Bayesian learning for limit-order book price prediction

Martin Magris, Mostafa Shabani, Alexandros Iosifidis

Department of Electrical and Computer Engineering
Machine Learning and Signal Processing Unit
Aarhus University, Denmark

Open Workshop "AI for Financial Portfolio Management"

June 30th, 2023

Introduction I

Econometrics:

- Applied statistics & probability theory & stochastics, for the study of economic phenomena (Ragnar, 1933).
- The probabilistic dimension is an innate and essential element in modeling.

Machine Learning (ML):

- Flexible, scalable and with proven predictive gains. (e.g. Dixon, Halperin, and Bilokon, 2020) gained much attention (e.g. Varian, 2014).
- Fit well the complexity of modern financial markets.

Introduction II

Yet:

- Traditional ML methods lack of the probabilistic dimension typical of econometric modelling.
- Business and financial applications are high-risk domains where quantifying the uncertainty of estimates and predictions is of utmost importance (Salinas et al., 2020; Makridakis, Hogarth, and Gaba, 2009).
- Enabling a probabilistic dimension in ML extends the set of available tools for e.g. model diagnostic, inference (Dixon, Halperin, and Bilokon, 2020).

Introduction III

Main point:

- Bayesian Deep Learning (DL) constitute a natural direction.
- Narrow the gap between the highly-probabilistic econometric practice, and the flexible non-parametric and highly non-linear and ML rationale.

This paper:

- First Bayesian DL econometric application in predicting mid-price movements in Limit Order Book (LOB) markets.
- Bayesian version of the Temporal Attention-augmented Bilinear network for a financial times-series classification.
- Investigate advantages/insights provided by the Bayesian approach.

Model, Bayesian Neural Network (BNN) I

- A BNN is any stochastic Artificial Neural Network (ANN) trained using Bayesian inference:

$$\begin{aligned} \boldsymbol{\theta} &\sim p(\boldsymbol{\theta}), \\ \boldsymbol{y} &= NN_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\epsilon}, \\ p(\boldsymbol{\theta}\,|\mathcal{D}) &= \frac{p(\mathcal{D}_{\boldsymbol{y}}|\mathcal{D}_{\boldsymbol{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} p(\mathcal{D}_{\boldsymbol{y}}|\mathcal{D}_{\boldsymbol{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta})} \propto p(\mathcal{D}_{\boldsymbol{y}}|\mathcal{D}_{\boldsymbol{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta}). \end{aligned}$$

- From the posterior distribution, the forecast's uncertainty is quantified as the marginal probability distribution of the output $\boldsymbol{y}_i$ for a certain input $\boldsymbol{x}_i$, through the predictive distribution:

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \mathcal{D}) = \int p(\boldsymbol{y}_i|\boldsymbol{x}_i, \theta)p(\theta|\mathcal{D})d\theta.$$

Model, BNN II

- For classification, the average model prediction approximates the relative probability of each class (c):

$$\hat{p}_{ic} \approx 1/N_s \sum_{n=1}^{N_s} p\left(y_i = c|\boldsymbol{x}_i, \theta^{(n)}\right), \qquad \theta^{(n)} \sim p(\theta|\mathcal{D}),$$

  with $n = 1, \ldots, N_s$.

- If the cost of giving a false positive is equal across all the classes, the final label is that of the most likely class:

$$\hat{y}_i = \max_c \hat{p}_{ic}.$$

Model, TABL I

The Temporal Attention-augmented Bilinear Network (TABL) (Tran et al., 2019) is a light-weight DL model, suited for multidimensional time-series forecasting:

$\rightarrow$ It maps a $D \times T$ input matrix $\mathbf{X}$ onto a $D' \times T'$ output matrix $\mathbf{Y}$.

