Chapter 8 - Hierarchical DSTMs: Implementation and Inference

Odd Kolbjørnsen & Geir Storvik

April 24 2017

Last time

Hierarchical Dynamical Spatio-Temporal Models

- Data in Process models
- Observation types
- Linear observations
- Kalman filter
- Kalman smoother
- nonlinear/non Gaussian

Bayesian approach: Also include model for parameters

Today

Methodology for inference in Hierarchical Dynamical Spatio-Temporal Models

- General Problem
- Sequential vs non sequential
- Kalman filter
- EM-algorithm
- MCMC
- Sequential Monte Carlo, particle filter
- INLA

Hierarchical model

- Model for $p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}_D]$
- Model for $p[\mathbf{Y}|\boldsymbol{\theta}_P]$
- Bayesian approach: Model for $p[\boldsymbol{\theta}_D,\boldsymbol{\theta}_P]$

Inference: Extract information about $\boldsymbol{\theta}$ and $\mathbf{Y}$ from $\mathbf{Z}$

Hierarchical model

- Model for $p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}_D]$
- Model for $p[\mathbf{Y}|\boldsymbol{\theta}_P]$
- Bayesian approach: Model for $p[\boldsymbol{\theta}_D,\boldsymbol{\theta}_P]$

Inference: Extract information about $\boldsymbol{\theta}$ and $\mathbf{Y}$ from $\mathbf{Z}$

Likelihood:

$$p[\mathbf{Z}|\boldsymbol{\theta}] = \int_{\mathbf{Y}} p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}_D]p[\mathbf{Y}|\boldsymbol{\theta}_P]d\mathbf{Y}$$

Bayesian posterior

$$p[\boldsymbol{\theta},\mathbf{Y}|\mathbf{Z}] = \frac{p[\boldsymbol{\theta},\mathbf{Y}]p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}]}{p[\mathbf{Z}]}$$

$$p[\mathbf{Z}] = \int_{\boldsymbol{\theta}} p[\mathbf{Z}|\boldsymbol{\theta}]p[\boldsymbol{\theta}]d\boldsymbol{\theta}$$

Hierarchical model

- Model for $p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}_D]$
- Model for $p[\mathbf{Y}|\boldsymbol{\theta}_P]$
- Bayesian approach: Model for $p[\boldsymbol{\theta}_D,\boldsymbol{\theta}_P]$

Inference: Extract information about $\boldsymbol{\theta}$ and $\mathbf{Y}$ from $\mathbf{Z}$
Likelihood:

$$p[\mathbf{Z}|\boldsymbol{\theta}] = \int_{\mathbf{Y}} p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}_D]p[\mathbf{Y}|\boldsymbol{\theta}_P]d\mathbf{Y}$$

Bayesian posterior

$$p[\boldsymbol{\theta},\mathbf{Y}|\mathbf{Z}] = \frac{p[\boldsymbol{\theta},\mathbf{Y}]p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}]}{p[\mathbf{Z}]}$$

$$p[\mathbf{Z}] = \int_{\boldsymbol{\theta}} p[\mathbf{Z}|\boldsymbol{\theta}]p[\boldsymbol{\theta}]d\boldsymbol{\theta}$$

How to obtain these quantities:

- Huge computational problem
- Very active research field
- Some general methods
- Software for specific (classes of) models

Sequential/non-sequential inference

- Data $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_T)$
- $p(\mathbf{Z}) = \prod_t p(\mathbf{Z}_t | \mathbf{Z}_1, ..., \mathbf{Z}_{t-1})$
- Sequential updating:
$$p(\boldsymbol{\theta}, \mathbf{Y}_{1:t} | \mathbf{Z}_{1:t}), \quad t = 1, 2, 3, ...$$
- Non-sequential updating
$$p(\boldsymbol{\theta}, \mathbf{Y}_{1:T} | \mathbf{Z}_{1:T})$$

