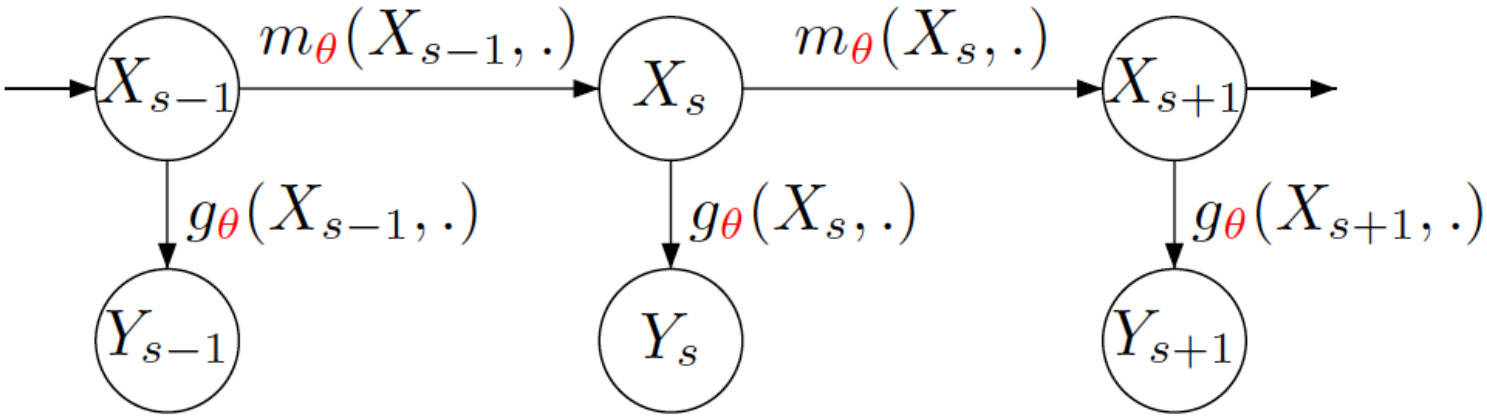
Online Parameter estimation in HMM

Sylvain Le Corff, Gersende Fort, Eric Moulines

Télécom ParisTech

April 19, 2012

Outline



- $\mathbf{Y} \overset{\text{def}}{=} \{\mathbf{Y}_t\}_{t\in\mathbb{Z}}$ is the observation process whereas $\mathbf{X} \overset{\text{def}}{=} \{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ are the hidden states.
- The distribution of the HMM is specified by
  1. A family of transition kernels $\{m_\theta\}_{\theta\in\Theta}$ on $\mathbb{X}\times\mathcal{B}(\mathbb{X})$ governing the transition of the hidden chain.
  2. A family of transition kernels $\{g_\theta\}_{\theta\in\Theta}$ on $\mathbb{X}\times\mathcal{B}(\mathbb{Y})$, the conditional likelihood of the observations.

Online estimation in HMM

1. Objective: estimate the parameter θ using maximum likelihood estimator (or a quasi-MLE if the distribution is misspecified !).

2. Requirements
   - No storage of the observations, no growing capacity memory.
   - Constant complexity per incoming observation.
   - The parameter are updated as each new observation.

3. Applications
   - Inference of partially observed Markov chains for very large data sets (e.g. proteomics, volatility from high-frequency data, etc...)
   - Online localization and mapping in robotics.

Outline

MLE in Missing data: the IID case

- (Curved) exponential family model.

$$p_\theta(x_t, y_t) = \exp\left(\langle s(x_t, y_t), \psi(\theta)\rangle - A(\theta)\right)$$

with respect to some σ-finite dominating measure.

- Explicit complete-data MLE.

$$S \mapsto \bar{\theta}(S) = \arg\max_\theta \langle S, \psi(\theta)\rangle - A(\theta)$$

is available in closed-form.

- IID Data.
(Y_t) is an iid. process with marginal π.
(not necessarily equal to $f_{\theta_\star}$.)

The (Usual) Expectation-Maximization Algorithm

k-th EM Iteration (with T observations).

E-Step

$$S_{T,k} = \frac{1}{n} \sum_{t=1}^{T} \mathbb{E}_{\theta_{k-1}} \left[s(X_t, Y_t) | Y_t \right] .$$

M-Step

$$\theta_k = \bar{\theta} \left(S_{T,k} \right) .$$

Can be fully reparameterized in the domain of sufficient statistics

$$S_{T,k} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\bar{\theta}(S_{T,k-1})} \left[s(X_t, Y_t) | Y_t \right] \stackrel{\text{def}}{=} \Phi_T(S_{T,k-1}) .$$

The Limiting EM Recursion

As T goes to infinity, the sequence of EM mappings (Φ_T) converges to a limit,

Sufficient Statistics Update

$$S_k = \mathbb{E}_\pi \left[\mathbb{E}_{\bar{\theta}(S_{k-1})} \left[s(X_0, Y_0) | Y_0 \right] \right] = \Phi_\infty(S_{k-1}) \, .$$

Parameter Update

$$\theta_k = \bar{\theta}(S_k) \, .$$

Some results

1. The Kullback-Leibler divergence between the marginal distribution f_{θ_k} and π, $\mathrm{KL}(\pi || f_{\theta_k})$ is monotonically decreasing with k.
2. Converge to $\{\theta : \nabla_\theta D(\pi | f_\theta) = 0\}$.

Outline

Online EM: Rationale

- Objective: find the roots of

$$\mathbb{E}_{\pi}\left[\mathbb{E}_{\bar{\theta}(S)}\left[s(X_0, Y_0)|\, Y_0\right]\right] - S = 0\,.$$

- Stochastic Approximation (or Robbins-Monro) setup.
- $\mathbb{E}_{\bar{\theta}(S)}\left[s(X_n, Y_n)|\, Y_n\right]$ is seen as a noisy observation of $\mathbb{E}_{\pi}\left[\mathbb{E}_{\bar{\theta}(S)}\left[s(X_0, Y_0)|\, Y_0\right]\right]$.

$$S_n = S_{n-1} + \gamma_n \left(\mathbb{E}_{\bar{\theta}(\boldsymbol{S_{n-1}})}\left[s(X_n, Y_n)|\, Y_n\right] - S_{n-1}\right)\,,$$

where (γ_n) is a sequence of decreasing positive stepsizes.

Online EM Algorithm

- Stochastic E-Step

$$S_n = (1 - \gamma_n) S_{n-1} + \gamma_n \mathbb{E}_{\theta_{n-1}} \left[s(X_n, Y_n) | Y_n \right] .$$

- M Step

$$\theta_n = \bar{\theta}(S_n) .$$

Practical Recommendations

- $\gamma_n = c/n^\alpha$ with $\alpha \in [0.7, 0.9]$.
- Don't do M for the first 10–20 obs.
- (optional) Use Polyak-Ruppert averaging.

