

Online regression analysis for streaming data

Carraro Enrico

PhD Course in Statistical Sciences
University of Padua

a.y. 2023/2024



Introduction

- Data arriving in batches
- Memory size
- Computational time
- Inference



Notation

- $\mathcal{D}_{ij} = \{y_{ij}, X_{ij}\}$: batch collected at t_j for unit i , $i = 1, \dots, m$, $j = 1, \dots$
- $\mathcal{D}_j = \{\mathcal{D}_{ij}\}_{i=1}^m$: batch collected at t_j for every unit
- $\mathcal{D}_{ib}^* = \{\mathcal{D}_{i1}, \dots, \mathcal{D}_{ib}\}$: cumulative dataset up to batch b for unit i
- $\mathcal{D}_b^* = \{\mathcal{D}_{ib}^*\}_{i=1}^m$: cumulative dataset up to batch b for every unit
- n_j : dimension of \mathcal{D}_j
- N_j : dimension of \mathcal{D}_j^*
- $\hat{\beta}_b$: estimator of beta computed only on batch b
- $\hat{\beta}_b^*$: estimator of beta computed on all the batches up to b
- $\tilde{\beta}_b$: renewable estimator computed on all the batches up to b



Independence

- Method proposed by Luo and Song (2020)
- Try to approximate Maximum Likelihood Estimator
- Based on score equation



Independence: procedure with 2 batches

- Find the MLE $\hat{\beta}_1$ such that $U_1(\mathcal{D}_1; \hat{\beta}_1) = 0$
- For the MLE $\hat{\beta}_2^*$ we have

$$U_1(\mathcal{D}_1; \hat{\beta}_2^*) + U_2(\mathcal{D}_2; \hat{\beta}_2^*) = 0$$

- Taylor expansion around $\hat{\beta}_1$

$$U_1(\mathcal{D}_1, \hat{\beta}_1) + J_1(\mathcal{D}_1, \hat{\beta}_1)(\hat{\beta}_1^* - \hat{\beta}_2^*) + U_2(\mathcal{D}_2, \hat{\beta}_2^*) + O_p\left(\|\hat{\beta}_1 - \hat{\beta}_2^*\|^2\right) = 0$$

where $J_1(\mathcal{D}_1, \hat{\beta}_1)$ is the observed information for batch 1

- Then the proposed estimator $\tilde{\beta}_2$ solves

$$J_1(\mathcal{D}_1, \hat{\beta}_1^*)(\hat{\beta}_1^* - \tilde{\beta}_2^*) + U_2(\mathcal{D}_2, \tilde{\beta}_2^*) = 0$$



Independence: procedure with b batches

- The estimator can be obtained via Newton-Raphson algorithm

$$\tilde{\beta}_b^{(r+1)} = \tilde{\beta}_b^{(r)} + \left\{ \tilde{J}_{b-1} + J_b \left(\mathcal{D}_b; \tilde{\beta}_{b-1} \right) \right\}^{-1} \tilde{U}_b^{(r)}.$$

where $\tilde{J}_b = \sum_{j=1}^b J_j(\tilde{\beta}_j)$.

- Dispersion parameter ϕ can be estimated by

$$\tilde{\phi}_b = \frac{N_{b-1} - p}{N_b - p} \tilde{\phi}_{b-1} + \frac{n_b - p}{N_b - p} \hat{\phi}_b$$

where $\hat{\phi}_b$ is the MLE of the single batch b .

- Then the distribution of $\tilde{\beta}_b$ is

$$\tilde{\beta}_b \sim N \left(\beta, \tilde{\phi}_b \tilde{J}_b^{-1} \right)$$



Dependence: State Space Model

- Method proposed by Luo and Song (2023) based on state-space models:

$$y_b = X_b\beta + Z_b\theta_b + \epsilon_b, \quad \epsilon_b \stackrel{\text{iid}}{\sim} N_{n_b}(0, \phi I)$$

$$\theta_{b+1} = B_b\theta_b + \xi_b, \quad \xi_b \stackrel{\text{iid}}{\sim} N_{n_b}(0, \delta I)$$

- B_b transition matrix ($B_b = \text{diag}(\rho_1, \dots, \rho_q)$) for an AR(1) process, where q is the dimension of Z_b
- Given θ_b , y_b is conditionally independent of the other y_j s
- The estimation of the fixed effects β is done through the marginal augmented log-likelihood, where the parameter θ_b is treated as missing data.