Methods

- Kalman filter - sequential
- Markov chain Monte Carlo - nonsequential
- Sequential Monte Carlo - sequential
- INLA - nonsequential
- Ensemble Kalman Filter - sequential
- Ensemble (Kalman) Smoother - nonsequentia (not in book)l

Model

$$\mathbf{Y}_t = \mathbf{M}_t\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \overset{ind}{\sim} N(0,\mathbf{Q}_t)$$

$$\mathbf{Z}_t = \mathbf{H}_t\mathbf{Y}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \overset{ind}{\sim} N(0,\mathbf{R}_t)$$

Aim: Calculate $p(\mathbf{Y}_t|\mathbf{Z}_{1:t})$. Enough with

$$\widehat{\mathbf{Y}}_{t|t} = E[\mathbf{Y}_t|\mathbf{Z}_{1:t}]$$

$$\mathbf{P}_{t|t} = \text{Var}[\mathbf{Y}_t|\mathbf{Z}_{1:t}]$$

Kalman filter

$$\mathbf{P}_{t|t-1} = \mathbf{M}_t\mathbf{P}_{t-1|t-1}\mathbf{M}_t^T + \mathbf{Q}_t \qquad \widehat{\mathbf{Y}}_{t|t-1} = \mathbf{M}_t\widehat{\mathbf{Y}}_{t-1|t-1}$$

$$\mathbf{S}_t = \mathbf{H}_t^T\mathbf{P}_{t|t-1}\mathbf{H}_t + \mathbf{R}_t$$

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t^T\mathbf{S}_t^{-1}$$

$$\mathbf{P}_{t|t} = [\mathbf{I} - \mathbf{K}_t\mathbf{H}_t]\mathbf{P}_{t|t-1} \qquad \widehat{\mathbf{Y}}_{t|t} = \widehat{\mathbf{Y}}_{t|t-1} + \mathbf{K}_t[\mathbf{Z}_t - \mathbf{H}_t\widehat{\mathbf{Y}}_{t|t-1}]$$

Likelihood:

$$L(\boldsymbol{\theta}) = p(\mathbf{Z};\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{Z}_t|\mathbf{Z}_{1:t-1};\boldsymbol{\theta})$$

Kalman filter (cont)

Parameter estimation:
Kalman filter give $p(\mathbf{Y}_t|\mathbf{Z}_{1:t})$ given parameters.

- $L(\boldsymbol{\theta}) = p(\mathbf{Z}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{Z}_t|\mathbf{Z}_{1:t-1};\boldsymbol{\theta})$
- This can be obtaind directly from the Kalman filter (exercise)
- Can optimize wrt $\boldsymbol{\theta}$ to obtain ML estimates
- Can also do Bayesian versions.

Properties

- Computationally very efficient, sequential (online) inference
- Can use alternative filters calculating $\mathbf{P}_{t|t}^{-1}$ and utilizing that this often is sparse
- Can be extended to nonlinear models through linear approximations

EM-algorithm

Maximum likelihood estimates in case of missing data (or latent variables) , and unknown parameters.

General formulation:
Iterate Expectation and Maximization until convergence...

- E-step: Calculate expectation of log likelihood $E(\ln L(\theta|Z))|Z_{obs}, \hat{\theta}^{(i-1)}) = q(\theta|\hat{\theta}^{(i-1)})$
- M-step: Find θ that maximizes $q(\theta|\hat{\theta}^{(i-1)})$, and call this $\hat{\theta}^{(i)}$

Expectation (E-step) is to "get rid of" the latent variables (or missing data) to get back to the standard maximum likelihood estimator (M-Step).

