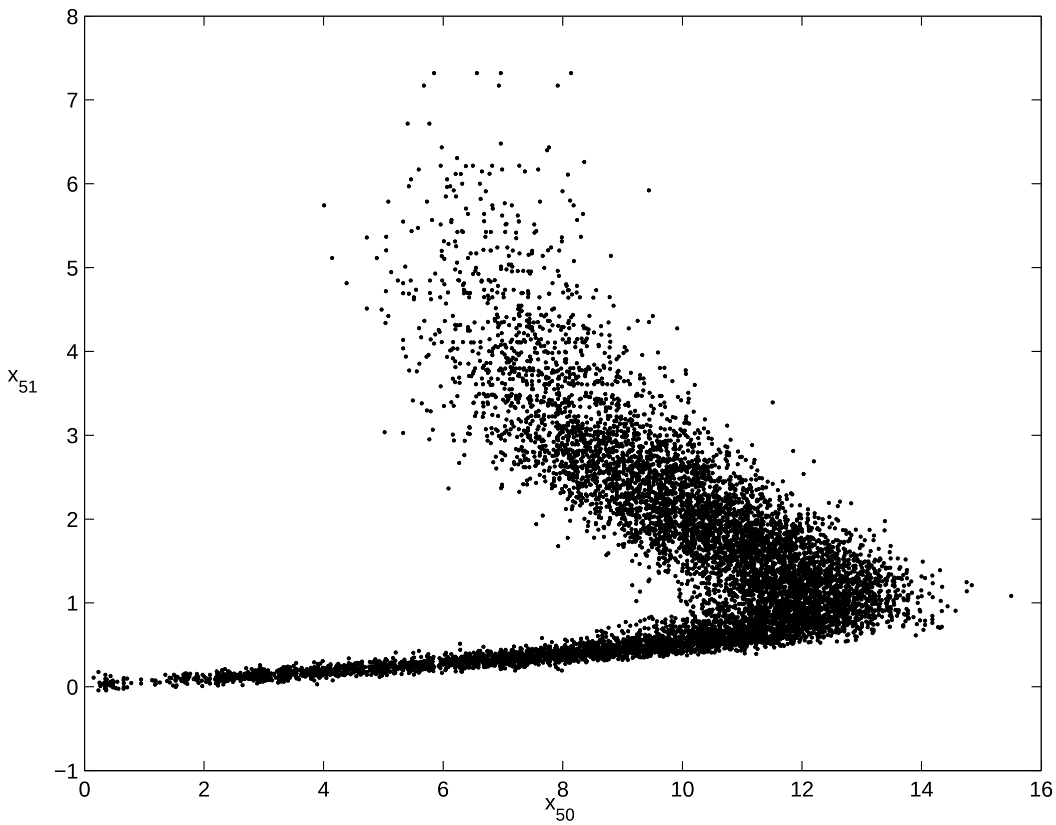

Figure 10: Scatter plot of points drawn from $p(x_{50:51}|y_{1:100})$.

