Inference for discrete latent time series processes

Bhattacharya, Arnab
Trinity College, Department of Computer Science and Statistics
02 College Green
Dublin 02,Ireland
E-mail: bhattaca@tcd.ie

Wilson, Simon P.
Trinity College, Department of Computer Science and Statistics
02 College Green
Dublin 02,Ireland
E-mail: simon.wilson@tcd.ie

Abstract

The problem of performing online inference when observations arrive sequentially in time is well known, and there exists a wide range of statistical literature devoted to this problem. The aim of this project is to develop a functional approximation method so as to perform real-time inference with sequential data which are dependent on some underlying latent variable. The new proposed method sequential INLA (SINLA) has derived its idea from Integrated Nested Laplace Approximation (INLA) (Rue., Martino & Chopin 2009), a fast Bayesian approximation technique where for a class of latent variable models the underlying latent variable follow a Gaussian Markov Random Field (GMRF) (Rue & Held 2005).

1 Introduction

Many real-life problems require estimation of unknown quantities from observations that arrive sequentially in time. Often in such circumstances one is interested in performing inference sequentially or 'on-line'. In this paper, we will be concentrating on estimating the posterior distribution of the parameters associated with this sequential process using Bayes' theorem. The sequential process will be modeled by a state-space approach, and the primary focus here will be a discrete-time formulation of the process. Examples include single and multiple target tracking, estimating digital communication signals, estimating of volatility of financial factors using for example stock market data. A state-space model in discrete time can be conveniently written in the form of equations

(1.1)

$$\mathbf{y}_t = f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t, \theta_1)$$

(1.2)

$$\mathbf{x}_t = g(\mathbf{x}_{t-1}, \mathbf{w}_t, \theta_2)$$

where $\mathbf{v}_t$ is the *observation error* and $\mathbf{w}_t$ is the *system error.* $\mathbf{u}_t$ is the exogenous output that is fully known. f and $\mathbf{v}_t$ fully specify the likelihood of observations $\mathbb{P}(y_t|x_t, \theta_t)$, while the transition density $\mathbb{P}(x_t|x_{t-1})$ is completely specified by g and $\mathbf{w}_t$. Θ is the set of hyperparameters $(\Theta_1, \Theta_2) \equiv \Theta$.

The Kalman filter (Kalman 1960) has been the most widely used algorithm to deal with a linear Gaussian system. Linearity of the model and Gaussian errors ensure exact expressions for the estimate of the state process. However, most real-world problems require nonlinearity and/or non-Gaussian errors. Extensions of Kalman filter exists that provide sub-optimal solutions to the problem. Extended Kalman filter (EKF) (Anderson & Moore 1979), unscented Kalman filter (UKF) (Julier, Uhlmann & Durrant-Whyte 1995), ensemble Kalman filter (G. 1994) and many more are examples of such extensions. An absolutely different outlook is taken by Monte Carlo based methods like Particle filters (Doucet & Gordon 2001). A third way is using grid based methods (Pole & West 1990). The *curse of dimsionality* prevented any further work on these type of filters. Our method uses a fast but accurate grid based method for making inference about the hyper parameters.

2 Sequential Bayesian Estimation of Θ

The optimal method to sequentially update the posterior density of Θ as new observations arrive is given here. Let the unobserved signal (hidden states) $\mathbf{X}_t, t \in \mathbb{N}$, $\mathbf{X}_t \in \mathcal{X}$ be a GMRF, with initial distribution being $\mathbb{P}(\mathbf{x}_0)$ and the transition equation $\mathbb{P}(\mathbf{x}_t|\mathbf{x}_{t-1})$. The observed variables $\mathbf{Y}_t, t \in \mathbb{N}$, $\mathbf{Y}_t \in \mathcal{Y}$ are assumed to be conditionally independent given the latent process $\mathbf{X}_t, t \in \mathbb{N}$ and Θ, and has a marginal distribution $\mathbb{P}(\mathbf{y}_t|\mathbf{x}_t, \theta)$.

