Recursive computation for HMMs and particle approximations

Olivier Cappé (ENST TSI / CNRS)

joint work with François LeGland (IRISA)

- Notations
- Filtering formulas
- Particle approximation
- Application to the estimation of the initial condition
- Conclusions

Notations for discrete-time HMM

Hidden state $(X_k)_{k \geq 0}$

Transition kernel Q **(also** $q\lambda$**, with** $p_0\lambda$ **for the distribution of the initial state, where** λ **is a dominating measure)**

Observations $(Y_k)_{k \geq 0}$ **and** $Y_{0:n} = (Y_k)_{0 \leq k \leq n}$

Observation pdf $g(y_k|x_k)$ **(or** $\psi_k(x_k)$ **when conditioning upon** $Y_{0:n}$**)**

Prediction distribution $\langle \mu_{n|n-1}, \phi \rangle = \mathrm{E}[\phi(X_n)|Y_{0:n-1}]$

Prediction distribution $\langle \mu_n, \phi \rangle = \mathrm{E}[\phi(X_n)|Y_{0:n}]$**, where**

$$\mu_{n+1|n} = \mu_n Q \qquad (prediction)$$

$$\mu_{n+1} = \frac{\psi_{n+1}}{\langle \mu_{n+1|n}, \psi_{n+1} \rangle} \mu_{n+1|n} \qquad (Bayes)$$

Starting point Most standard texts on Hidden Markov Models (eg. Rabiner's 1989 tutorial, McDonald & Zucchini's 1997 monograph) ignore a remarkable observation about HMMS:

- The intermediate quantity of the EM algorithm
- The gradient of the log-likelihood (score)
- The Hessian of the log-likelihood (observed information)

and more generally any function that may be written as

$$A_n = \mathrm{E}\left[\sum_{k=1}^{n} f_k(X_k, X_{k-1}, Y_k) \middle| Y_{0:n} \right] \tag{1}$$

can be computed recursively in t (ie. without resorting to "Forward-Backward" smoothing)

Zeitouni & Dembo (IT, 1989), Campillo & LeGland (SPA, 1989), Elliot, Aggoun & Moore (1994) + applications to Gaussian state space models, eg. Charalambous & Logothetis (CDC, 1998)

EM

$$Q_{(\mathrm{EM})}(\theta;\hat{\theta}) = \mathrm{E}^{\hat{\theta}}\Bigg[\log p_0^{\theta}(X_0) + \log g^{\theta}(Y_0|X_0) + \sum_{k=1}^{n}\left(\log g^{\theta}(Y_k|X_k) + \log q^{\theta}(X_{k-1},X_k)\right)\Bigg|Y_{0:n}\Bigg]$$

In exponential families, it suffices to compute $\mathrm{E}^{\hat{\theta}}\left[\sum_{k=1}^{n} f_k(X_k, X_{k-1}, Y_k)\middle|\, Y_{0:n}\right]$ (with functions that do not depend upon θ

Gradient of the log-likelihood (Fisher formula)

$$\nabla_{\theta}\log p^{\theta}(Y_{1:n}) = \mathrm{E}^{\theta}\Bigg[\nabla_{\theta}\log p_0^{\theta}(X_0) + \nabla_{\theta}\log g^{\theta}(Y_0|X_0) + \sum_{k=1}^{n}\left(\nabla_{\theta}\log g^{\theta}(Y_k|X_k) + \nabla_{\theta}\log q^{\theta}(X_{k-1},X_k)\right)\Bigg|Y_{0:n}\Bigg]$$

What's the trick ? (for a slightly simplified form of (1))

Define the signed measures $w_{n|n-1}$ and w_n such that

$$\langle w_{n|n-1}, \phi \rangle = \mathrm{E}\left[\phi(X_n) \sum_{k=1}^{n} f_k(X_k) \middle| Y_{0:n-1} \right]$$

$$\langle w_n, \phi \rangle = \mathrm{E}\left[\phi(X_n) \sum_{k=1}^{n} f_k(X_k) \middle| Y_{0:n} \right] \qquad \text{so that } A_n = \langle w_{n|n-1}, 1 \rangle$$

Prediction

$$\begin{aligned}
\langle w_{n+1|n}, \phi \rangle &= \mathrm{E}\big[\phi(X_{n+1}) \sum_{k=0}^{n+1} f_k(X_k) \big| Y_{0:n}\big] \\
&= \mathrm{E}\big[\phi(X_{n+1}) f_{n+1}(X_{n+1}) \big| Y_{0:n}\big] \\
&\quad + \mathrm{E}\big[\underbrace{\mathrm{E}\big[\phi(X_{n+1}) \sum_{k=0}^{n} f_k(X_k) \big| X_{0:n}, Y_{0:n}\big]}_{(Q\phi)(X_n) \sum_{k=0}^{n} f_k(X_k)} \big| Y_{0:n}\big] \\
&= \langle \mu_{n+1|n}, f_{n+1}\phi \rangle + \langle w_n, Q\phi \rangle
\end{aligned}$$

What's the trick ? (cont.)

Bayes

$$\mathrm{P}(dx_{0:n+1} | y_{0:n+1}) = \frac{g(y_{n+1} | x_{n+1})}{\int g(y_{n+1} | x_{n+1}) \, \mathrm{P}(dx_{n+1} | y_{0:n+1})} \, \mathrm{P}(dx_{0:n+1} | y_{0:n})$$

So that

$$w_{n+1} = \frac{\psi_{n+1}}{\langle \mu_{n+1|n}, \psi_{n+1} \rangle} w_{n+1|n}$$

(same relation as for the predictor to filter update)

In summary,

$$w_{n+1|n} = w_n Q + f_{n+1} \mu_{n+1|n} \qquad (\textit{prediction})$$

$$w_{n+1} = \frac{\psi_{n+1}}{\langle \mu_{n+1|n}, \psi_{n+1} \rangle} w_{n+1|n} \qquad (\textit{Bayes})$$

$$= \frac{\psi_{n+1}}{\langle \mu_{n+1|n}, \psi_{n+1} \rangle} w_n Q + f_{n+1} \mu_{n+1}$$

to be compared with

$$\mu_{n+1|n} = \mu_n Q$$

$$\mu_{n+1} = \frac{\psi_{n+1}}{\langle \mu_{n+1|n}, \psi_{n+1} \rangle} \mu_{n+1|n}$$

(recall that $A_n = \langle w_m, 1 \rangle$)

Comments

These recursions can be implemented

For finite state spaces,

$$w_n(x_n) = \sum_{k=1}^{n} \sum_{x_k} f_k(x_k)\, \mathrm{P}(X_k = x_k, X_n = x_n | Y_{0:n})$$

Warning: Computing w_n is $O(\#X^2 \times n)$ but there are many such statistics of interest: $\mathbb{I}_{\{i\}}(x_s)$ ($\#X - 1$ of them), $\mathbb{I}_{\{i\}}(x_{s-1})\mathbb{I}_{\{j\}}(x_s)$ ($\#X \times (\#X - 1)$ of these)...

In the Gaussian case,
with quadratic f_k

Since $w_n, w_{n|n-1} \ll \mu_n, \mu_{n|n-1}$ it is natural to approximate $w_{n|n-1}$ by $1/p \sum_{i=1}^p \rho_n^i \delta_{\xi^i_{n|n-1}}$ when $\mu_{n|n-1}$ is approximated as $1/p \sum_{i=1}^p \delta_{\xi^i_{n|n-1}}$

Proposed algorithm:

Prediction

1. $\tau^1_{n+1|n}, \ldots, \tau^p_{n+1|n} \sim \mathrm{Mult}(w_n^1, \ldots, w_n^p)$
2. $\xi^1_{n+1|n}, \ldots, \xi^p_{n+1|n}$ indep. $\sim Q(\xi_{n|n-1}^{\tau^1_{n+1|n}}, \cdot), \ldots, Q(\xi_{n|n-1}^{\tau^p_{n+1|n}}, \cdot)$
3. $\rho^i_{n+1} = \rho_n^{\tau^i_{n+1|n}} + f_{n+1}(\xi^i_{n+1|n})$ for $i = 1, \ldots, p$

Correction

$$\omega^i_{n+1} = \frac{\psi_{n+1}(\xi^i_{n+1|n})}{\sum_{j=1}^p \psi_{n+1}(\xi^j_{n+1|n})}$$

for $i = 1, \ldots, p$

Conditionally on $\mathcal{F}_n = \sigma\{(Y_k)_{k\geq 0}, (\xi^{1:p}_{k|k-1})_{0\leq k\leq n}, (\omega^{1:p}_k)_{0\leq k\leq n}\}$, $(\xi^i_{n+1|n}, \rho^i_{n+1})$ for $i = 1, \ldots, p$ are iid and satisfy

$$\mathrm{E}[\rho^i_{n+1}\phi(\xi^i_{n+1|n})|\mathcal{F}_n] = \sum_{i=1}^{p} \left[\rho^i_n \omega^i_n (Q\phi)(\xi^i_{n|n-1}) + \omega^i_n (Q f_{n+1}\phi)(\xi^i_{n|n-1})\right]$$

$$= \langle Q^*(\sum_{i=1}^{p} \rho^i_n \omega^i_n \delta_{\xi^i_{n|n-1}}), \phi\rangle$$

$$+ \langle f_{n+1} Q^*(\sum_{i=1}^{p} \omega^i_n \delta_{\xi^i_{n|n-1}}), \phi\rangle$$

Different types of applications

1. Cumulative sums (EM, gradient of log-likelihood...)
2. Fixed point smoothing ($f_k = 0$ for $k \geq l$)
3. Not in the form shown in (1) (increment, tangent filter...)

Type 1: Gradient of the log-likelihood

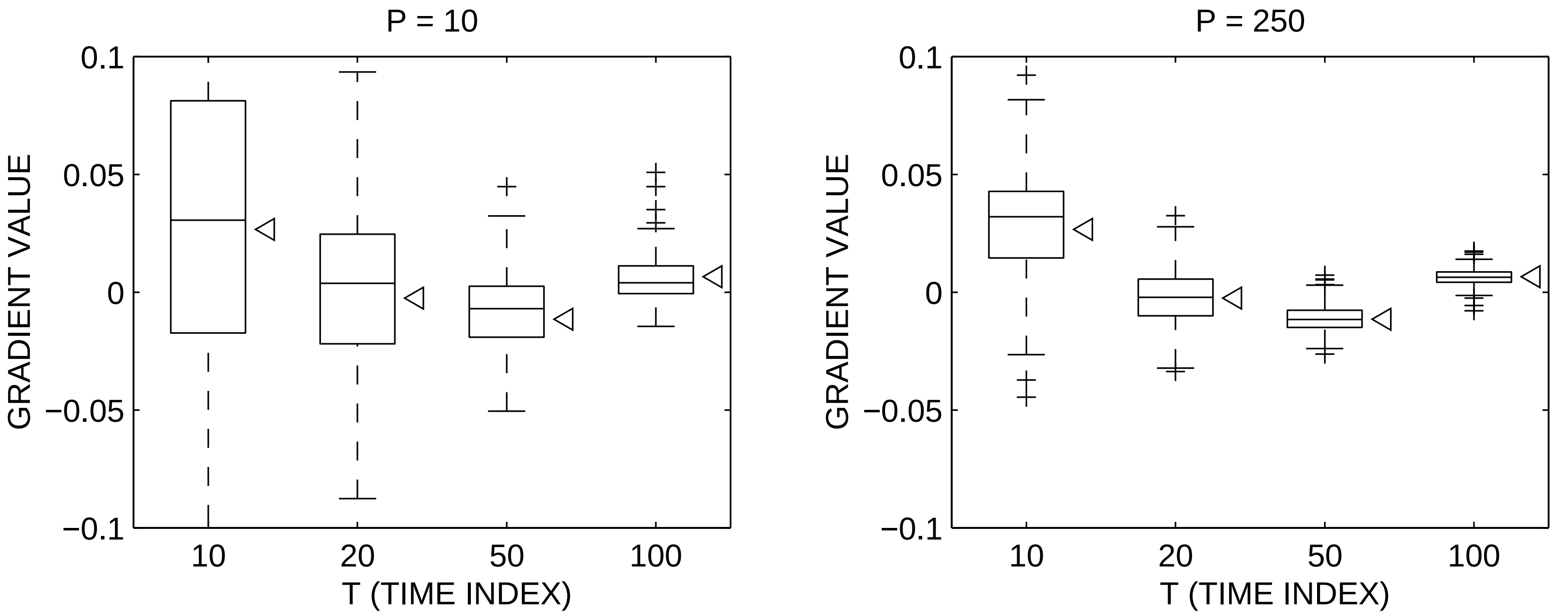


Figure 1: Box and whiskers plots summarizing 200 independent runs of the proposed algorithm compared with exact computations (triangles): $1/n\nabla_\theta \log p^\theta(Y_{1:n})$ for different combination of p and T (ie. n).

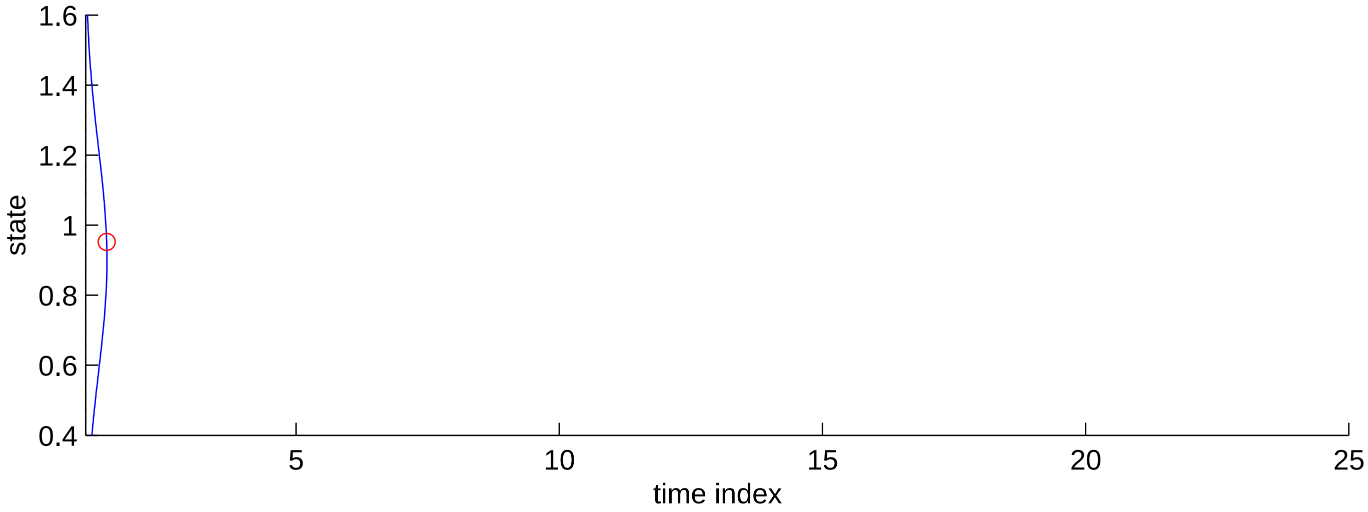
Model: AR(1): $\mu = 0.9, \gamma = 0.95, \sigma^2 = 0.01$
+ noise: $\eta^2 = 0.02 = (\sigma^2/(1-\gamma^2))/5$

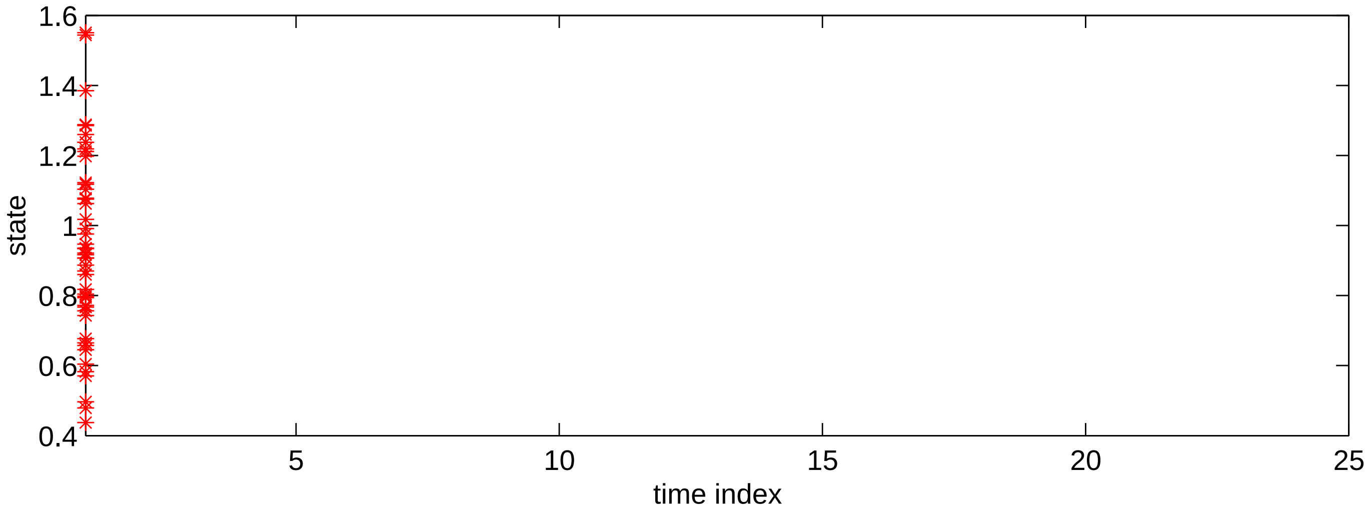
Type 2: Histogram of smoothed initial distribution

If $f_1(x_1, x_0, y_1) = \mathbb{I}_{(a,b]}(x_0)$**,**

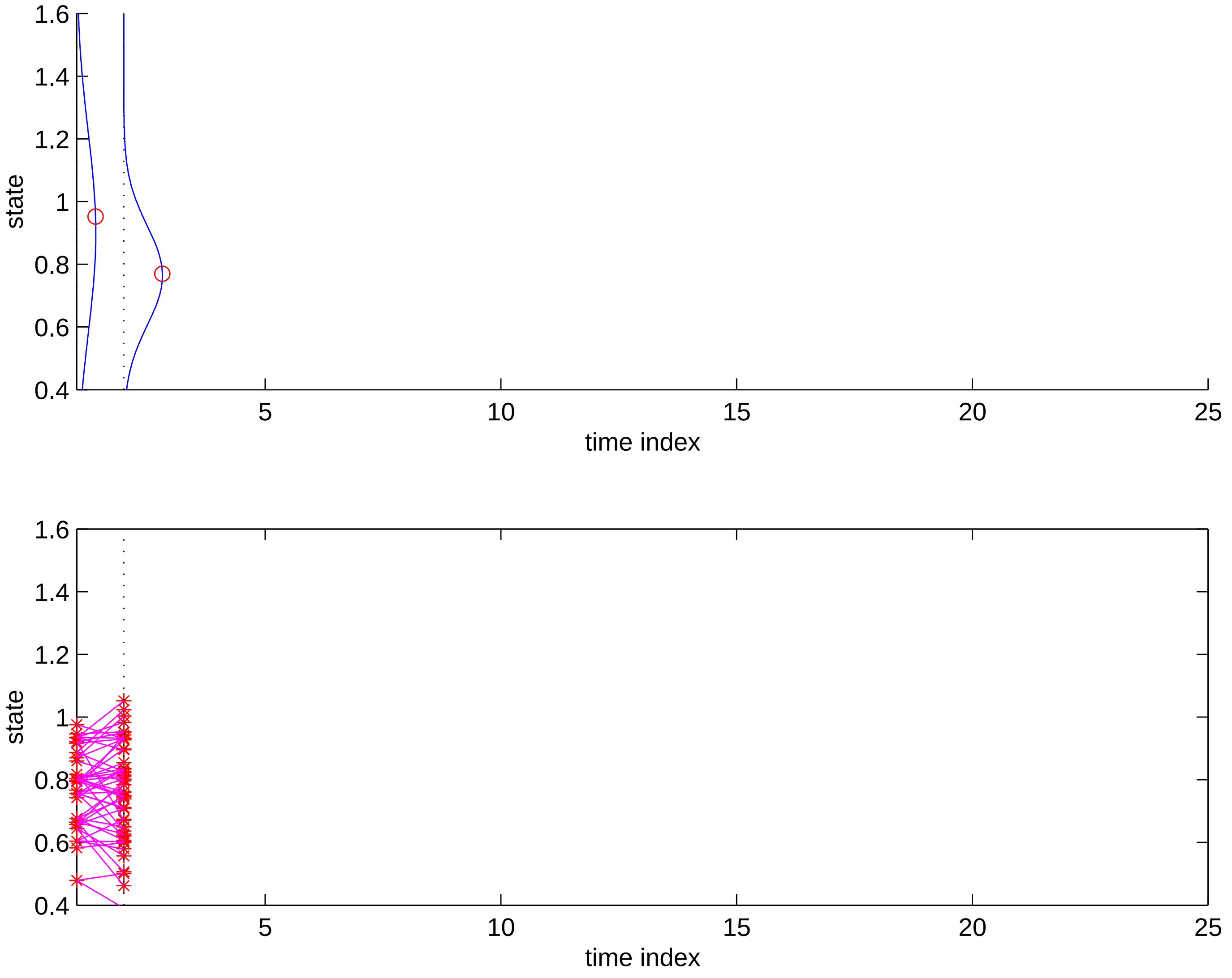
ρ_n^i **equals 1 if "ancestor" of the** i**th particle** $\xi_{n|n-1}^i$ **is in** $(a, b]$**, 0 otherwise**

$\hat{A}_n$ **Number of particle whose ancestor is in** $(a, b]$

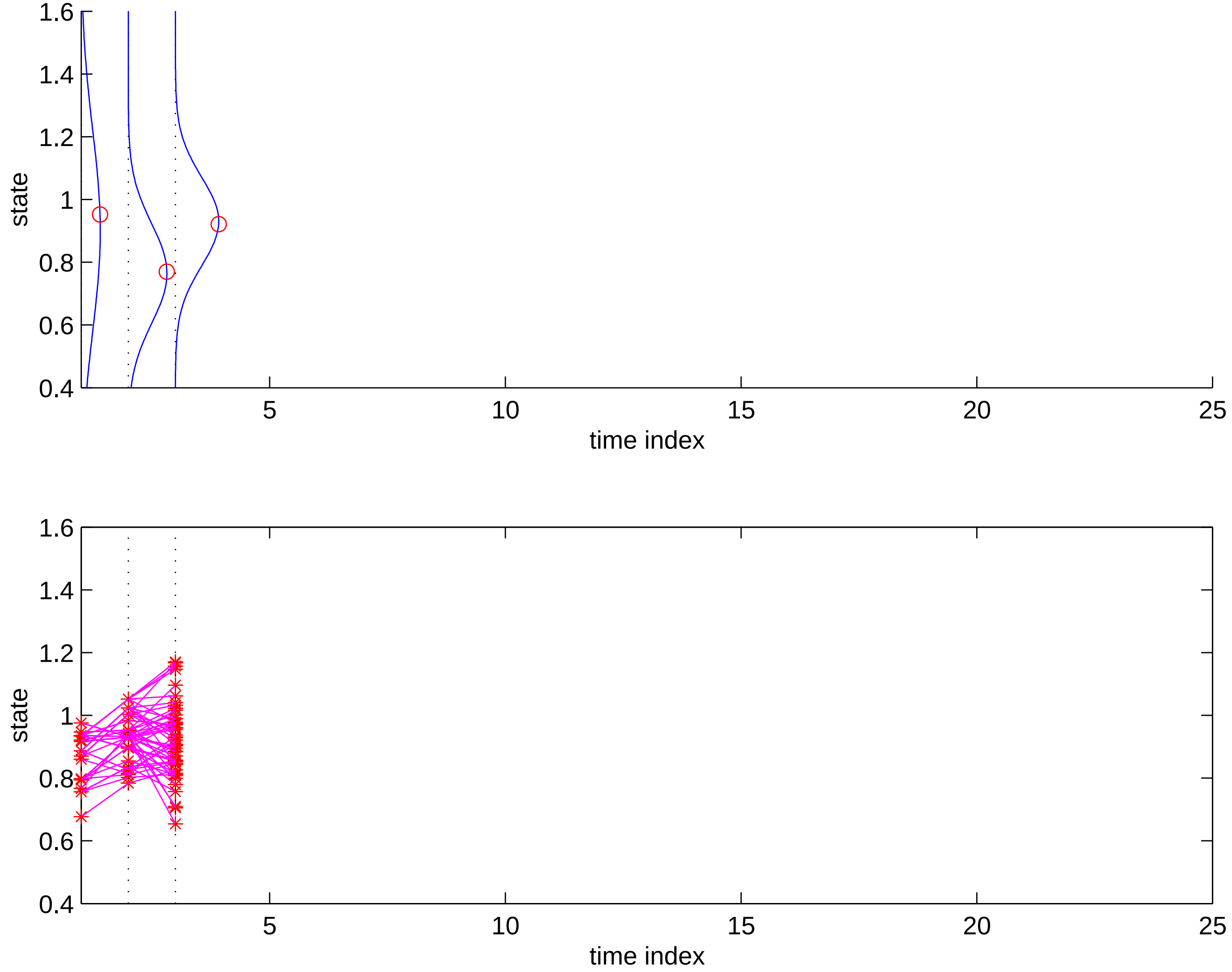




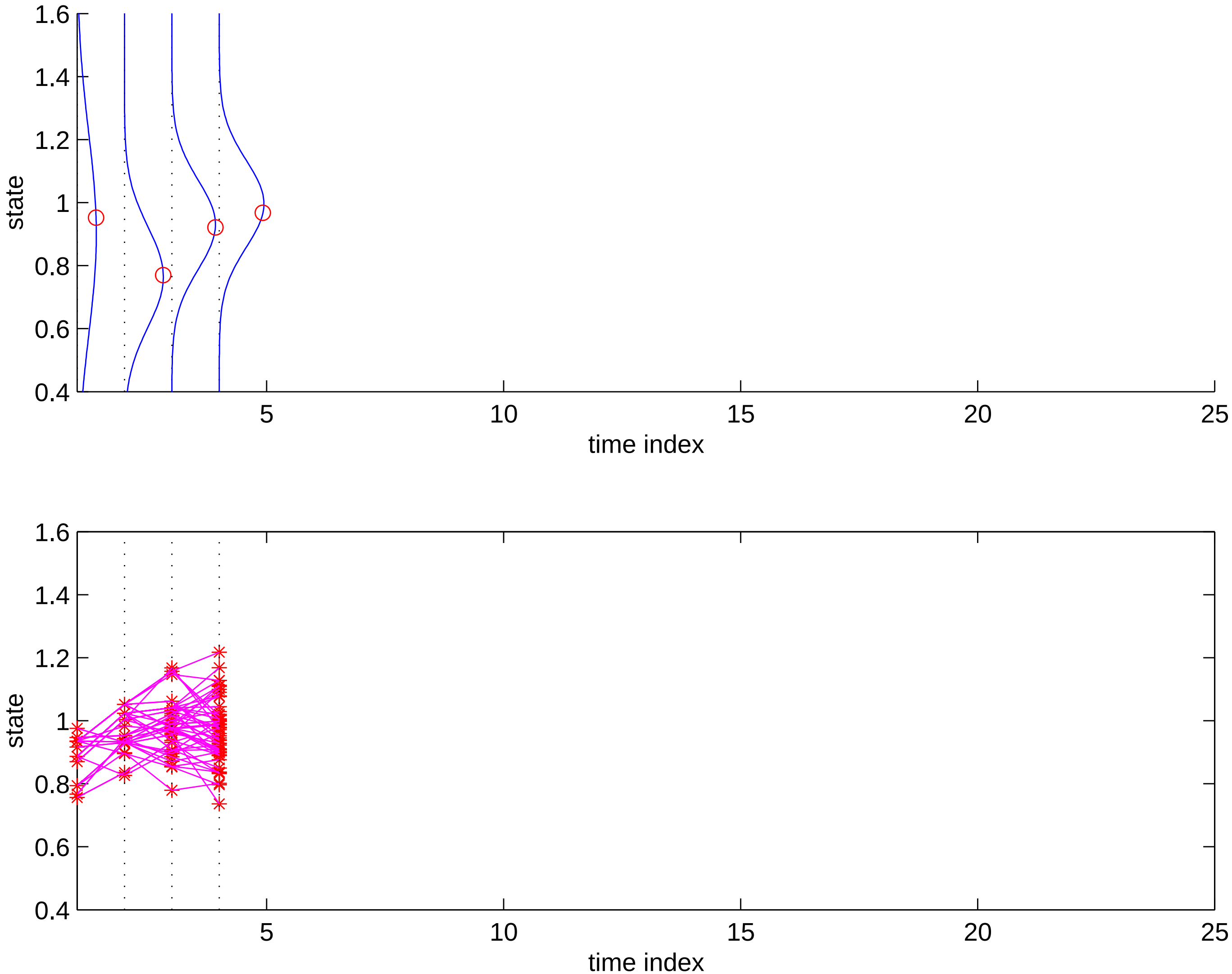
Predictive densities and evolution of the ancestor tree



