Ensemble-based data assimilation method for count data

中野　慎也　　モデリング研究系　教授

Variational data assimilation with count data

We denote the system state at time t_k as $\boldsymbol{x}_k \in \mathrm{R}^d$ and the observation obtained from $\boldsymbol{x}_k$ as $\boldsymbol{y}_k \in \mathrm{R}^n$. The temporal evolution of the state given the numerical simulation model is written as

$$\boldsymbol{x}_k = \boldsymbol{f}_k(\boldsymbol{x}_{k-1}) \tag{1}$$

for $k = 1, \ldots, K$.

We consider cases where each element of an observation vector $\boldsymbol{y}_k$ follows a Poisson distribution and derive an ensemble-based variational method for such cases. The likelihood of $\boldsymbol{x}_k$ is

$$P(\boldsymbol{y}_k|\boldsymbol{x}_k) = \prod_{j=1}^{n} \frac{h_{kj}(\boldsymbol{x}_k)^{y_{kj}}}{y_{kj}!} \exp[-h_{kj}(\boldsymbol{x}_k)], \tag{2}$$

where y_{kj} denotes the j-th element of the vector $\boldsymbol{y}_k$ and the function h_{kj} provides a prediction of y_{kj} from $\boldsymbol{x}_k$.

We assume that all the elements of the state vector $\boldsymbol{x}_k$ are non-negative and that the inequality $h_{kj}(\boldsymbol{x}_k) \geq 0$ is satisfied for each j. To ensure that each element $x_{0\ell}$ $(\ell = 1, \ldots, \mathrm{d})$ is positive, we represent $x_{0,\ell}$ as follows

$$\xi_\ell = \log x_{0\ell}.$$

Combining ξ_ℓ for all ℓ into a vector $\boldsymbol{\xi}$, the posterior distribution of $\boldsymbol{\xi}$ is

$$p(\boldsymbol{\xi}|\boldsymbol{y}_{1:K}) \propto \frac{1}{\sqrt{(2\pi)^d|\mathbf{P}_\xi|}} \exp\left(-\frac{1}{2}\left[\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_b\right]^{\mathrm{T}} \mathbf{P}_\xi^{-1}\left[\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_b\right]\right) \times \prod_{k=1}^{K}\prod_{j=1}^{n} \frac{h_{kj}(\boldsymbol{x}_k)^{y_{kj}}}{y_{kj}!} \exp[-h_{kj}(\boldsymbol{x}_k)].$$

The log posterior density becomes

$$\log p(\boldsymbol{\xi}|\boldsymbol{y}_{1:K}) = -\frac{1}{2}\left[\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_b\right]^{\mathrm{T}} \mathbf{P}_\xi^{-1}\left[\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_b\right] + \sum_{k=1}^{K}\sum_{j=1}^{n}[y_{kj}\log h_{kj}(\boldsymbol{x}_k) - \log y_{kj}! - h_{kj}(\boldsymbol{x}_k)] + \text{const.} \tag{3}$$

When we define the following composite function:

$$z_{kj}(\xi) = h_{kj} \circ \boldsymbol{f}_k \circ \boldsymbol{f}_{k-1} \circ \cdots \circ \boldsymbol{f}_1 \circ \boldsymbol{e}(\boldsymbol{\xi}) (= h_{kj}(\boldsymbol{x}_k)), \tag{4}$$

where $\boldsymbol{e}$ denotes a d-dimensional function which satisfies

$$\boldsymbol{e}(\boldsymbol{\xi}) = \begin{pmatrix} \exp(\xi_1) \\ \vdots \\ \exp(\xi_d) \end{pmatrix}.$$

To maximise $\log p(\boldsymbol{\xi}|\boldsymbol{y}_{1:K})$, we minimise the following objective function:

$$J_P(\boldsymbol{\xi}) = \frac{1}{2}\left[\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_b\right]^{\mathrm{T}} \mathbf{P}_\xi^{-1}\left[\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_b\right] + \sum_{k,j}[\log y_{kj}! + z_{kj}(\boldsymbol{\xi}) - y_{kj}\log z_{kj}(\boldsymbol{\xi})]. \tag{5}$$

Approximation with an ensemble

We approximate the objective function J_P using a result of ensemble simulation. We generate N different values for $\boldsymbol{\xi}$, $\{\boldsymbol{\xi}_m^{(1)}, \ldots, \boldsymbol{\xi}_m^{(N)}\}$ to satisty