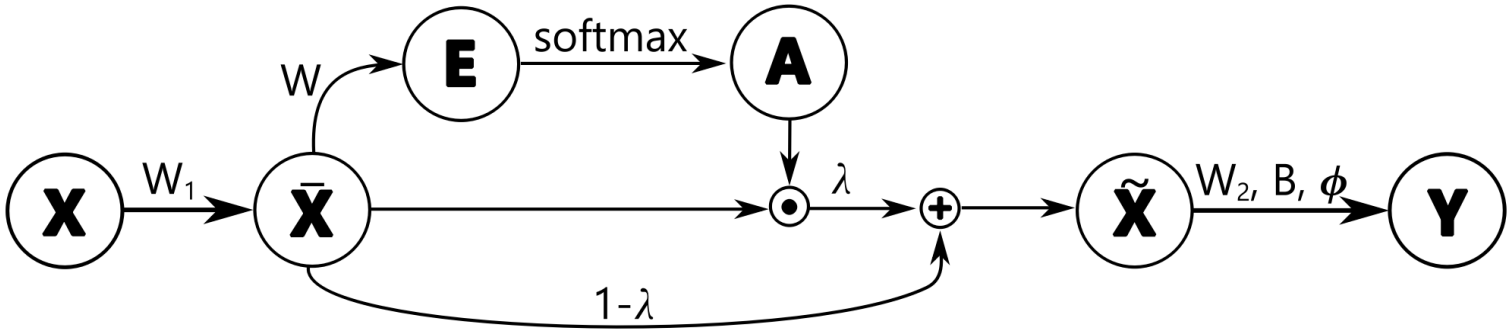
How:

i Operates a projection of the temporal dimension of the input to a $D' \times T$ feature space modeling the dependence on the first mode while preserving the temporal order of the features.

ii Learns the relative importance of the temporal instances producing an attention mask.

iii A learnable scalar drives the mixture of the temporal and non-temporal features.

Model, TABL II



This is achieved by:

$$\bar{\mathbf{X}} = \mathbf{W}_1 \mathbf{X}$$
$$\mathbf{E} = \bar{\mathbf{X}} \mathbf{W}$$
$$a_{ij} = \exp(e_{ij}) / \sum_{k=1}^{T} \exp(e_{ik})$$
$$\tilde{\mathbf{X}} = \lambda(\bar{\mathbf{X}} \odot \mathbf{A}) + (1 - \lambda)\bar{\mathbf{X}}$$
$$\mathbf{Y} = \phi(\tilde{\mathbf{X}} \mathbf{W}_2 + \mathbf{B})$$

Bayesian TABL I

- Parameter vector $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{B}, \lambda\}$.
- Tackle the Bayesian inference problem under a (mean-field, fixed) Variational Inference framework:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{I}/\alpha), \qquad q(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathrm{diag}\left(\boldsymbol{\sigma}^2\right)\right),$$

  where $\alpha > 0$, $\boldsymbol{\mu} \in \mathbb{R}^P$, $\boldsymbol{\sigma}^2 \in \mathbb{R}^P$, P the number of the parameters.
- The variational parameters $\left(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\right)$ are obtained by optimizing:

$$\mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\right) = \sum_{i=1}^{N} \mathbb{E}_q \left[\log p(\mathcal{D}|\boldsymbol{\theta})\right] + \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right]. \tag{1}$$

- Maximized with the gradient-based optimization:

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \rho_t \nabla_{\boldsymbol{\mu}} \mathcal{L}_t \quad \text{and} \quad \boldsymbol{\sigma}_{t+1} = \boldsymbol{\sigma}_t + \delta_t \nabla_{\boldsymbol{\sigma}} \mathcal{L}_t.$$

Bayesian TABL II, NGVI

We use the natural-gradient VI approach of Khan and Lin, 2017:

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \beta_t(g(\boldsymbol{\theta}_t) + \tilde{\alpha}\boldsymbol{\mu}_t)/(\mathbf{s}_{t+1} + \tilde{\alpha}),$$
$$\mathbf{s}_{t+1} = (1-\beta_t)\mathbf{s}_t + \beta_t \text{diag}(\mathbf{H}(\boldsymbol{\theta}_t)).$$