EM-algorithm, Kalman filter

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \overset{ind}{\sim} N(0, \mathbf{Q})$$

$$\mathbf{Z}_t = \mathbf{H}_t\mathbf{Y}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \overset{ind}{\sim} N(0, \mathbf{R})$$

Unknown parameters: $\mathbf{M}, \mathbf{R}, \mathbf{Q}$, and $\boldsymbol{\mu}_Y$. Known parameter: $\mathbf{H}_t$

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \overset{ind}{\sim} N(0, \mathbf{Q})$$

$$\mathbf{Z}_t = \mathbf{H}_t\mathbf{Y}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \overset{ind}{\sim} N(0, \mathbf{R})$$

Unknown parameters: $\mathbf{M}, \mathbf{R}, \mathbf{Q}$, and $\boldsymbol{\mu}_Y$. Known parameter: $\mathbf{H}_t$

Start with an initial guess on the parameters: $\mathbf{M}^{(0)}, \mathbf{R}^{(0)}, \mathbf{Q}^{(0)}, \boldsymbol{\mu}_Y^{(0)}$
Then iterate:

- E-step: Run a Kalman smoother (Filter-smoother/ Forward-backward) with current parameter estimates $\mathbf{M}^{(i)}, \mathbf{R}^{(i)}, \mathbf{Q}^{(i)}, \boldsymbol{\mu}_Y^{(i)}$ to obtain updated state estimates (estimate of latent process), i.e the estimates of $\mathbf{Y}_{1:T}$

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \overset{ind}{\sim} N(0, \mathbf{Q})$$
$$\mathbf{Z}_t = \mathbf{H}_t\mathbf{Y}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \overset{ind}{\sim} N(0, \mathbf{R})$$

Unknown parameters: $\mathbf{M}, \mathbf{R}, \mathbf{Q}$, and $\boldsymbol{\mu}_Y$. Known parameter: $\mathbf{H}_t$

Start with an initial guess on the parameters: $\mathbf{M}^{(0)}, \mathbf{R}^{(0)}, \mathbf{Q}^{(0)}, \boldsymbol{\mu}_Y^{(0)}$
Then iterate:

- E-step: Run a Kalman smoother (Filter-smoother/ Forward-backward) with current parameter estimates $\mathbf{M}^{(i)}, \mathbf{R}^{(i)}, \mathbf{Q}^{(i)}, \boldsymbol{\mu}_Y^{(i)}$ to obtain updated state estimates (estimate of latent process), i.e the estimates of $\mathbf{Y}_{1:T}$
- M-step: Assume the state space (latent process) is observed, and use them in the maximum likelihood estimation together with $\mathbf{Z}_{1:T}$ to get updated parameter estimates of $\mathbf{M}^{(i+1)}, \mathbf{R}^{(i+1)}, \mathbf{Q}^{(i+1)}, \boldsymbol{\mu}_Y^{(i+1)}$.

Markov chain Monte Carlo (MCMC)

Interest in $\hat{g} = E[g(\mathbf{Y}, \boldsymbol{\theta})|\mathbf{Z}]$
Monte Carlo

- Sample $\{(\mathbf{Y}^{(s)}, \boldsymbol{\theta}^{(s)}\}$ from $p[\mathbf{Y}, \boldsymbol{\theta}|\mathbf{Z}]$
- Approximate $\hat{g}$ by

$$\frac{1}{S}\sum_{s=1}^{S} g(\mathbf{Y}^{(s)}, \boldsymbol{\theta}^{(s)})$$

Markov chain Monte Carlo (MCMC)

Interest in $\hat{g} = E[g(\mathbf{Y}, \boldsymbol{\theta})|\mathbf{Z}]$
Monte Carlo

- Sample $\{(\mathbf{Y}^{(s)}, \boldsymbol{\theta}^{(s)}\}$ from $p[\mathbf{Y}, \boldsymbol{\theta}|\mathbf{Z}]$
- Approximate $\hat{g}$ by
$$\frac{1}{S}\sum_{s=1}^{S} g(\mathbf{Y}^{(s)}, \boldsymbol{\theta}^{(s)})$$
- Difficult to sample from $p[\mathbf{Y}, \boldsymbol{\theta}|\mathbf{Z}]$
- MCMC: Tool for complex simulation
- Very general (we have looked at Gibbs sampler)
- Non-sequential (offline) inference
- Some general software (Winbugs), SLOWWWW
- Usually need to implement from scratch to make it efficient
- Time-consuming both in implementation time and running time
- Separate courses for this