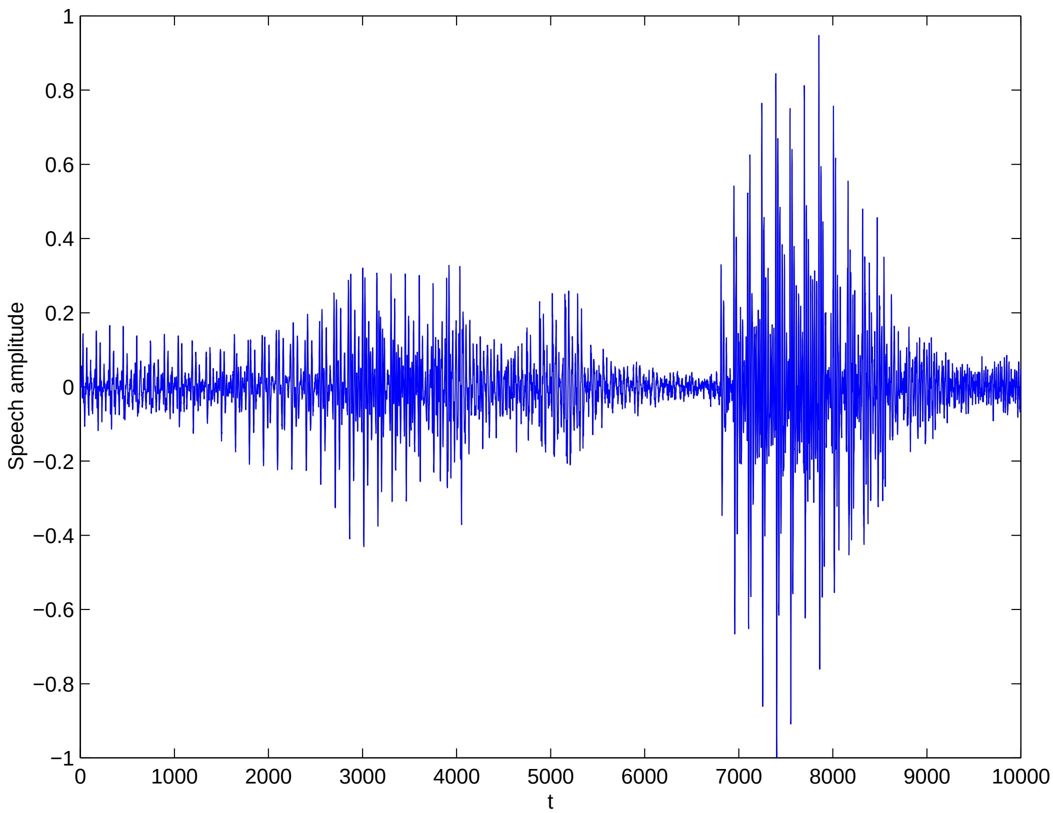

Figure 11: Speech data. 0.62s of a US male speaker saying the words '...rewarded by...'. Sample rate 16kHz, resolution 16-bit, from the TIMIT speech database