Outline

The EM Algorithm for HMMs

1. The EM update with T observations is now

$$S_{T,k} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\bar{\theta}(S_{T,k-1})} \left[s(X_{t-1}, X_t, Y_t) | Y_{1:T} \right] .$$

2. Dependence of the conditional expectation on the future values $Y_{1:T}$.

- Problem 1: how to compute additive functional recursively in time ?
- Problem 2: how to adapt the parameters within such framework ?

The limiting EM for HMMs

- An iteration of the EM algorithm writes

$$S_{T,k} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\bar{\theta}(s_{T,k-1})} \left[s(X_{t-1}, X_t, Y_t) | Y_{1:T} \right] .$$

- Assuming that,
  - (Y_t) is an ergodic process with distribution π,
  - some form of forgetting properties for the HMM model,

  the limiting EM recursion becomes (as $T \to \infty$)

$$S_k = \mathbb{E}_{\pi} \left[\mathbb{E}_{\bar{\theta}(s_{k-1})} \left[s(X_{-1}, X_0, Y_0) | Y_{-\infty:\infty} \right] \right] .$$

- Idea: develop a sequential algorithm allowing to approximate the limiting EM !

- The M-step is performed on blocks of observations $Y_{T_k:T_{k+1}}$, for an appropriately chosen sequence of time instants $\{T_k, k \geq 1\}$.
- The parameters are kept constant while accumulating the information brought by the observations $Y_{T_k:T_{k+1}}$.

Algorithm

1. Block n
   - From $T_{n-1}+1$ to T_n, compute recursively

$$\bar{S}_{\tau_n}(\theta_{n-1}, \mathbf{Y}) = \frac{1}{\tau_{n-1}} \mathbb{E}_{\theta_n} \left[\sum_{t=1}^{\tau_n} S(X_{t-1}, X_t, \mathbf{Y}_{t+T_{n-1}}) \middle| \mathbf{Y}_{T_{n-1}+1:T_{n-1}+\tau_n} \right] .$$

2. Parameter update:
   - $\theta_n \stackrel{\text{def}}{=} \bar{\theta}[\bar{S}_{\tau_n}(\theta_{n-1}, \mathbf{Y})]$.

- The M-step is performed on blocks of observations $Y_{T_k:T_{k+1}}$, for an appropriately chosen sequence of time instants $\{T_k, k \geq 1\}$.
- The parameters are kept constant while accumulating the information brought by the observations $Y_{T_k:T_{k+1}}$.

Algorithm

1. Block n
   - From $T_{n-1}+1$ to T_n, compute recursively

$$\bar{S}_{\tau_n}(\theta_{n-1}, \mathbf{Y}) = \frac{1}{\tau_{n-1}} \mathbb{E}_{\theta_n} \left[\sum_{t=1}^{\tau_n} S(X_{t-1}, X_t, \mathbf{Y}_{t+T_{n-1}}) \middle| \mathbf{Y}_{T_{n-1}+1:T_{n-1}+\tau_n} \right] .$$

   - Compute

$$\Sigma_n \stackrel{\text{def}}{=} \left(1 - \frac{\tau_n}{T_n}\right) \Sigma_{n-1} + \frac{\tau_n}{T_n} \bar{S}_{\tau_n}(\theta_{n-1}, \mathbf{Y}) .$$

2. Parameter update:
   - $\theta_n \stackrel{\text{def}}{=} \bar{\theta}[\bar{S}_{\tau_n}(\theta_{n-1}, \mathbf{Y})]$.
   - $\widetilde{\theta}_n \stackrel{\text{def}}{=} \bar{\theta}[\Sigma_n]$.

Outline

Online computation of additive functionals

- Consider the following additive functional:

$$\bar{S}_T = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} S(X_{t-1}, X_t, \mathbf{Y}_t)\middle|\mathbf{Y}_{1:T}\right] .$$

- By the tower property of the conditional expectation,

$$\bar{S}_T = \mathbb{E}\left[\rho_T(X_T)|\mathbf{Y}_{1:T}\right] = \phi_T[\rho_T] .$$

where ϕ_T is the filtering distribution at time T and

$$\rho_T(x_T) \stackrel{\text{def}}{=} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} S(X_{t-1}, X_t, \mathbf{Y}_t)\middle|\mathbf{Y}_{1:T}, x_T\right] .$$

Online computation of additive functionals

Decompose

$$\frac{1}{T}\sum_{t=1}^{T} S(X_{t-1}, X_t, Y_t) = \left(1 - \frac{1}{T}\right) \frac{1}{T-1} \sum_{t=1}^{T-1} S(X_{t-1}, X_t, Y_t) + \frac{1}{T} S(X_{T-1}, X_T, Y_T) \,.$$

Then, use that $X_{0:T}|Y_{0:T}$ is a Markov chain

$$\rho_T(x_T) = \left(1 - \frac{1}{T}\right) B_{T|T-1}[x_T, \rho_{T-1}] + \frac{1}{T} B_{T|T-1}[x_T, S(\cdot, x_T, \mathbf{Y}_T)] \,,$$

Here $B_{T|T-1}$ is the backward Markov transition kernel

$$B_{T|T-1}(x_T, \mathrm{d}x_{T-1}) \stackrel{\text{def}}{=} \frac{\phi_{T-1}(\mathrm{d}x_{T-1}) m(x_{T-1}, x_T)}{\int \phi_{T-1}(\mathrm{d}x_{T-1}) m(x_{T-1}, x_T)} \,.$$

where ϕ_{T-1} is the filtering distribution at time $T-1$.

The computations can be carried out forward in time !

Online computation for additive functional

- This sequential computation can be done only when it is possible to obtain an explicit expression for the filter:
  1. Linear Gaussian models.
  2. HMM with finite state-spaces.

- In the online framework, sequential Monte Carlo methods (aka. particle filter) are appealing:
  1. these methods are easy to implement and to tweak (as long as the dimension of the hidden space is not too large).
  2. these methods are amenable to parallel computations.