Sequential Monte Carlo (SMC)

- Kalman filter: Calculate $p[\mathbf{Y}_t|\mathbf{Z}_{1:t},\boldsymbol{\theta}_D]$ analytically
- SMC: Approximate $p[\mathbf{Y}_t|\mathbf{Z}_{1:t},\boldsymbol{\theta}_D]$ by Monte Carlo samples
- Utilize samples from time $t-1$ when sampling at time t
- Differ from MCMC in performing simulations sequentially
- Very efficient in low dimensions of $\mathbf{Y}_t$, slow for high dimensions
- Active field, much progress made continuously!

Particle filter (An SMC)

"Sampling based Kalman filter".

Initial forecast:
Sample $\mathbf{Y}_0^{(l)} \sim p(\mathbf{y}_0)$, for $l = 1, ..., L$

Iterate for all time steps:

- Forecast step is sampling:
  Sample $\tilde{\mathbf{Y}}_t^l \sim p(\mathbf{y}_t|\mathbf{Y}_{t-1}^{(l)})$, for $l = 1, ..., L$
  Set $\tilde{\mathbf{Y}}_{0:t}^{(l)} = [\tilde{\mathbf{Y}}_{0:(t-1)}^{(l)} \tilde{\mathbf{Y}}_t^{(l)}]$
- Filter step is importance re-sampling:
  * Evaluate the importance weight (i.e. likelihood) $w_t^{(l)} = p(\mathbf{Z}_t|\tilde{\mathbf{Y}}_t^{(l)})$
  * Resample with replacement L particles $[\mathbf{Y}_{0:t}^{(l)}$,l=1,...,L] from $\tilde{\mathbf{Y}}_{0:t}^{(l)}$ using the importance weigth for resampling i.e. the probability of sampling particle k (at time t) is $w_t^{(k)}/\Sigma_l w_t^{(l)}$

Integrated nested Laplace approximation (INLA)

Interest in calculation of

$$L(\boldsymbol{\theta}) = \int_{\mathbf{Y}} p[\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}_D] p[\mathbf{Y}|\boldsymbol{\theta}_P] d\mathbf{Y} = \int_{\mathbf{Y}} e^{f(\mathbf{Y})} d\mathbf{Y}$$

Integrated nested Laplace approximation (INLA)

Interest in calculation of

$$L(\boldsymbol{\theta}) = \int_{\mathbf{Y}} p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}_D]p[\mathbf{Y}|\boldsymbol{\theta}_P]d\mathbf{Y} = \int_{\mathbf{Y}} e^{f(\mathbf{Y})}d\mathbf{Y}$$

Approximate (using $\widehat{\mathbf{Y}} = \text{argmax}_{\mathbf{Y}} f(\mathbf{Y}), \mathbf{f}'(\widehat{\mathbf{Y}}) = \mathbf{0}$)

$$\begin{aligned}
f(\mathbf{Y}) \approx& \hat{f}(\mathbf{Y})\\
=& f(\widehat{\mathbf{Y}}) + \mathbf{f}'(\widehat{\mathbf{Y}})(\mathbf{Y}-\widehat{\mathbf{Y}}) + \frac{1}{2}(\mathbf{Y}-\widehat{\mathbf{Y}})^T\mathbf{f}''(\widehat{\mathbf{Y}})(\mathbf{Y}-\widehat{\mathbf{Y}})\\
=& f(\widehat{\mathbf{Y}}) + \frac{1}{2}(\mathbf{Y}-\widehat{\mathbf{Y}})^T\mathbf{f}''(\widehat{\mathbf{Y}})(\mathbf{Y}-\widehat{\mathbf{Y}})
\end{aligned}$$