Where,

- The objective (1) is expressed in terms of the standard MLE objective $f(\boldsymbol{\theta}) = -1/N \sum_{i=1}^{N} \log p(\mathcal{D}_i|\boldsymbol{\theta})$.
- The gradients of (1) with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ now involve the gradient $g(\boldsymbol{\theta})$ and Hessian $\mathbf{H}(\boldsymbol{\theta})$ of $f(\boldsymbol{\theta})$.
- $\boldsymbol{\theta}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_t, \text{diag}\left(\boldsymbol{\sigma}_t^2\right)\right), \quad \boldsymbol{\sigma}_t^2 = [N(\mathbf{s}_t + \tilde{\alpha})]^{-1}$, $\tilde{\alpha} = \alpha/N$.

Bayesian TABL III, VOGN

- Non-negativity of the Hessian is granted by the following approximation:

$$\nabla^2_{\theta_j\theta_j} f(\boldsymbol{\theta}) \approx \frac{1}{M} \sum_{i \in \mathcal{M}} \left[\nabla_{\theta_j} f_i(\theta)\right]^2 := \hat{h}_j(\boldsymbol{\theta}).$$

- Then the VOGN update for $\mathbf{s}_t$ reads:

$$\mathbf{s}_{t+1} = (1 - \beta_t)\mathbf{s}_t + \beta_t \hat{\boldsymbol{h}}(\boldsymbol{\theta}_t).$$

- Good empirical performance, is of practical feasibility on large datasets and relatively simple to implement over existing libraries (Osawa et al., 2019).

Data I

Finnish LOB dataset (Ntakaris et al., 2019):

- NASDAQ Nordic Helsinki exchange from June 1 to June 14, 2010 ($\approx$ 4.5 million events).
- 144-dimensional feature vectors, at each epoch.
- 75%-10%-15% split for training, validation and test sets, same setup as Tsantekidis et al., 2017.

Train and compare VOGN, ADAM, MC Dropout, and SGD.