Particle approximation of the additive functional

- ϕ_t is approximated by weighted samples $\{(\xi_t^i, \omega_t^i)\}_{i=1}^N$:

$$\phi_t^N[h] = \sum_{i=1}^N \omega_t^i h(\xi_t^i) \; .$$

- The Backward kernel can be approximated at the current particle locations

$$B_{t|t-1}^N(\xi_t^i, \mathrm{d}x_{T-1}) \stackrel{\mathrm{def}}{=} \frac{\phi_{t-1}^N(\mathrm{d}x_{t-1}) m(x_{t-1}, \xi_t^i)}{\int \phi_{t-1}^N(\mathrm{d}x_{t-1}) m(x_{t-1}, \xi_t^i)} \; .$$

- The functions ρ_t can then be computed at all particle locations (the computational cost grows like N^2; algorithm with linear complexity may be derived, but do not proceed entirely forward in time)

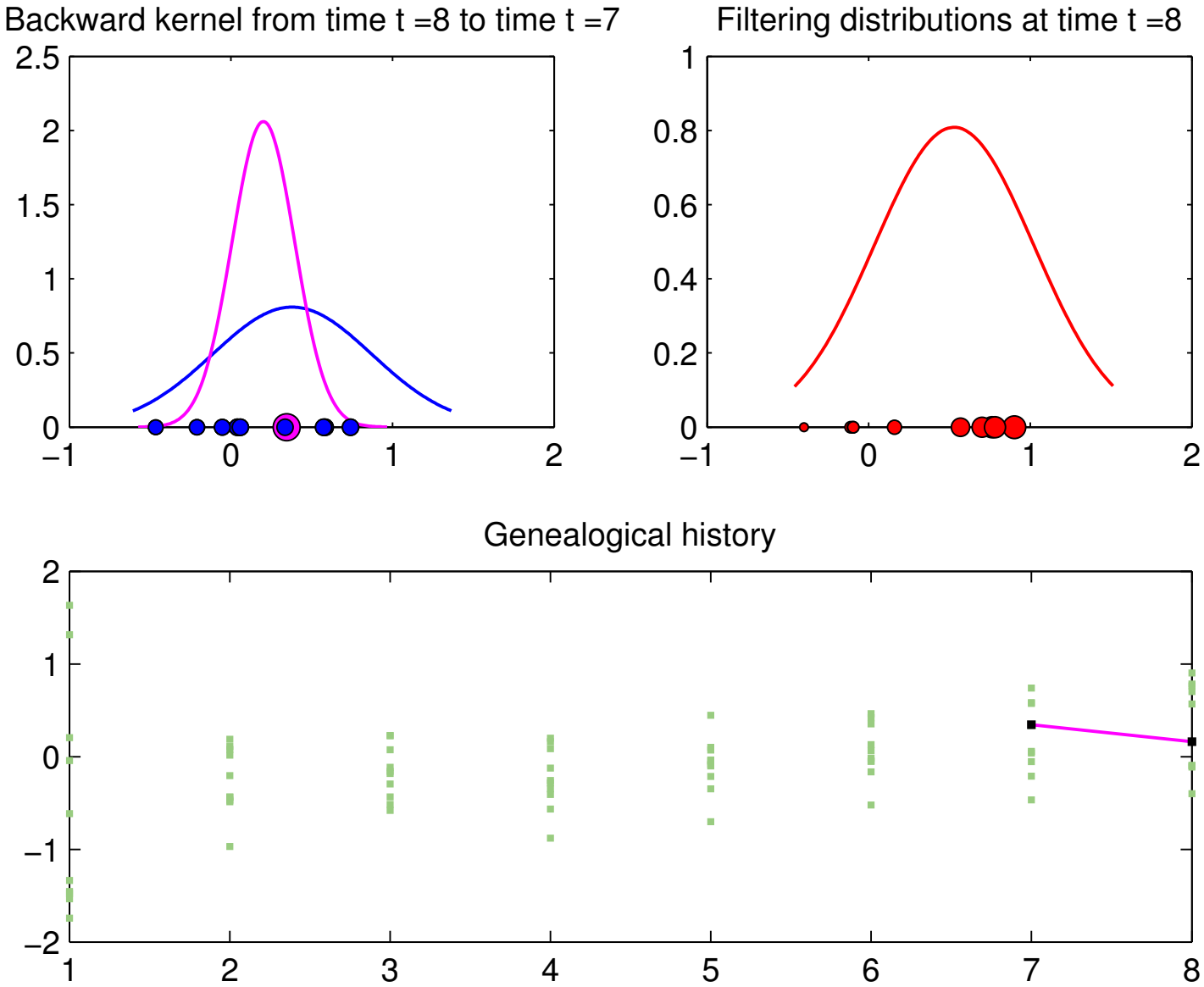
$$\rho_t^N(\xi_t^i) = B_{T|T-1}^N \left[\xi_t^i, \left(1 - \frac{1}{t}\right) \rho_{t-1}^N(\cdot) + \frac{1}{t} S(\cdot, \xi_t^i, \mathbf{Y}_t) \right] \; .$$

