Second order Techniques for Learning Time-series With Structural Breaks

Presented by Team 2:
Sergazin Iskander
Zaharov Artem
Shatalov Andrey
Borschev Aristarkh
Keseli Timur

Paper’s author: Takayuki Osogami

Challenges & approaches in time series forecasting

Issues with time series:

- Non-stationarity of data
- Structural breaks

Gradual Data forgetting drawbacks:

- Constant forgetting rate
- Appropriate regularisation

Novel Techniques

Paper’s regulazer overcomes optimisation and transformation challenge with $O(n^2)$ complexity.

Proposed model:

“Following best hyper forgetting rate”;

Adaptively tunes hyperparameters;

Regularization is invariant to linear transformations.

Metamorphosis

Consider linear model: $f_{\boldsymbol{\theta}}(\mathbf{x}) \equiv \boldsymbol{\theta}^\top \mathbf{x}$

Weighted Mean Squared Error(WMSE):

$$L_t(\boldsymbol{\theta}) \equiv \frac{1}{2} \sum_{d=0}^{t-1} \gamma^d \left(f_{\boldsymbol{\theta}}(\mathbf{x}_{t-d}) - y_{t-d}\right)^2$$

Suggestions & Efficient Computations

Minimizer: $\theta_{t+1} = H_t^{-1} g_t$

where g_t is the negative gradient at the origin and H_t is the Hessian. They are calculated as:

$$\mathbf{g}_t \equiv \sum_{d=0}^{t-1} \gamma^d \, \mathbf{x}_{t-d} \, y_{t-d} = \gamma \, \mathbf{g}_{t-1} + \mathbf{x}_t \, y_t$$

$$\mathbf{H}_t \equiv \sum_{d=0}^{t-1} \gamma^d \, \mathbf{x}_{t-d} \, \mathbf{x}_{t-d}^\top = \gamma \, \mathbf{H}_{t-1} + \mathbf{x}_t \, \mathbf{x}_t^\top .$$

H_t^{-1} computed in $O(n^2)$ using Sherman-Morrison.

Advancements in Regularization Techniques

Shortcomings of standard L2:

$$\dot{L}_t(\boldsymbol{\theta}) = L_t(\boldsymbol{\theta}) + (\lambda/2)\,||\boldsymbol{\theta}||_2^2.$$

Cannot be computed recursively

Proposed regularizer and its advantages:

$$\tilde{L}_t(\boldsymbol{\theta}) = L_t(\boldsymbol{\theta}) + (\lambda/2)\,||\boldsymbol{\theta}||_{\mathbf{H}_t}^2 \qquad ||\theta||_{\mathbf{H}_t}^2 \equiv \theta^\top \mathbf{H}_t\,\theta.$$

Properties:

$O(n^2)$ complexity & invariance to linear transformations.

Following best Hyper Forgetting Rate

Train ~30 model with different γ and λ

↓

Choose the eta with the lowest CSE:

$$\mathrm{CSE}_t^{(i)}(\eta) \equiv \sum_{d=0}^{t-1} \eta^d \left(\hat{y}_{t-d}^{(i)} - y_{t-d}\right)^2 \tag{8}$$

$$= \eta\, \mathrm{CSE}_{t-1}^{(i)} + (\hat{y}_t^{(i)} - y_t)^2. \tag{9}$$

↓

Model corresponding to the lowest CSE is chosen to be the best for t+1 step.

Experiments to answer the following questions

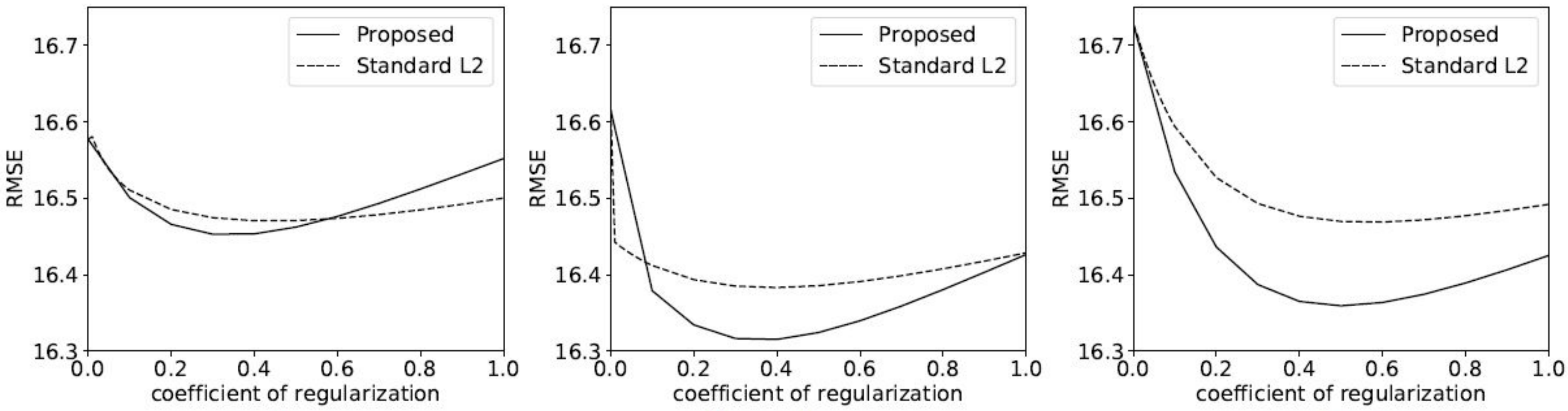
1. How does the proposed regularization compare against the L2 regularization?
2. Can the proposed algorithm adaptively tune hyperparameters?
3. How does the proposed approach compare against existing methods for predicting non-stationary time series?