Data II

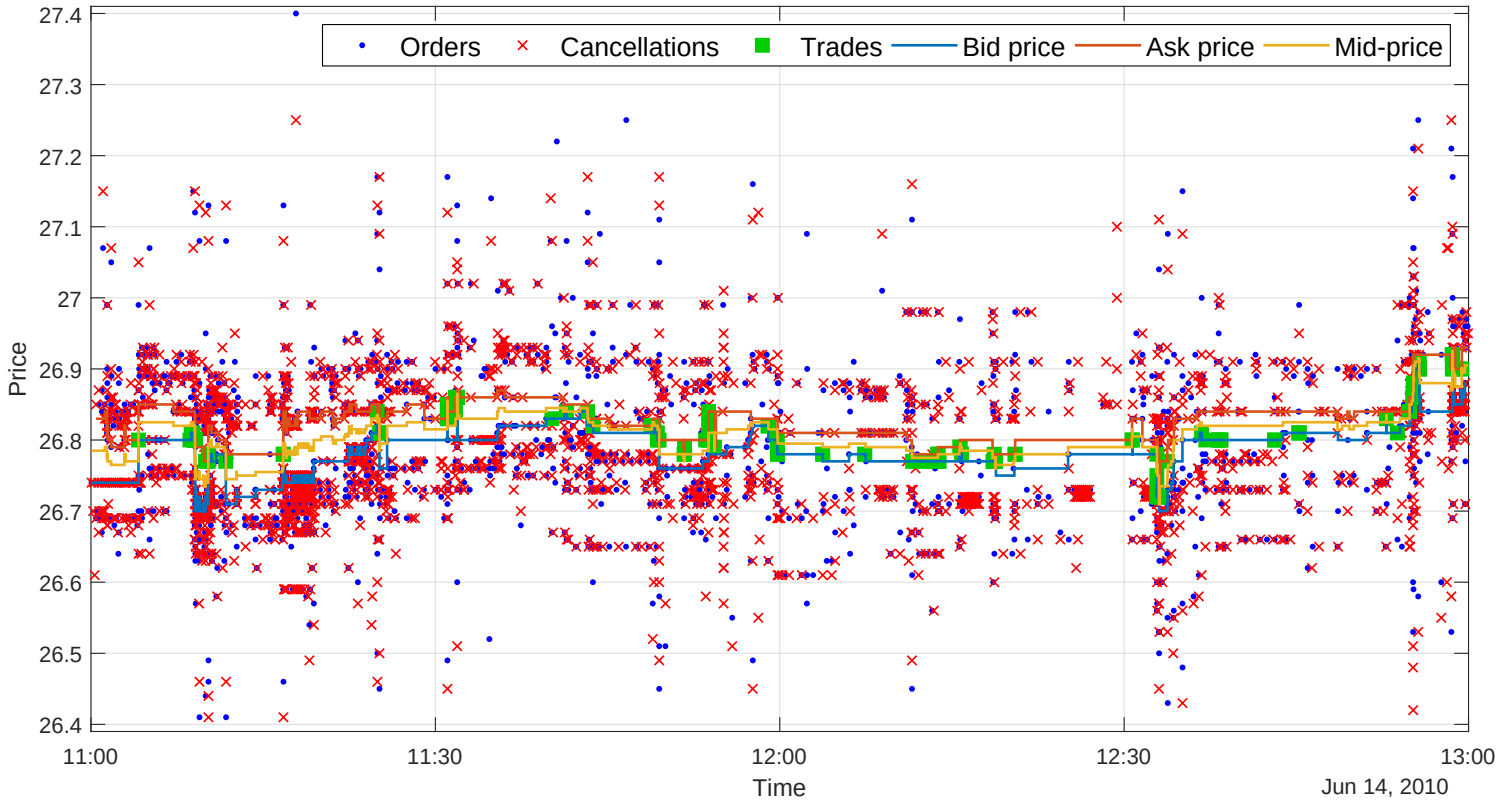


Figure: Snapshot of the LOB data (ticker: KESKO B.)