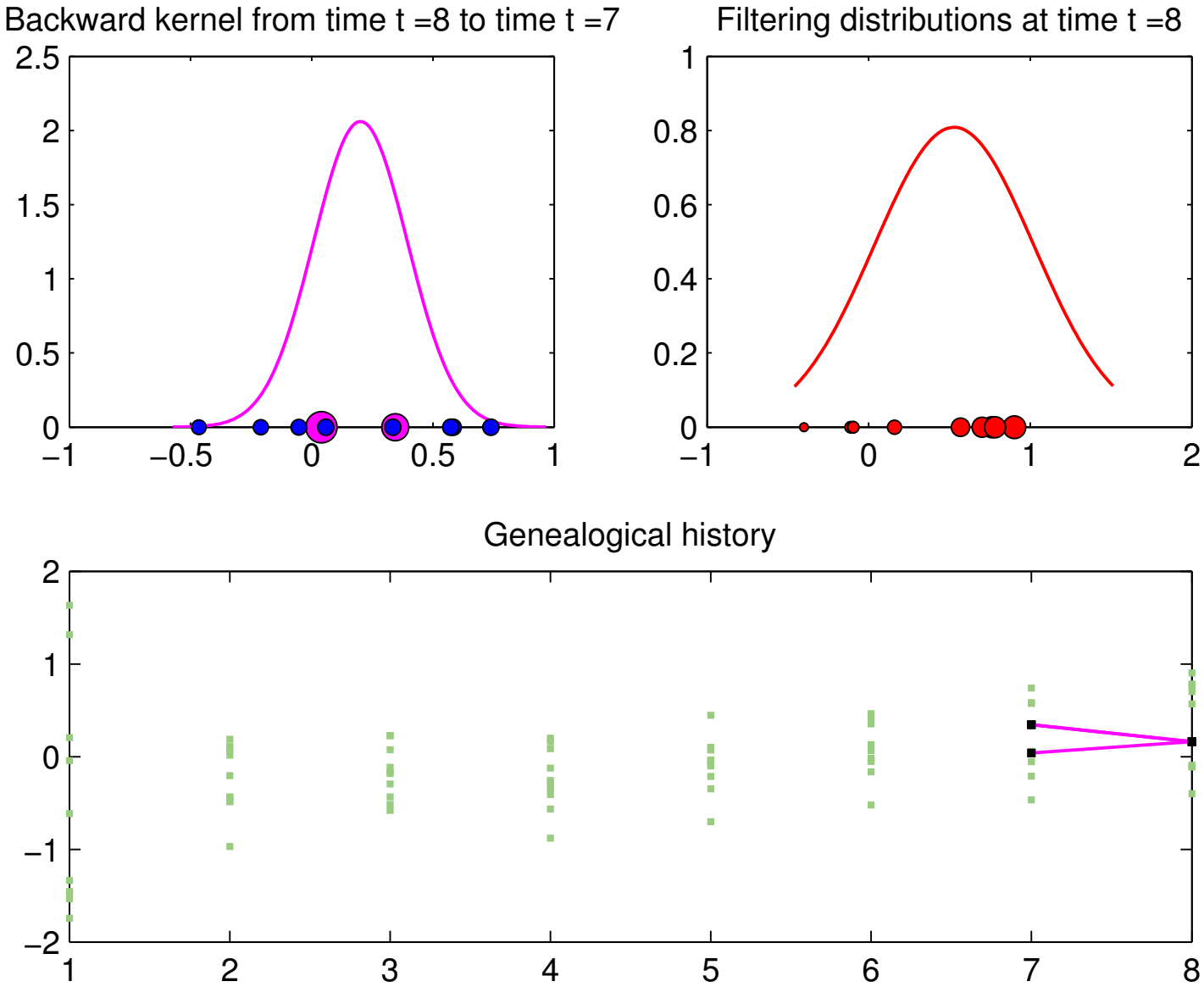
Particle filtering in action

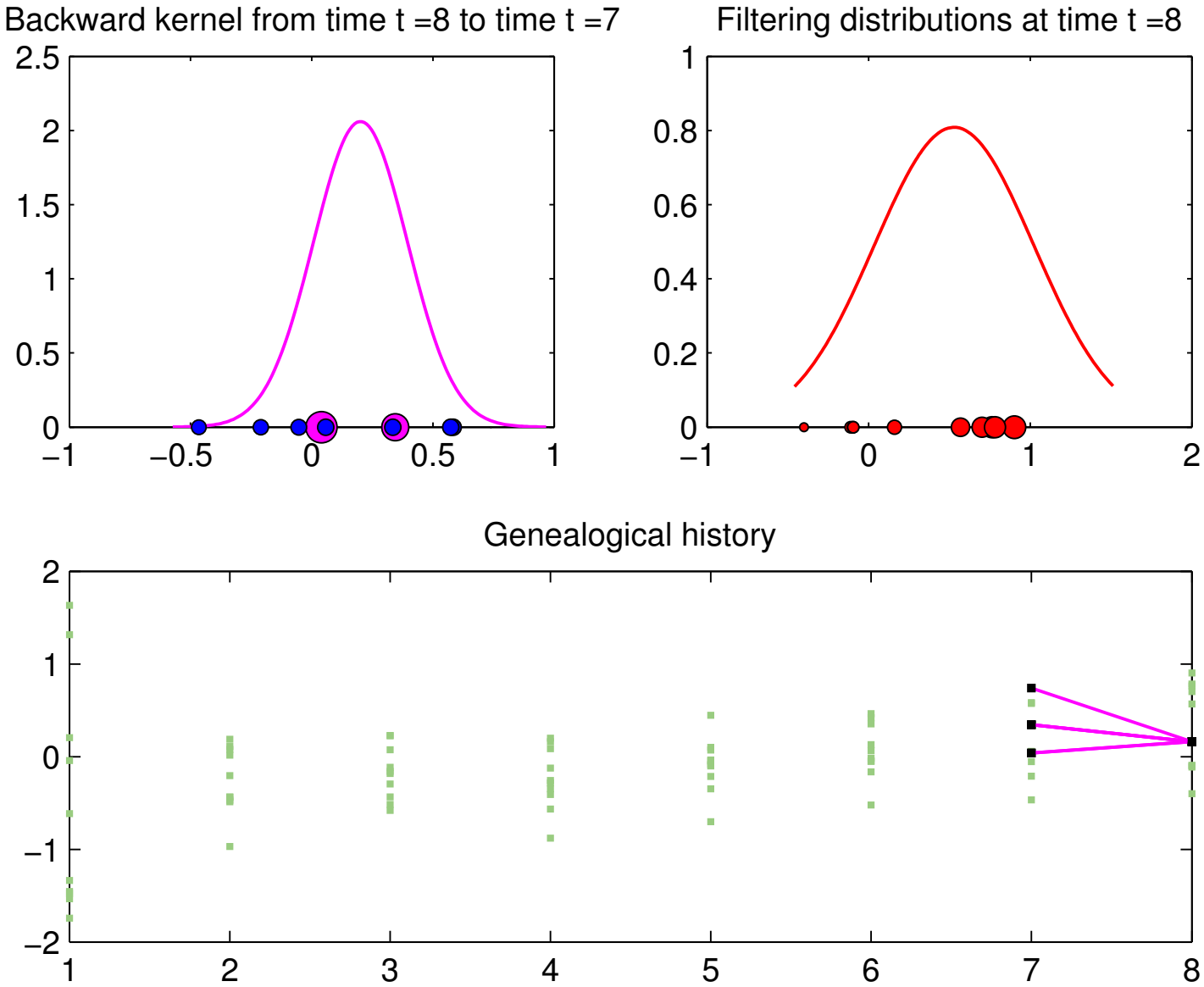
1. Computation of ϕ_t^N with Y_t and ϕ_{t-1}^N.
2. Computation of $\{\rho_t^N(\xi_t^i)\}_{i=1}^N$ with Y_t and ϕ_t^N.