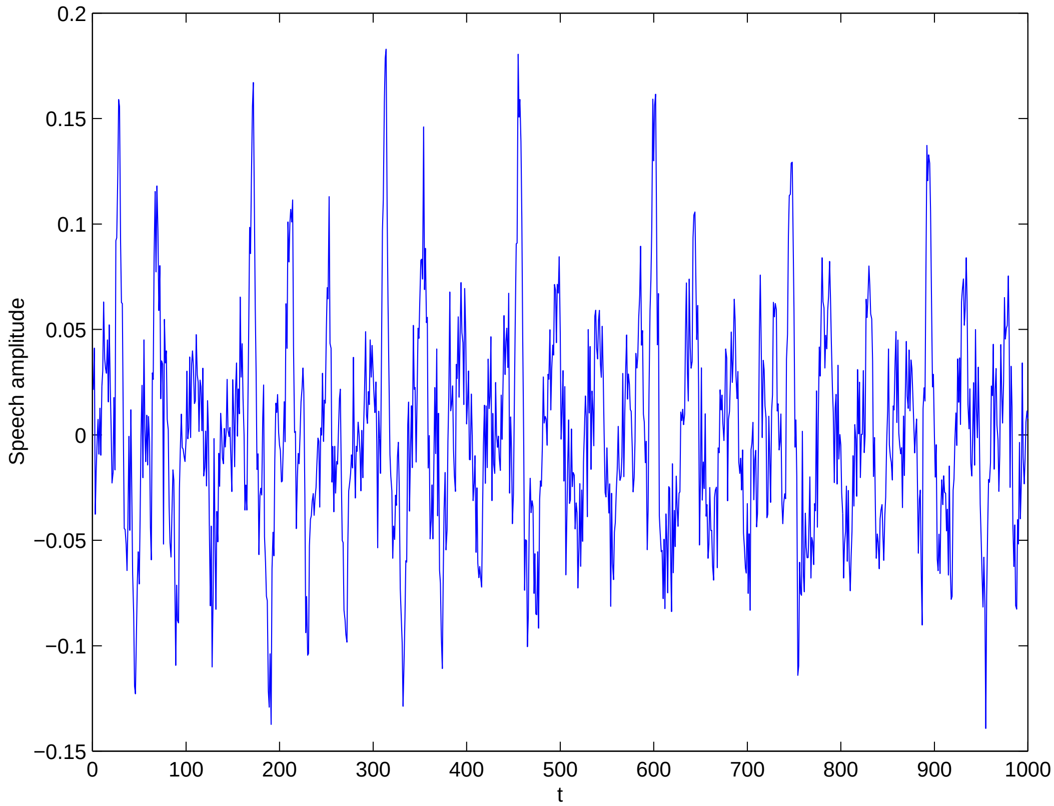

Figure 12: Noisy speech data - 1st 1000 data points

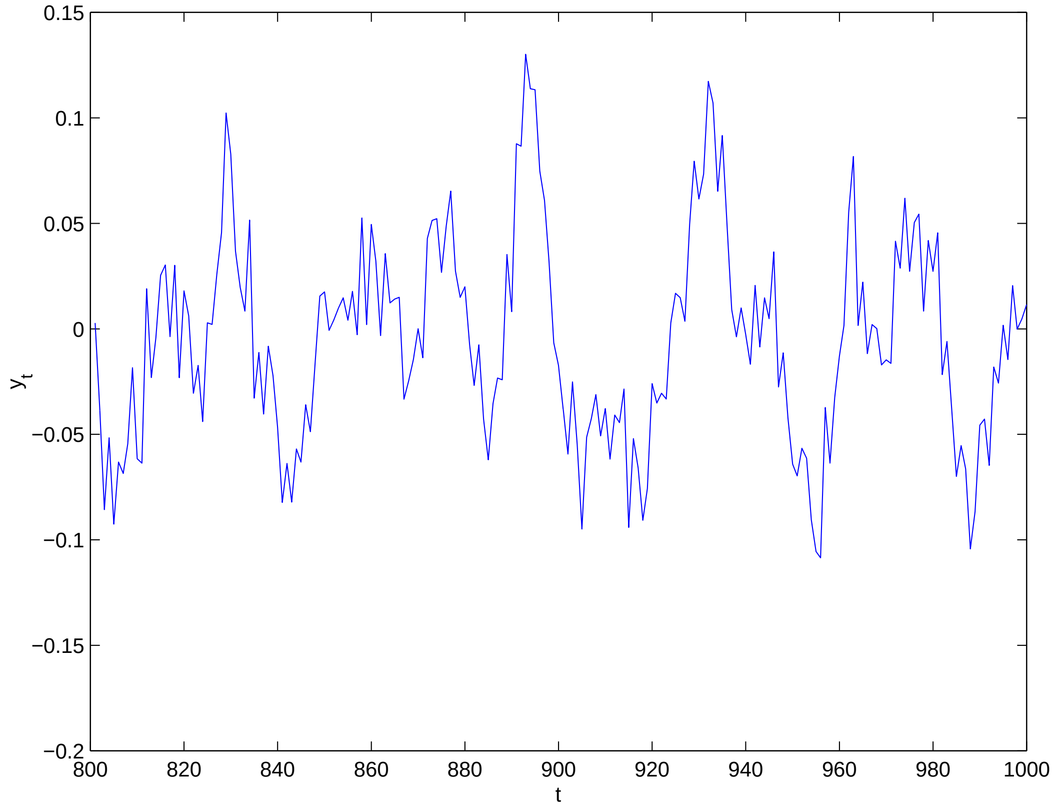

Figure 13: Noisy speech, t=801,...,1000

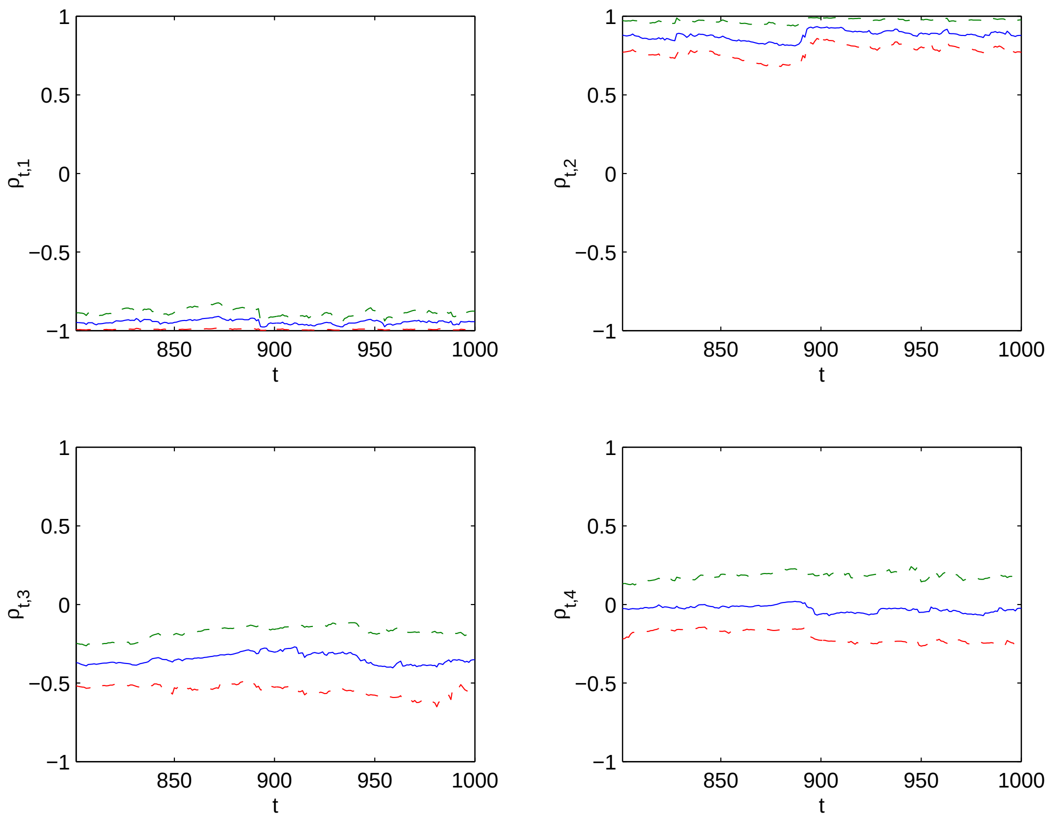

Figure 14: Posterior mean and 5/95 percentiles for filtered TV-PARCOR coefficient estimates