We make use of Bayes Law and the structure of the state-space models, and factor the posterior density in the following recursive form,

$$\begin{aligned}\mathbb{P}\left(\theta|\mathbf{y}_{1:t}\right) &= \frac{\mathbb{P}\left(\theta, \mathbf{y}_{1:t}\right)}{\mathbb{P}\left(\mathbf{y}_{1:t}\right)}\\ &= \frac{\mathbb{P}\left(\theta, \mathbf{y}_t, \mathbf{y}_{1:t-1}\right)}{\mathbb{P}\left(\mathbf{y}_{1:t}\right)}\end{aligned}$$

(2.1)

$$= \frac{\mathbb{P}\left(\theta, \mathbf{y}_{1:t-1}\right)}{\mathbb{P}\left(\mathbf{y}_{1:t}\right)} \frac{\mathbb{P}\left(\mathbf{y}_t, \mathbf{y}_{1:t-1}, \theta\right)}{\mathbb{P}\left(\mathbf{y}_{1:t-1}, \theta\right)}$$

(2.2)

$$= \frac{\mathbb{P}\left(\theta|\mathbf{y}_{1:t-1}\right) \mathbb{P}\left(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \theta\right)}{\mathbb{P}\left(\mathbf{y}_t|\mathbf{y}_{1:t-1}\right)}$$

(2.3)

$$= \frac{\mathbb{P}\left(\theta|\mathbf{y}_{1:t-1}\right)}{\mathbb{P}\left(\mathbf{y}_t|\mathbf{y}_{1:t-1}\right)} \frac{\mathbb{P}\left(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \mathbf{x}_t, \theta\right) \mathbb{P}\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}, \theta\right)}{\mathbb{P}\left(\mathbf{x}_t|\mathbf{y}_{1:t}, \theta\right)}$$

(2.4)

$$= \frac{\mathbb{P}\left(\theta|\mathbf{y}_{1:t-1}\right)}{\mathbb{P}\left(\mathbf{y}_t|\mathbf{y}_{1:t-1}\right)} \left(\left. \frac{\mathbb{P}\left(\mathbf{y}_t|\mathbf{x}_t, \theta\right) \mathbb{P}\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}, \theta\right)}{\mathbb{P}\left(\mathbf{x}_t|\mathbf{y}_{1:t}, \theta\right)} \right|_{\mathbf{x}_t = \mathbf{x}_t^*(\theta)} \right)$$

where $\mathbf{x}_t^*(\theta)$ is some estimate of $\mathbf{x}_t$ which allows the probabilities to be calculated at that value.

The above recursion formula, involving multi-dimansional integrals are only tractable for linear Guassian systems, with Gaussian priors for the hyperparameters. The terms $\mathbb{P}\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}, \theta\right)$ and $\mathbb{P}\left(\mathbf{x}_t|\mathbf{y}_{1:t}, \theta\right)$ have very complicated forms themselves. One can use the Kalman filter or extensions of Kalman Filter and other methods like Expectation Propagation (Minka 2001) to compute the estimate $\mathbf{x}_t^*(\theta)$. INLA is used on the first few observations to make the initial grid for the parameters.

2.1 Dynamic Grid

First and foremost, the most important thing about dynamic grid should be mentioned. It is advisable to try and create a situation where one doesn't need to add on new grid points at every time point. It slows the algorithm down, since a single extra grid increases the calculation exponentially. Hence one should make the initial grid using INLA quite wide.

2.2 Internal new grid point

For each of the parameters we look at their marginal densities to determine the new grid points. Some criteria can be defined based on the approximate densities of successive grid points which determines the requirement of a new point in between. The Euclidean distance between the density of two successive points can be one such criterion. A linear interpolation scheme is used to calculate the density of a new point between two existing points. It is less accurate than fitting a curve which makes more sense in interpolating for a density but it is very easy to calculate and hence makes the algorithm very fast.