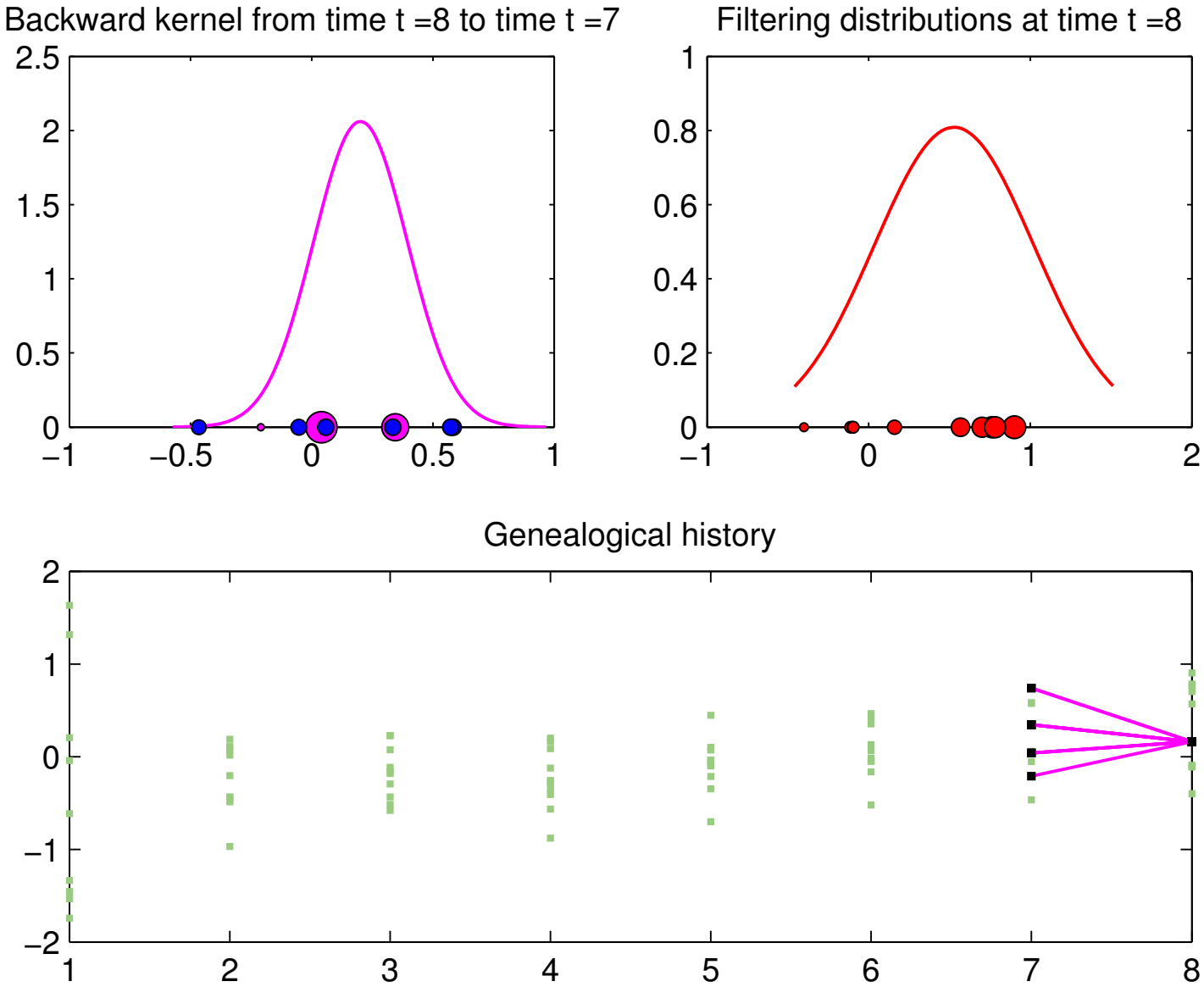
For each particle ξ_t^i, weights $\{\widetilde{\omega}_{t-1}^{i,j} = \omega_{t-1}^j m(\xi_{t-1}^j, \xi_t^i)\}_{j=1}^N$ are computed to match the target kernel.

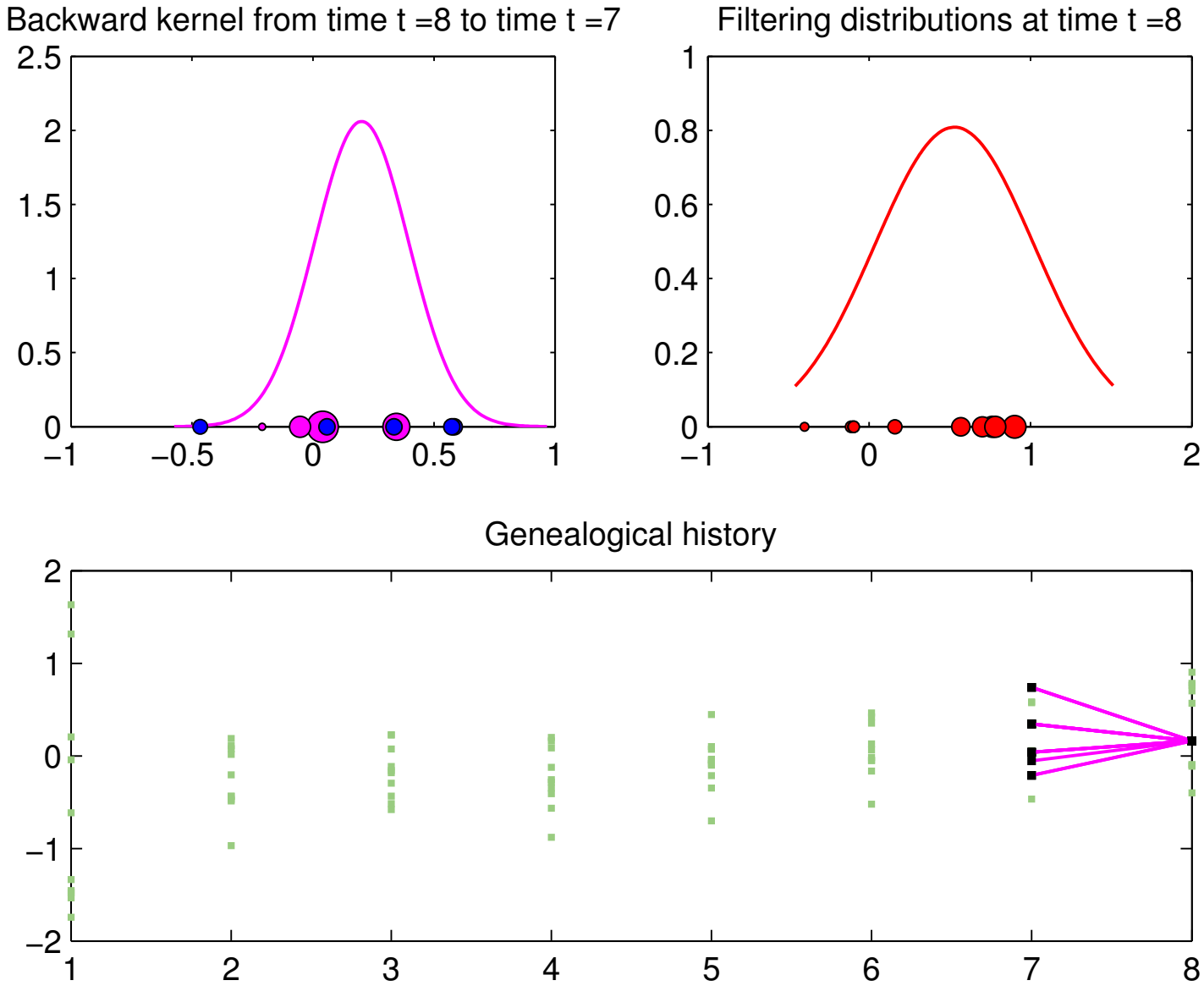
$$B_{t|t-1}^N(\xi_t^i, \mathrm{d}x_{T-1}) = \frac{\sum_{j=1}^N \omega_{t-1}^j m(\xi_{t-1}^j, \xi_t^i) \delta_{\xi_{t-1}^j}(\mathrm{d}x_{t-1})}{\sum_{j=1}^N \omega_{t-1}^j m(\xi_{t-1}^j, \xi_t^i)}$$
$$= \sum_{j=1}^N \frac{\widetilde{\omega}_{t-1}^{i,j}}{\sum_{k=1}^N \widetilde{\omega}_{t-1}^{i,k}} \delta_{\xi_{t-1}^j}(\mathrm{d}x_{t-1}) \, .$$