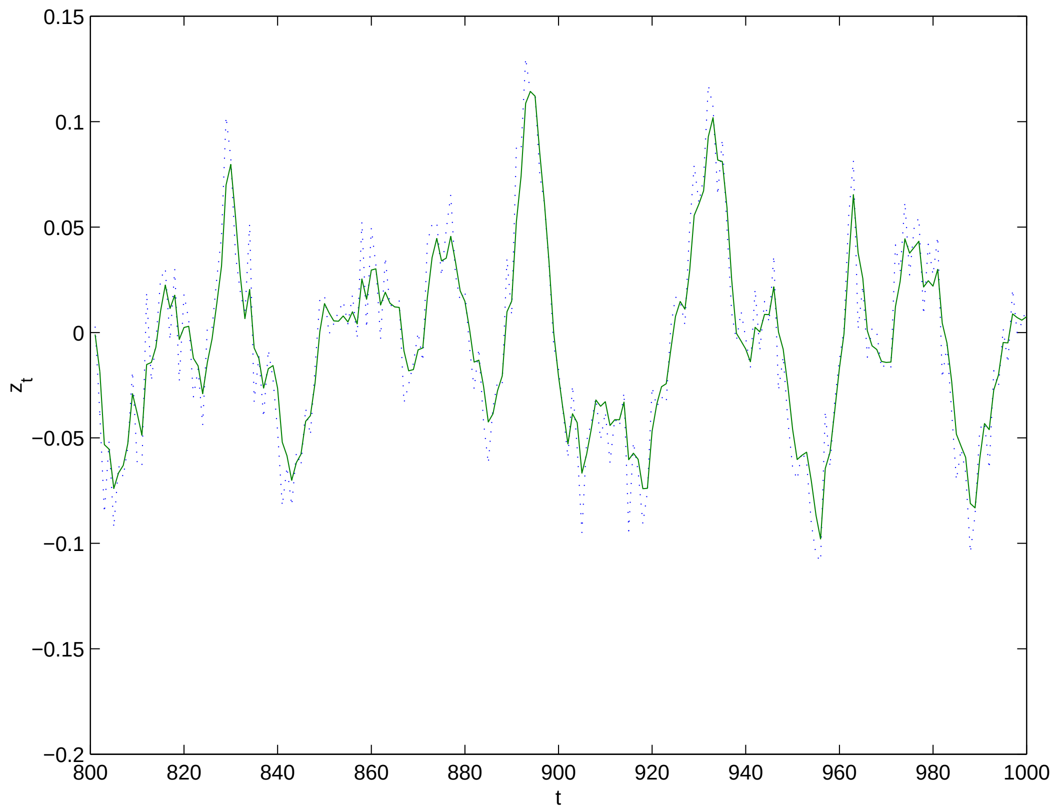

Figure 15: Filtered speech - estimated posterior mean (solid), noisy data (dotted)