Dependence: State Space Model

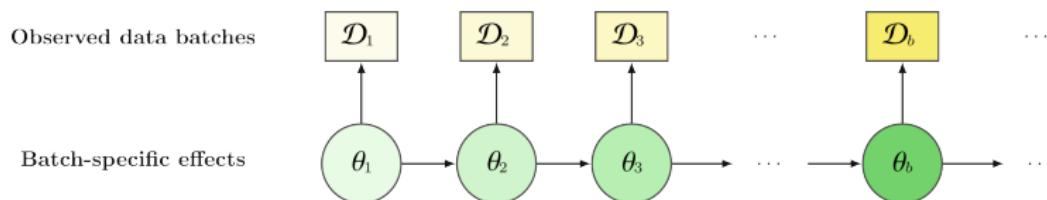


Figure: A structure for a hierarchical dynamic system. Data batches $\{\mathcal{D}_b, b \geq 1\}$ are generated from a state-space model with common fixed effect β and batch-specific latent effects θ_b governed by a Markov process.



Dependence: State Space Model

- $\hat{\theta}_b$ is estimated via Expectation-Maximization algorithm, where Kalman filter is used in the Expectation step
- $\tilde{\beta}_{b-1}$ is updated to $\tilde{\beta}_b$ solving the unbiased aggregated Kalman estimating equation
- The estimators of ϕ , ρ and δ are updated through a weighted mean between their values after batch $b - 1$ and their moments estimate for batch b
- the distribution of $\tilde{\beta}_b$ is

$$\tilde{\beta}_b \sim N\left(\beta, (\tilde{S}_b^\top \tilde{V}_b^{-1} \tilde{S}_b)^{-1}\right)$$

where both \tilde{V}_b and \tilde{S}_b depend of the estimated mean square error



Dependence: Weighted Generalized Estimating Equation

- Method proposed by Luo et al. (2023)
- Autoregressive structure where

$$\text{cor}(Y_{ij}, Y_{ik}) = \alpha^{|t_j - t_k|}, \quad \alpha \in (-1, 1)$$

- The starting point is to think for the estimator of β to the solution of the weighted generalized estimating equation

$$\psi_b^*(\beta, \alpha; \{\mathcal{D}_{ij}^*\}_{i=1}^m) = \sum_{i=1}^m D_i^\top \Sigma_i^{-1} W_b (y_i - \mu_i) = 0$$

where $D_i = \Delta_\beta \mu_i$, $W_b = \text{diag}\{W_{bj}\}_{j=1}^b$ is a weighting matrix with $W_{bj} = q^{t_b - t_j} I_{n_j}$ for $0 < q < 1$, $\Sigma = \text{cov}(y_i | X_i) \propto A_i^{1/2} R(\alpha) A_i^{1/2}$, $A_i = \text{diag}\{v(\mu_{i,kj})\}_{k,j=1}^{n_j, b}$, $v(\cdot)$ is a known variance function and $R(\alpha)$ is a working correlation matrix.



Dependence: Weighted Generalized Estimating Equation

- Avoid estimation of α by approximating $R^{-1}(\alpha)$
- Sparse structure

$$R^{-1}(\alpha) \approx \gamma_1 M_1 + \gamma_2 M_2$$

where γ_s , $s = 1, 2$ are unknown constants, $M_1 = I_{N_b}$ and M_2 is a matrix with 1 on the two main off-diagonals and 0 elsewhere.

