An introduction to the DLM and GDLM.1.0 package

Yiping Dou

ydou@stat.ubc.ca

Department of Statistics

University of British Columbia

Supervisor: James V. Zidek

Co–supervisor: Nhu D. Le

Outline

- Introduction
- An Implementation of DLM to Ozone Study
- Results and Conclusions for the Ozone Study
- Prior Settings for the Hyperparameters by the Discount Factors
- GDLM.1.0 Package and Example 1 in Its DEMO
- Lab Exercises

Introduction

- Space-time field: underlying process varies in space and time
  - spatio-temporal (ST) data: dependence structures between pollutants, temporally and spatially.
- Probs:
  - non-stationary process: correlations? heterogeneity?
  - parameter uncertainty
  - curse of dimensionality
  - computational inefficiency

Space-Time (ST) Process

- Space-Time (ST) modelling:

$$\begin{aligned} Z(s_i, t) &= Y(s_i, t) + \epsilon(s_i, t) \\ Y(s_i, t) &= \mu(s_i, t) + \omega(s_i, t), \end{aligned}$$

  where $\epsilon \sim WN(0, .)$, ω : mean 0, ST process, $s_i \in D$, $i = 1, \ldots, n$ & $t = 1, \ldots, T$.
- Definitions:
  - stationary? non-stationary?
  - isotropic? anisotropic?
  - separable? non-separable?

DLM Theory

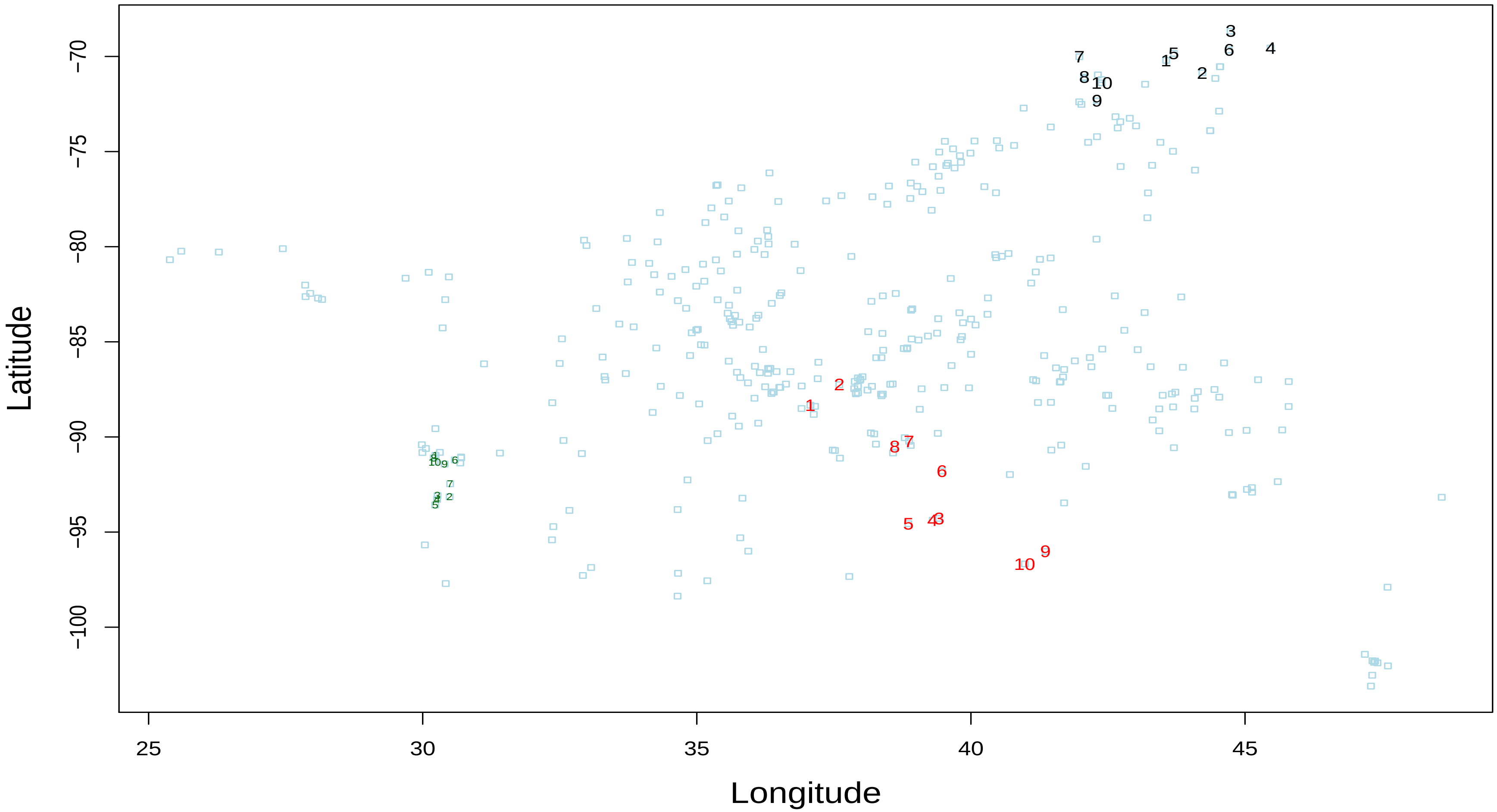
- DLM:

$$
\begin{aligned}
Y_t &= F_t' x_t + \nu_t \\
x_t &= G_t x_{t-1} + \omega_t,
\end{aligned}
$$

  where $\nu_t \sim N[0, V_t]$, $\omega_t \sim N[0, W_t]$, for $t = 1, \ldots, T$.
- Initial information: $(x_0|D_0) \sim N[m_0, C_0]$.
- Forward-filtering-backward-sampling (FFBS):
  - for $t \geq 1$, $(x_{t-1}|y_{1:t-1}, \theta) \rightarrow (x_t|y_{1:t-1}, \theta) \rightarrow (y_t|y_{1:t-1}, \theta) \rightarrow (x_t|y_{1:t}, \theta)$;
  - for $0 \leq k \leq T-1$, $(x_{T-k}|y_{1:T}, \theta)$.

Ozone Study: AIRS Dataset

Figure 1. Ozone Study: AIRS dataset in the USA.



Ozone Study: EDA

- Explanatory Data Analysis of Ozone:
  - measured in *ppb*
  - May 14 - Sept. 11, 1995 (120 days)
  - 375 monitoring stations
  - 3 Clusters of sites are chosen:
    - 24-, 12- hour cycles: periodicity v.s. spectrum
    - $\sqrt{\text{Ozone}}$: normality assumption

Ozone Study: DLM setting

- DLM form I:

$$
\begin{aligned}
\mathbf{Y_t} &= \mathbf{1}_n'\beta_t + S_{1t}(a_1)\alpha_{\mathbf{1t}} + S_{2t}(a_2)\alpha_{\mathbf{2t}} + \nu_{\mathbf{t}} \\
\beta_t &= \beta_{t-1} + \omega_t^{\beta} \\
\alpha_{\mathbf{jt}} &= \alpha_{\mathbf{j,t-1}} + \omega_{\mathbf{t}}{}^{\alpha_j},
\end{aligned}
$$

where $S_{jt}(a_j) = \cos(\pi tj/12) + a_j \sin(\pi tj/12)$, $j = 1, 2$; $\nu_t \sim N[0, \sigma^2 V_\lambda]$; $\omega_t^{\beta} \sim N[0, \sigma^2\tau_y^2]$, $\omega_t^{\alpha_j} \sim N[0, \sigma^2\tau_j^2 V_{\lambda_j}]$, $j = 1, 2$; $V_\lambda = \exp(-V/\lambda)$.

Ozone Study: DLM setting (Cont'd)

- DLM form II:

$$\begin{aligned} \mathbf{Y_t} &= \mathbf{F_t'x_t} + \nu_\mathbf{t} \\ \mathbf{x_t} &= \mathbf{x_{t-1}} + \omega_\mathbf{t}, \end{aligned}$$

  where $x_t' = (\beta_t, \alpha_{1t}', \alpha_{2t}')'$; $\omega_t \sim N[0, \sigma^2 W]$,
  $W =$ block diag $\{\tau_y^2, \tau_1^2, V_{\lambda_1}, \tau_2^2, V_{\lambda_2}\}$.