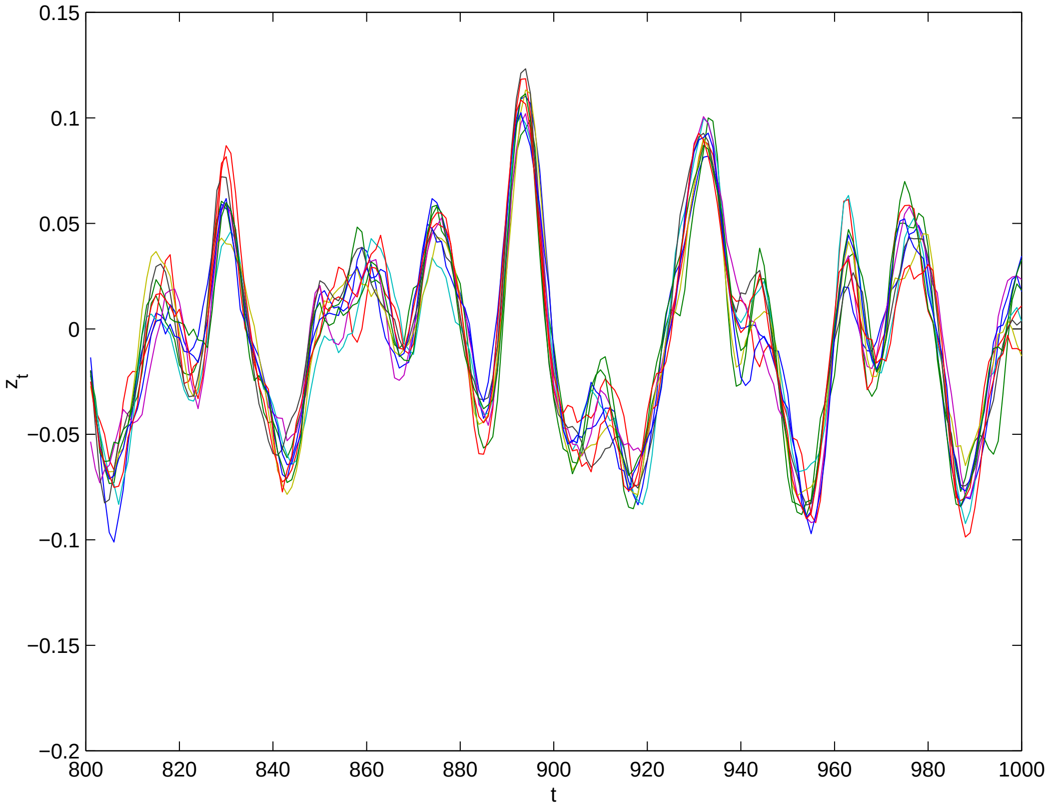

Figure 16: 10 realizations from the smoothed signal process

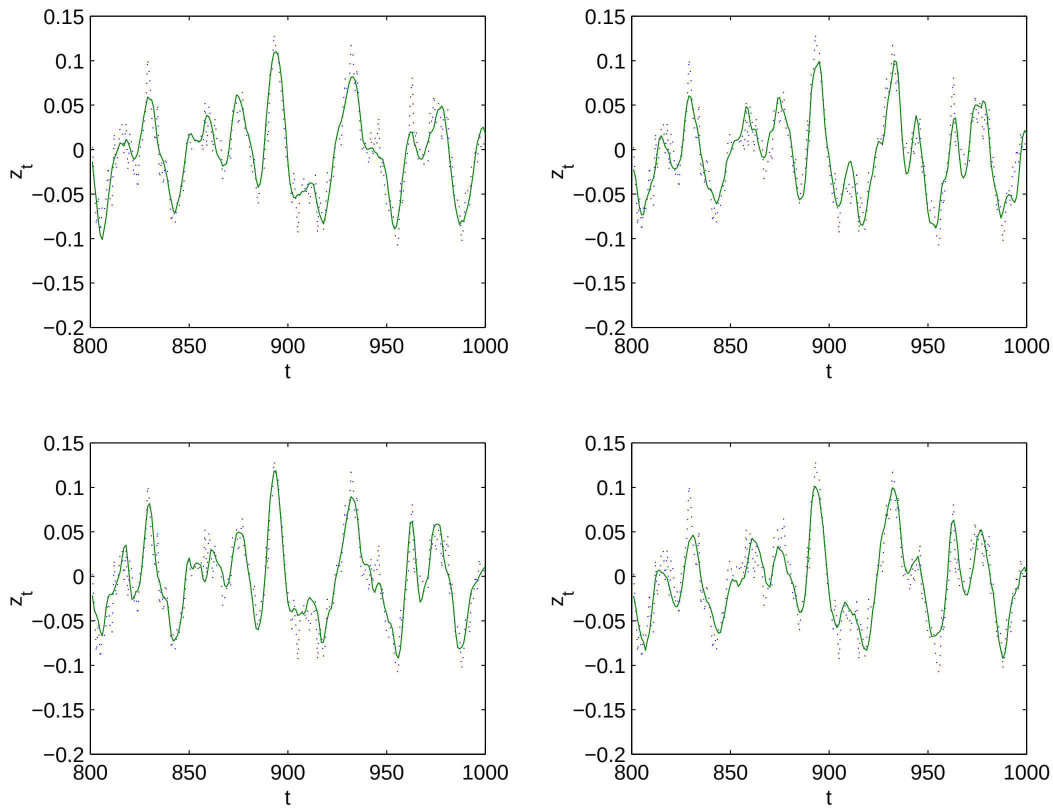

Figure 17: 4 realizations from the smoothing density for the signal process (noisy data shown dotted in each case)

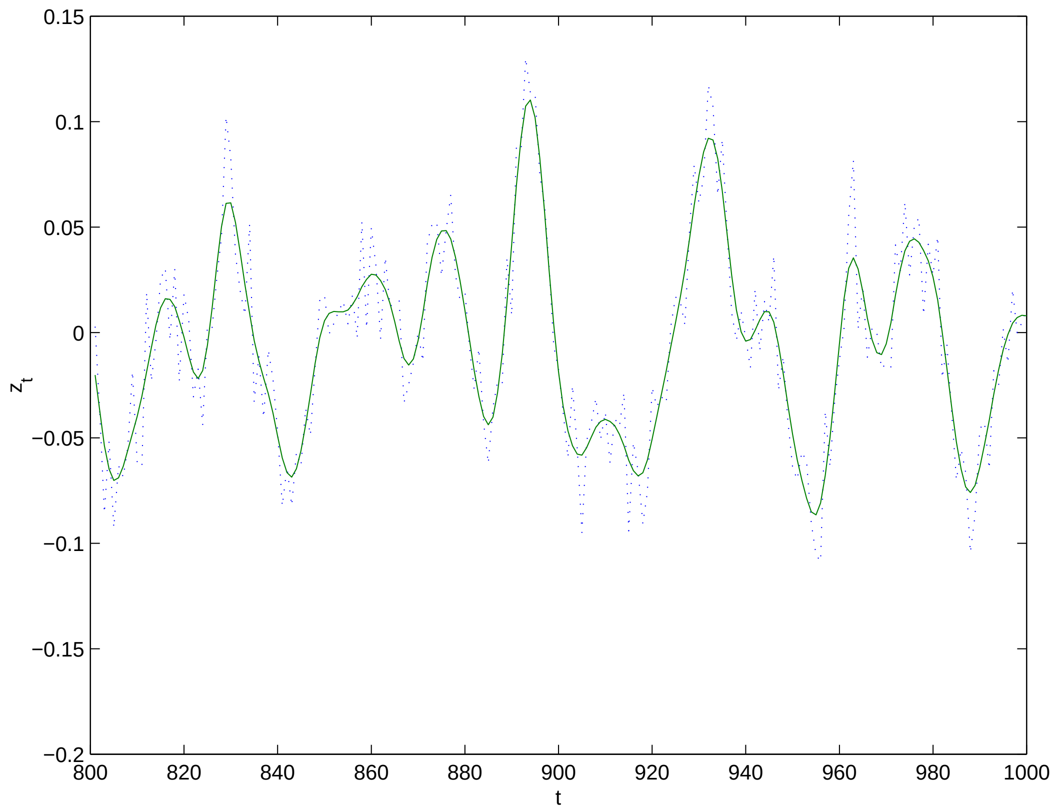

Figure 18: Average of 10 realizations from the smoothing density (noisy data shown dotted)