Learning Curves I

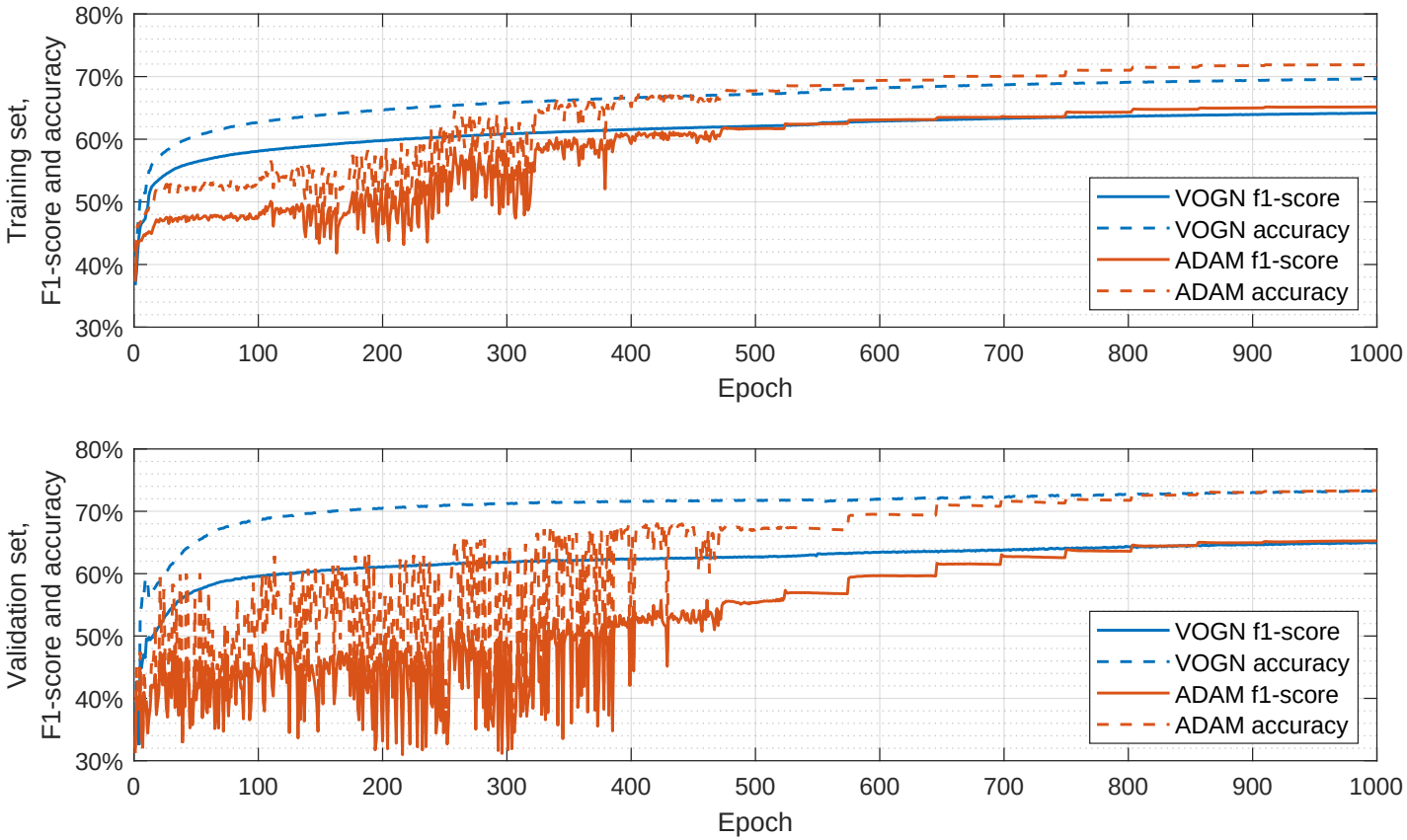


Figure: Learning curves for VOGN and ADAM on the training and validation sets (upper and lower panel, respectively).