Interest in calculation of

$$L(\boldsymbol{\theta}) = \int_{\mathbf{Y}} p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}_D]p[\mathbf{Y}|\boldsymbol{\theta}_P]d\mathbf{Y} = \int_{\mathbf{Y}} e^{f(\mathbf{Y})}d\mathbf{Y}$$

Approximate (using $\widehat{\mathbf{Y}} = \text{argmax}_{\mathbf{Y}} f(\mathbf{Y}), \mathbf{f}'(\widehat{\mathbf{Y}}) = \mathbf{0}$)

$$\begin{aligned} f(\mathbf{Y}) \approx & \hat{f}(\mathbf{Y}) \\ =& f(\widehat{\mathbf{Y}}) + \mathbf{f}'(\widehat{\mathbf{Y}})(\mathbf{Y}-\widehat{\mathbf{Y}}) + \frac{1}{2}(\mathbf{Y}-\widehat{\mathbf{Y}})^T\mathbf{f}''(\widehat{\mathbf{Y}})(\mathbf{Y}-\widehat{\mathbf{Y}}) \\ =& f(\widehat{\mathbf{Y}}) + \frac{1}{2}(\mathbf{Y}-\widehat{\mathbf{Y}})^T\mathbf{f}''(\widehat{\mathbf{Y}})(\mathbf{Y}-\widehat{\mathbf{Y}}) \end{aligned}$$

which gives (with $d = \dim(\mathbf{Y})$)

$$L(\boldsymbol{\theta}) \approx e^{f(\widehat{\mathbf{Y}})}\frac{(2\pi)^{d/2}}{|-\mathbf{f}''(\widehat{\mathbf{Y}})|^{1/2}}$$

Interest in calculation of

$$L(\boldsymbol{\theta}) = \int_{\mathbf{Y}} p[\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}_D]p[\mathbf{Y}|\boldsymbol{\theta}_P]d\mathbf{Y} = \int_{\mathbf{Y}} e^{f(\mathbf{Y})}d\mathbf{Y}$$

Approximate (using $\widehat{\mathbf{Y}} = \text{argmax}_{\mathbf{Y}} f(\mathbf{Y}), \mathbf{f}'(\widehat{\mathbf{Y}}) = \mathbf{0}$)

$$\begin{aligned} f(\mathbf{Y}) \approx& \hat{f}(\mathbf{Y}) \\ =& f(\widehat{\mathbf{Y}}) + \mathbf{f}'(\widehat{\mathbf{Y}})(\mathbf{Y}-\widehat{\mathbf{Y}}) + \frac{1}{2}(\mathbf{Y}-\widehat{\mathbf{Y}})^T\mathbf{f}''(\widehat{\mathbf{Y}})(\mathbf{Y}-\widehat{\mathbf{Y}}) \\ =& f(\widehat{\mathbf{Y}}) + \frac{1}{2}(\mathbf{Y}-\widehat{\mathbf{Y}})^T\mathbf{f}''(\widehat{\mathbf{Y}})(\mathbf{Y}-\widehat{\mathbf{Y}}) \end{aligned}$$

which gives (with $d = \dim(\mathbf{Y})$)

$$L(\boldsymbol{\theta}) \approx e^{f(\widehat{\mathbf{Y}})} \frac{(2\pi)^{d/2}}{|-\mathbf{f}''(\widehat{\mathbf{Y}})|^{1/2}}$$

INLA

- Nested Laplace approximations
- Utilize sparsity in precision matrices

INLA

- Require latent field to be Gaussian
- Great flexibility with respect to observation processes
- Extremely fast compared to MCMC
- The number of models it can cover is increasing frequently
- R overhead makes it relatively easy to use
- Only give marginal distributions, not simultaneous ones.