- In this setting $\hat{\beta}_b^* = \arg \min_{\beta} Q_b^*(\beta)$ solves

$$S_b^* \left(\hat{\beta}_b^* \right)^T \left\{ V_b^* \left(\hat{\beta}_b^* \right) \right\}^{-1} U_b^* \left(\hat{\beta}_b^* \right) = 0$$

where

$$Q_b^*(\beta) = U_b^*(\beta)^T \{ V_b^*(\beta) \}^{-1} U_b^*(\beta),$$

$$U_b^*(\beta) = \sum_{i=1}^m \begin{pmatrix} D_i^T A_i^{-1/2} M_1 A_i^{-1/2} W_b(y_i - \mu_i) \\ D_i^T A_i^{-1/2} M_2 A_i^{-1/2} W_b(y_i - \mu_i) \end{pmatrix},$$

$V_b^*(\beta) = \sum_{i=1}^m U_b^*(\beta) U_b^*(\beta)^T$ is the sample covariance matrix of $U_b^*(\beta)$, and $S^*(\beta)$ is the negative gradient of $U^*(\beta)$



Dependence

- For just two batches of dimension 2 and 3 M_2 can be divided as follows

$$M_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} = \left(\begin{array}{cc|ccc} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{array} \right) = \begin{pmatrix} M_{21} & B_1 \\ B_2 & M_{22} \end{pmatrix},$$

- It is possible to show that $U_b^*(\beta)$ can be decomposed into estimating functions for within-batch dependencies through $U_{ij}(\beta)^{(2)}$ and between batch dependencies, through $U_{i,j,j+1}(\beta)$ and $U_{i,j+1,j}(\beta)$
- For the update only batch b and the last observation of the historical data are required



Dependence: Weighted Generalized Estimating Equation

- To obtain an estimator that does not depend on historical data it is necessary to take the first order expansion of the terms $U_{b-1}(\hat{\beta}_b^*)$ and $S_{b-1}(\hat{\beta}_b^*)$ around $\hat{\beta}_{b-1}^*$ to obtain \tilde{S}_{b-1} and \tilde{U}_{b-1} , whereas
$$\tilde{V}_b = \sum_{i=1}^m \tilde{U}_{ib} \tilde{U}_{ib}^\top$$
- The estimator $\tilde{\beta}$ is the solution to the incremental estimating equation

$$\tilde{S}_b^\top \tilde{V}_b^{-1} \tilde{U}_b = 0$$

- The solution can be found via Newton-Raphson

$$\tilde{\beta}_b^{(r+1)} = \tilde{\beta}_b^{(r)} + \left\{ \tilde{S}_b^{(r)\top} \left(\tilde{V}_b^{(r)} \right)^{-1} \tilde{S}_b^{(r)} \right\}^{-1} \tilde{S}_b^{(r)\top} \left(\tilde{V}_b^{(r)} \right)^{-1} \tilde{U}_b^{(r)}$$



Dependence: Weighted Generalized Estimating Equation

- The optimal value for the weighting parameter is

$$q_b^{\text{opt}} = \underset{q \in \mathcal{C}_q}{\operatorname{argmin}} U_b(\tilde{\beta}_b, q)^\top \left\{ V_b(\tilde{\beta}_b, q) \right\}^{-1} U_b(\tilde{\beta}_b, q).$$

where \mathcal{C}_q is the candidate set

- the distribution of $\tilde{\beta}_b$ is

$$\tilde{\beta}_b \sim N\left(\beta, (\tilde{S}_b^\top \tilde{V}_b^{-1} \tilde{S}_b)^{-1}\right)$$



References

-  Luo, Lan and Peter X.-K. Song (2020). "Renewable Estimation and Incremental Inference in Generalized Linear Models with Streaming Data Sets". en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.1, pp. 69–97.
-  — (2023). "Multivariate Online Regression Analysis with Heterogeneous Streaming Data". en. In: *Canadian Journal of Statistics* 51.1, pp. 111–133.
-  Luo, Lan, Jingshen Wang, and Emily C Hector (2023). "Statistical Inference for Streamed Longitudinal Data". In: *Biometrika* 110.4, pp. 841–858.