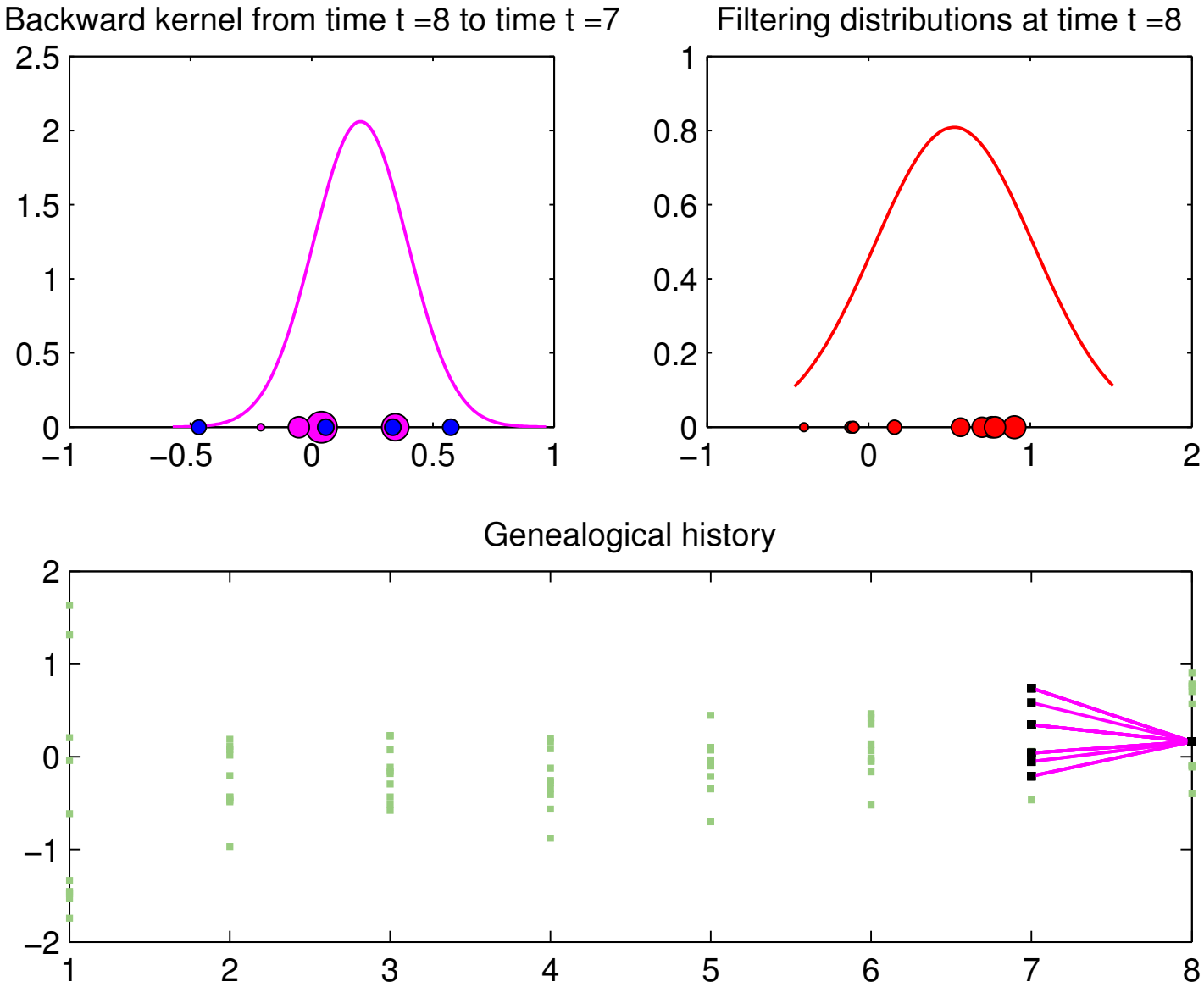

Backward kernel from time t =8 to time t =7
Filtering distributions at time t =8
Genealogical history