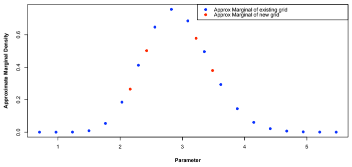

Figure 1: Approximate marginal distribution of one of the parameters after interpolating the density for four new points.

2.3 External new grid point

The need to add more point to the grid on the edges of the existing ones is also quite complicated. The joint distribution of the parameters tends to be "mobile" at the start of the iterations, until with time it settles down into a grid which holds its place. This is because INLA is being used to get the initial grid based on a small set of data, the mode of the density generally may not be close to the "true" value. Note that linear extrapolation may work on a nonlinear curve at an extremely "local" level. A very short term extrapolation should be a safe method then. At every time point, whenever a situation arrives when the grid needs to be moved, a single (or a couple) of grid point is added at a distance which is same as the one between the nearest two grid points.

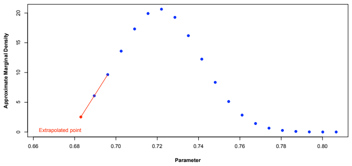

Figure 2: The extrapolation is done to a single point.

3 Results

This section provides various simulated examples of the SINLA approach with models varying from linear Gaussian to further generalizations using fixed (over time) parameter θ.

3.1 Linear Gaussian model

A very standard linear Gaussian state space model is used to simulate data here. We tried to simulate an experiment which can be described by the following: a single radio antenna is transmitting at a fixed frequency and the signal is being received simultaneously at several spatially distinct nodes. We assume the observable data to be noisy tri-variate realizations of a latent signal which follows an

$AR(1)$ model. The complete model thus has the form

(3.1)

$$\mathbf{y}_t = x_{t-1}\underline{\mathbf{1}} + \eta_t$$

(3.2)

$$x_t = \phi x_{t-1} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma^2_{err})$ and $\eta_t \sim \mathcal{MVN}(\mathbf{0}, \Sigma)$. The entries of the covariance matrix Σ are of the following type

$$\Sigma_{ii} = \sigma^2_{obs} \quad \Sigma_{ij} = \sigma^2_{bs} \exp^{-rd(i,j)}$$

where $r > 0$ and $d(i,j)$ is a measure of distance between nodes i and j. The set of unknown hyperparameters is thus $\mathbf{\Theta} = \{\phi, \sigma^2_{obs}, \sigma^2_{err}\}$ or equivalently taking the precision parameters $\Theta = \{\phi, \rho_{Obs}, \rho_{Sys}\}$.

Data was generated by fixing the values at $\phi = 0.35, \rho^2_{Obs} = 250, \rho^2_{Sys} = 28.5$, the value of r at $2/3$ and the distance values were set at $d(1,2) = 1$, $d(1,3) = 3$ and $d(2,3) = \sqrt{10}$. At each time point, the approximate mode of the parameters are plotted. An approximate 95% confidence bound is also plotted. Figure 3 shows us the true value of the parameters and the plot of the approximate mode and intervals.