Experiment 1 (L2 vs paper's regularization)

Overall results are quite promising:

- Effectiveness depends on particular data and our model occasionally outperforms L2 regularisation
- Our algorithm works in $O(n^2)$ time with the expected effect of regularisation

Experiment 1: performance comparison



(a) Order 2 (b) Order 3 (c) Order 4

Experiment 2: adaptiveness of the model

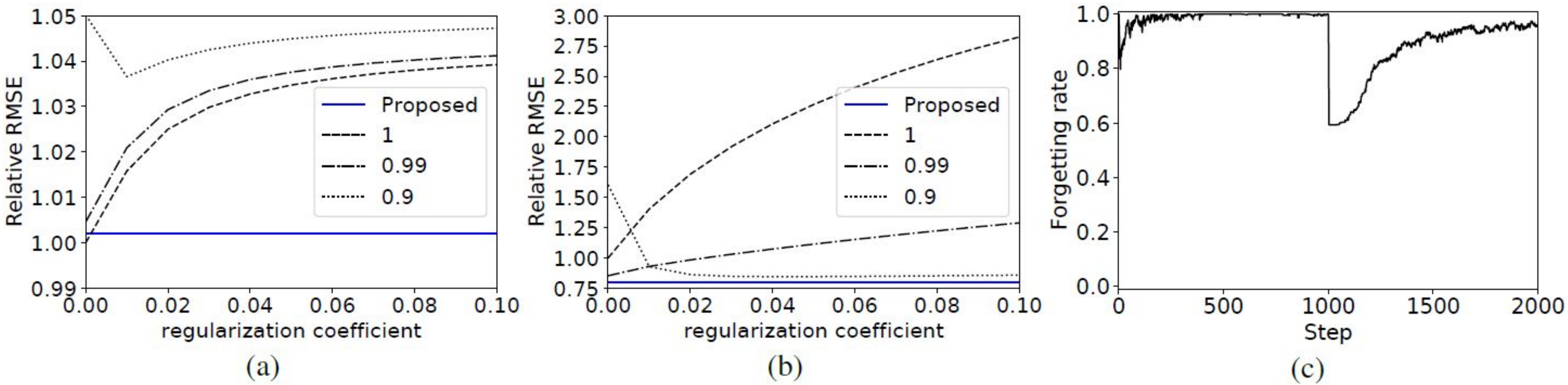
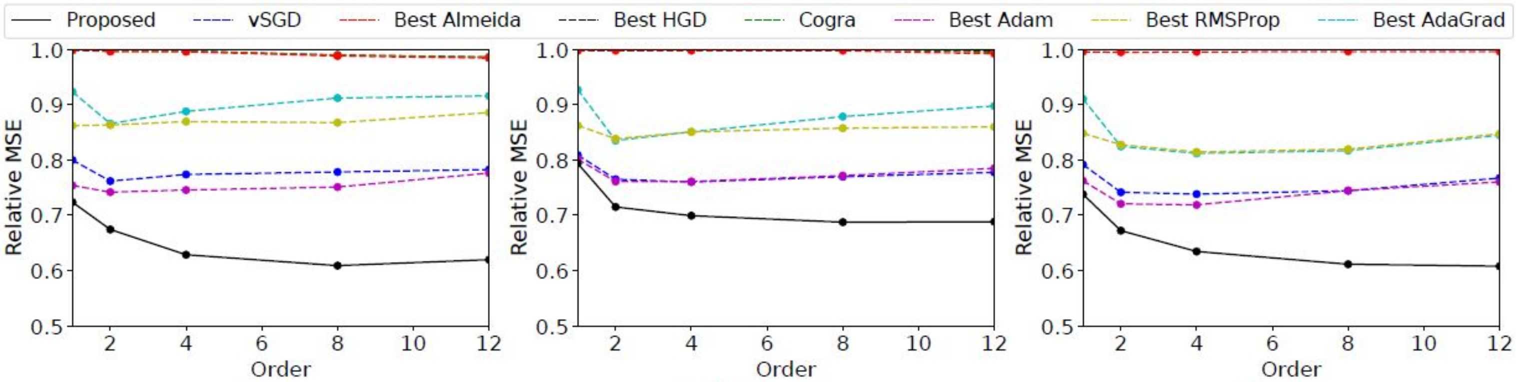


Figure 2: The results of the experiments with the synthetic time-series with a change point at $t = 1,000$ (averaged over 30 runs): RMSE during the 100 steps immediately (a) before and (b) after the change point; (c) the forgetting rate used by Algorithm 1 at each step. The regularization-coefficient used by Algorithm 1 at each step is shown in Figure 7 of Osogami (2020).

Experiment 3: compassion with other models



(a) SPX (b) Nikkei 225 (c) DAX

Figure 3: MSE of the predicted absolute daily return of various financial indices (as indicated in each column), relative to the naïve prediction of no change. Algorithm 1 (solid curve) is compared against seven baselines with the optimal choice of hyperparameters (dashed curves). Results with FTSE 100 and SSEC are shown in Figure 12 of Osogami (2020). Error bars are shown in Figure 8-12 of Osogami (2020).

Conclusion

1. Proposed regularisation outperforms L2

2. Algorithm 1 can tune hyperparameters after a changepoint

3. Algorithm 1 outperforms baselines in predicting financial time series.

Extra outcome

- Compared the performance of two SOTA models with the proposed algorithm.
- Used an OIL Brent dataset (1987 - 2022).

Algorithm	RMSE	Runtime (20 iters in sec)
Proposed Algorithm	2.578	0.04
ARIMA	2.214	149
Prophet	12.589	70

Thank you for your attention!