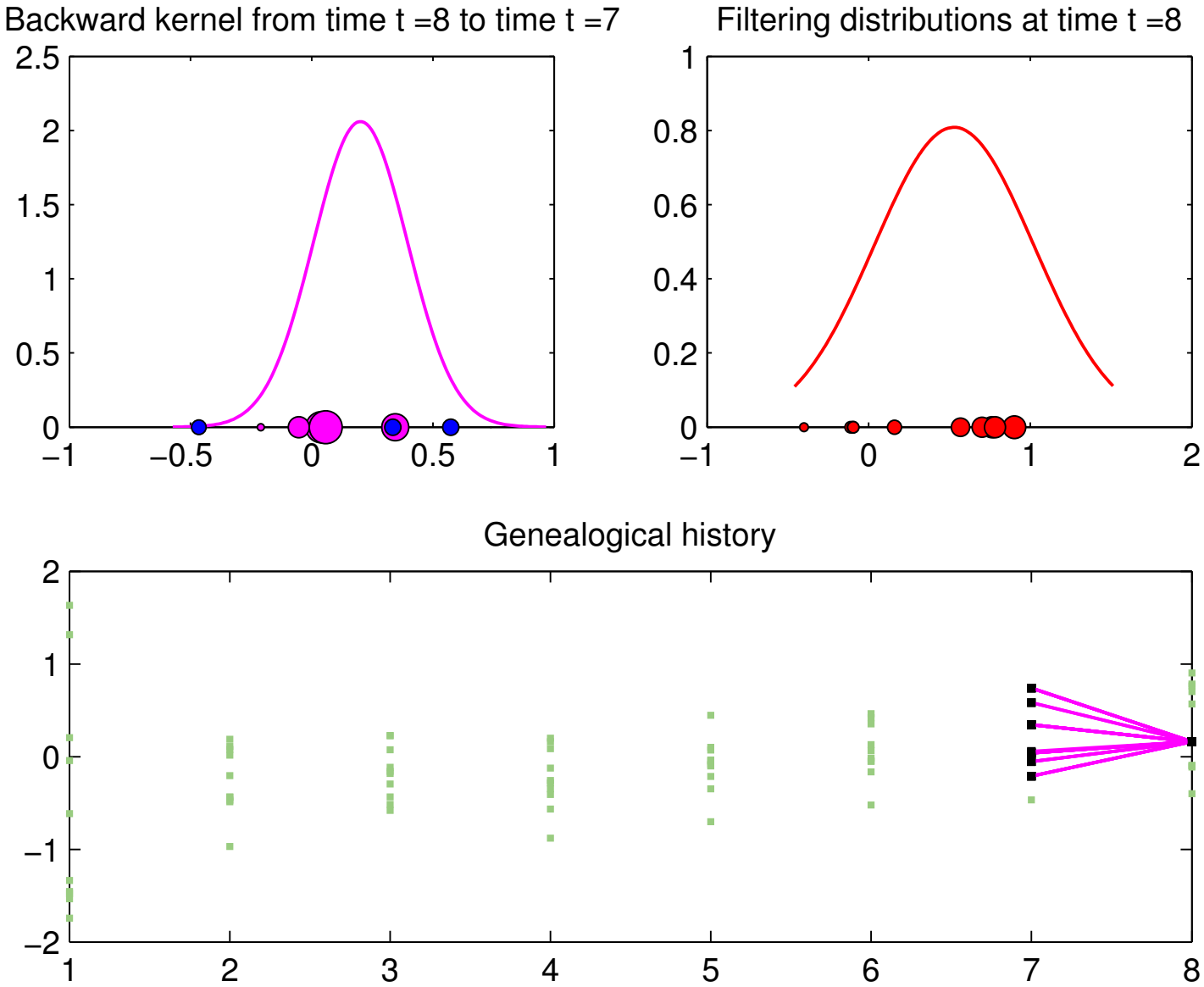

Backward kernel from time t =8 to time t =7
Filtering distributions at time t =8
Genealogical history

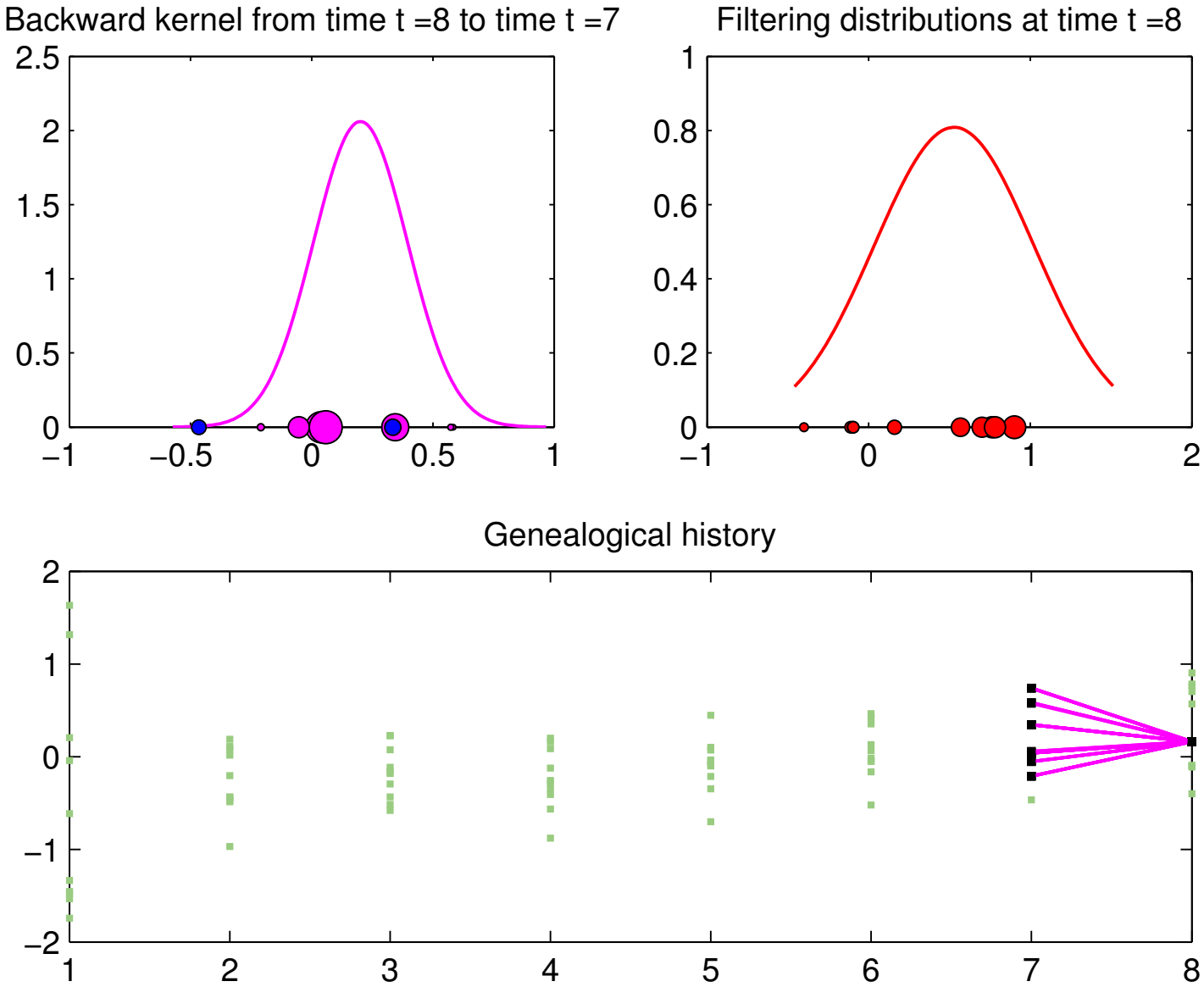

Backward kernel from time t =8 to time t =7
Filtering distributions at time t =8
Genealogical history


Backward kernel from time t =8 to time t =7
Filtering distributions at time t =8
Genealogical history


Backward kernel from time t =8 to time t =7
Filtering distributions at time t =8
Genealogical history


Backward kernel from time t =8 to time t =7
Filtering distributions at time t =8
Genealogical history


Backward kernel from time t =8 to time t =7
Filtering distributions at time t =8
Genealogical history


Backward kernel from time t =8 to time t =7
Filtering distributions at time t =8
Genealogical history

Consider the following stochastic volatility model (SVM):

$$\begin{cases} X_{t+1} = \widetilde{\phi} X_t + \sigma U_t \ , \\ Y_t = \beta \mathrm{e}^{\frac{X_t}{2}} V_t \ , \end{cases}$$

where $X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\widetilde{\phi}^2}\right)$, U_t and V_t are i.i.d. $\mathcal{N}(0,1)$.

- Data sampled using $\phi = 0.8$, $\sigma^2 = 0.2$ and $\beta^2 = 1$.
- Runs started with $\phi = 0.1$, $\sigma^2 = 0.6$ and $\beta^2 = 2$.