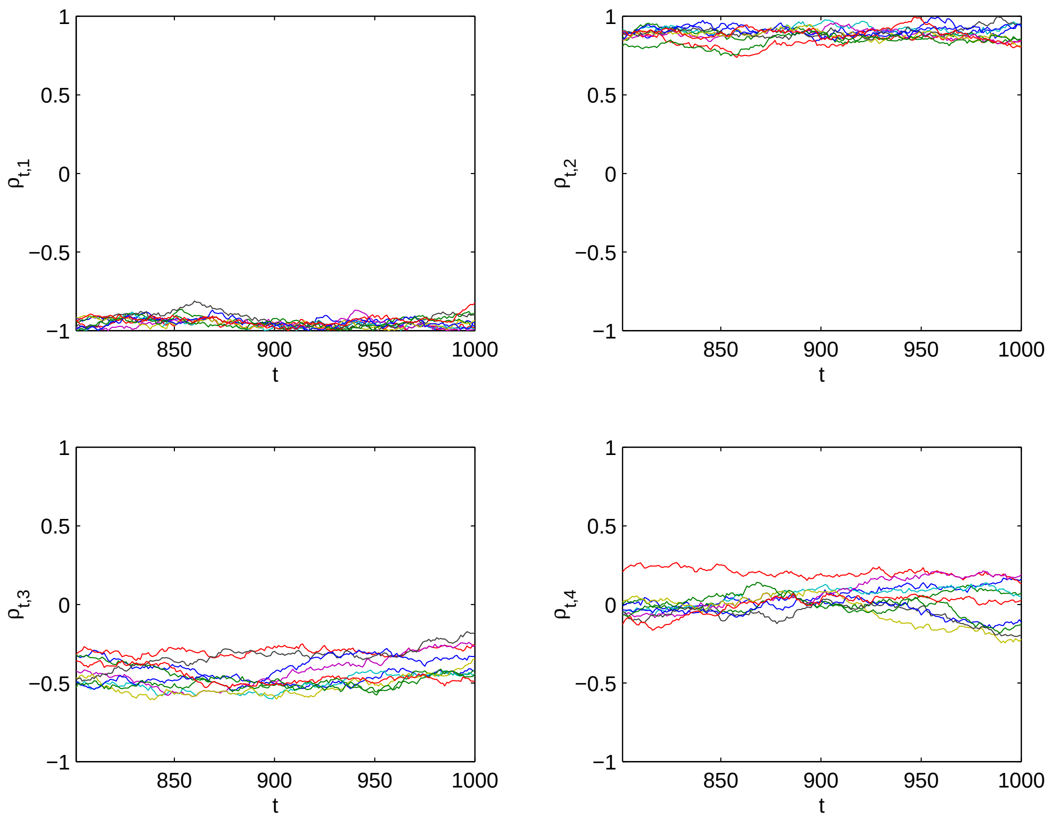

Figure 19: 10 realizations from the smoothing density for the TV-PARCOR coefficients.