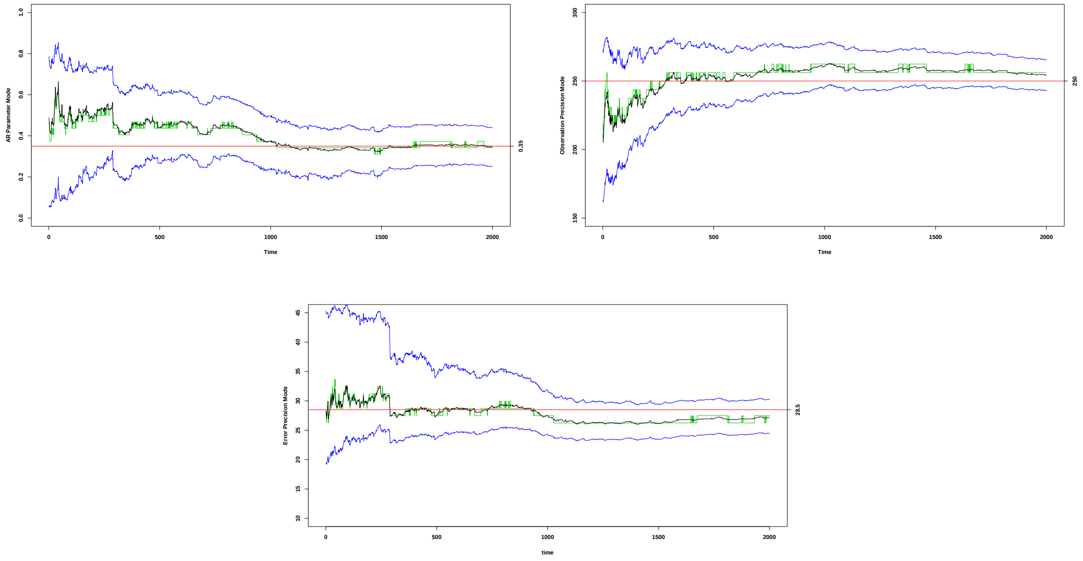

Figure 3: The solid green line shows the "mode" of the parameters. The dotted blue line is the marginal mode and the solid blue lines indicate the probability bounds.

The method seems to be working quite well here as all the three parameters stay within the bounds. The approximate mode converges to the true values in about 1000 observations. The fact that implementation of INLA on the first 20 observations provided us with very good starting values for the grid also helped this case. A constant grid was sufficient in this case.

3.2 Nonlinear Gaussian model

A nonlinear model with additive gaussian errors was tried as our next simulated example. Data has been generated from a model with nonlinearity in the observation equation,

(3.3)

$$y_t = \theta x^2_t + v_t$$

(3.4)

$$x_{t+1} = 4 + \phi x_t + sin(\omega \pi t) + w_t$$

where $v_t \sim \mathcal{N}(0, \sigma^2_{Obs})$, $w_t \sim \mathcal{N}(0, \sigma^2_{Sys})$ and ω is assumed to be known. The hyperparameters are given by the vector $\Psi \equiv \left(\phi, \theta, \sigma^2_{State}, \sigma^2_{Obs}\right)$. The values assigned to these hyperparameters were $(0.7, 2, 0.35, 0.0001)$ respectively and the value of ω was set at 1.718. A reparameterisation of the variance parameters was done to felicitate computer arithmetic operations. Thus now we have $\Psi \equiv (\phi, \theta, \rho_{State}, \rho_{Obs})$ and their actual values are $(0.7, 2, 2.87, 10000)$. The approximate mode and confidence intervals we computed the same way as explained before. For each parameter the approximate mode of the distribution defined on the grid seems to have converged approximately to the true value for almost all the parameters, except for the parameter ρ_{State}, precision of the state equation error. The observation precision looks like it stays constant over time but in reality it is making very small moves.