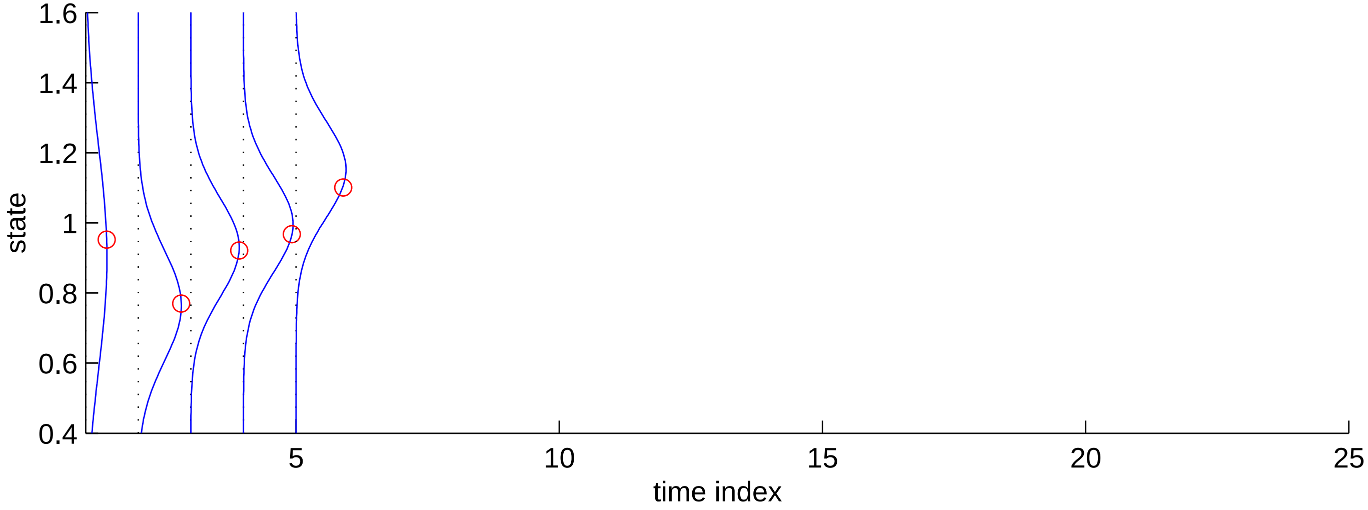
Predictive densities and evolution of the ancestor tree

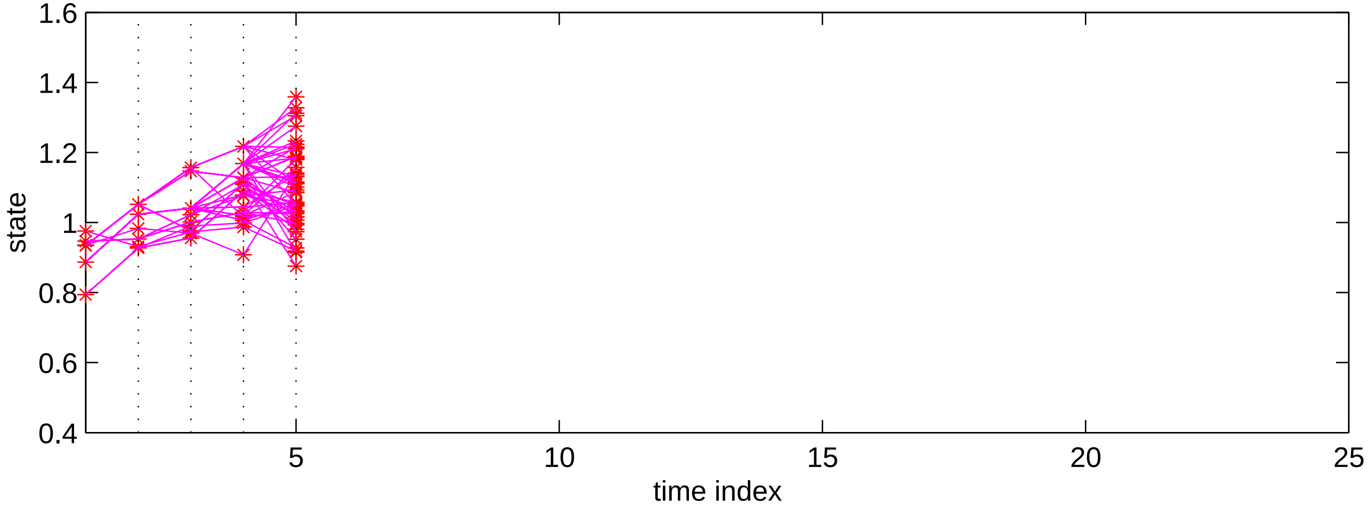


Predictive densities and evolution of the ancestor tree

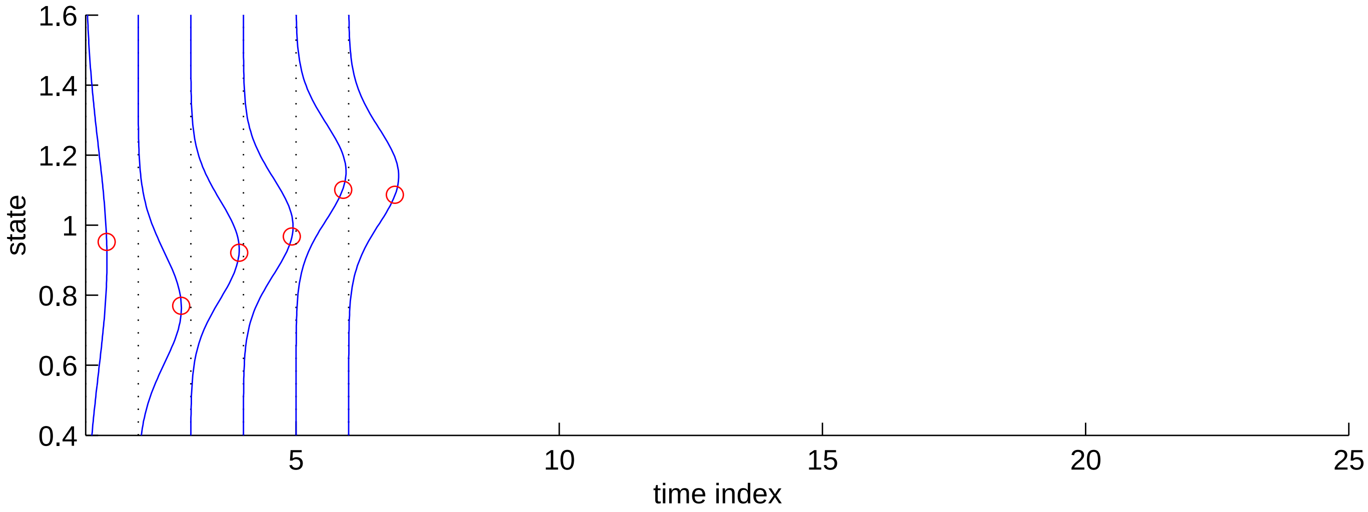


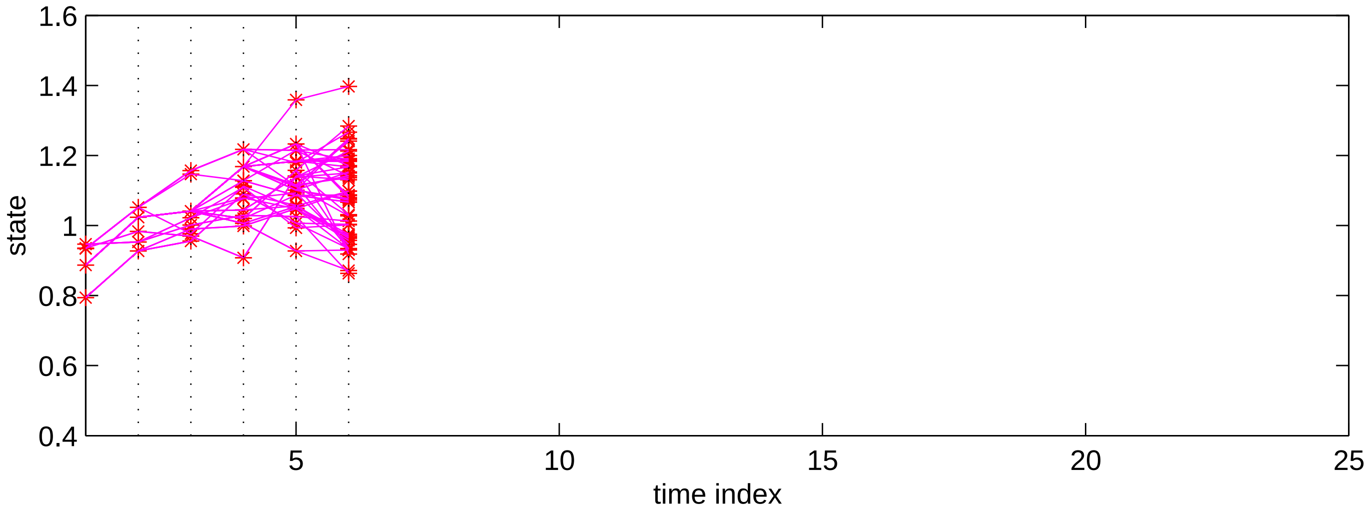
Predictive densities and evolution of the ancestor tree