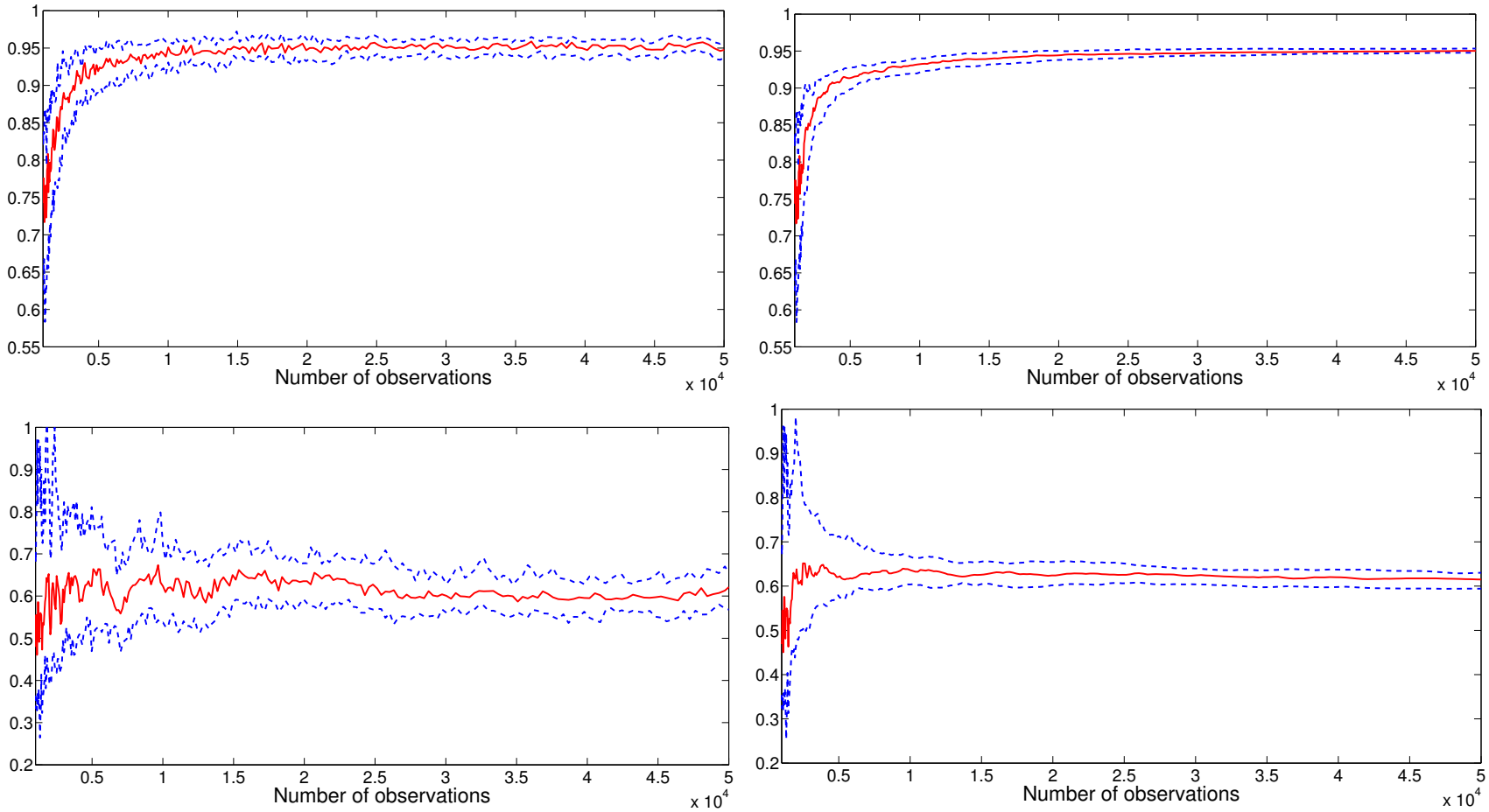


Figure: **Estimation of ϕ, σ^2 and β^2 without (left) and with (right) averaging. Each graph represents the empirical median (bold line) and first and last quartiles (dotted line) over 50 independent Monte Carlo runs. The averaging procedure is started after 1500 observations.**

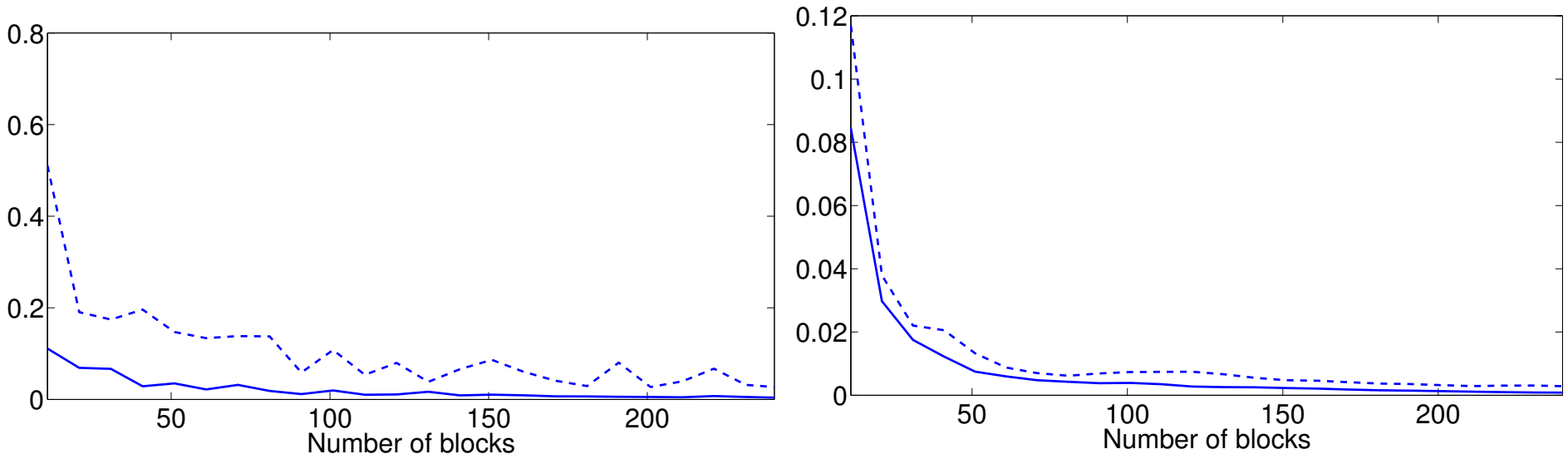


Figure: **Empirical variance of the estimation of β^2 with P-BOEM (top) and its averaged version (bottom) when $N_n = \sqrt{\tau_n}$ (dotted line) and when $N_n = \tau_n$ (bold line).**

Results on online EM procedures.

- Convergence of the limiting EM to the stationary points of the limiting log-likelihood.
- Control of the fluctuation of the Monte Carlo approximation on each block.
- Averaging procedure leads to an optimal rate of convergence in L_p.