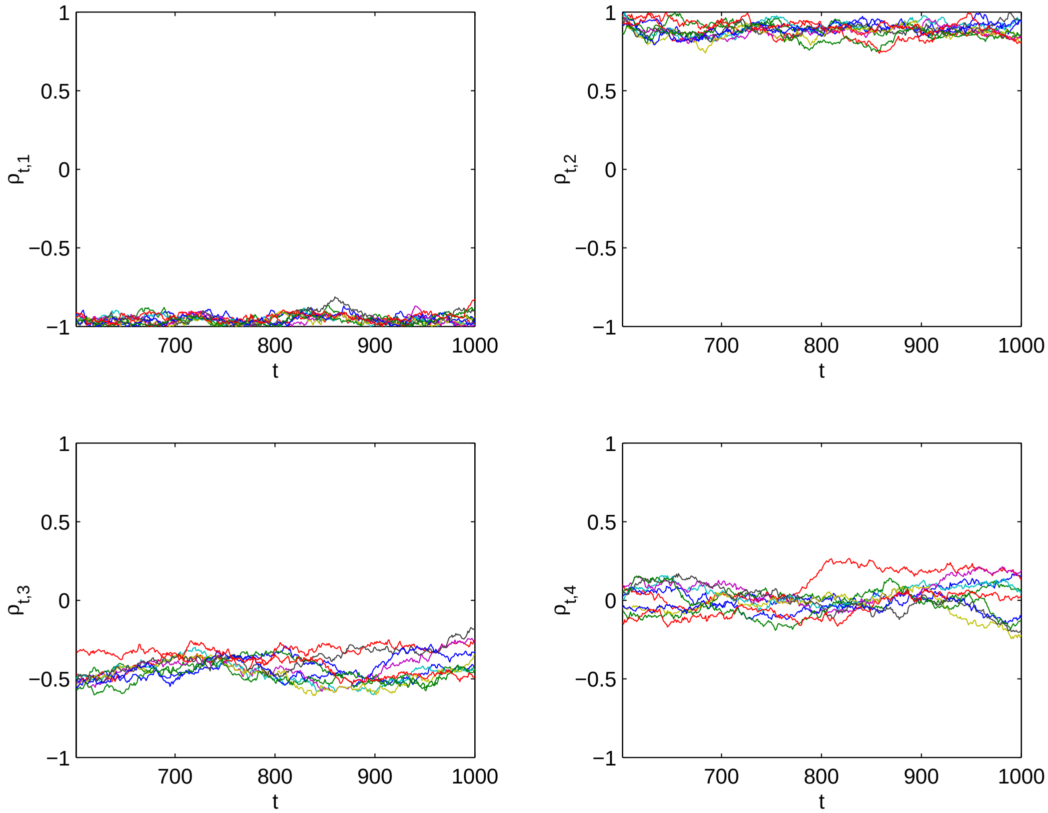

Figure 20: 10 realizations from the smoothing density for the TV-PARCOR coefficients - proposed simulation smoothing method ($N = 2000$)

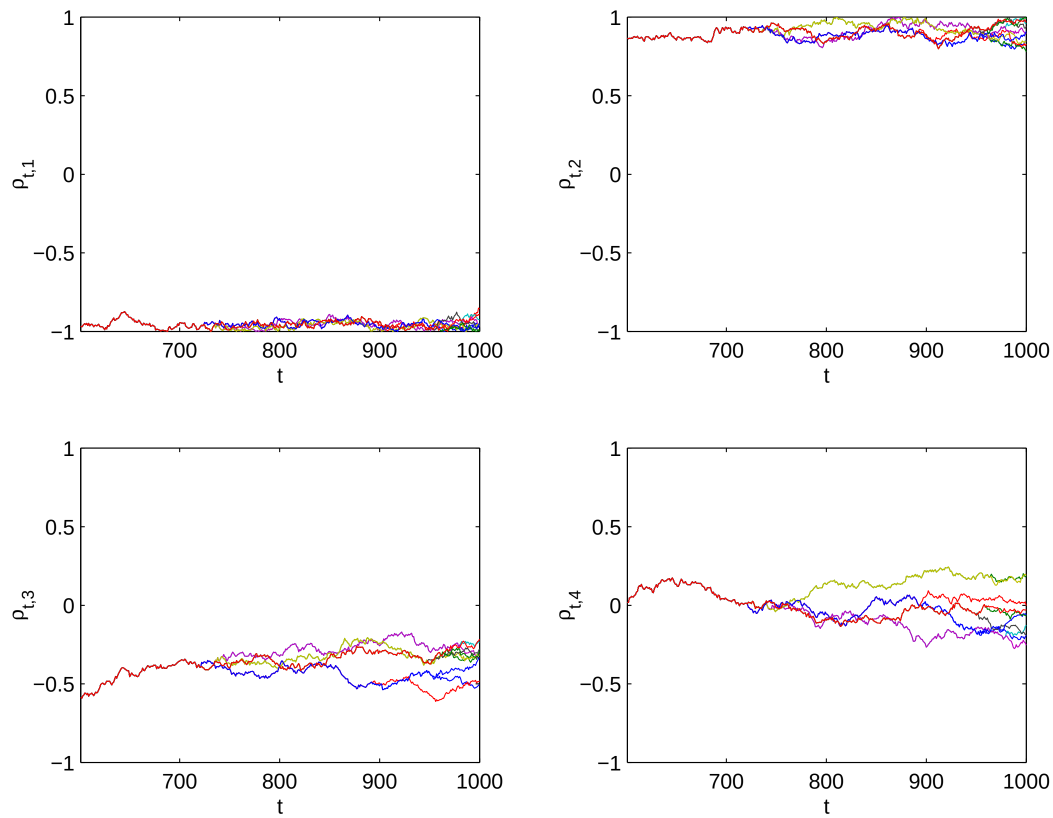

Figure 21: 10 realizations from the smoothing density for the TV-PARCOR coefficients - standard trajectory-based method ($N = 2000$)