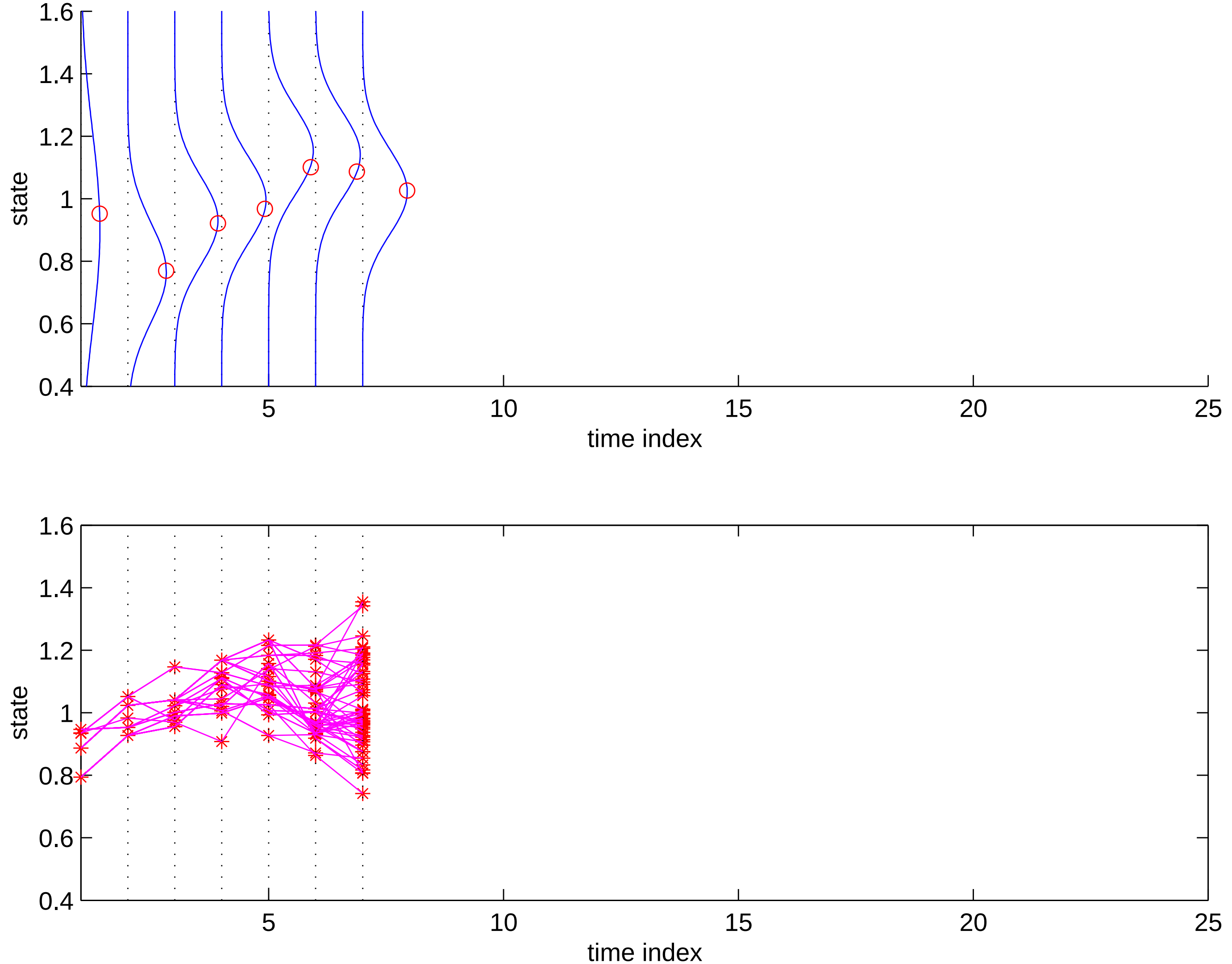


Predictive densities and evolution of the ancestor tree

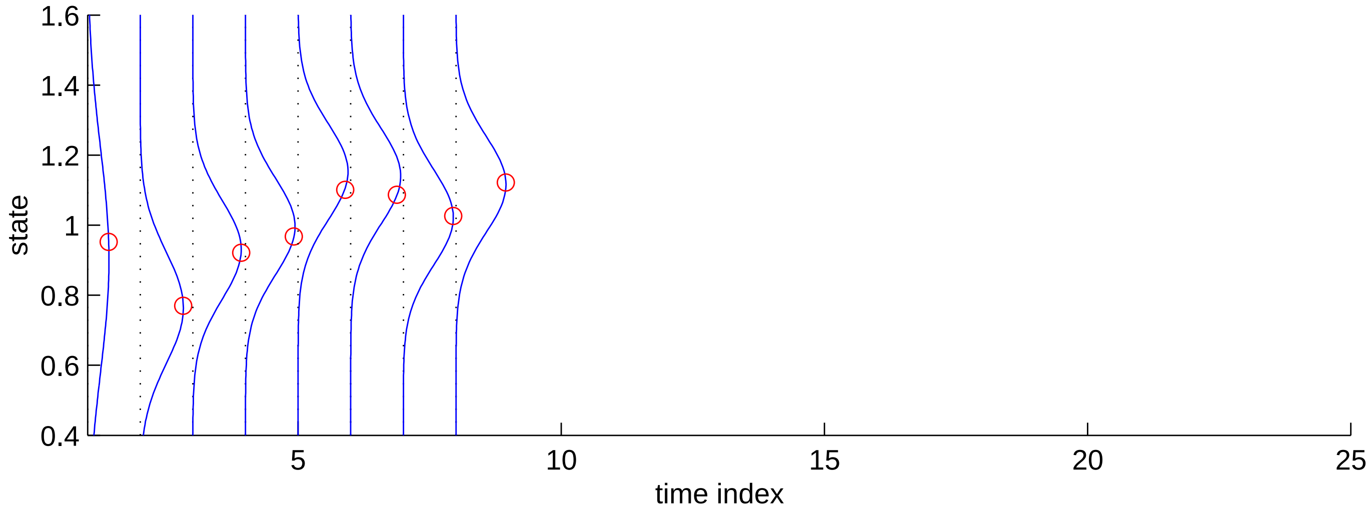


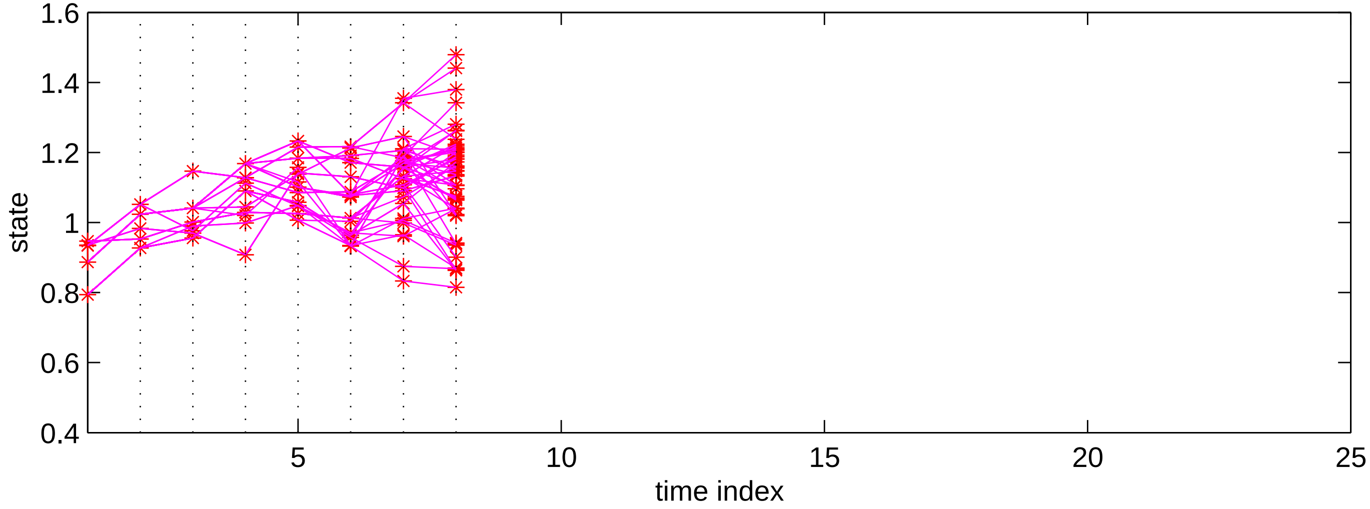


Predictive densities and evolution of the ancestor tree

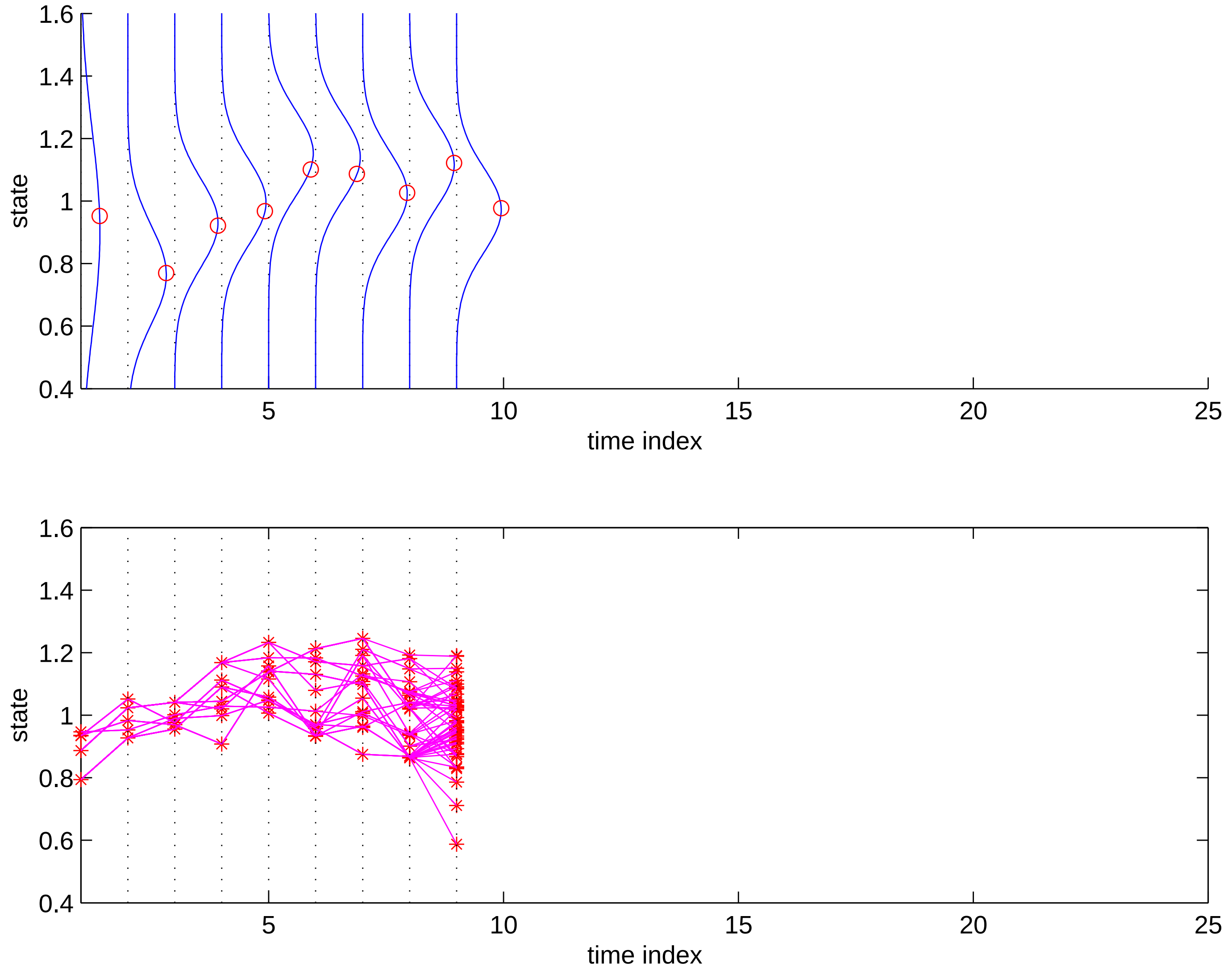


Predictive densities and evolution of the ancestor tree

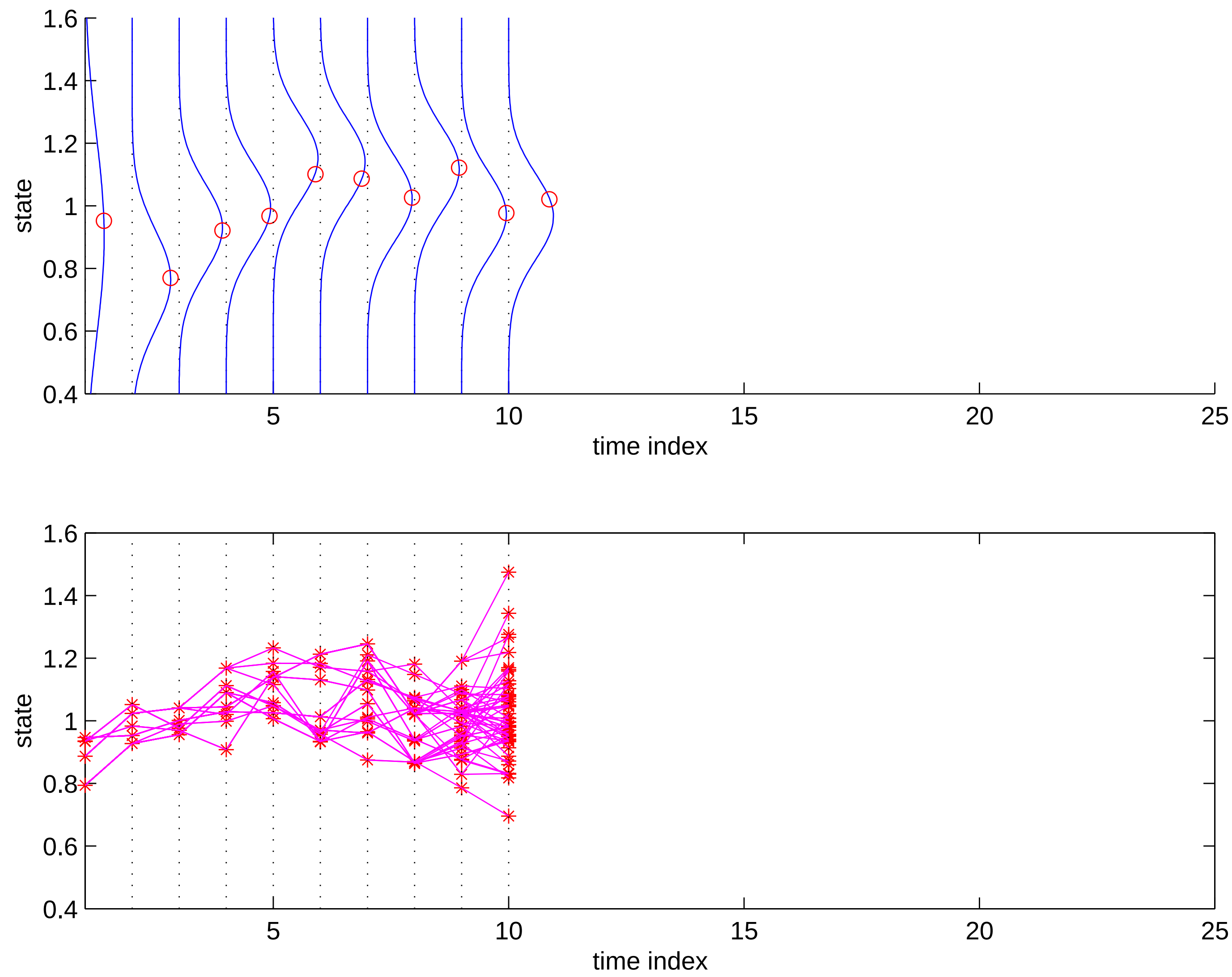




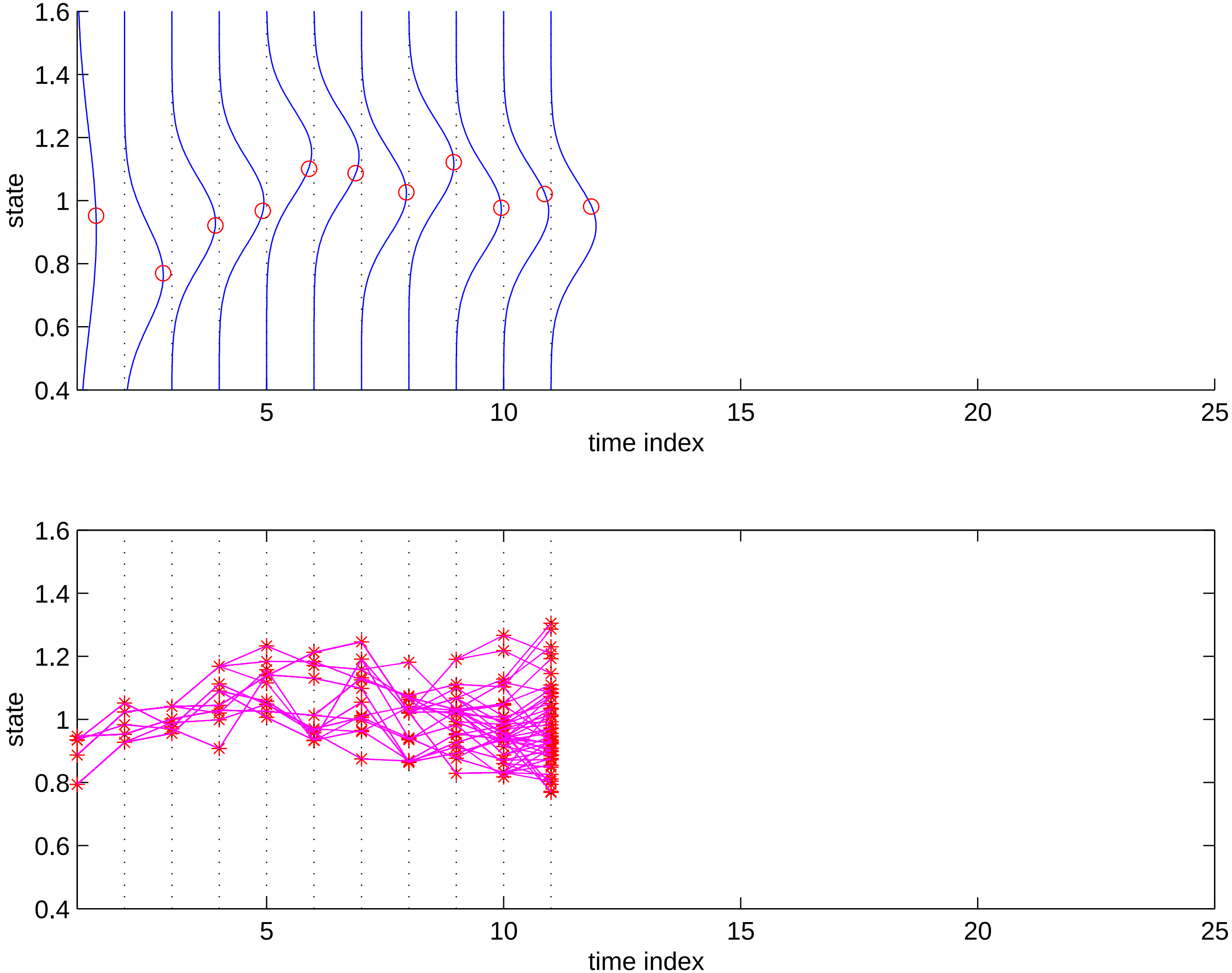
Predictive densities and evolution of the ancestor tree



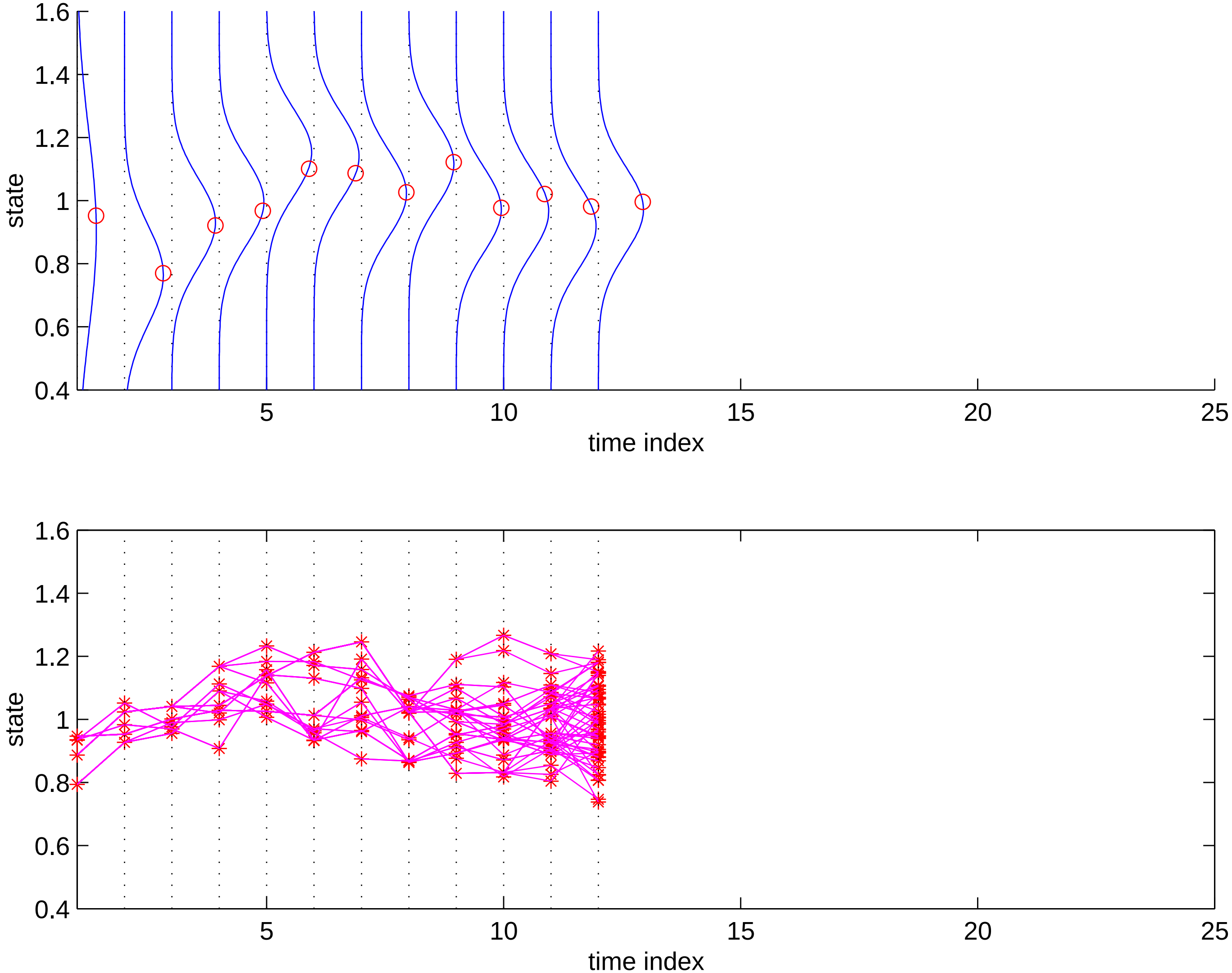
Predictive densities and evolution of the ancestor tree



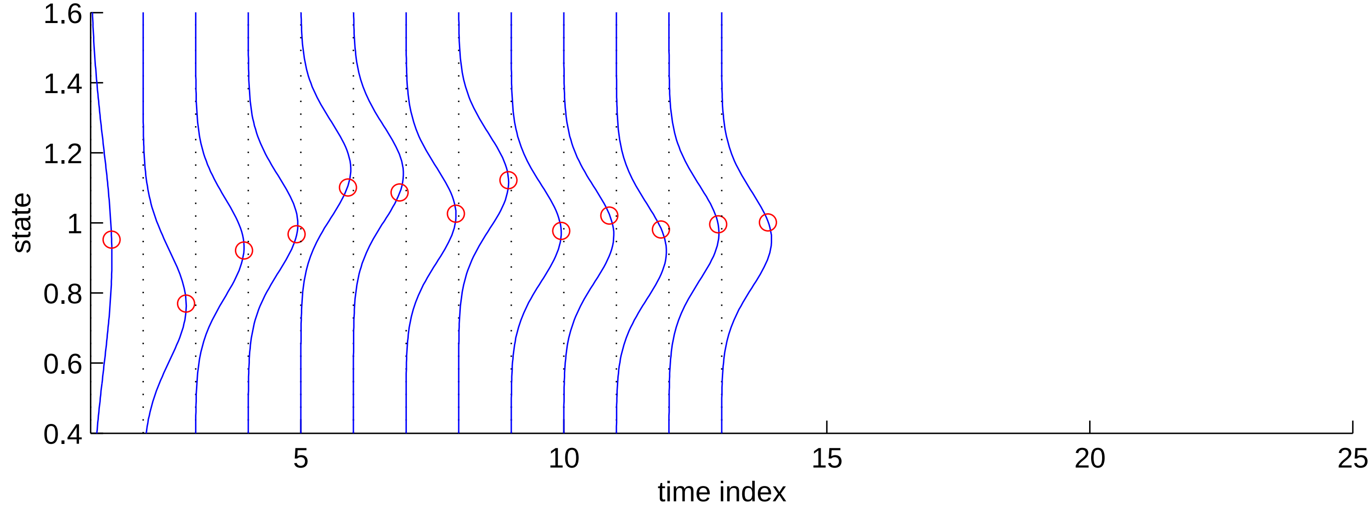
Predictive densities and evolution of the ancestor tree