Learning Curves II

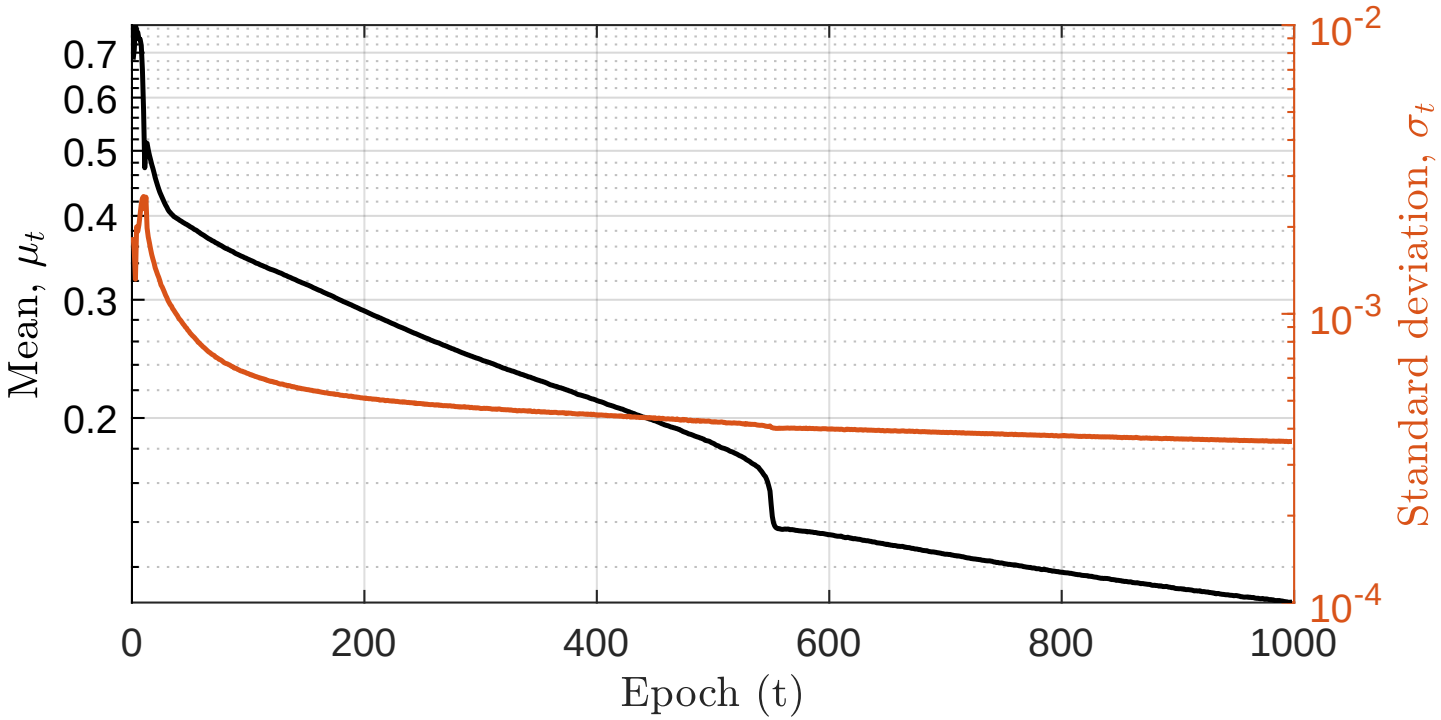


Figure: Learning of the variational parameters for TABL's mixing coefficient λ.