- Initial information: $(x_0|D_0, \theta) \sim N[m_0, \sigma^2 C_0]$.
- Advantages: It
  - reflects the time-dependent structure of data;
  - captures the spatial interaction structure of sites;
  - captures the periodicities of data.

Ozone Study: MCMC Sampling

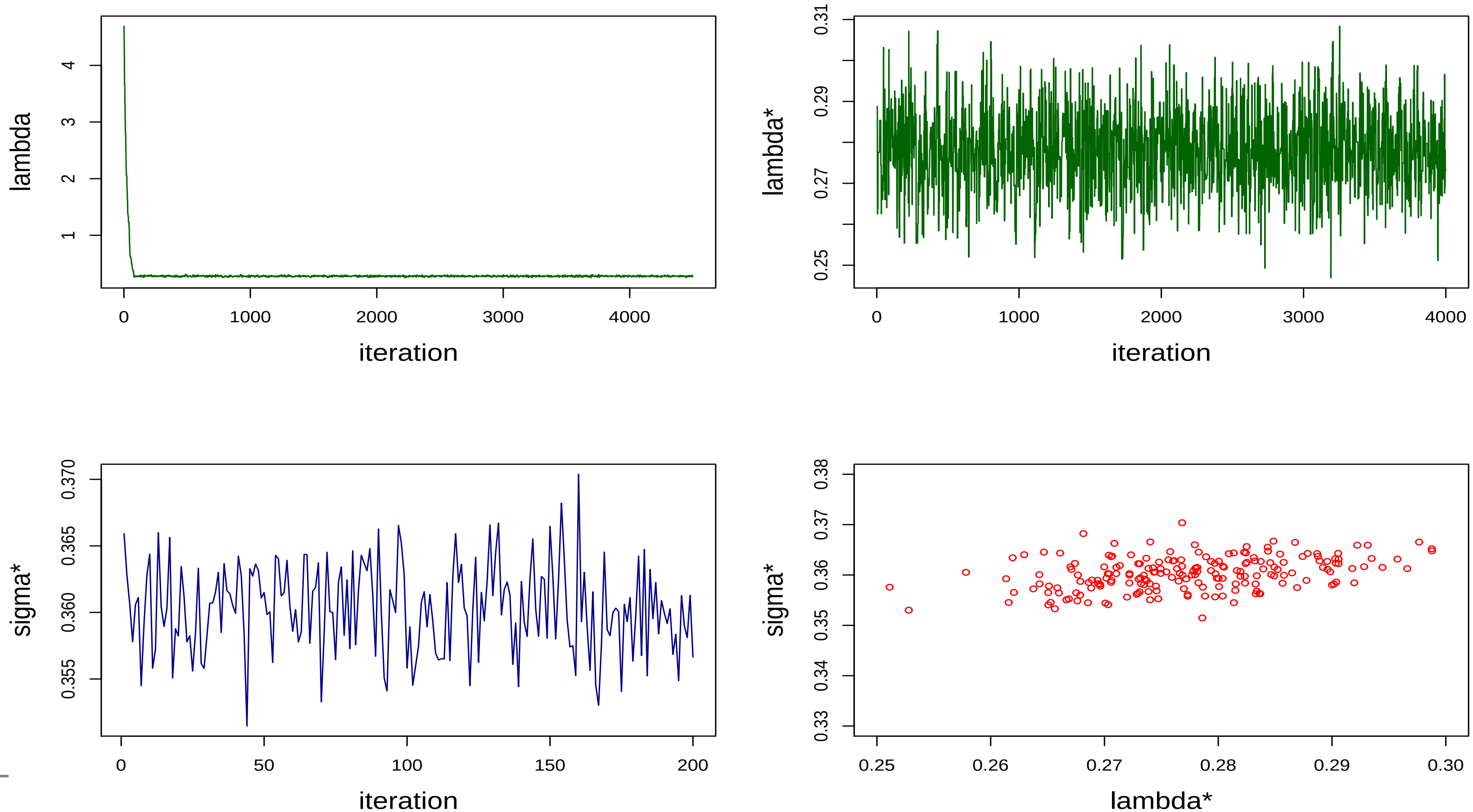
- Fix τ_y^2, τ_1^2, λ_1, τ_2^2 and λ_2, consider the joint posterior density:
$$p(x_{1:T}, \lambda, \sigma^2, a_1, a_2 | y_{1:T}).$$
- Sampling each component iteratively until convergence:
  - Sampling from $p(x_{1:T}|\lambda, \sigma^2, y_{1:T})$:
    - $p(\lambda|y_{1:T})$: Metropolis-Hasting algorithm
    - $p(\sigma^2|\lambda, y_{1:T}) \sim IG$: given $p(\sigma^2) \sim IG$
    - $p(x_{1:T}|\lambda, \sigma^2, y_{1:T})$: FFBS method