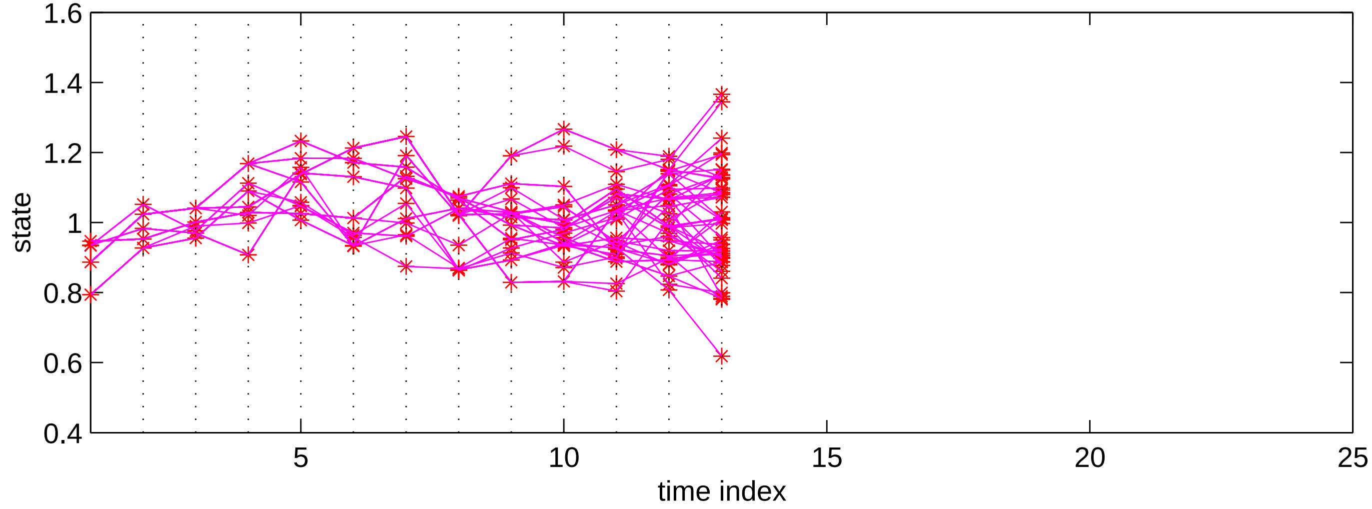


Predictive densities and evolution of the ancestor tree

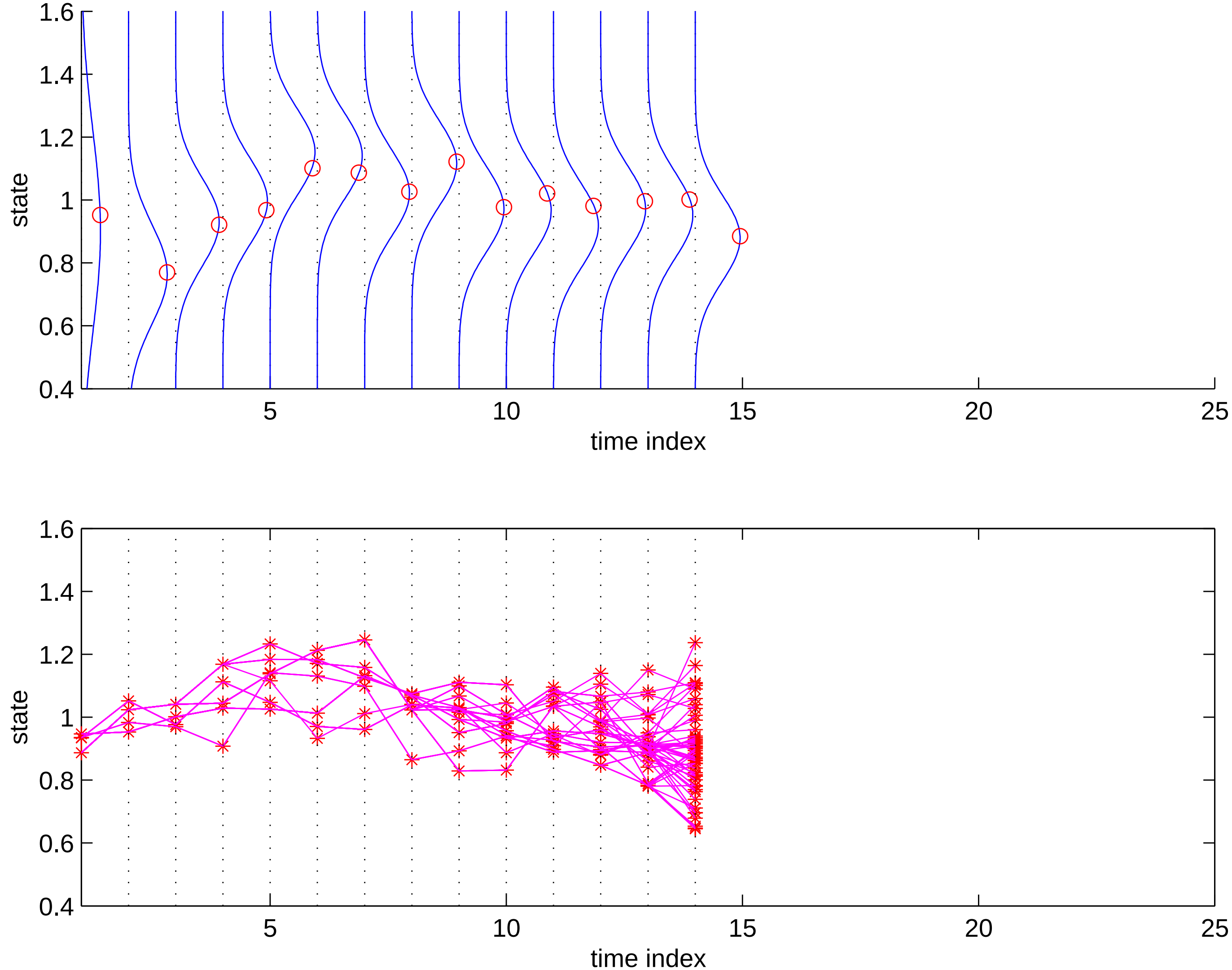


Predictive densities and evolution of the ancestor tree

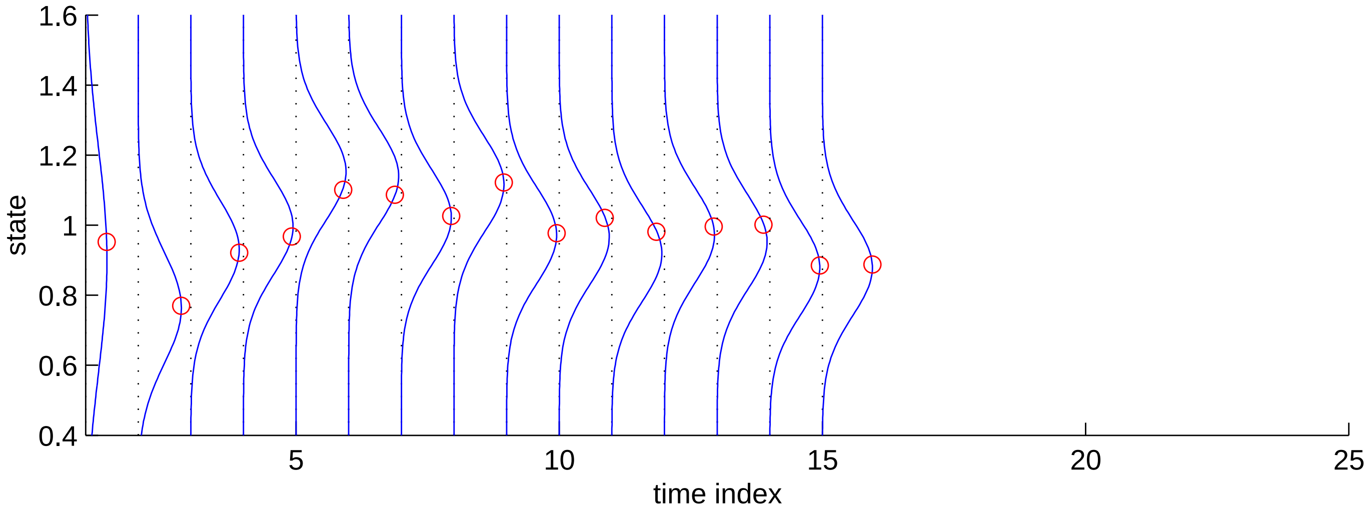


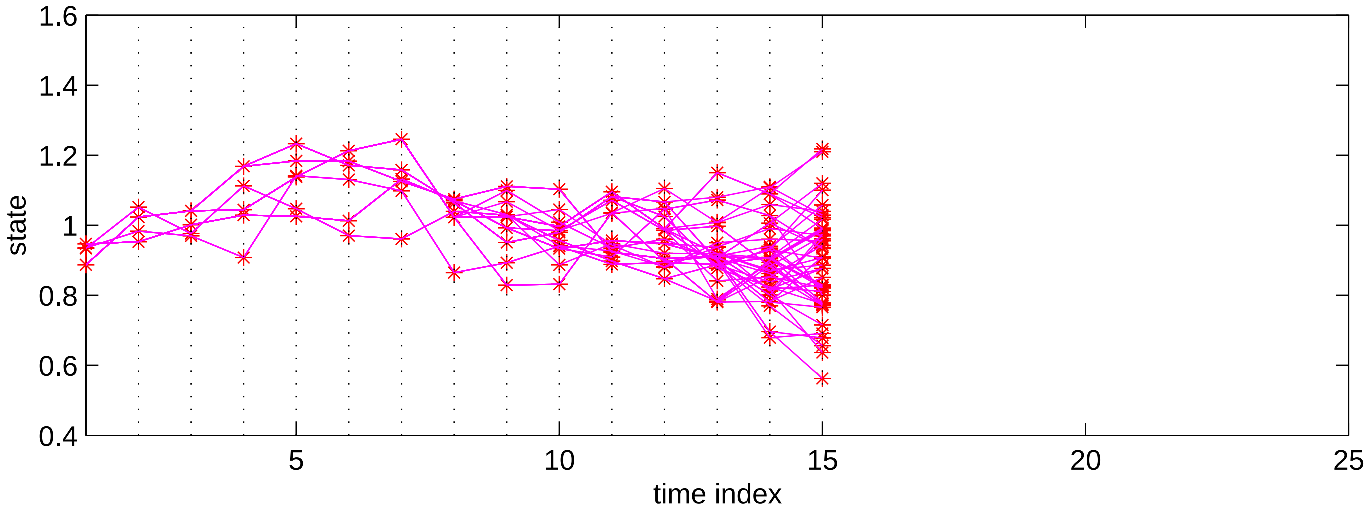


Predictive densities and evolution of the ancestor tree

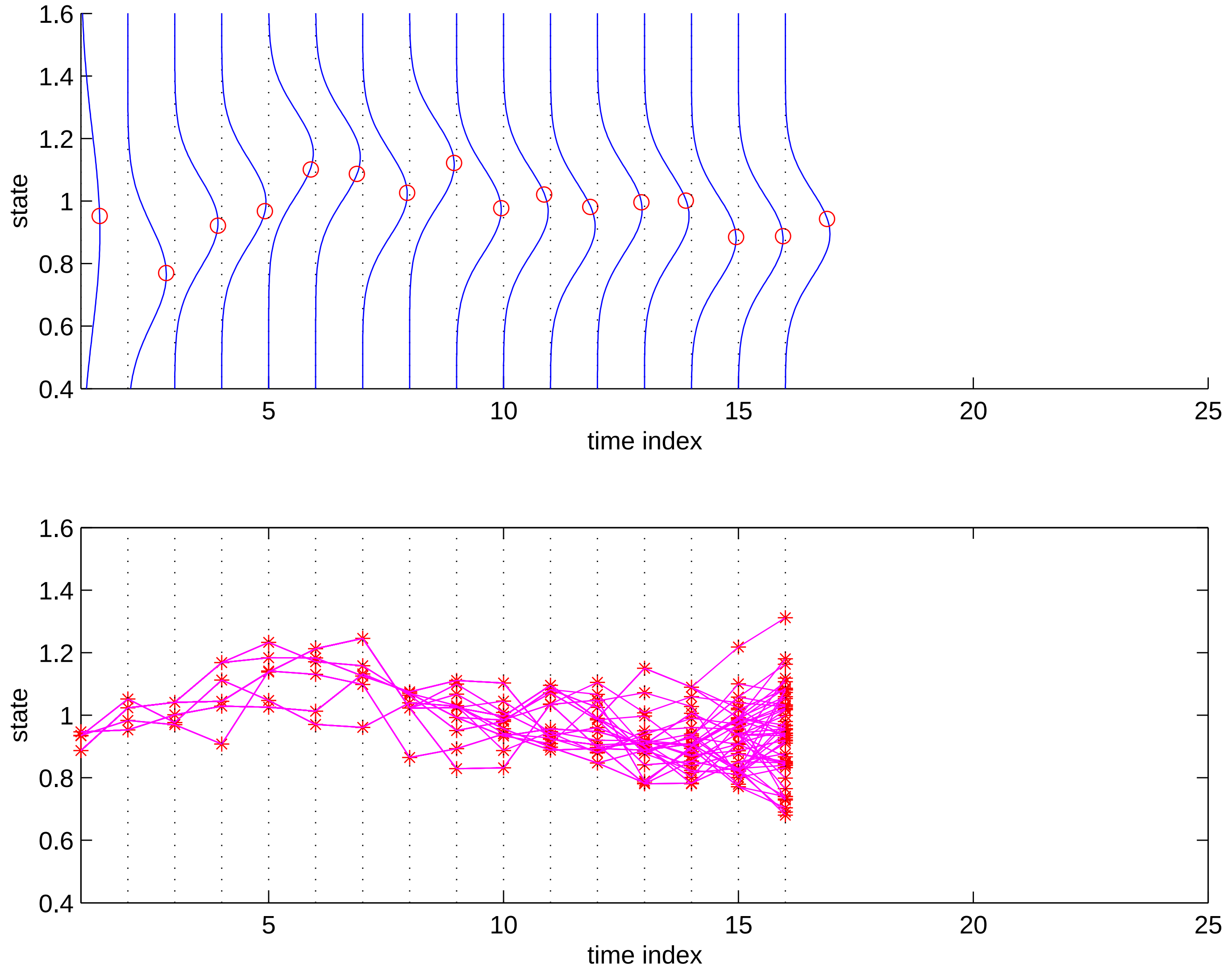


Predictive densities and evolution of the ancestor tree

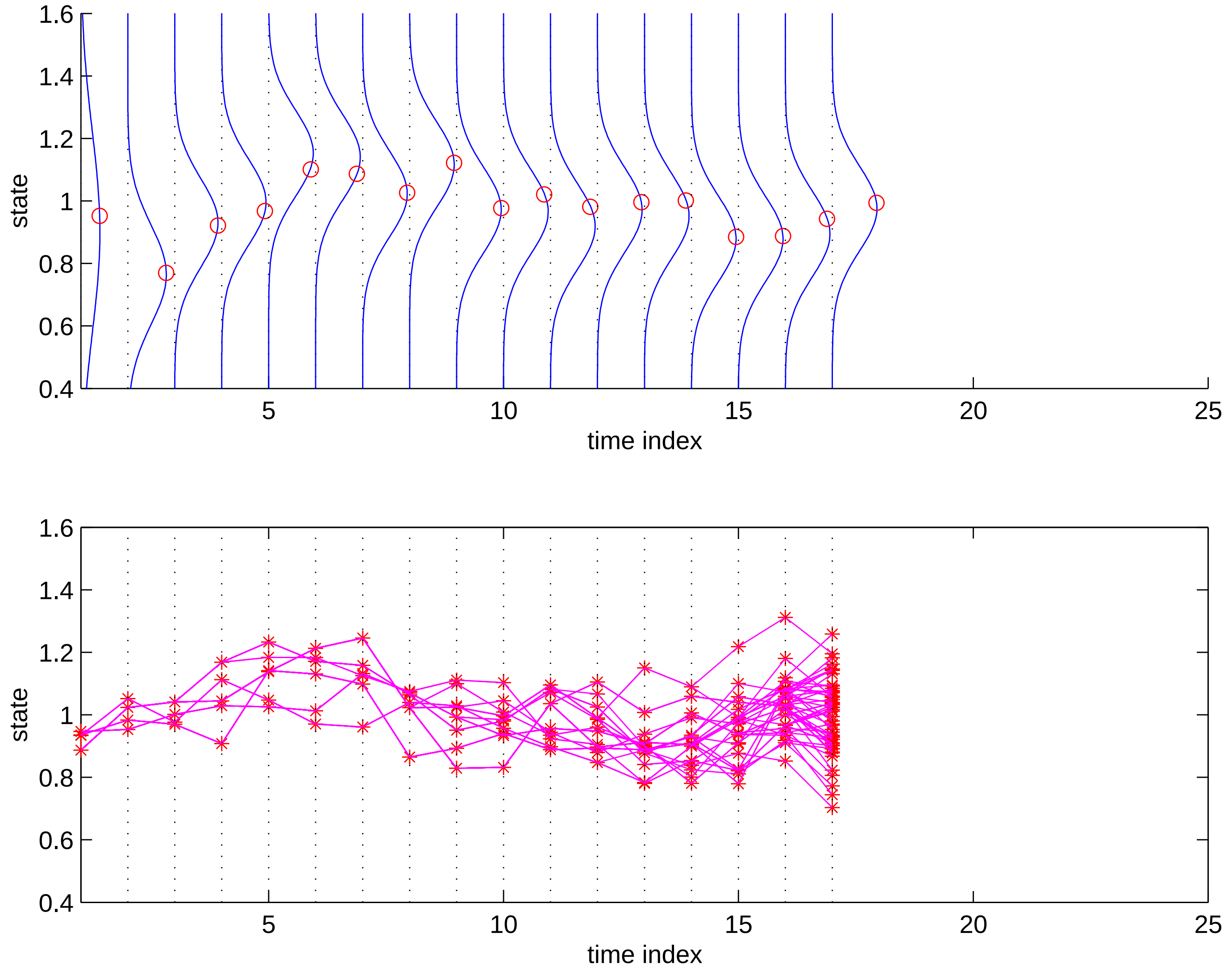




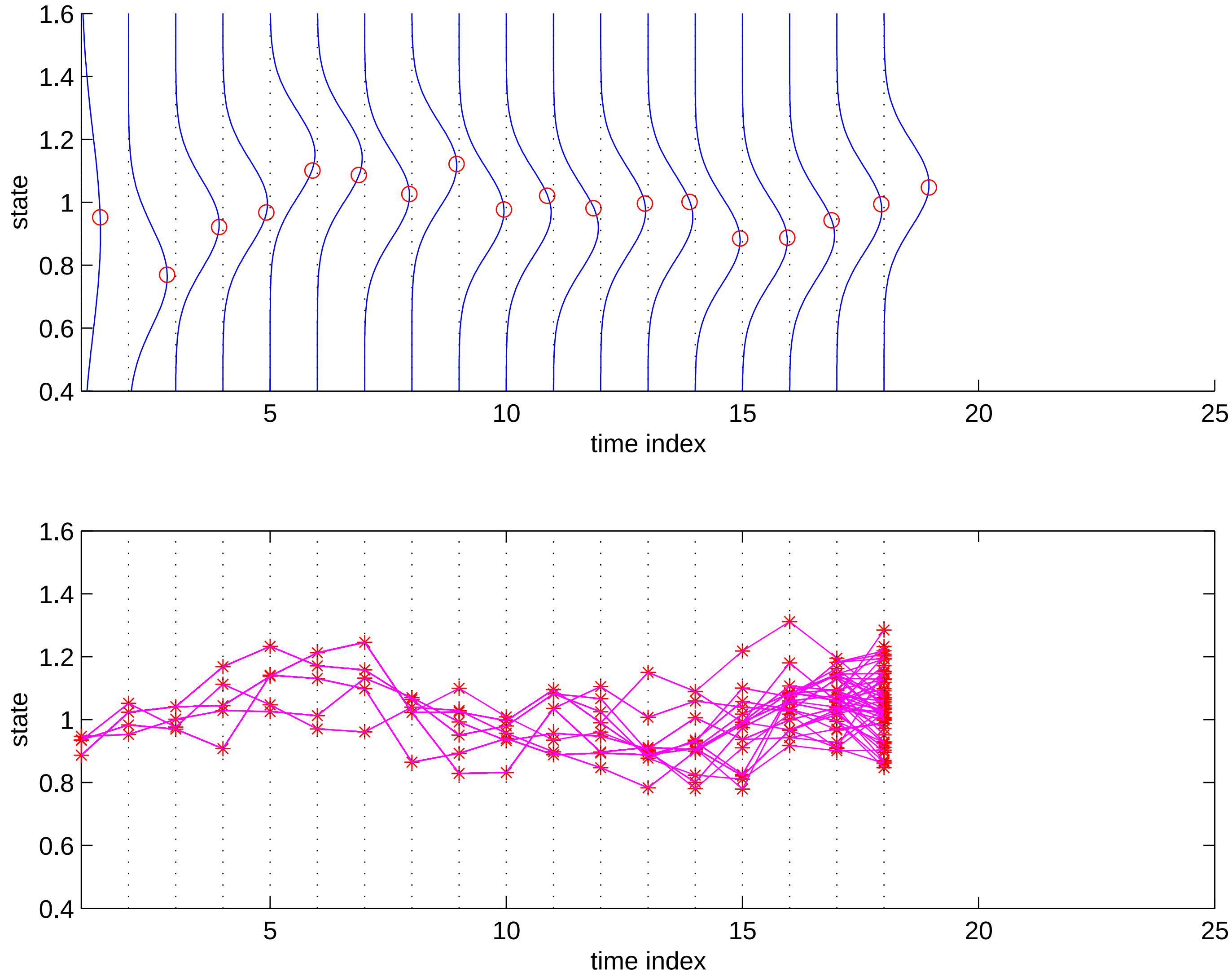
Predictive densities and evolution of the ancestor tree



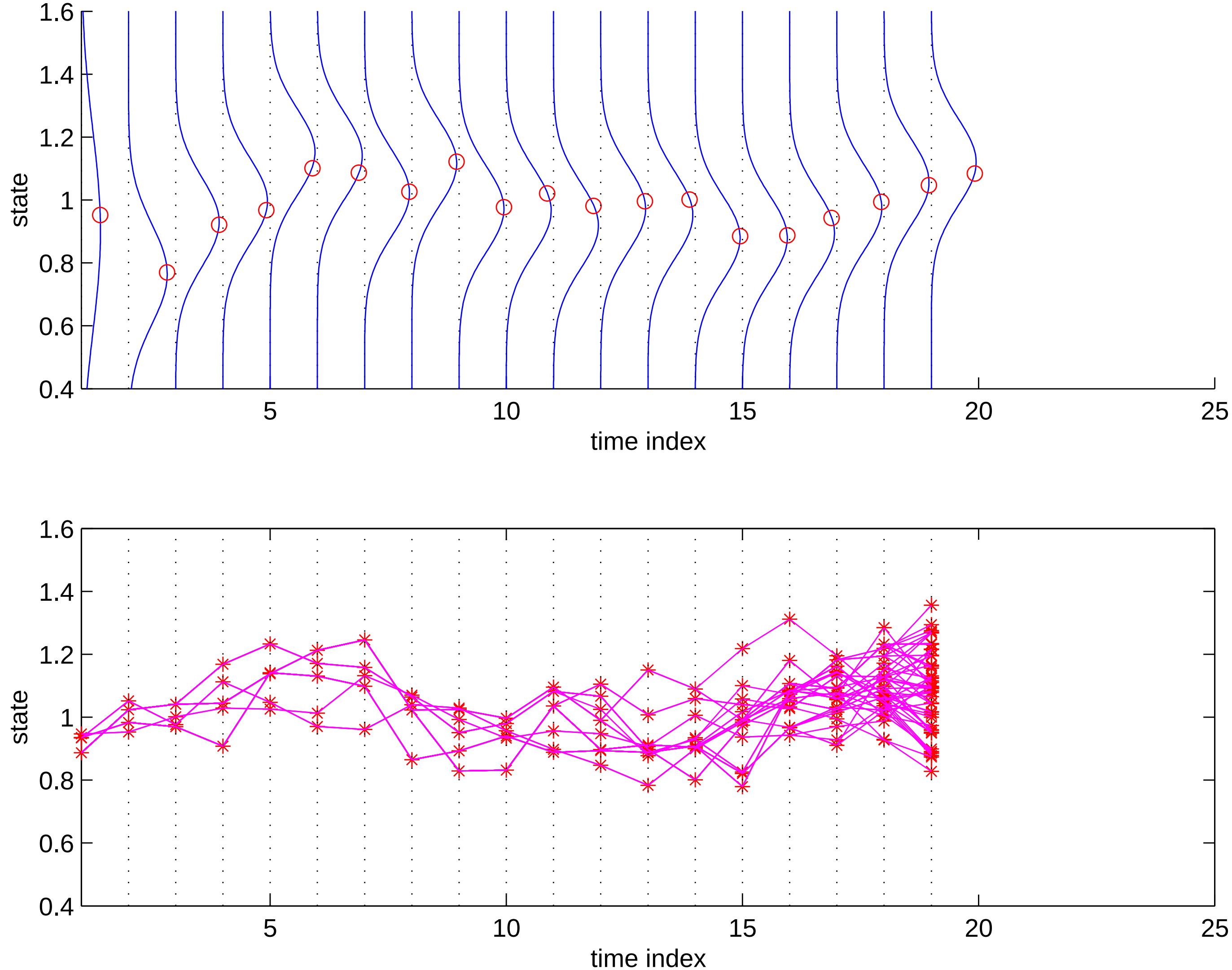
Predictive densities and evolution of the ancestor tree



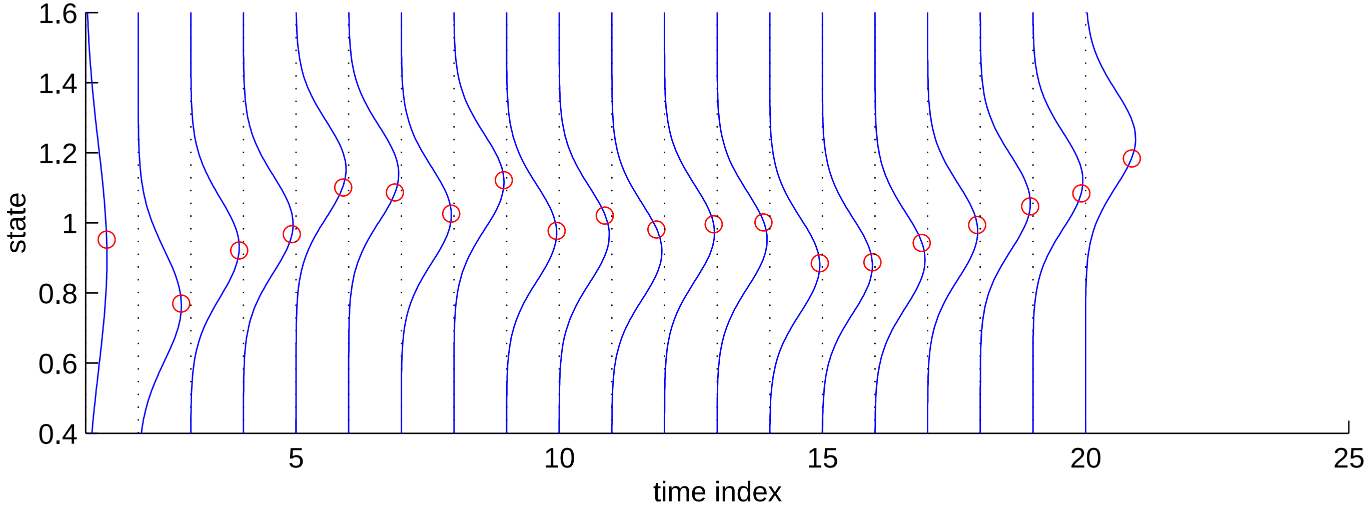
Predictive densities and evolution of the ancestor tree

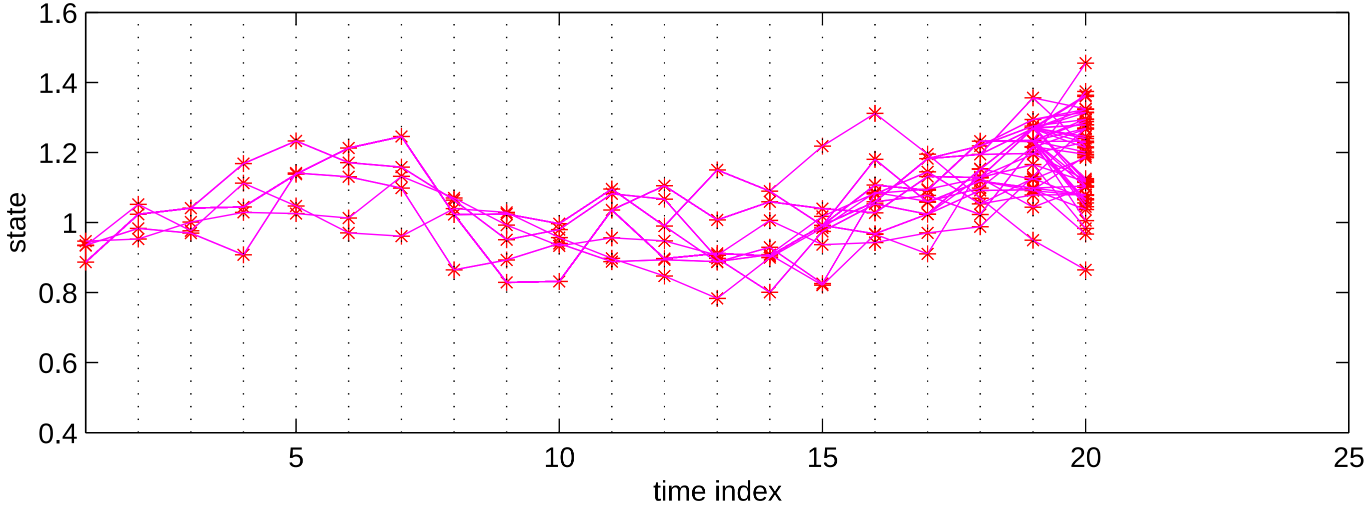


Predictive densities and evolution of the ancestor tree

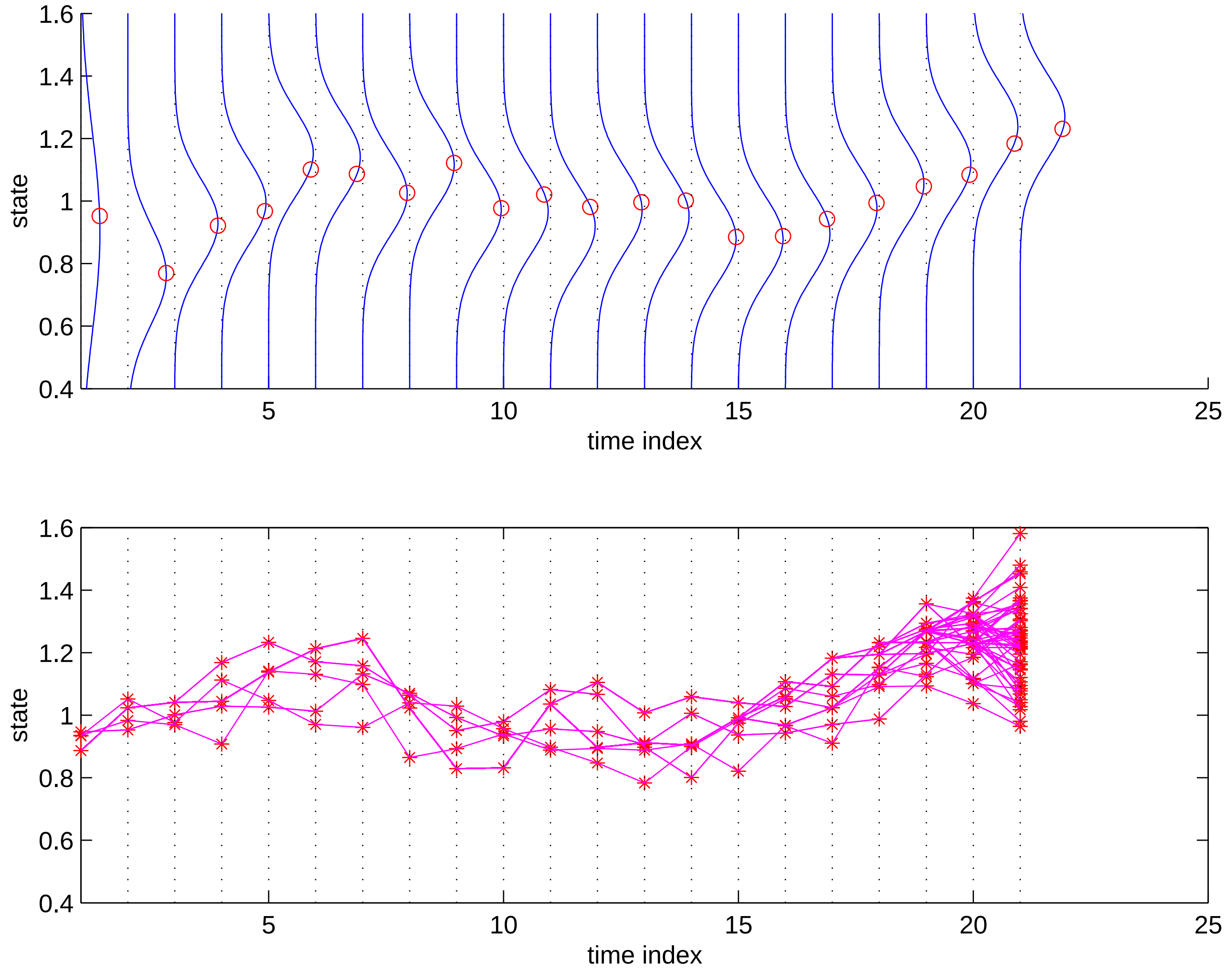


Predictive densities and evolution of the ancestor tree

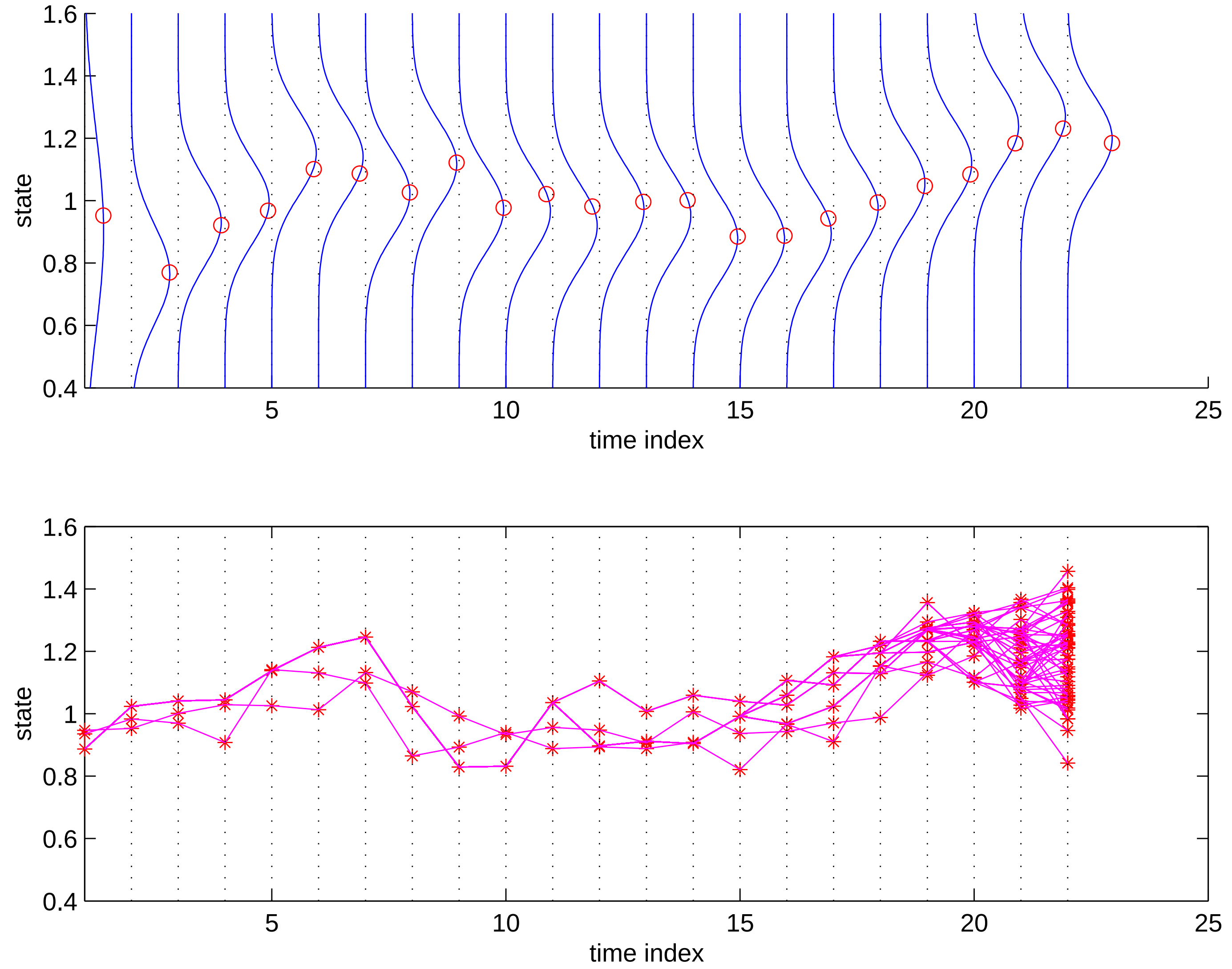




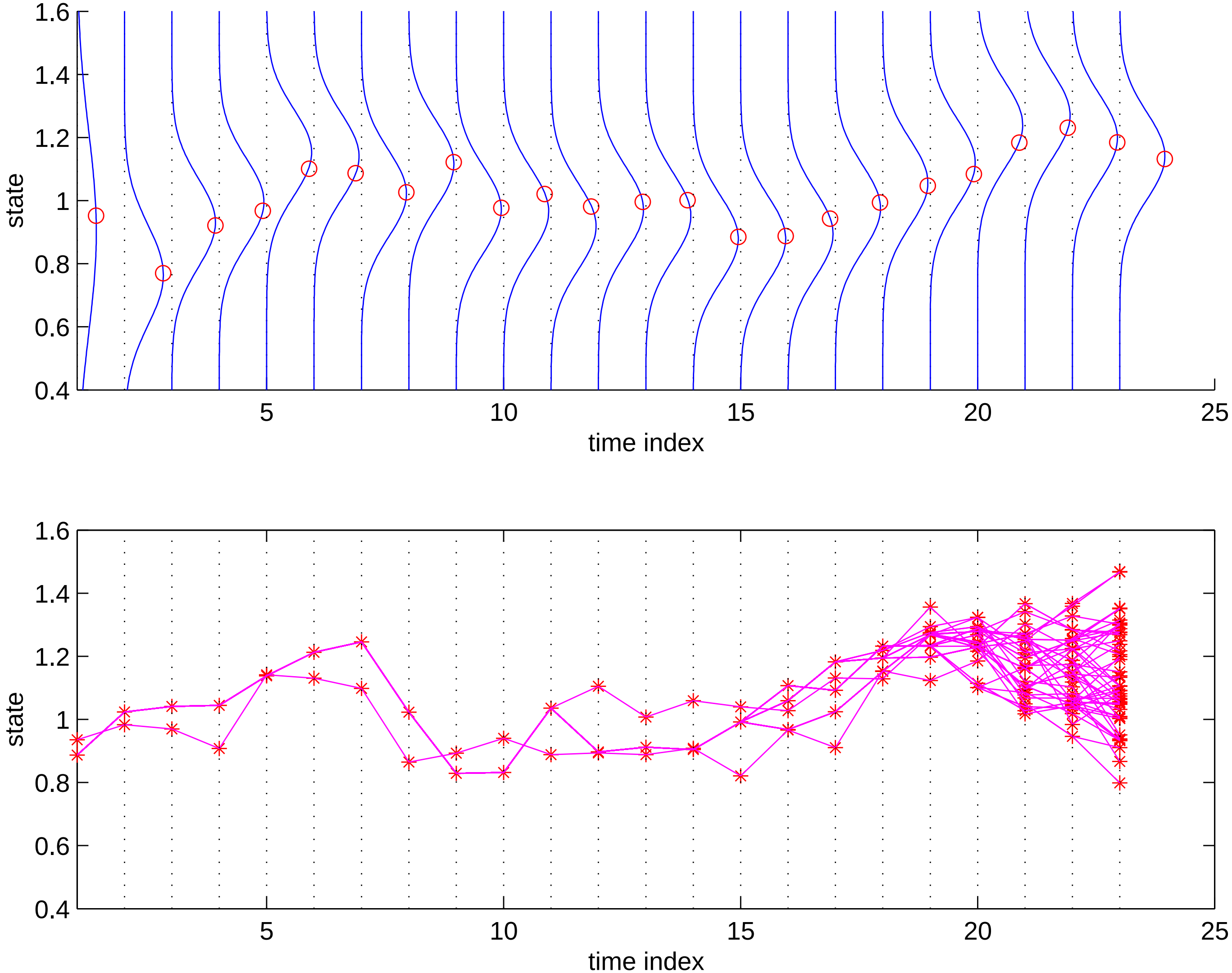
Predictive densities and evolution of the ancestor tree