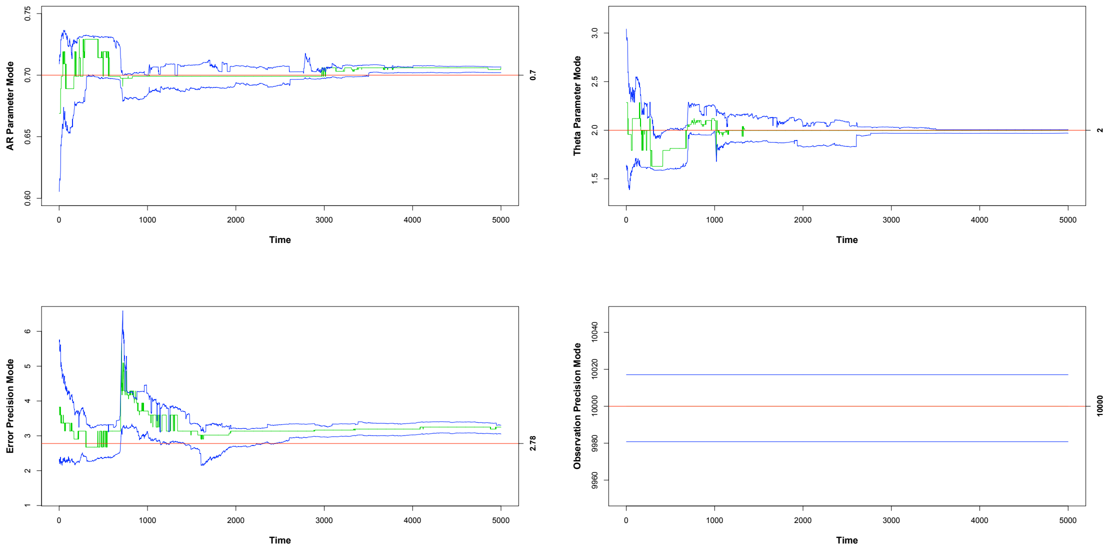

Figure 4: Plot of approximate mode and 95% confidence intervals that we get through the UKF based method based on 5000 observations.

3.3 Non-Gaussian model

Data was generated from a model which assumed that the observations came from a Poisson distribution.The state process was assumed to be an AR(1) process with some mean.

(3.5)

$$y_t \sim Poisson(e^x_t)$$

(3.6)

$$x_{t+1} = 2 + \phi x_t + \eta_t$$

where $\eta_t \sim \mathcal{N}(0, \sigma^2_{Sys})$. Our method was applied to an approximate form of this model with additive Gaussian error

(3.7)

$$y_t = e^x_t + \epsilon_t$$

(3.8)

$$x_{t+1} = 2 + \phi x_t + \eta_t$$

where one more unknown parameter was introduced through $\epsilon_t \sim \mathcal{N}(0, \sigma^2_{Obs})$. Since our parameter of interest are ϕ, σ^2_{Sys} or equivalently ϕ, ρ_{Sys} we will be monitoring the convergence of them. The real values of the parameters are 0.35 and 0.1 respectively. The two parameters did not quite converge to the correct parametric values. Thats understandable in some sense since we are using a wrong model here. Also since we are using UKF to estimate the state process, it has to be accurate in its estimation of the state process.

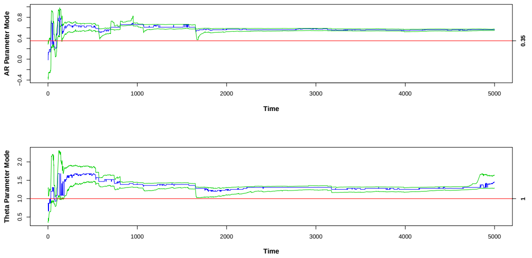

Figure 5: Plot of approximate mode and 95% confidence intervals that we get through the UKF based method based on 5000 observations.

4 Conclusion

The SINLA approach based on these simulated examples proves to be adequately accurate and if one can keep the number of grid points to a minimum then is very fast. Further work is to be done in models with non Gaussian error and also spatio-temporal models.

References

Anderson, B. D. & Moore, J. B. (1979), *Optimal Filtering*, Prentice Hall.

Doucet, A. de Freitas, N. & Gordon, N. (2001), *Sequential Monte Carlo Methods in Practise*, Springer.

G., E. (1994), 'Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics', *Journal of Geophysical Research* **99**(C5), 10143–10162.

Julier, S. J., Uhlmann, J. K. & Durrant-Whyte, H. F. (1995), A new approach for filtering nonlinear systems, *in* 'Proceedings of the American Control Conference.'.

Kalman, R. E. (1960), 'A new approach to linear filtering and prediction problem', *Transactions of the ASME – Journal of Basic Engineering* (82 (Series D)), 35–45.

Minka, T. P. (2001), A family of algorithms for approximate A family of algorithms for approximate Bayesian inference, PhD thesis, MIT.

Pole, A. & West, M. (1990), 'Efficient bayesian learning in non-linear dynamic models', *Journal of Forecasting* **9**(2), 119–136.

Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Application*, Chapman & Hall.

Rue., H., Martino, S. & Chopin, N. (2009), 'Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations', *Journal of the Royal Statistical Society Series B* **71(2)**, 1–35.