Ozone Study: MCMC Sampling (cont.)

- - Sampling from $p(a_1, a_2 | x_{1:T}, \lambda, \sigma^2, y_{1:T})$, given $p(a_1, a_2) \sim BN$.
- Interpolation at an unknown site s : Given α^s_{t-1}, y^s_{t-1},
  - Sampling from $p(\alpha^s_{jt} | \alpha^s_{j,t-1}, a_1, a_2, x_t, \lambda, \sigma^2, y_t), j = 1, 2;$
  - Sampling from $p(y^s_t | \alpha^s_{1t}, \alpha^s_{2t}, a_1, a_2, x_t, \lambda, \sigma^2, y_t)$.

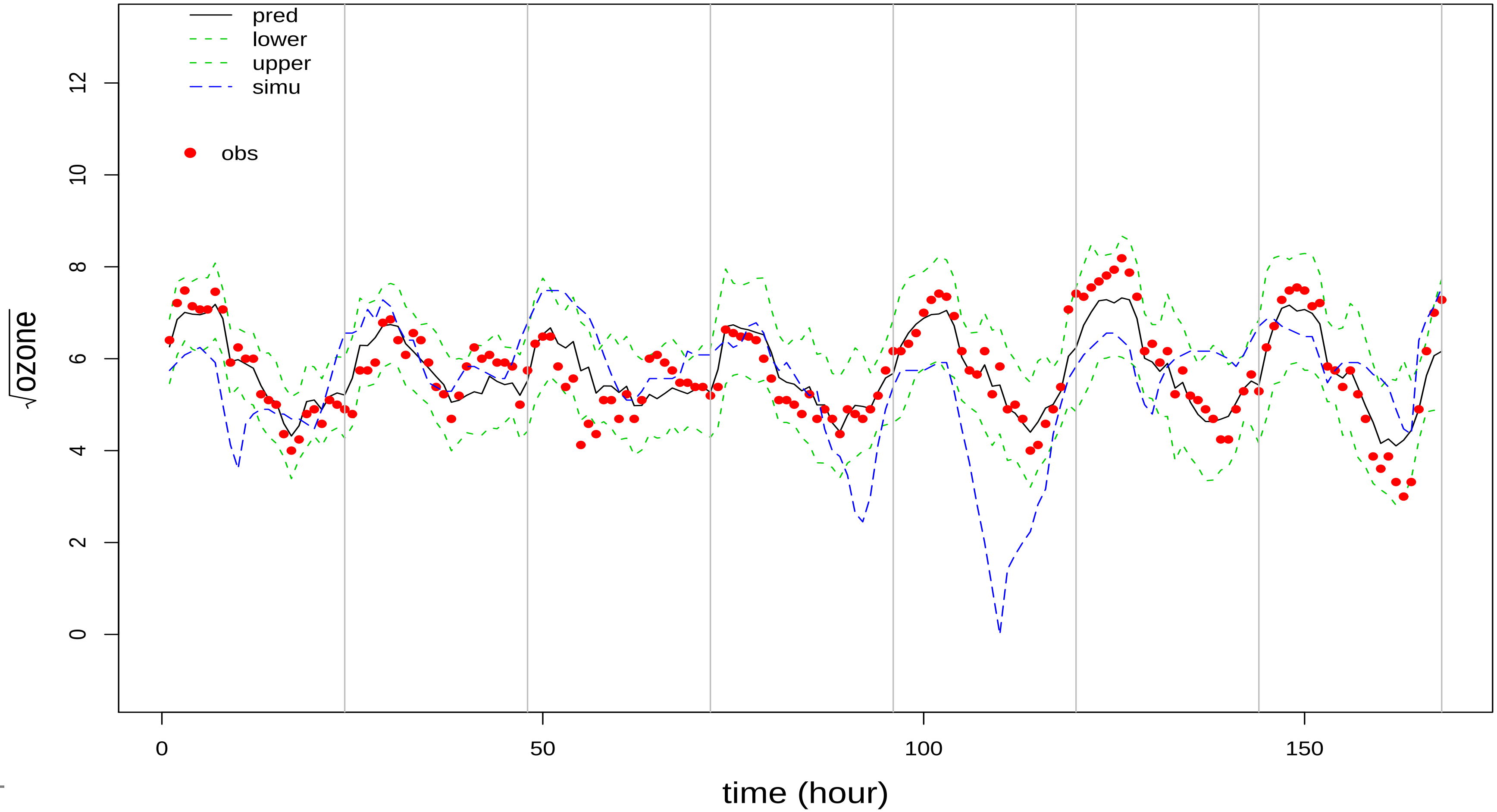
Ozone Study: MCMC Results

Figure 2. Upper panel: (left) λ by M-H ; (right) λ^* after 550 burn in period; Lower panel: (left) σ^{2*} sample from $p(\sigma^2|\lambda, y_{1:T})$; (right) λ^* v.s. σ^{2*}.



Ozone Study: Interpolation

Figure 3. Interpolation result at Center 2 in Cluster 2 sites, 1^{st} week during the study period.



Ozone Study: Model Assessment

- Cross-validation (CV) study:

Site	1	2	3	4	5	6	7	8	9	10
CP(%)	95.6	98.4	97.4	97.2	96.3	96.9	99.3	93.9	96.9	98.6

- Mean Square predicted error (MSPE):

Site	1	2	3	4	5	6	7	8	9	10
$\text{MSPE}_S(\times 10^4)$	1.5	8.9	0.6	1.0	0.7	0.5	2.9	1.0	1.3	0.6
$\text{MSPE}_p(\times 10^3)$	4.1	1.9	2.8	2.8	2.5	1.9	0.4	4.9	0.3	2.3
Ratio	3.74	4.79	2.12	3.70	2.93	2.55	81.13	2.10	36.96	2.48

Ozone Study: Conclusions

- Major achievements:
  - Develop C programs to solve the computational inefficiency
  - Implement the DLM in modelling conditionally Gaussian space-time fields

Ozone Study: Conclusions (cont.)

- Pros:
  - Bayesian hierarchical model + Gaussian framework
  - Good performance: better than MAQSIP?
- Cons:
  - Isotropic & stationary: Yes? No?
  - Computational burden
  - Estimating hyperparameters? Trial-and-error?

Prior setting for the hyperparameters

- Discount factor in the first order polynomial model
- Relationship between the discount factors and the hyperparameters
- How to set them?

GDLM.1.0 package

- Developed about one year ago
- Successfully in spatial interpolation in Dou, Le, & Zidek (2007)
- Use C and R codes to fast the computational speed
- Run in Linux, Unix and Windows
- URL: http://enviro.stat.ubc.ca
- Probs:
  - Curse of dimensionality
  - Computational burden

GDLM.1.0: one example

- Download the GDLM.1.0.tar.gz or GDLM.1.0.zip
- Install GDLM.1.0
- Run “Example1.txt” in C:/GDLM.1.0/DEMO
- Check the results – you can find some graphs in “Output”

Summary

- DLM and its modelling
- Ozone study using the DLM and GDLM.1.0 package
- Example on using GDLM.1.0 package

Reference

- West & Harrison, 1997
- Huerta, Sanso, & Stroud, 2004
- Dou, Le, & Zidek, 2007

Lab Exercises

- Experiment:
  - split into small teams
  - each team select one set of data
  - each team running GDLM.1.0 using daily, two-days data: record time
  - put your running results on the board

Lab Exercises

- Real data experiment:
  - select one real data set
  - set the DLM in the GDLM.1.0
  - run GDLM.1.0 and check your results