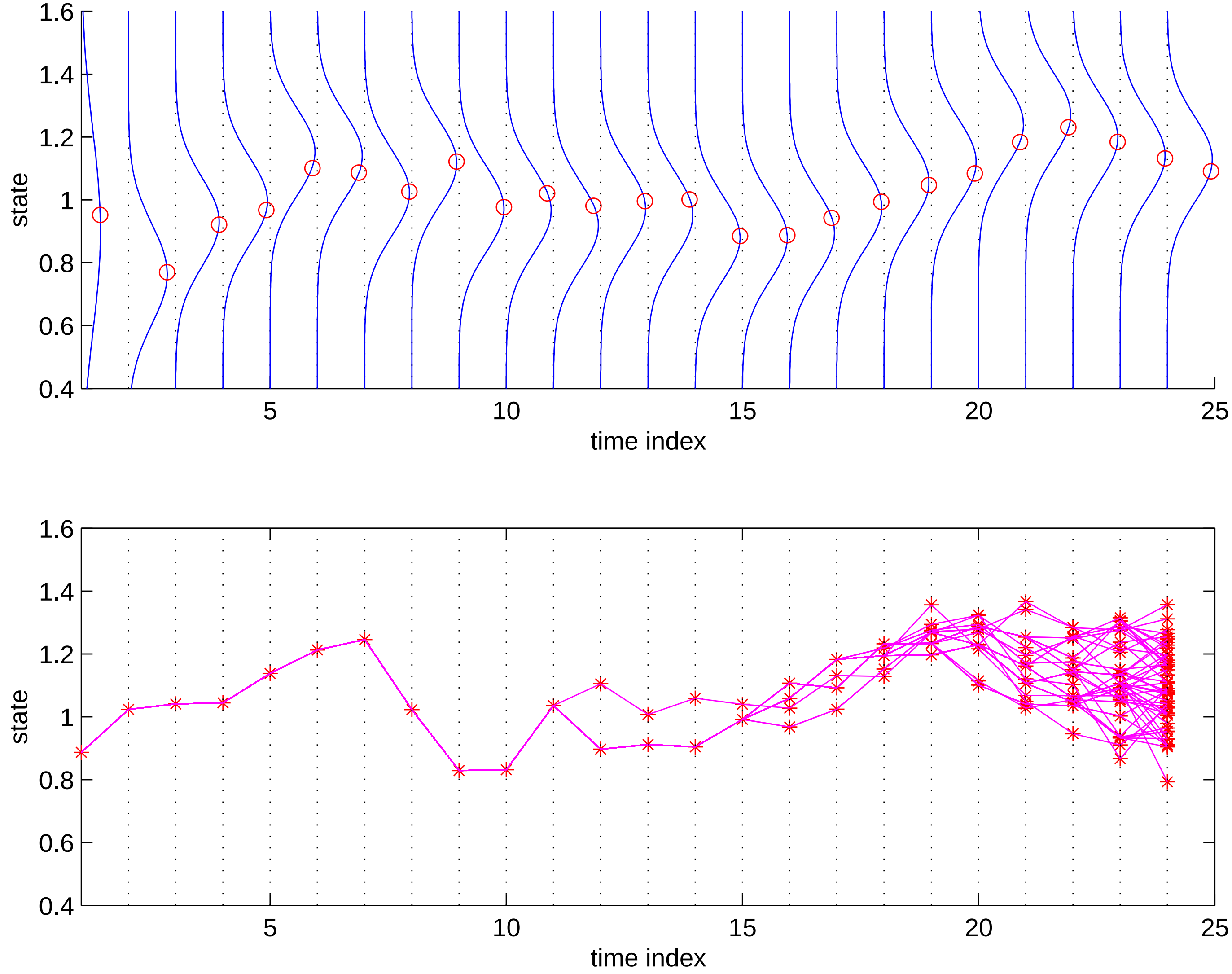
Predictive densities and evolution of the ancestor tree



Predictive densities and evolution of the ancestor tree



Predictive densities and evolution of the ancestor tree



Predictive densities and evolution of the ancestor tree

For $k \geq 0$,

$$w_{n+1} = \frac{\psi_{n+1}}{\langle \mu_{n+1|n}, \psi_{n+1} \rangle} w_n Q$$

and thus $w_n / \langle w_n, 1 \rangle$ satisfies the same recursion as μ_n. More precisely,

$$\begin{aligned}
\langle w_n, \phi \rangle &= \mathrm{E}\left[f(X_0)\phi(X_n) \middle| Y_{0:n} \right] \\
&= \mathrm{E}\left[\mathrm{E}\left[f(X_0) \middle| X_n, Y_{0:n} \right] \phi(X_n) \middle| Y_{0:n} \right] \\
&= \mathrm{E}\left[f(X_0) \middle| Y_{0:n} \right] \mathrm{E}\left[\phi(X_n) \middle| Y_{0:n} \right] \\
&\qquad + \mathrm{E}\left[\underbrace{\left(\mathrm{E}\left[f(X_0) \middle| X_n, Y_{0:n} \right] - \mathrm{E}\left[f(X_0) \middle| Y_{0:n} \right] \right)}_{|\cdot| \leq \|f\| \rho^n} \phi(X_n) \middle| Y_{0:n} \right]
\end{aligned}$$

where $\rho = 1 - \epsilon^2$ assuming that $\epsilon \pi \leq Q(x, \cdot) \leq \epsilon^{-1} \pi$

Thus for large n, $w_n \approx A_n \mu_n$

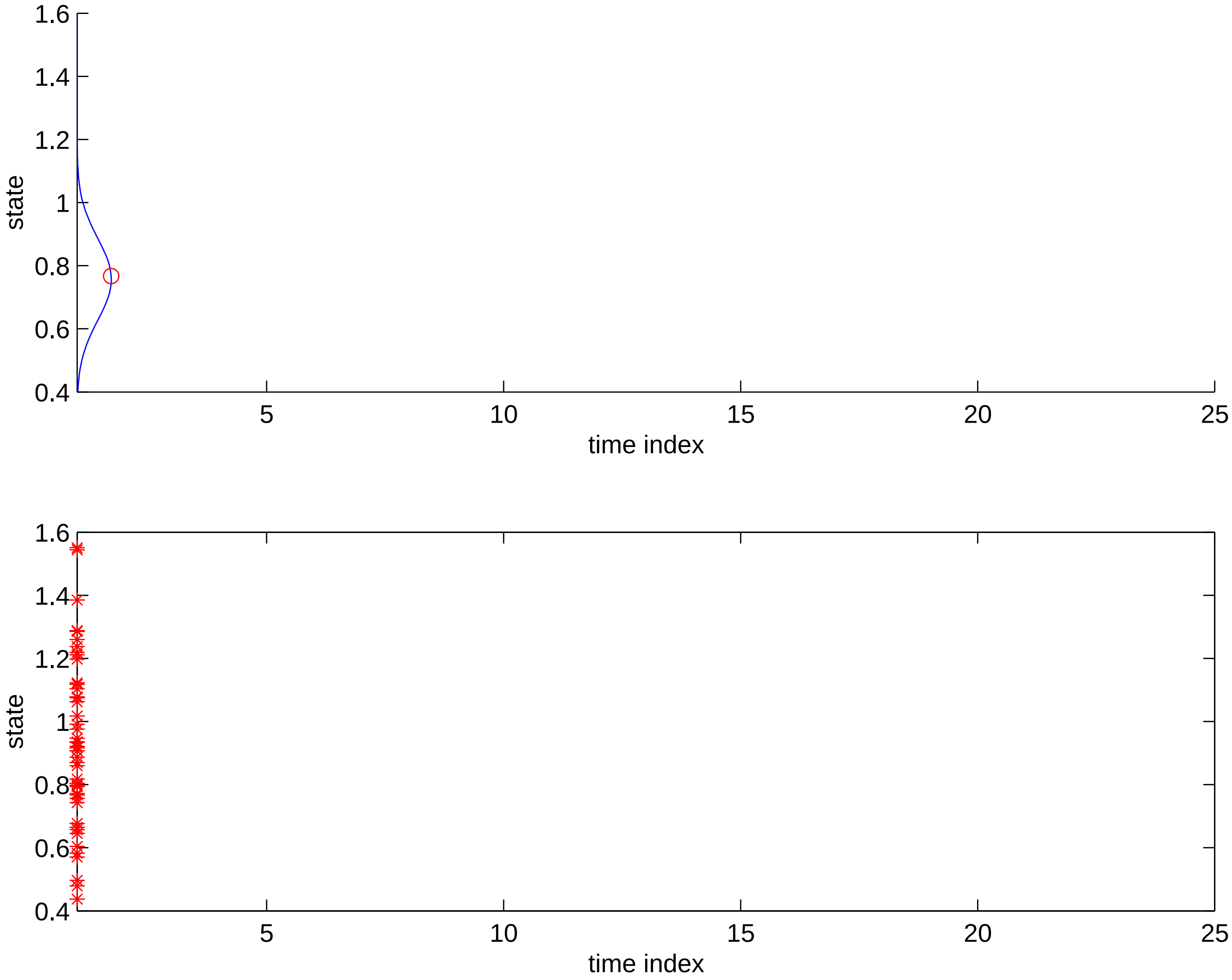
Suggestion

When updating from n to $n+1$ use a resampling scheme which compromises between $\hat{\mu}_n = \sum_{i=1}^{p} \omega_n^i \xi_{n|n-1}^i$ and $\hat{w}_n = \sum_{i=1}^{p} \rho_n^i \omega_n^i \xi_{n|n-1}^i$

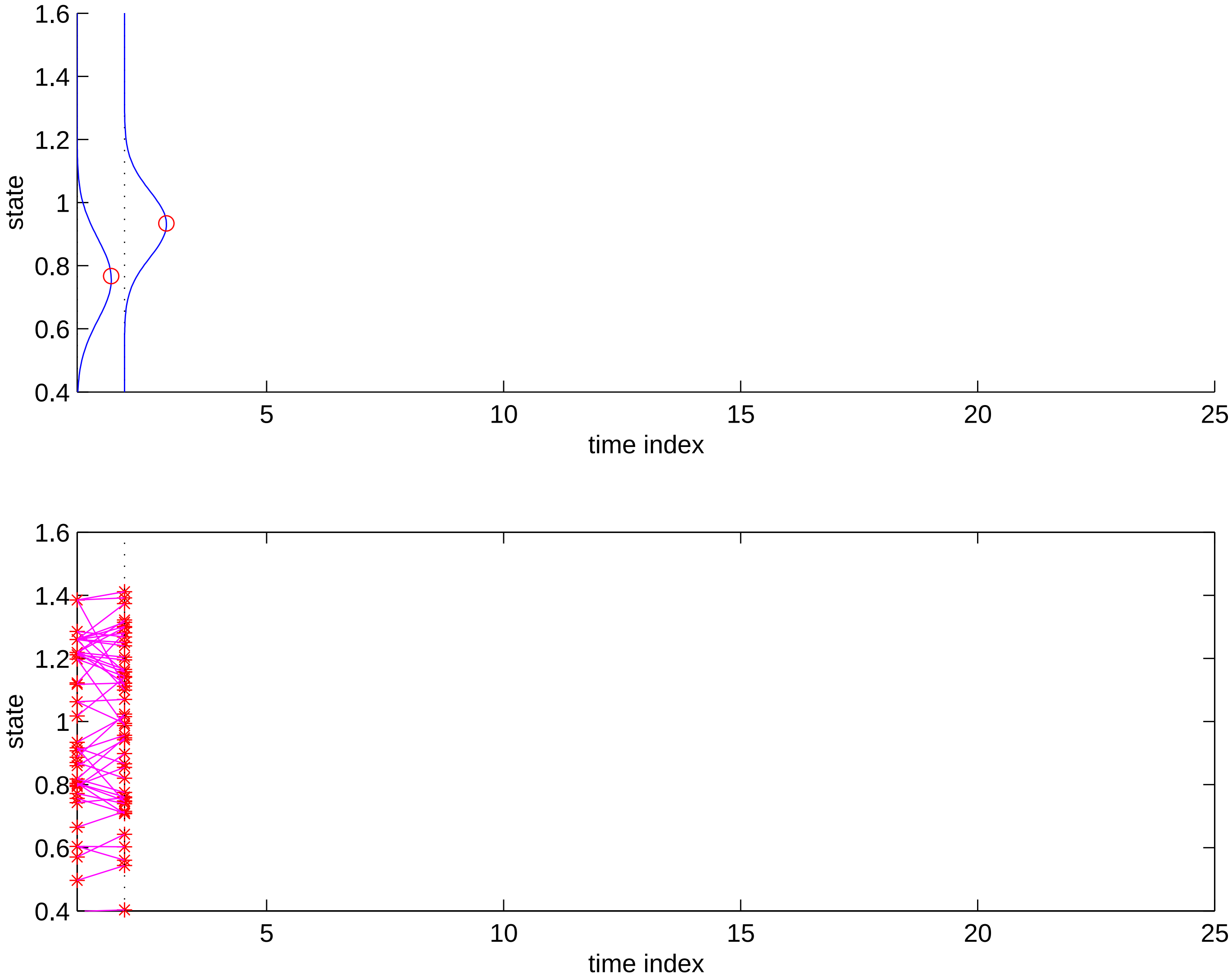
For example

$$\tilde{\omega}_i = \frac{1}{2}\left(\omega_n^i + \frac{\rho_n^i}{\sum_{j=1}^{p} \rho_n^j}\right)$$

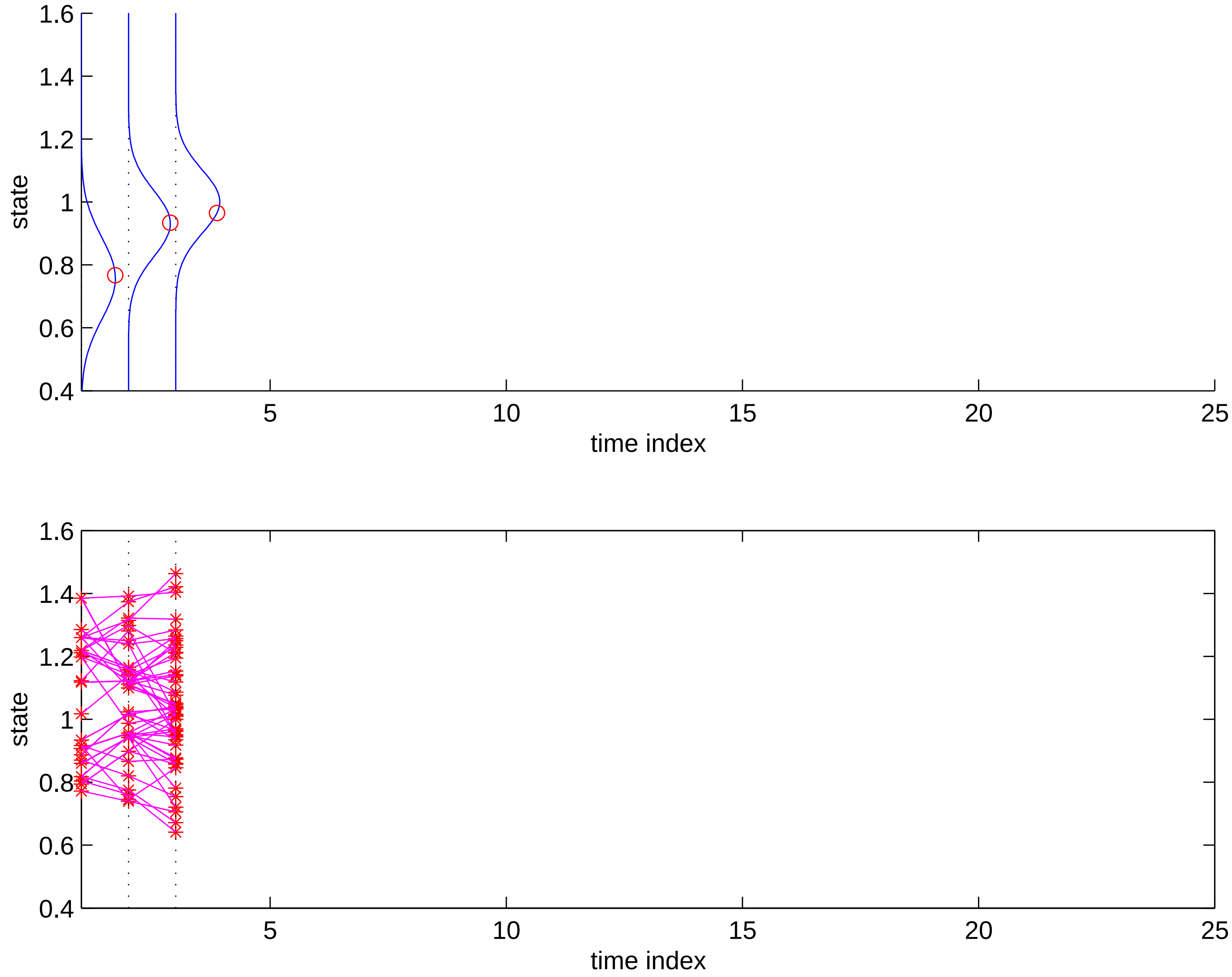
and use Bayesian importance sampling correction for $\omega_{n+1}^{1:p}$ and $\rho_{n+1}^{1:p}$



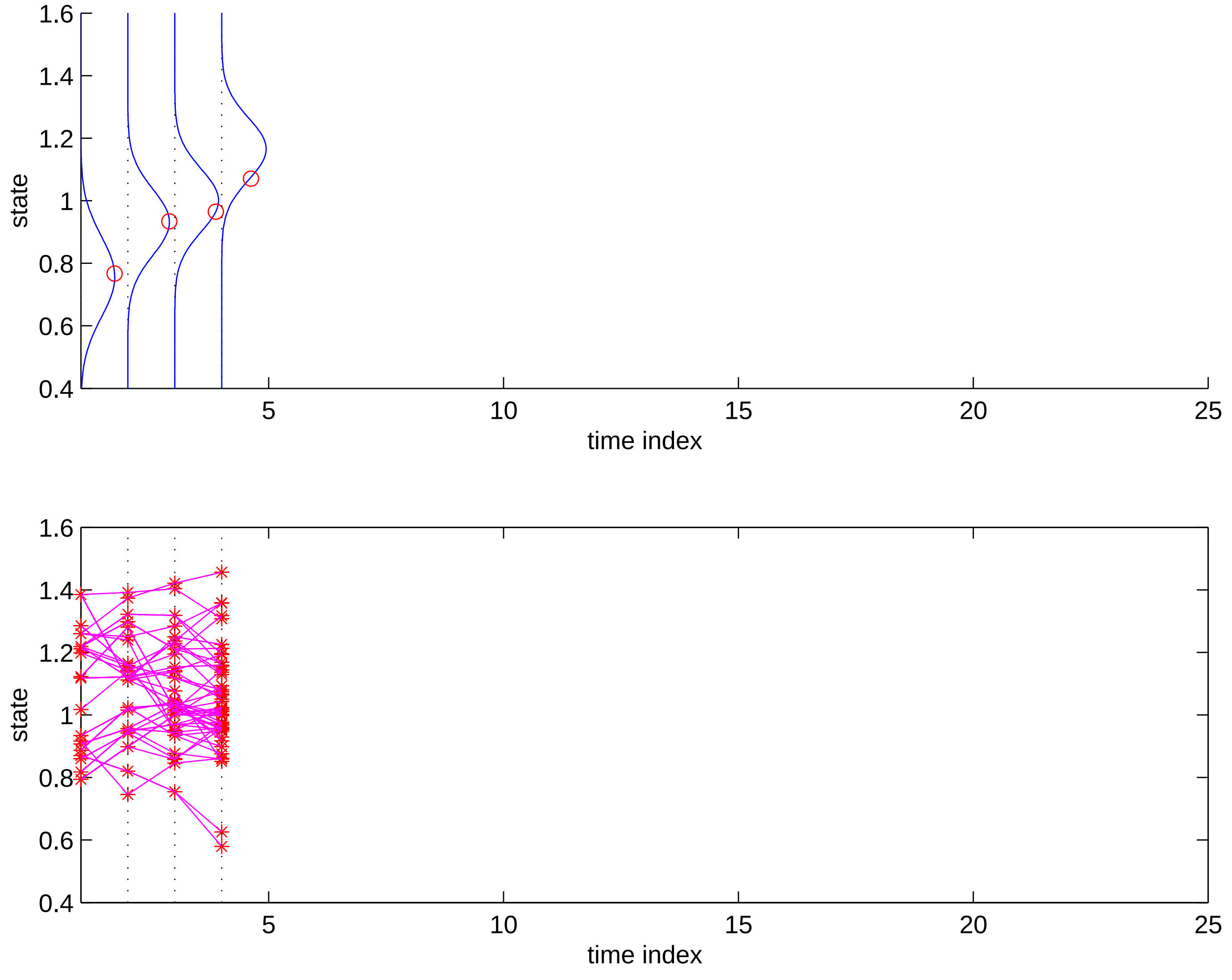
Filtering densities and evolution of the ancestor tree



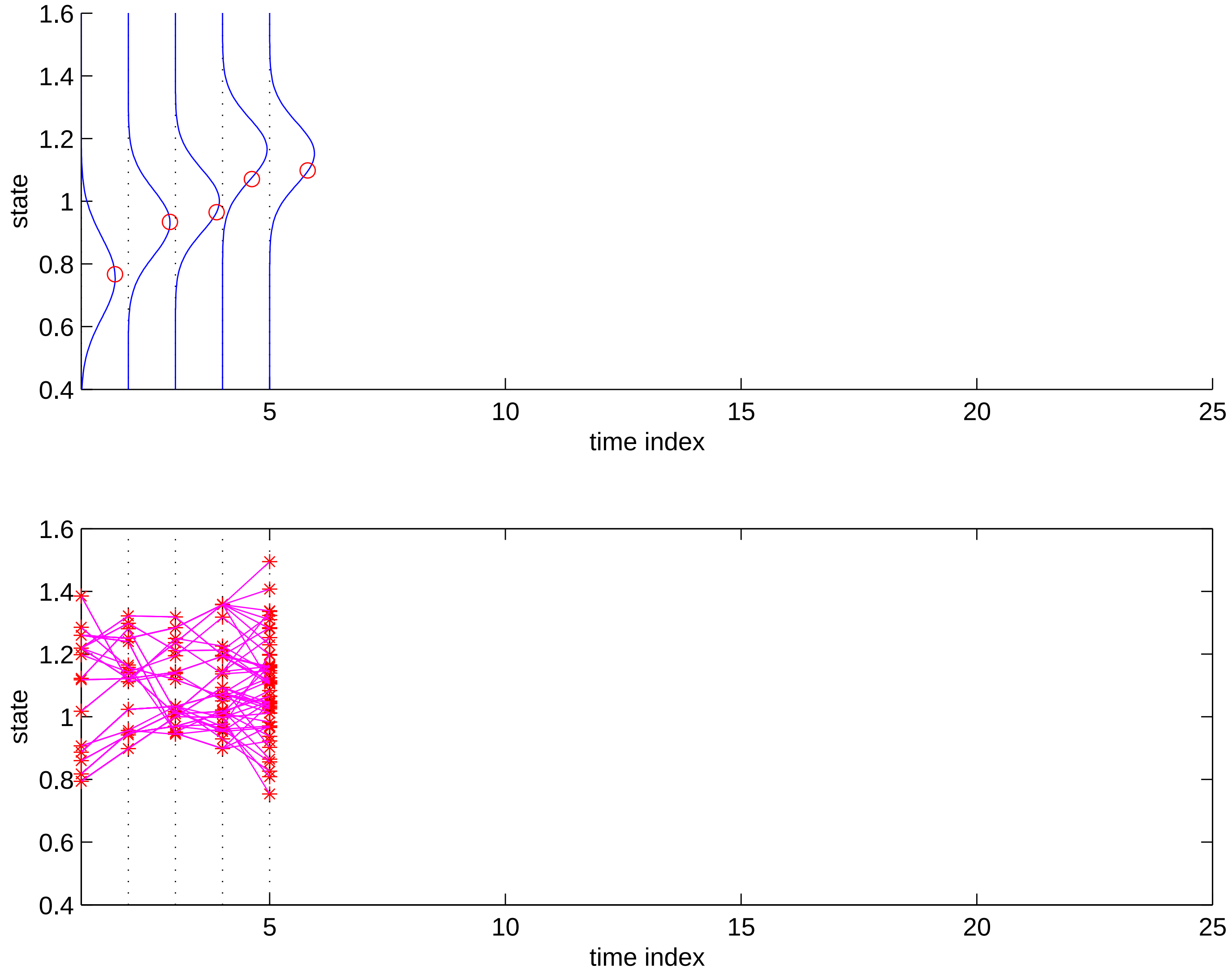
Filtering densities and evolution of the ancestor tree



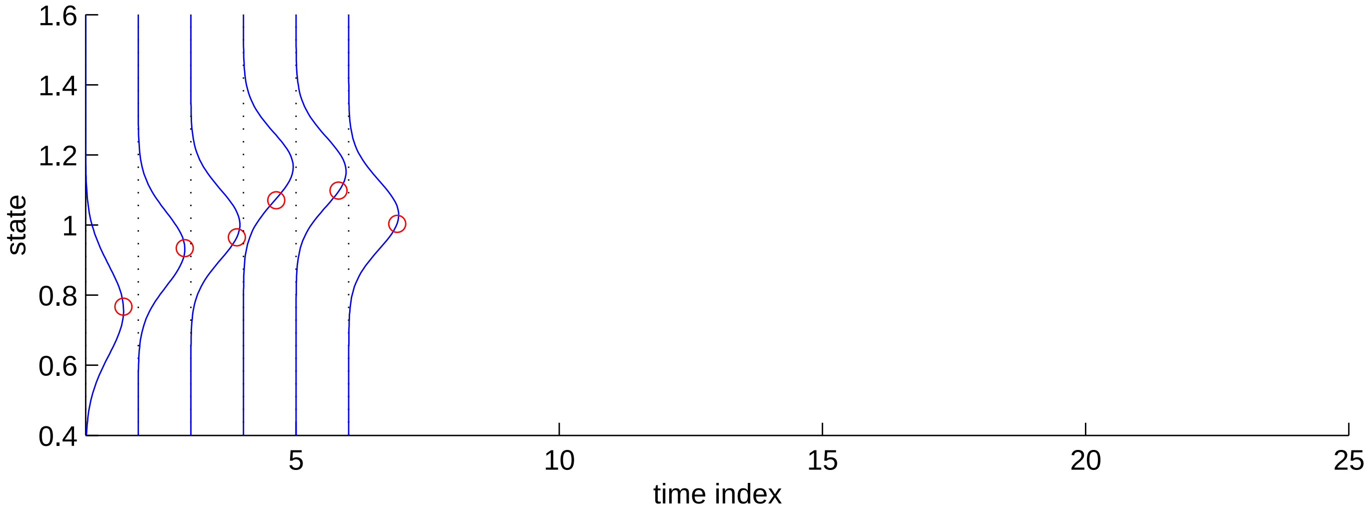
Filtering densities and evolution of the ancestor tree