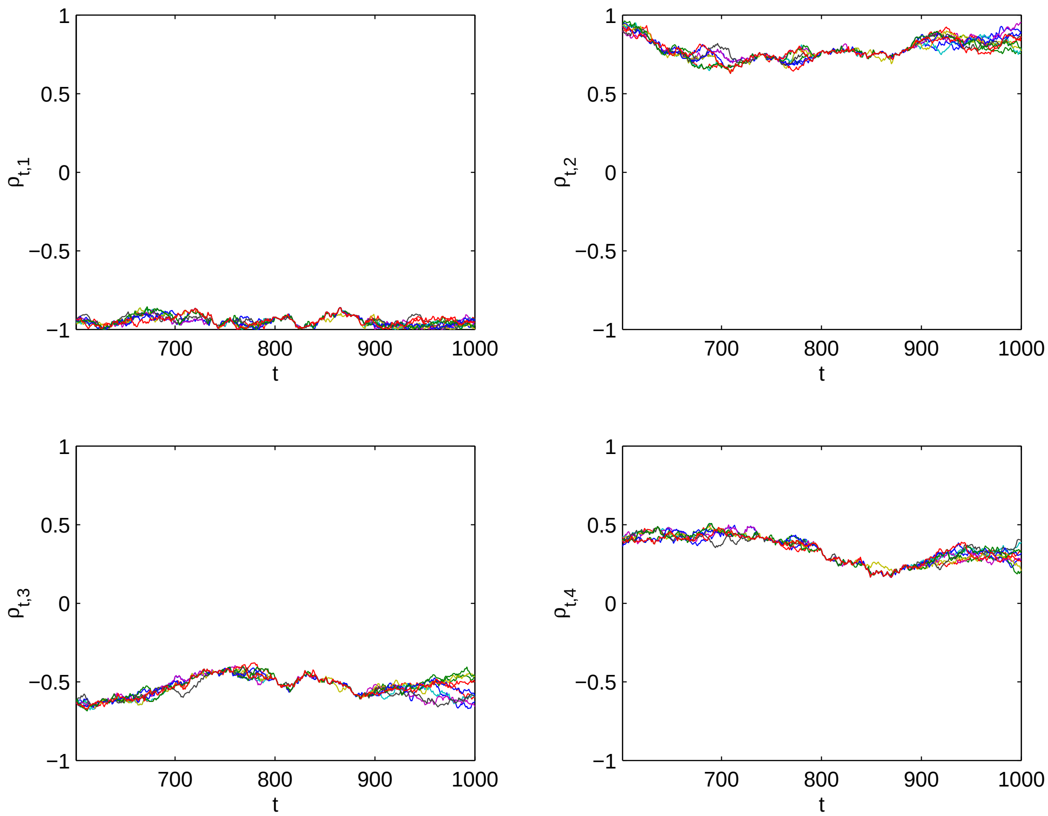

Figure 22: 10 realizations from the smoothing density for the TV-PARCOR coefficients - proposed simulation smoothing method ($N = 500$)

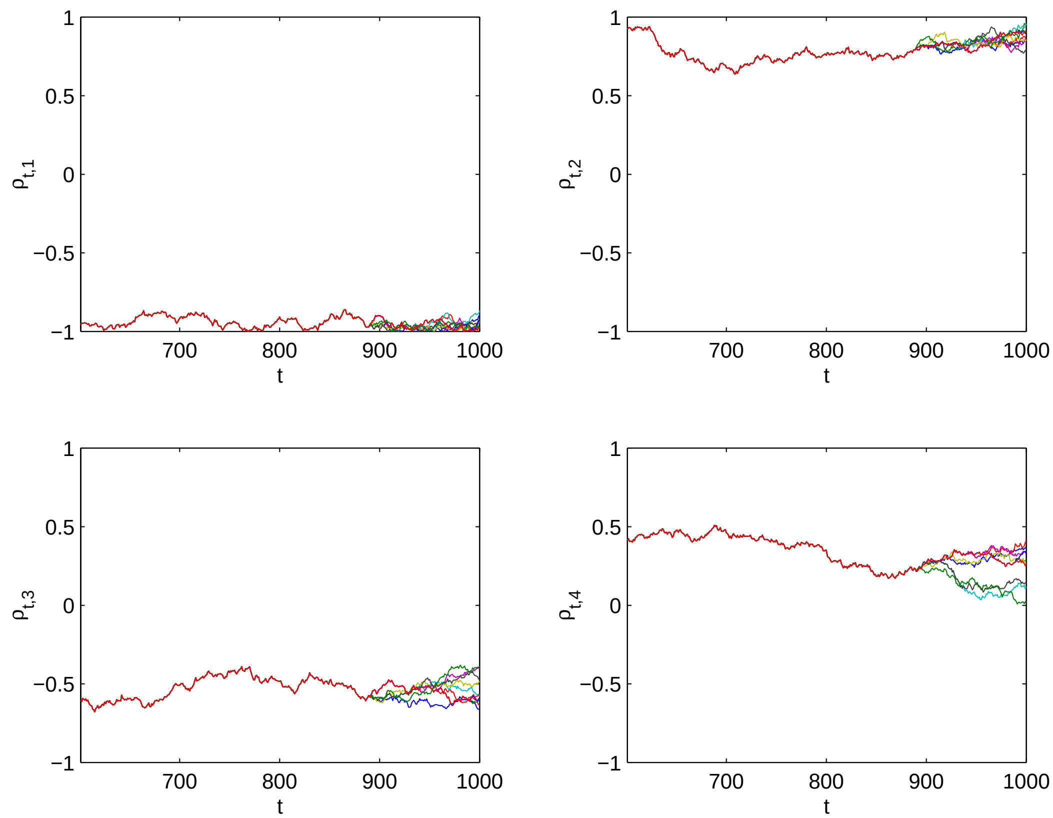

Figure 23: 10 realizations from the smoothing density for the TV-PARCOR coefficients - standard trajectory-based method ($N = 500$)