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\xi}_m^{(i)} = \hat{\boldsymbol{\xi}}_{m-1}. \tag{6}$$

where $\hat{\boldsymbol{\xi}}_{m-1}$ is the $(m-1)$-th estimate. We introduce a $d \times N$ matrix:

$$\boldsymbol{\Xi}_m = \left(\boldsymbol{\xi}_m^{(1)} - \hat{\boldsymbol{\xi}}_{m-1} \quad \cdots \quad \boldsymbol{\xi}_m^{(N)} - \hat{\boldsymbol{\xi}}_{m-1}\right),$$

and represent $\boldsymbol{\xi}$ as follows:

$$\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}_{m-1} + \boldsymbol{\Xi}_m \boldsymbol{w}_m. \tag{7}$$

We obtain $z_{kj}(\boldsymbol{\xi}_m^{(\mathrm{i})})$ for $i = 1, \ldots, N$ for each time step k by running the ensemble simulation. Defining an N-dimensional row vector $\boldsymbol{\zeta}_{kj,m}^{\mathrm{T}}$ as

$$\boldsymbol{\zeta}_{kj,m}^{\mathrm{T}} = \left(z_{kj}(\boldsymbol{\xi}_m^{(1)}) - z_{kj}(\hat{\boldsymbol{\xi}}_{m-1}) \quad \cdots \quad z_{kj}(\boldsymbol{\xi}_m^{(N)}) - z_{kj}(\hat{\boldsymbol{\xi}}_{m-1})\right),$$

we can approximate $z_{kj}(\boldsymbol{\xi})$ as

$$z_{kj}(\boldsymbol{\xi}) \approx z_{kj}(\bar{\boldsymbol{\xi}}_b) + \boldsymbol{\zeta}_{kj,m}^{\mathrm{T}} \boldsymbol{w}_m. \tag{8}$$

Ensemble-based variational method for count data

The objective function J_P in Eq. (5) can thus be approximated as

$$J_P(\xi) \approx \frac{1}{2}\left(\hat{\boldsymbol{\xi}}_{m-1} - \bar{\boldsymbol{\xi}}_b + \boldsymbol{\Xi}_m \boldsymbol{w}_m\right)^{\mathrm{T}} \mathbf{P}_\xi^{-1}\left(\hat{\boldsymbol{\xi}}_{m-1} - \bar{\boldsymbol{\xi}}_b + \boldsymbol{\Xi}_m \boldsymbol{w}_m\right) + \sum_{k,j}\left[\log y_{kj}! + z_{kj}(\bar{\boldsymbol{\xi}}_b) + \boldsymbol{\zeta}_{kj,m}^{\mathrm{T}} w_m - y_{kj}\log\left(z_{kj}(\bar{\boldsymbol{\xi}}_b) + \boldsymbol{\zeta}_{kj,m}^{\mathrm{T}} \boldsymbol{w}_m\right)\right].$$

We minimise an objective function with a regularisation term:

$$\hat{J}_P(\xi) = \frac{\sigma_m^2}{2} w_m^{\mathrm{T}} w_m + \frac{1}{2}\left(\hat{\boldsymbol{\xi}}_{m-1} - \bar{\boldsymbol{\xi}}_b + \boldsymbol{\Xi}_m w_m\right)^{\mathrm{T}} \mathbf{P}_\xi^{-1}\left(\hat{\boldsymbol{\xi}}_{m-1} - \bar{\boldsymbol{\xi}}_b + \boldsymbol{\Xi}_m \boldsymbol{w}_m\right) + \sum_{\mathrm{k,j}}\left[\log y_{kj}! + z_{kj}(\bar{\boldsymbol{\xi}}_b) + \boldsymbol{\zeta}_{kj,m}^{\mathrm{T}} \boldsymbol{w}_m - y_{kj}\log\left(z_{kj}(\bar{\boldsymbol{\xi}}_b) + \boldsymbol{\zeta}_{kj,m}^{\mathrm{T}} \boldsymbol{w}_m\right)\right]. \tag{9}$$