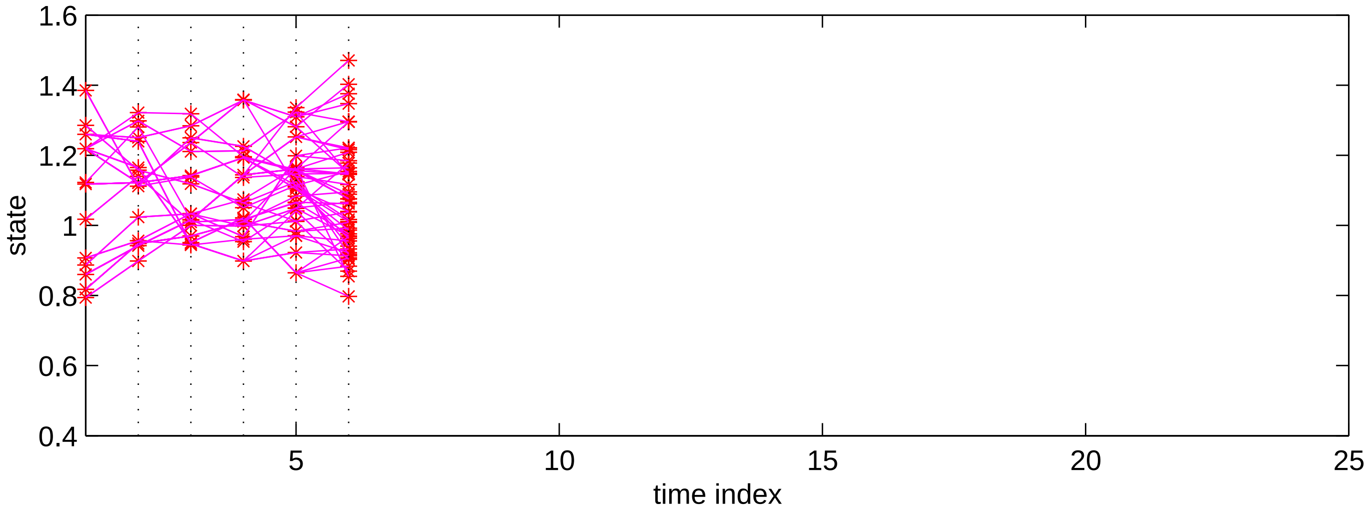


Filtering densities and evolution of the ancestor tree

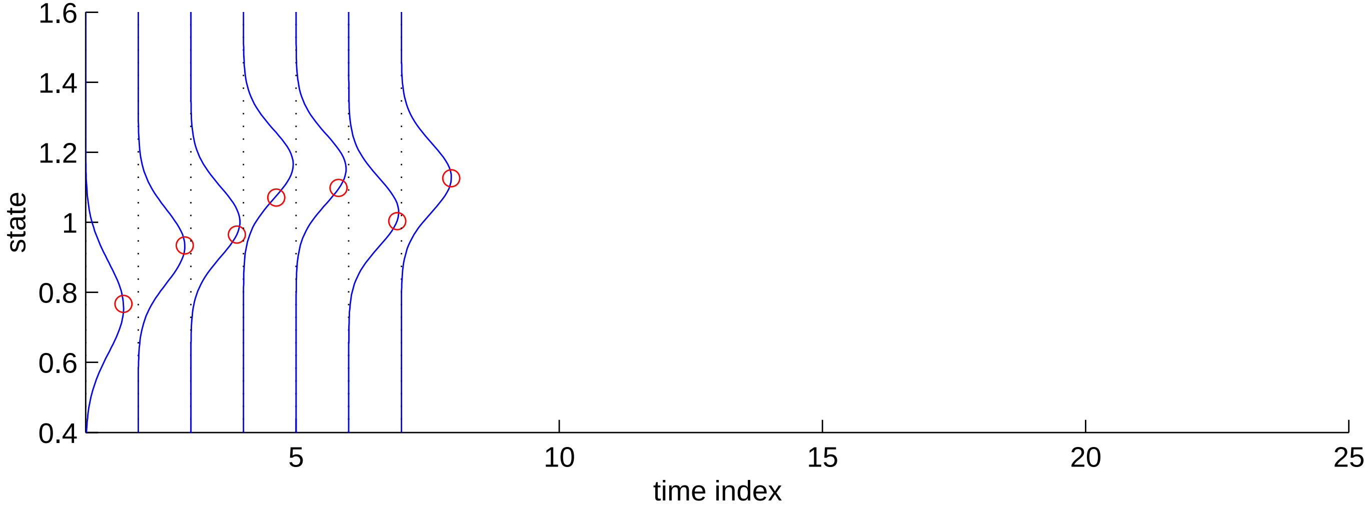


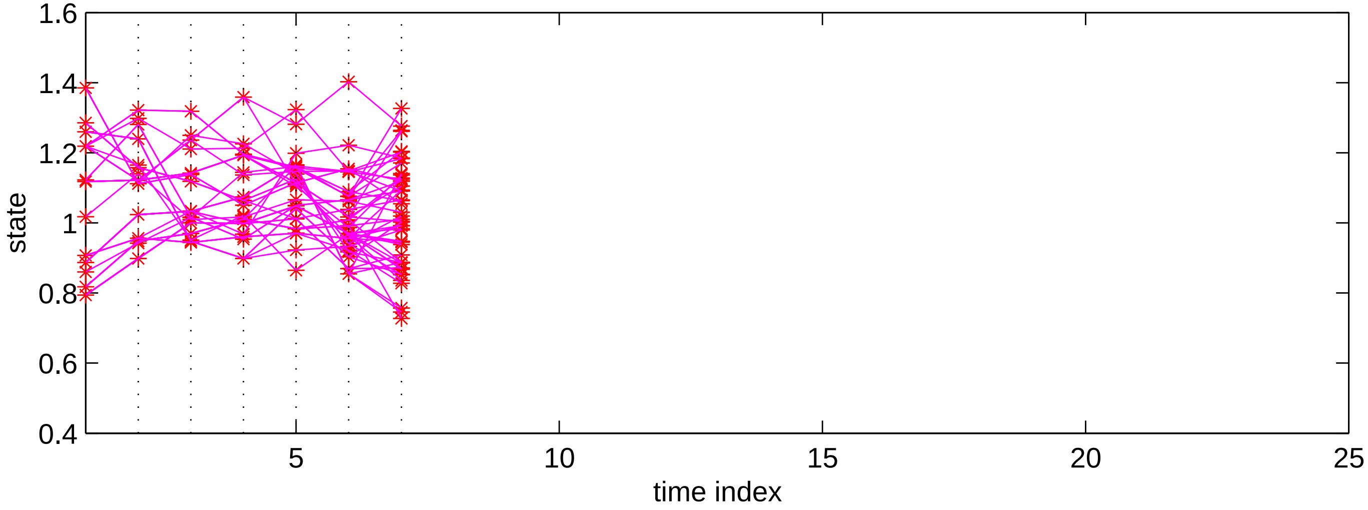
Filtering densities and evolution of the ancestor tree



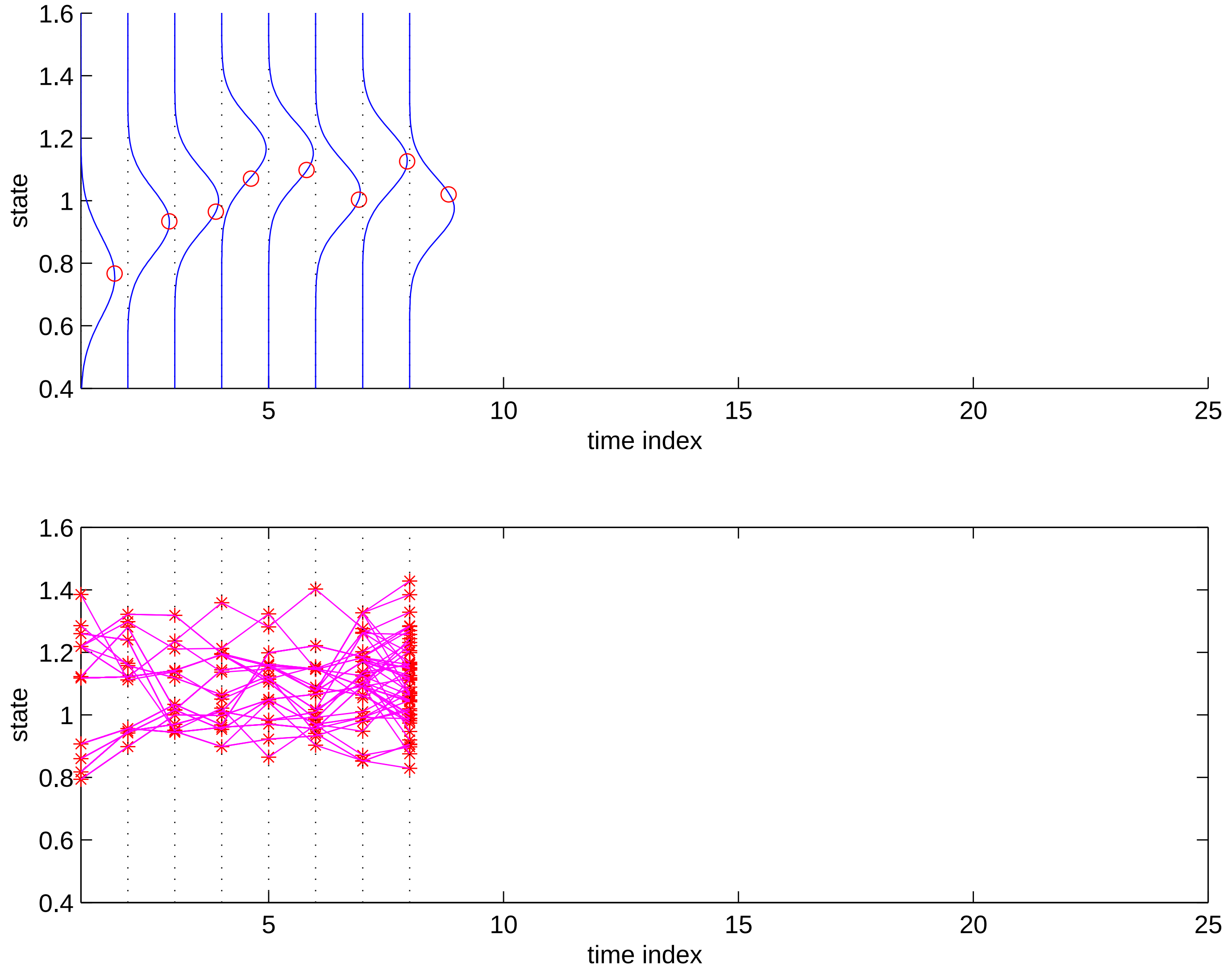


Filtering densities and evolution of the ancestor tree

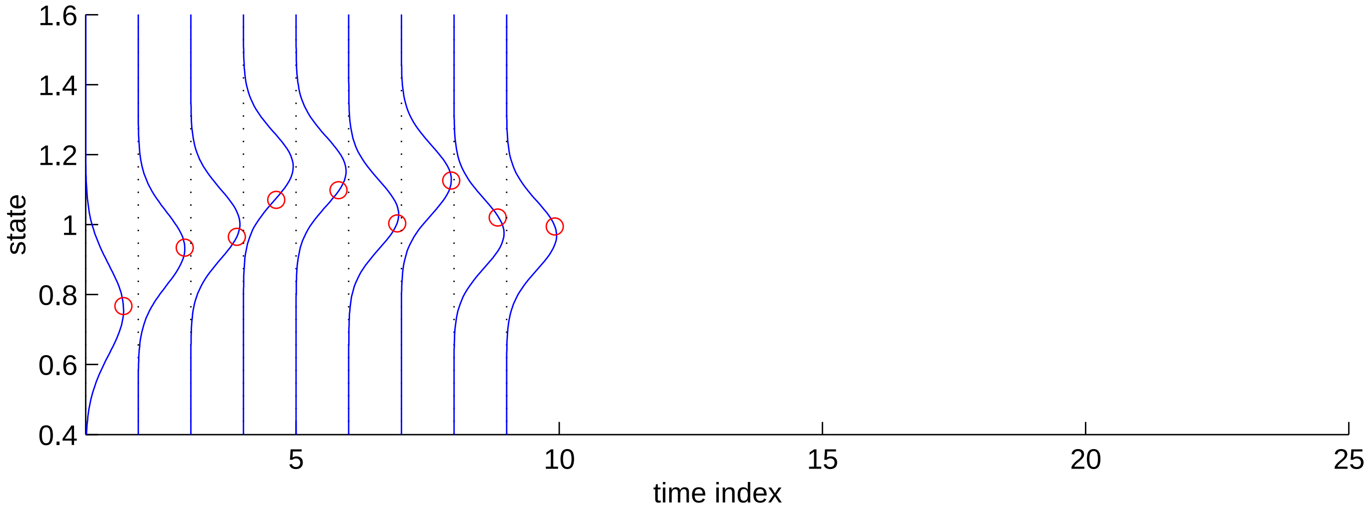


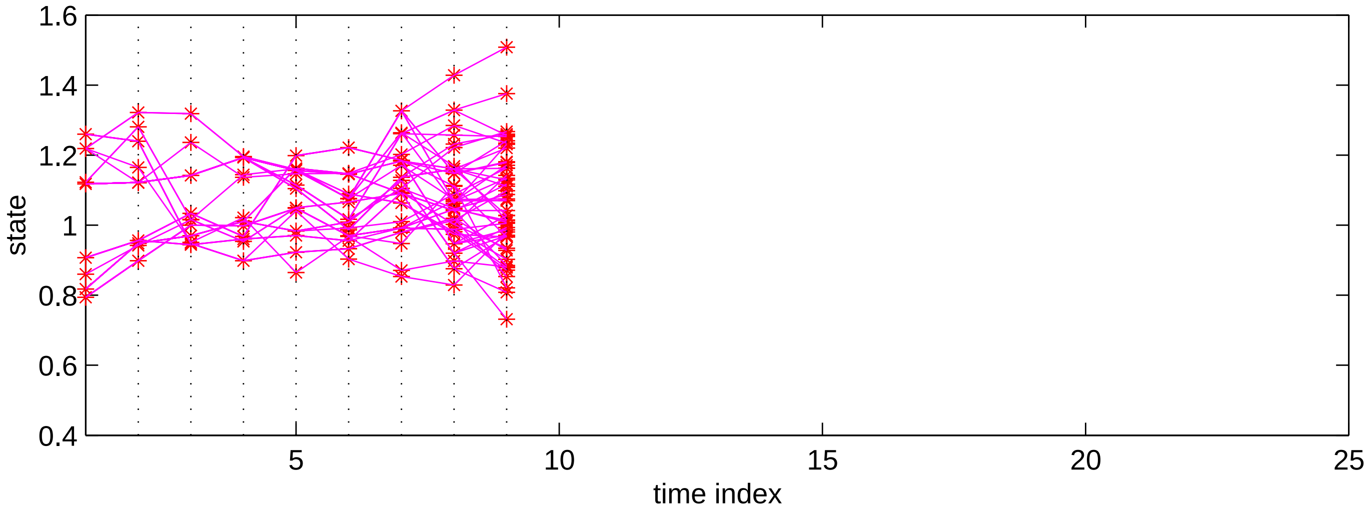


Filtering densities and evolution of the ancestor tree

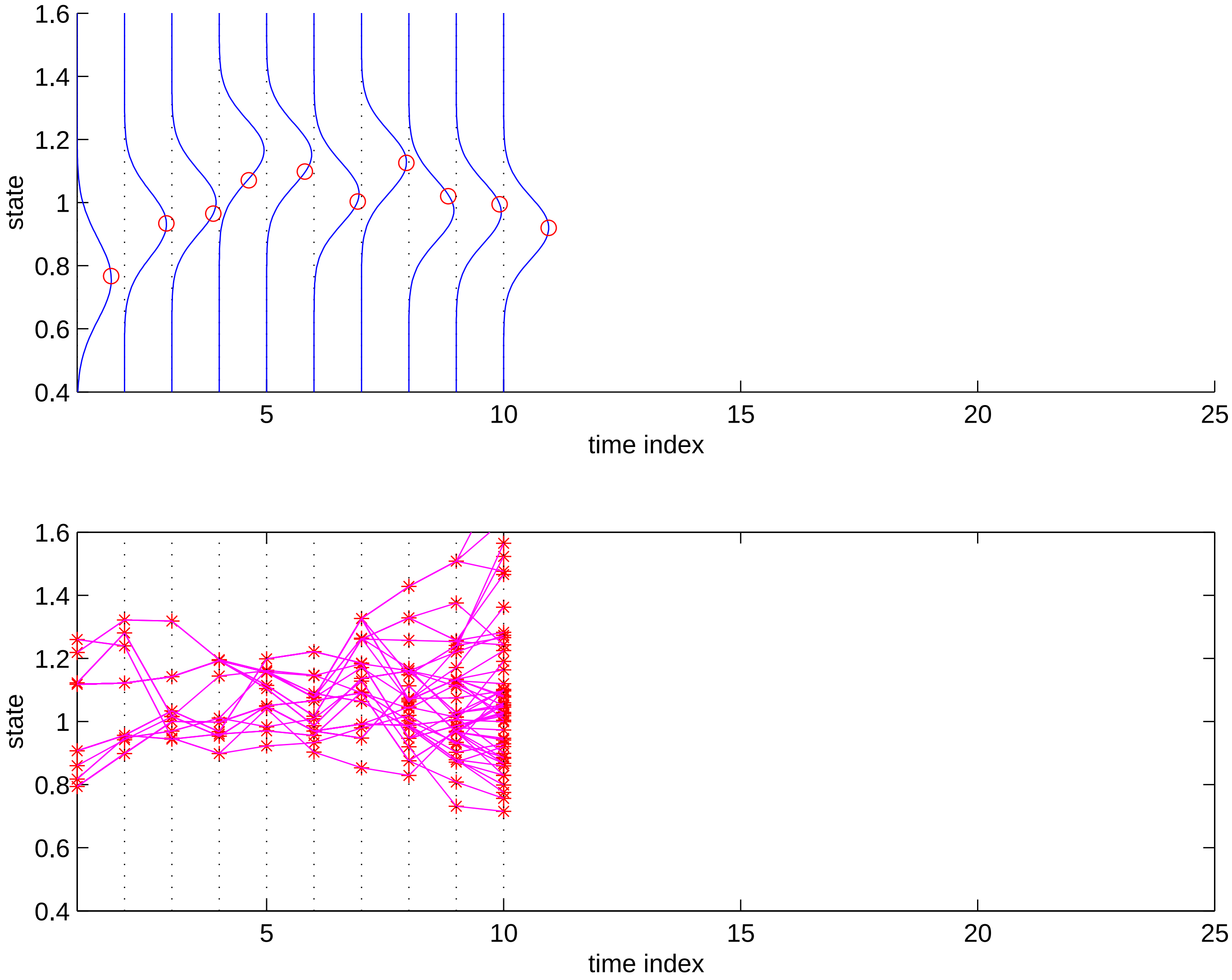


Filtering densities and evolution of the ancestor tree

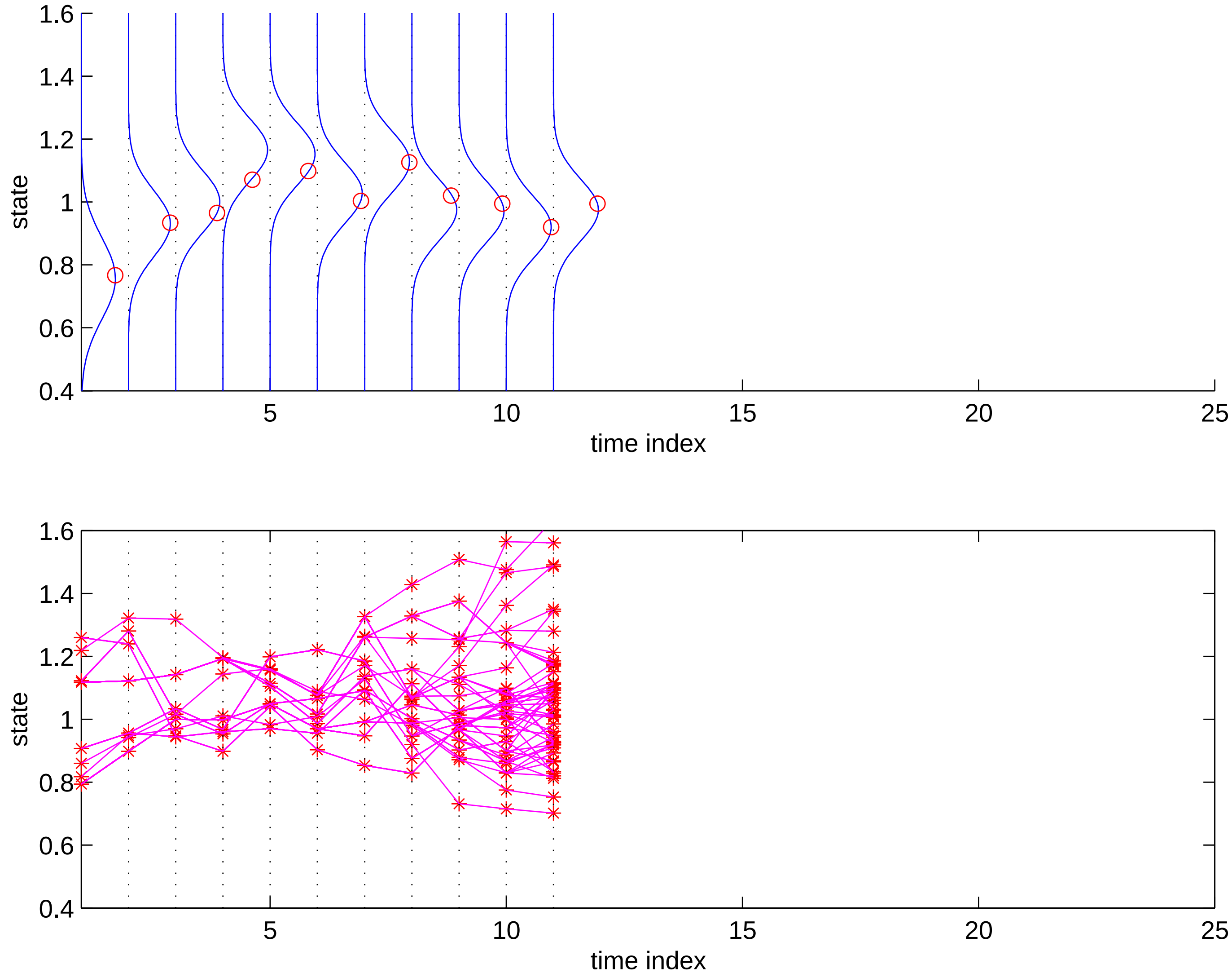




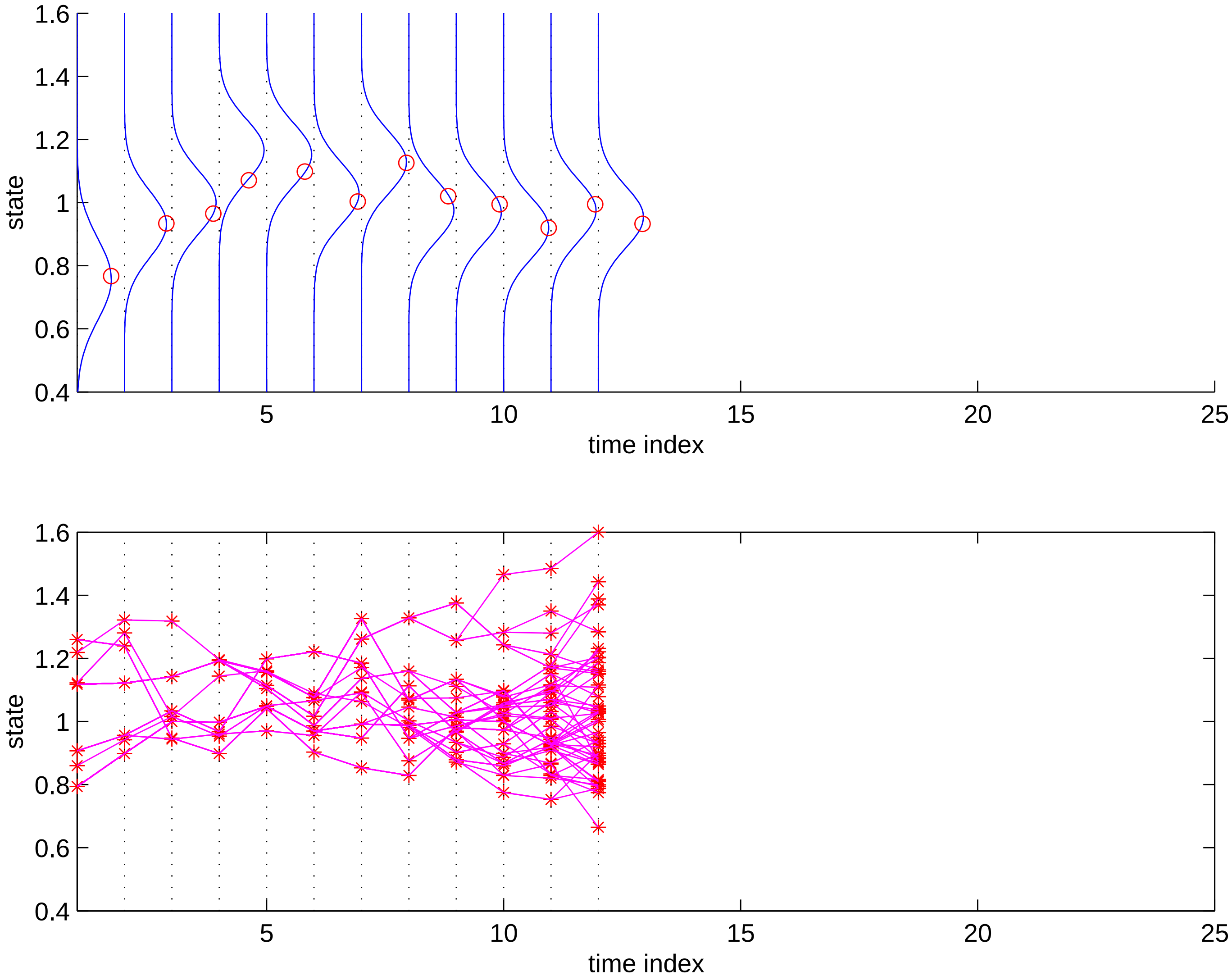
Filtering densities and evolution of the ancestor tree