Predictive distribution, interpreting predictive probabilities

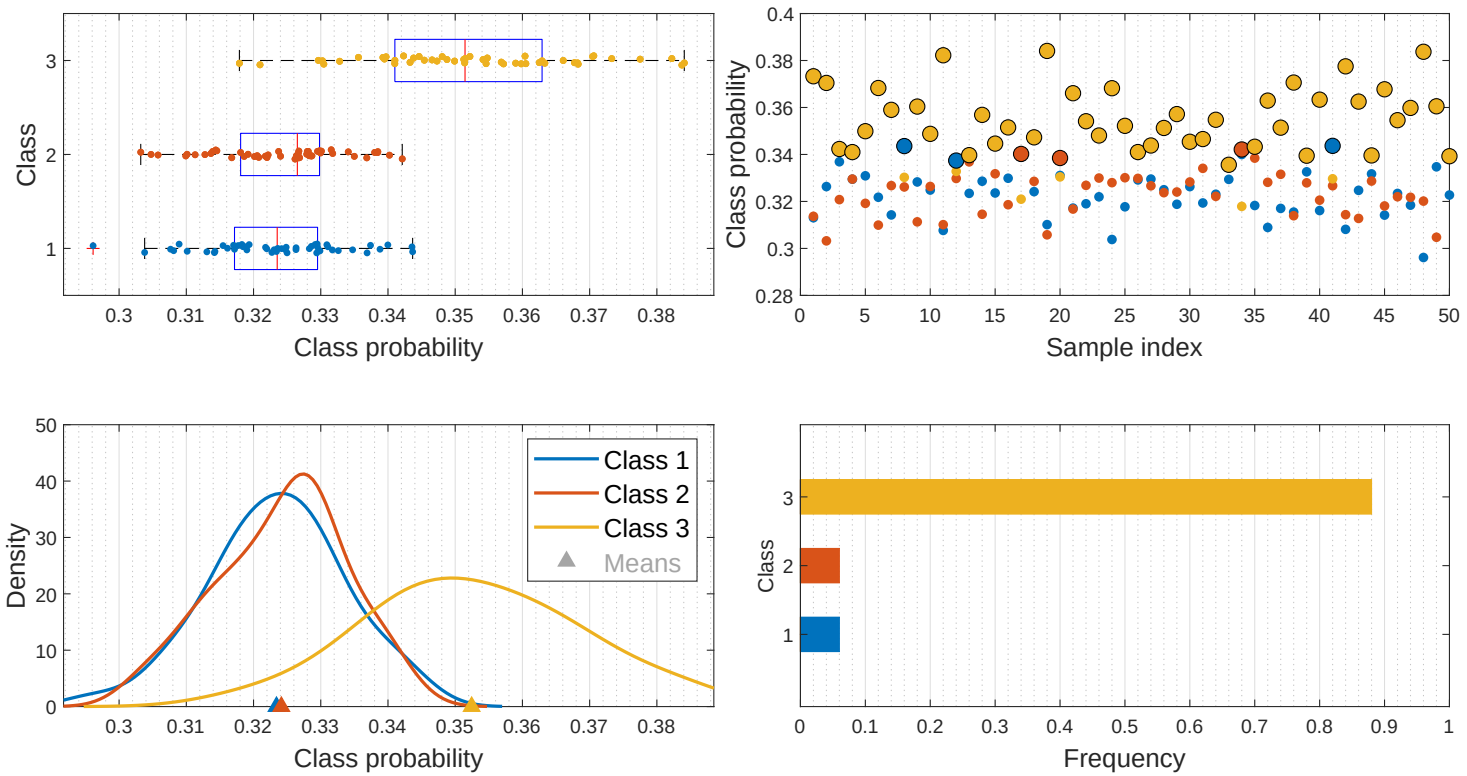
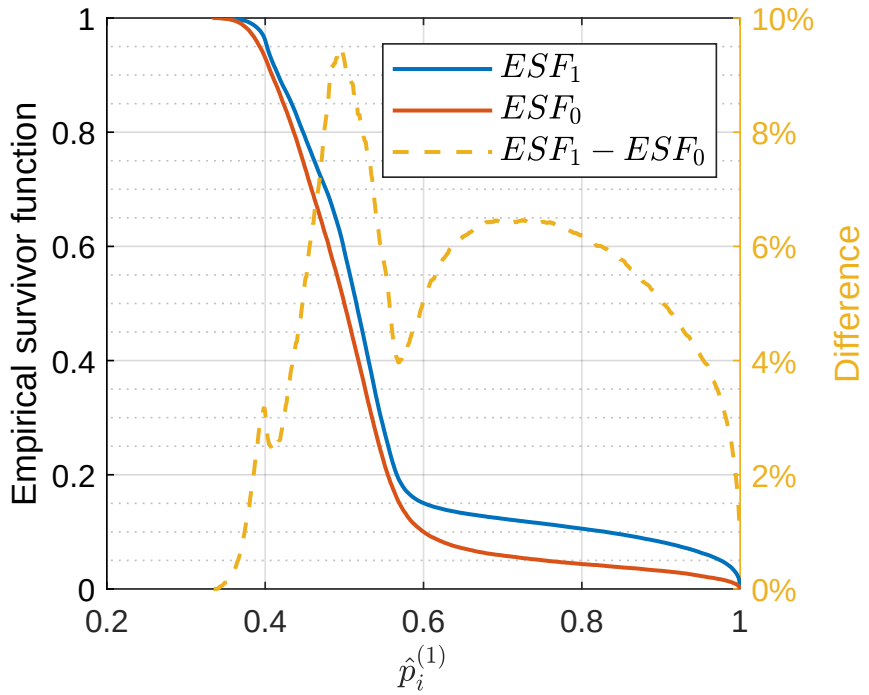


Figure: Class-probabilities and forecasts for a typical test example. Top-left, panel: box-plots of class-probabilities. Bottom-left, panel: kernel density estimates and means of class-probabilities. Top-right panel: class-probabilities per class, highlighting those of maximum probability. Bottom-right panel: histogram of forecasts' labels.

Predictive probability for the maximum-probability class

- Low-to-mild confidence is quite common, $\hat{p}_i^{(1)} > 0.6 \approx 10\% - 15\%$.
- High-confidence is even rarer, $\hat{p}_i^{(1)} > 0.9 \approx 7\% - 9\%$.
- ESFs do not cross, and the difference is positive:
  $\rightarrow$ for the same (or greater) level of confidence, the number of correctly classified samples is on average 5% higher for the correctly classified samples than the miss-classified ones.