At the m-th iteration, the optimal $\boldsymbol{w}_m$ is obtained by minimising $\hat{J}_P$ as follows:

$$\hat{\boldsymbol{w}}_m = -\left(\sigma_m^2 \mathbf{I}_N + \boldsymbol{\Xi}_m^{\mathrm{T}} \mathbf{P}_\xi^{-1} \boldsymbol{\Xi}_m + \sum_{k,j}\left[\frac{y_{kj}}{z_{kj}(\bar{\boldsymbol{\xi}}_b)^2}\boldsymbol{\zeta}_{kj,m}\boldsymbol{\zeta}_{kj,m}^{\mathrm{T}}\right]\right)^{-1} \times \left(\boldsymbol{\Xi}_m^{\mathrm{T}} \mathbf{P}_\xi^{-1}\left[\hat{\boldsymbol{\xi}}_{m-1} - \bar{\boldsymbol{\xi}}_b\right] + \sum_{k,j}\left[\boldsymbol{\zeta}_{kj,m} - \frac{y_{kj}}{z_{kj}(\bar{\boldsymbol{\xi}}_b)}\boldsymbol{\zeta}_{kj,m}\right]\right). \tag{10}$$

The m-the estimate of $\boldsymbol{\xi}$ is obtained as

$$\hat{\boldsymbol{\xi}}_m = \hat{\boldsymbol{\xi}}_{m-1} + \boldsymbol{\Xi}_m \hat{\boldsymbol{w}}_{m+1}, \tag{11}$$

and therefore the estimate of $\boldsymbol{x}_0$ at the m-th iteration becomes

$$\hat{\boldsymbol{x}}_{0,m} = \boldsymbol{e}(\hat{\boldsymbol{\xi}}_m). \tag{12}$$

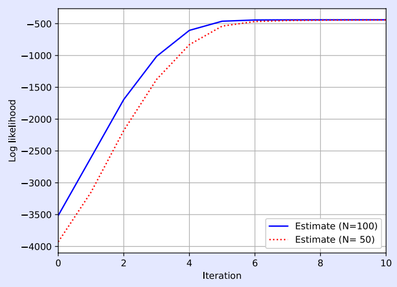
Experiment with a one-dimensional fluid model

We consider the one-dimensional advection and diffusion equation:

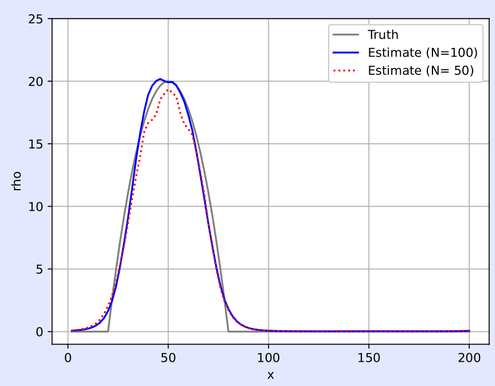
$$\frac{\partial \rho}{\partial t} = -u\frac{\partial \rho}{\partial x} + v\frac{\partial^2 \rho}{\partial x^2}, \tag{13}$$

where ρ is the fluid density, t indicates the time, x indicates the location of one-dimensional space, u is the velocity, and v is the diffucion coefficient.

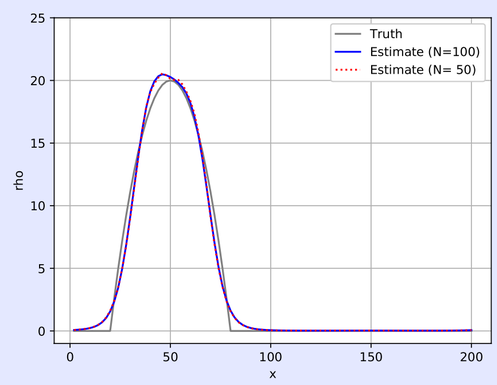
- The range of x is $0 \leq x < 200$ and a periodic boundary condition $\rho(t, x = 200) = \rho(t, x = 0)$ is imposed for any t. We also assume $u = 2$ and $v = 2$.
- The resolution is $\Delta x = 2$ in space and $\Delta t = 0.2$ in time.
- The Lax–Wendroff scheme is used for calculating the temporal evolution.
- The observations were obtained with a spatial resolution of $2\Delta x$ and a time interval of $20\Delta t$ until $t = 100\Delta t = 20$. Each of the data follows the Poisson distribution with a mean of $\rho(t, x)$.



The log likelihood of the estimate at each iteration. The solid line shows the result with the ensemble size $N = 100$ and the dotted line shows the result with $N = 50$.



The estimated after 6 iterations.



The estimated after 10 iterations.

The solid black line indicates the result with $N = 100$ and the dotted black line indicates the result with $N = 50$. The true initial state is shown with the solid gray line.

See Nakano (2003), Proc. Inst. Statist. Math., v. 70, p. 235 (in Japanese) for detail.