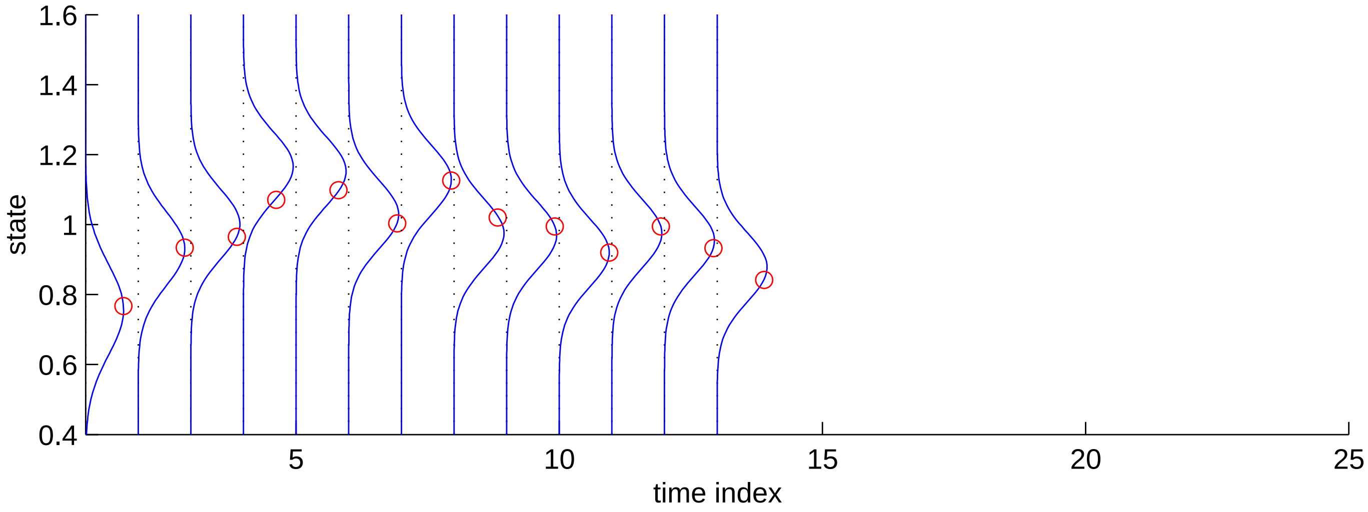
Filtering densities and evolution of the ancestor tree

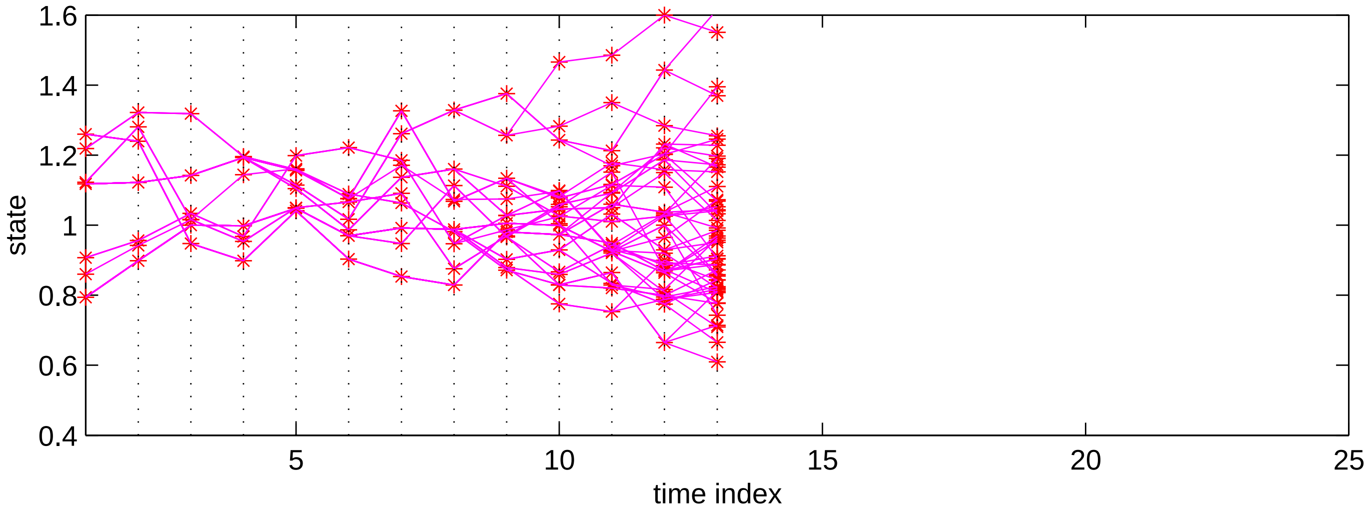


Filtering densities and evolution of the ancestor tree

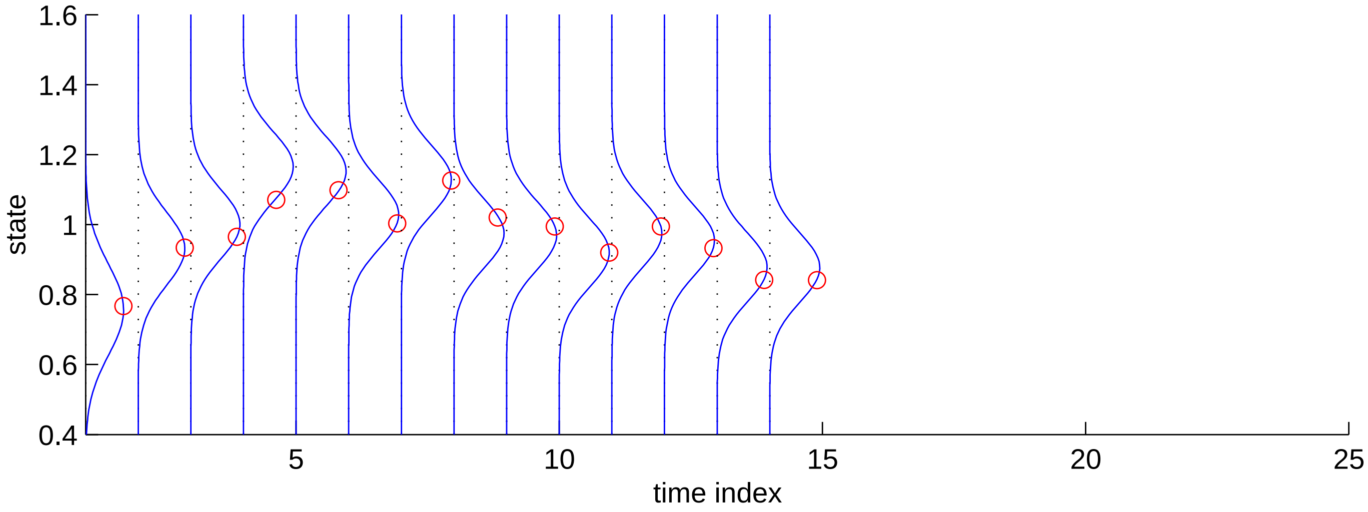


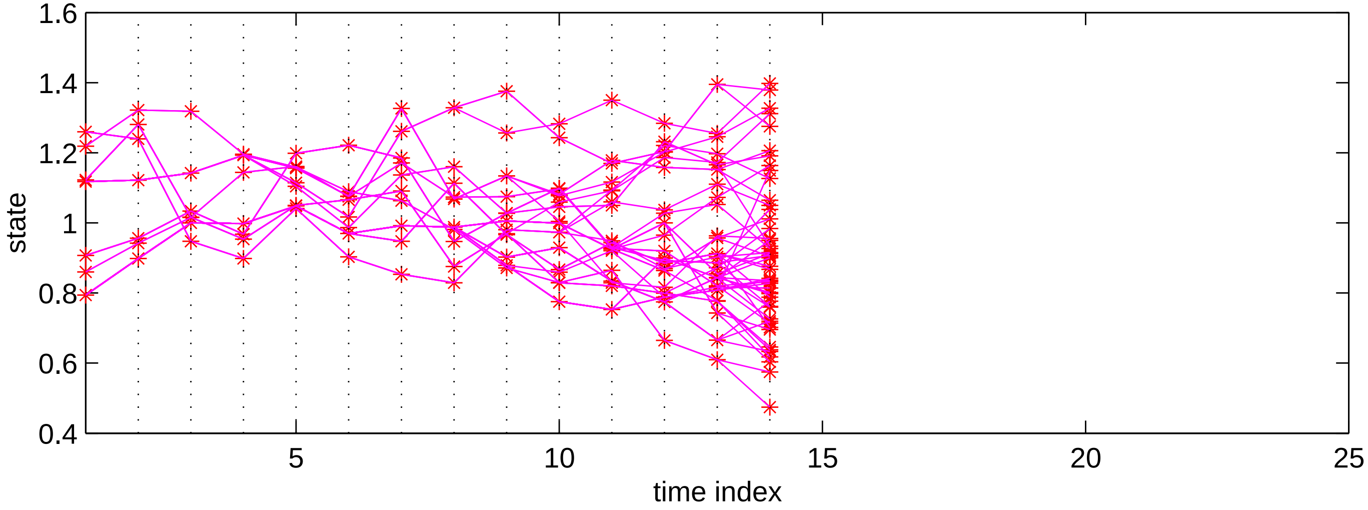
Filtering densities and evolution of the ancestor tree



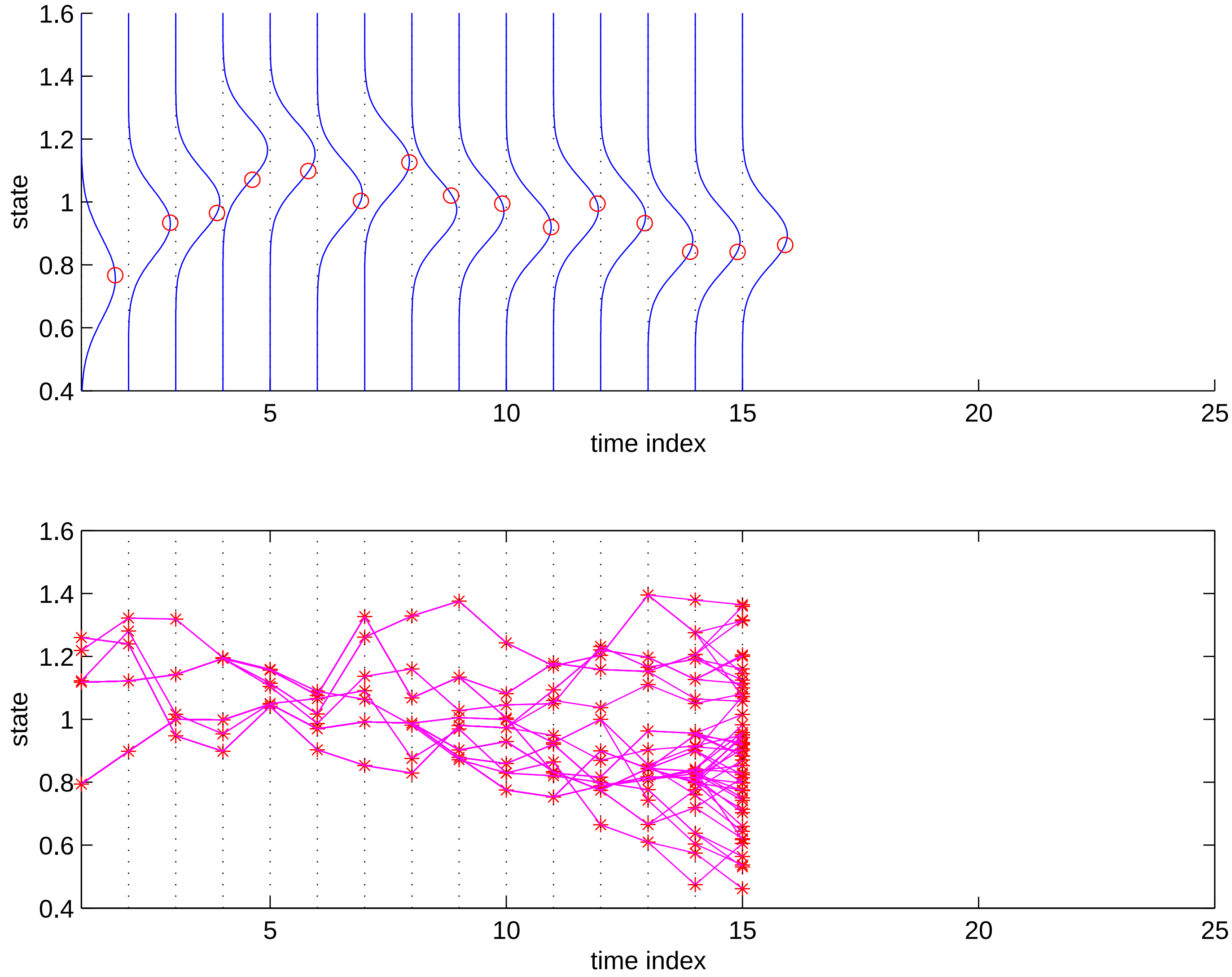


Filtering densities and evolution of the ancestor tree

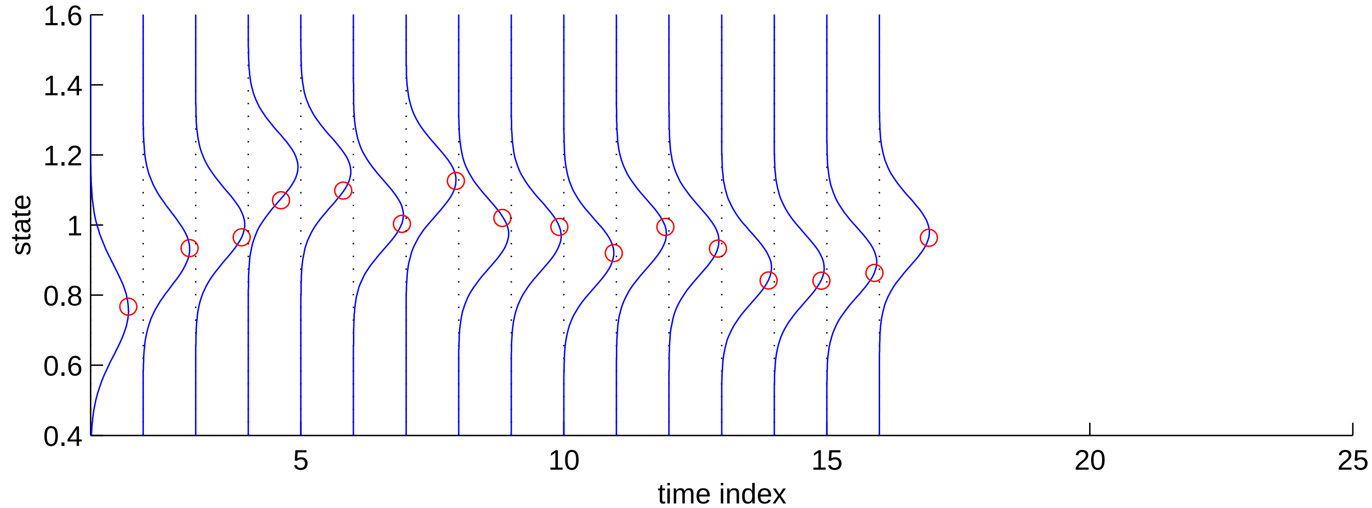


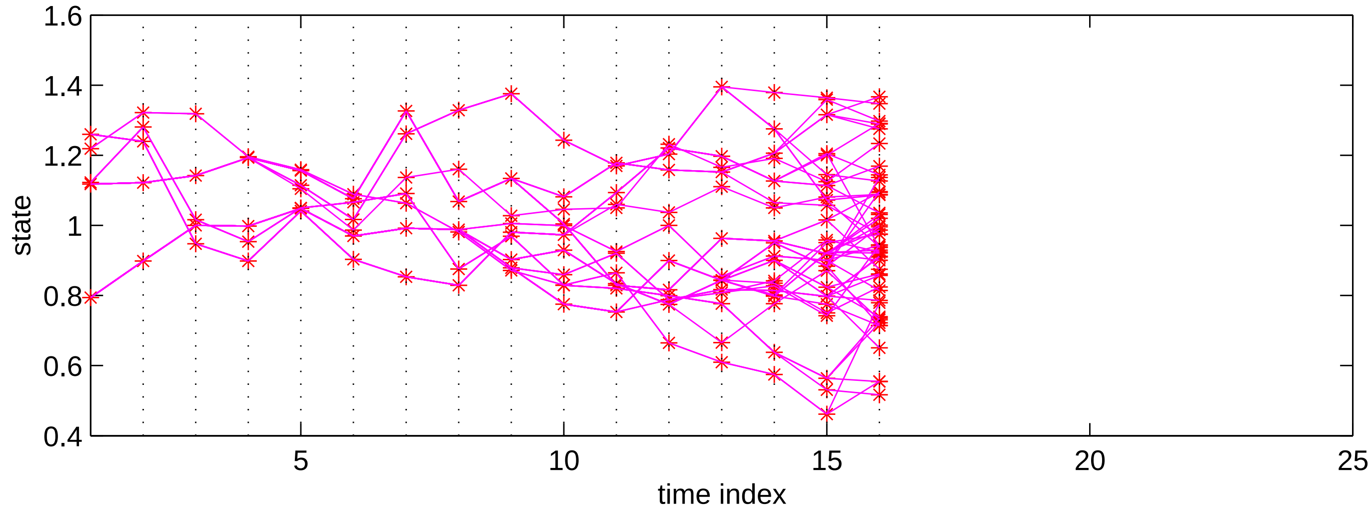


Filtering densities and evolution of the ancestor tree

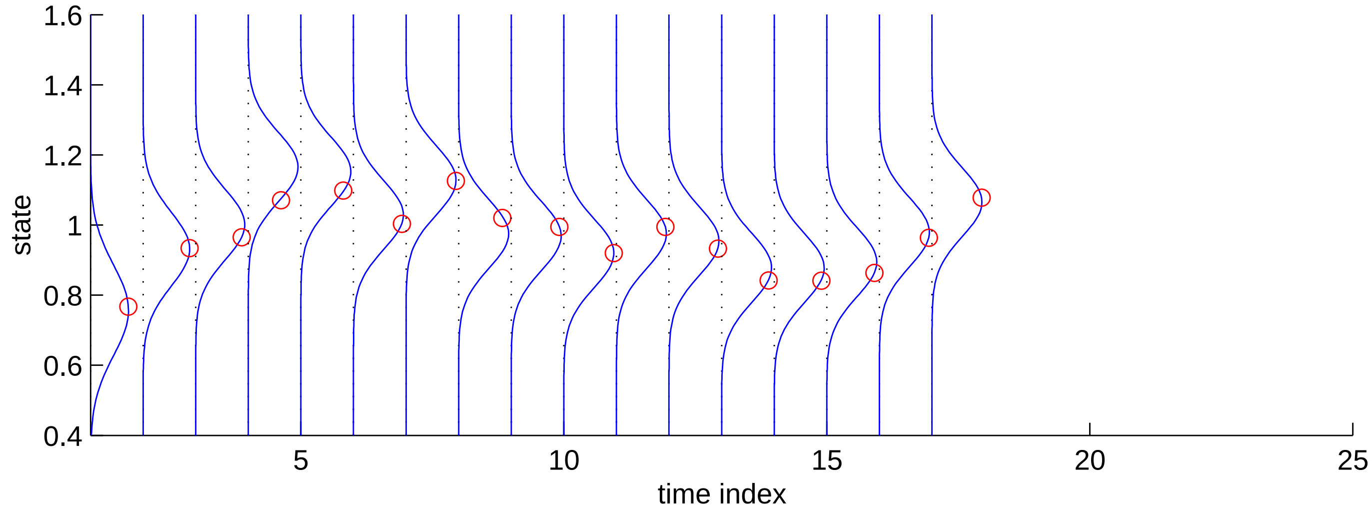


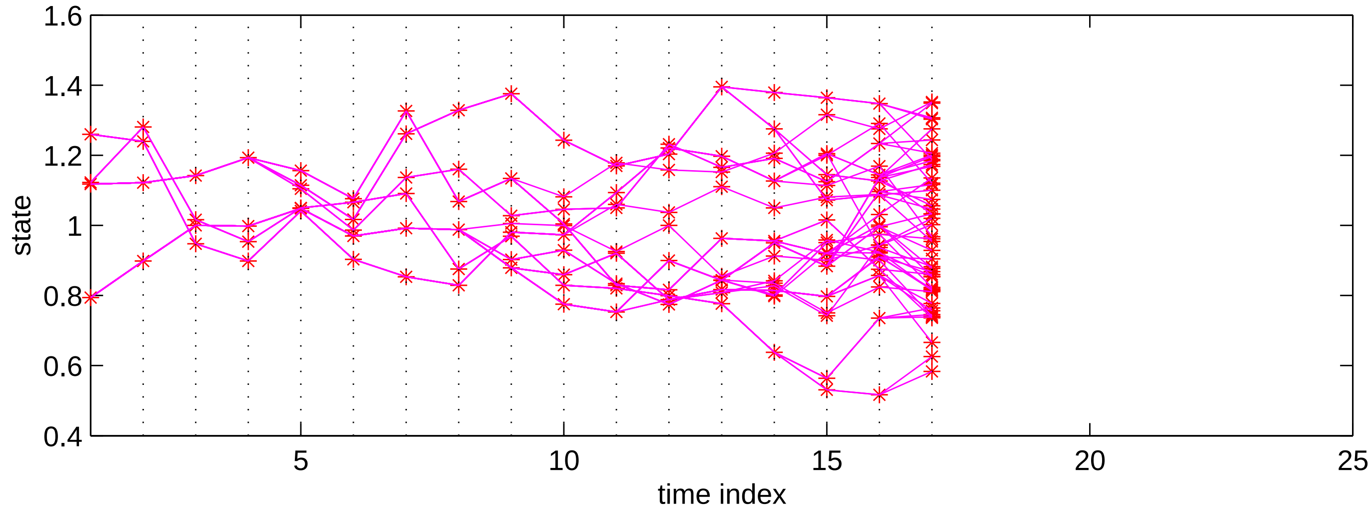
Filtering densities and evolution of the ancestor tree



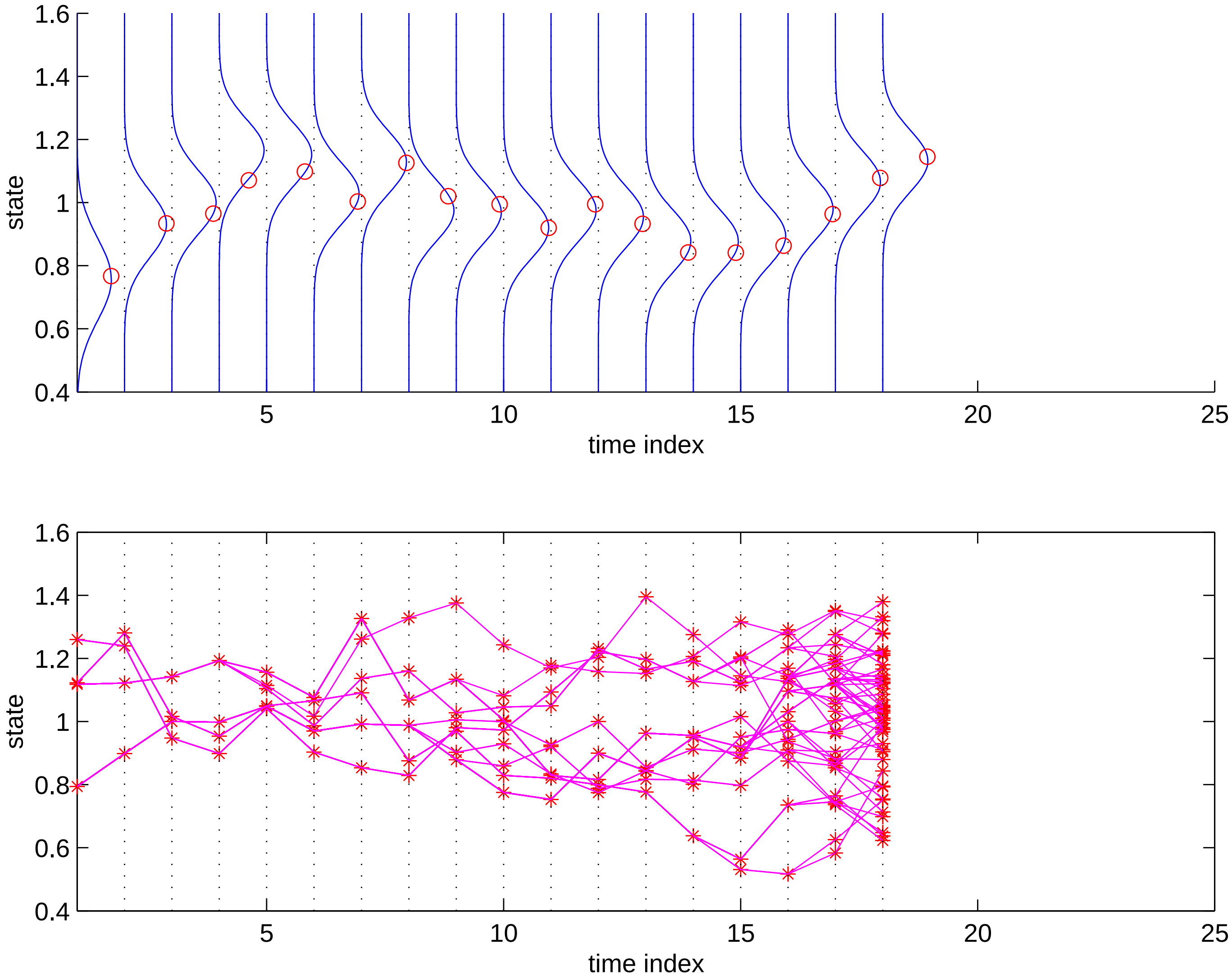


Filtering densities and evolution of the ancestor tree

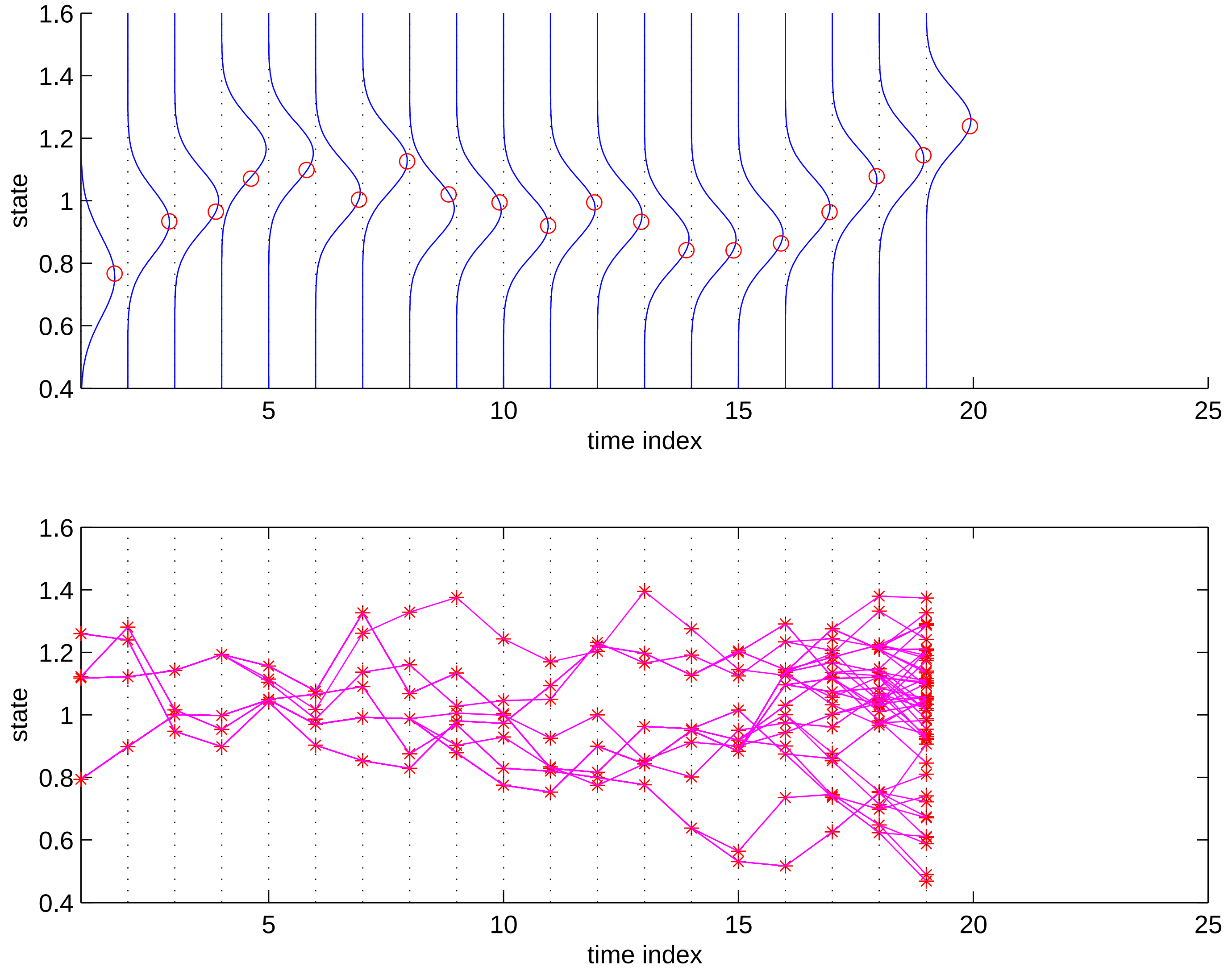




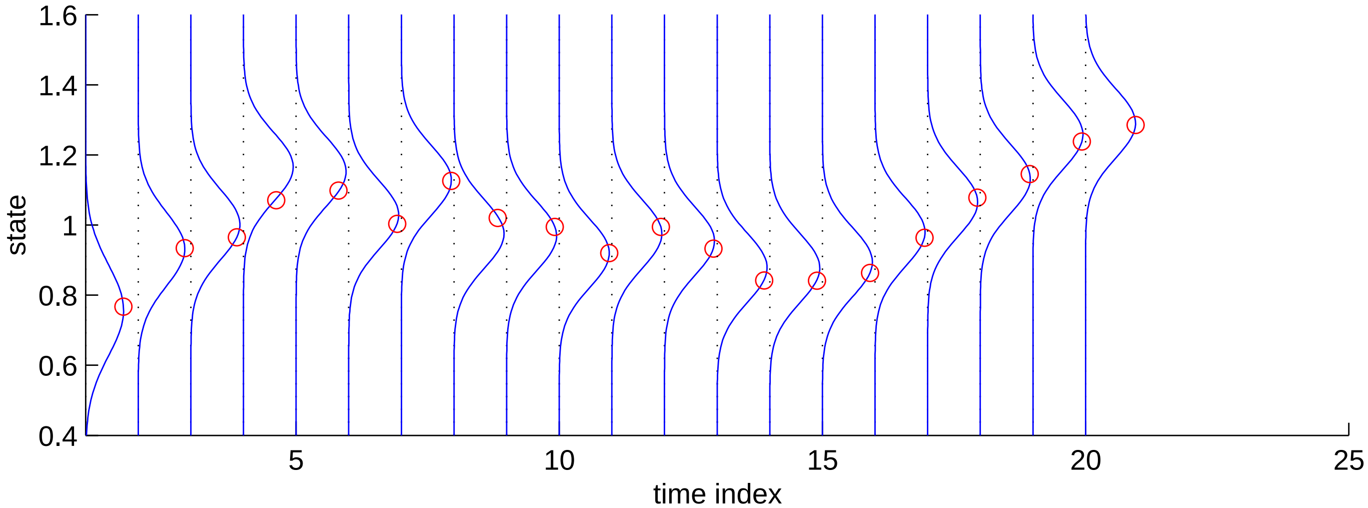
Filtering densities and evolution of the ancestor tree

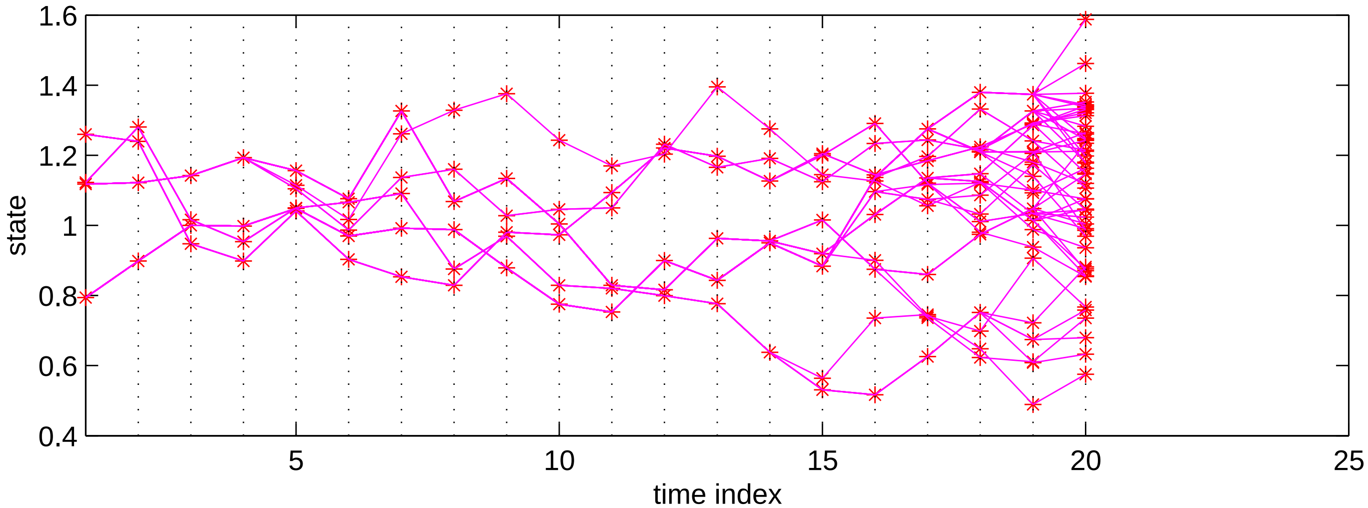


Filtering densities and evolution of the ancestor tree

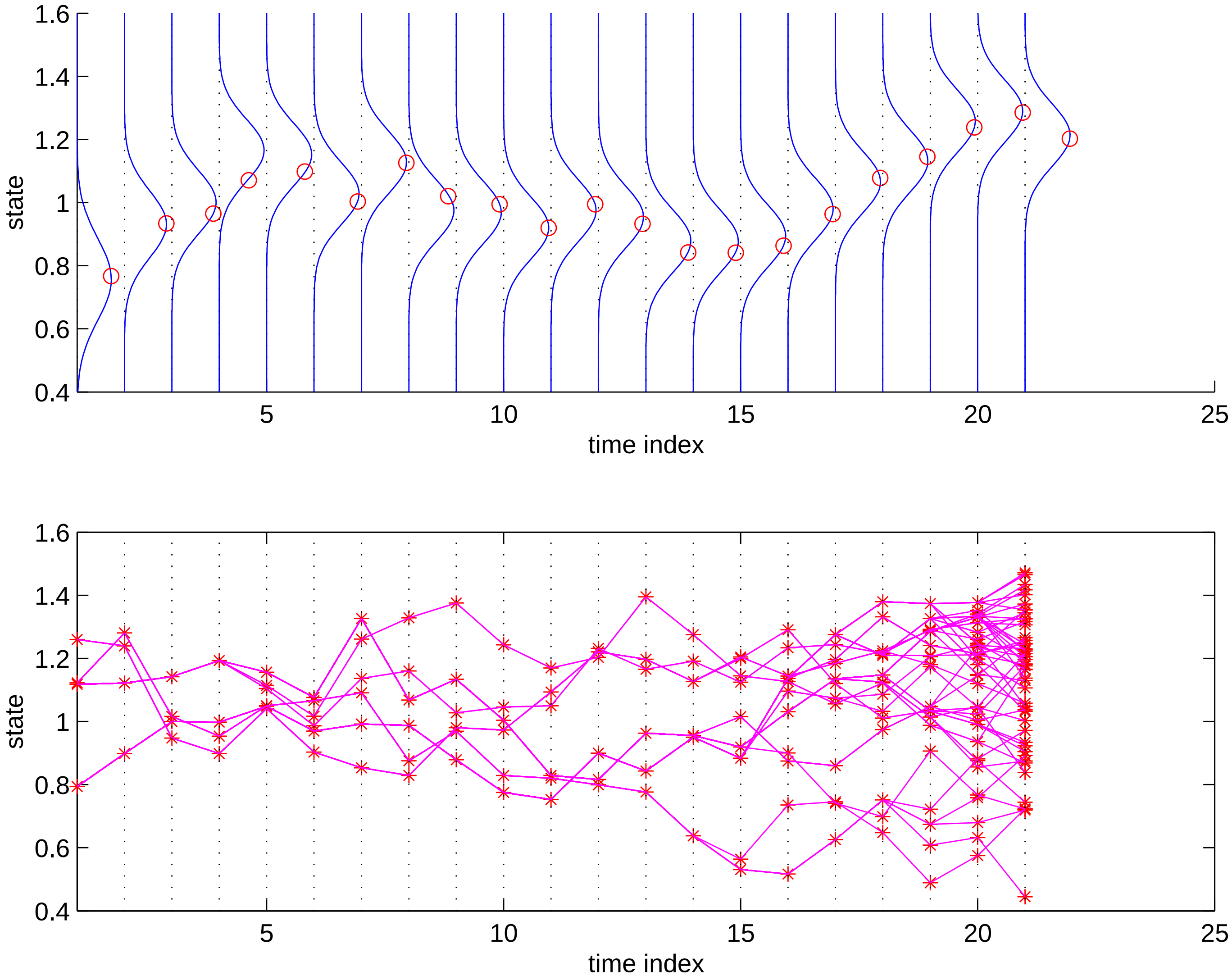


Filtering densities and evolution of the ancestor tree

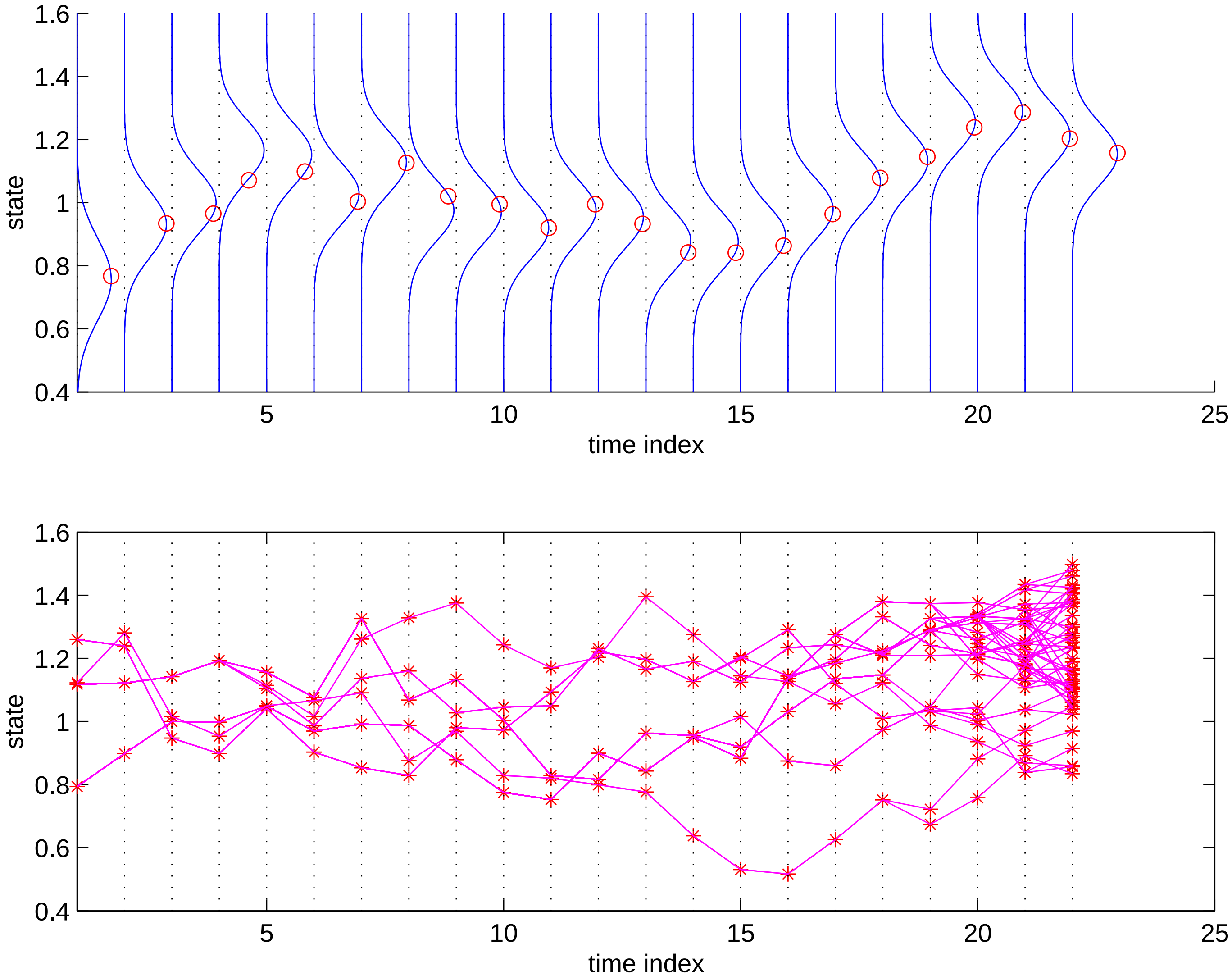




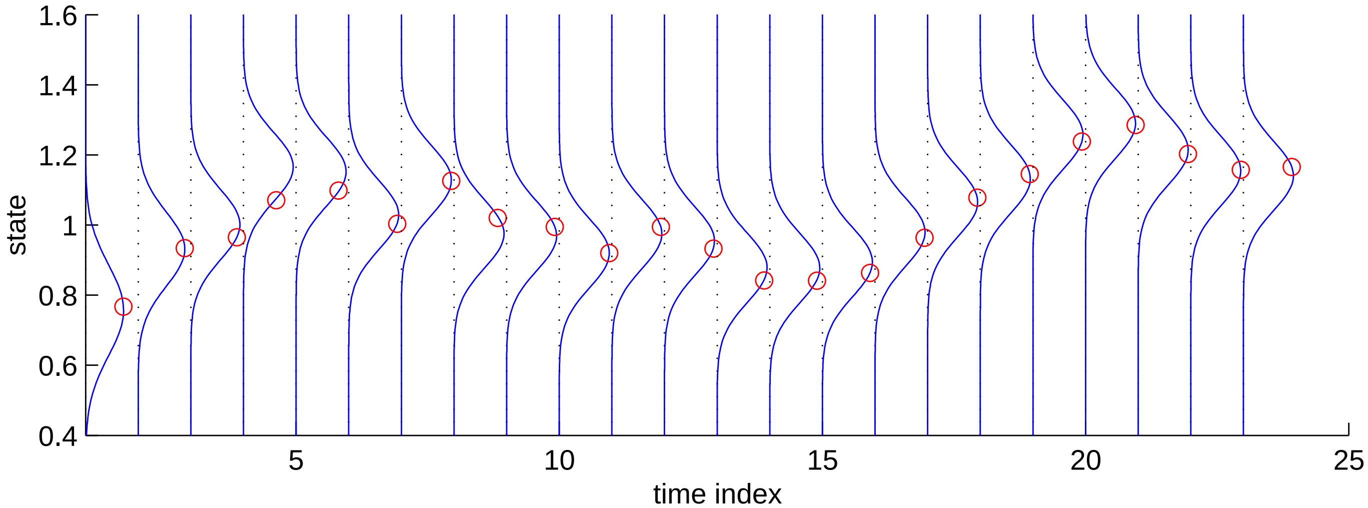
Filtering densities and evolution of the ancestor tree

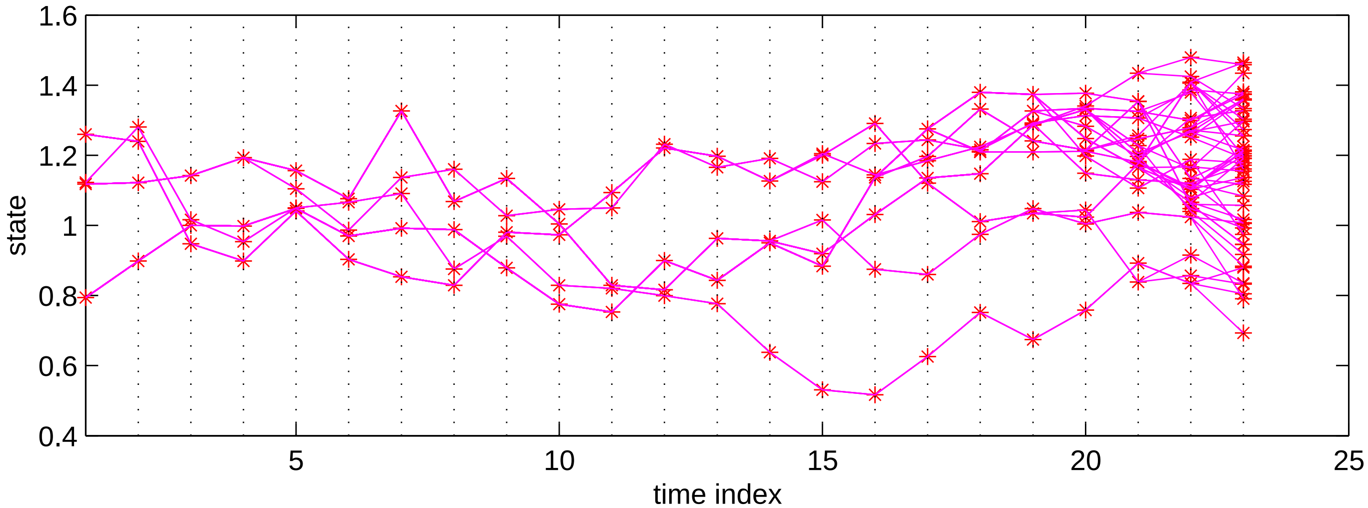


Filtering densities and evolution of the ancestor tree

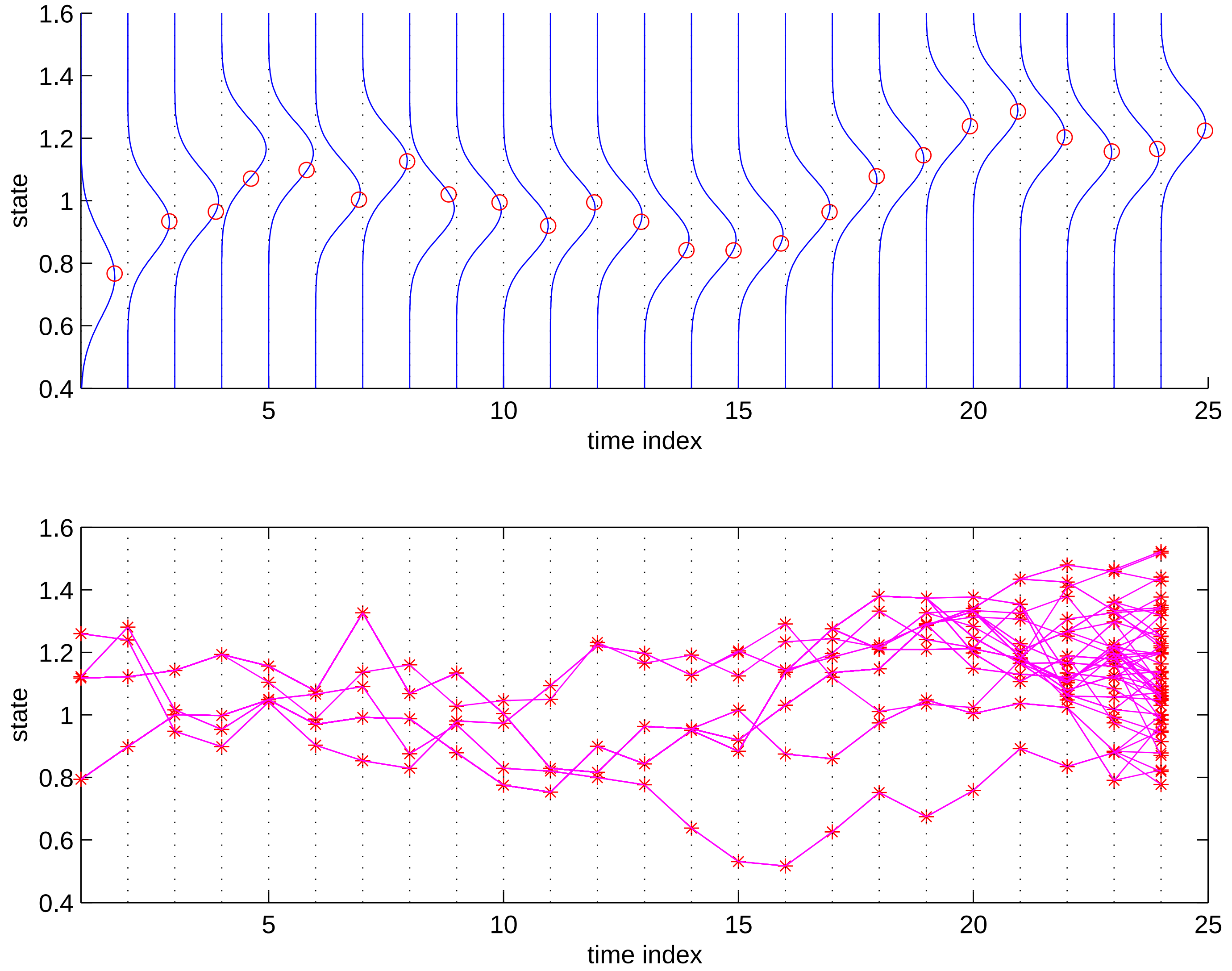


Filtering densities and evolution of the ancestor tree





Filtering densities and evolution of the ancestor tree



Filtering densities and evolution of the ancestor tree

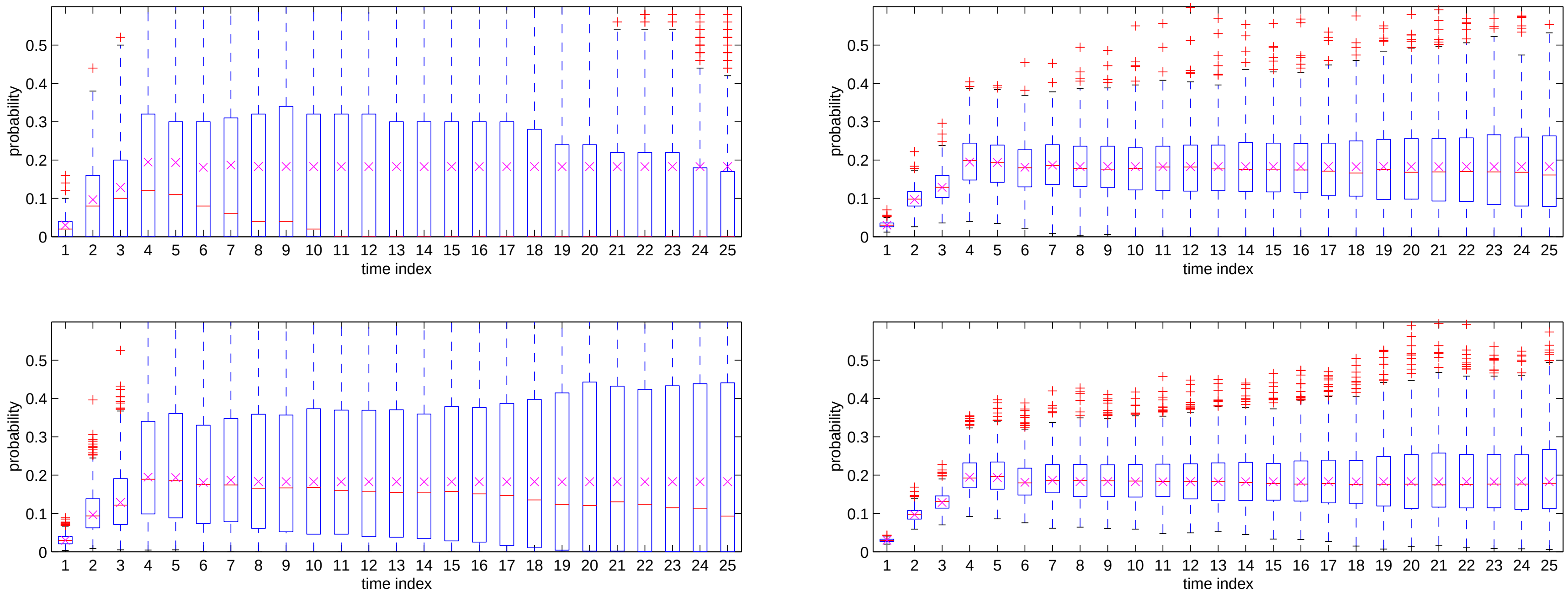


Figure 2: Box and whiskers plots (500 draws) with, left, $p = 50$ and, right, $p = 500$ particles (top: original smoother, bottom: modified one)

Conclusions

Suggestions made during the workshop:

- $A_n = \mathrm{E}\left[\sum_{k=0}^{n} f_k(X_k) \middle| Y_{0:n}\right]$ is the most general form since the state of the system may be chosen as (X_{n-1}, X_n, Y_n)

- The proposed particle estimator is equivalent to

$$\sum_{i=1}^{p} \omega_i \left(\sum_{k=0}^{n} f_k(\xi^i_{n|n-1,k}) \right)$$

  where $\xi^i_{n|n-1,0:n}$ is the path associated with the ith particle at time n