Distribution of the scores I

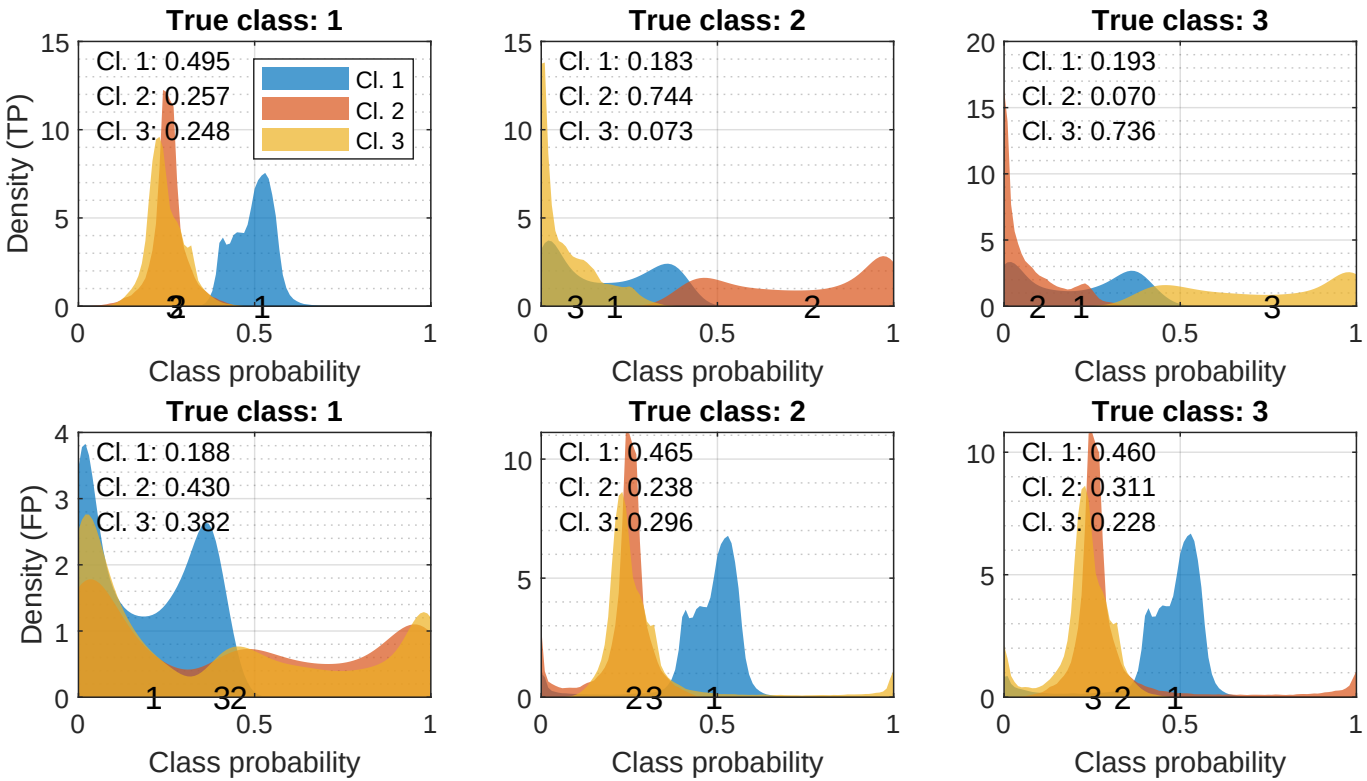


Figure: Distribution of VOGN's predictive probabilities. Top row: distribution of the class-probabilities for correctly-classified labels. Bottom row: distribution of the class-probabilities for miss-classified labels.

Distribution of the scores II

True positives:

- When the model is correct, uncertainty on classes 2 and 3 is much lower than the stationary-price case from the others.
- When the model is correct about class-1 assignments, its confidence is somewhat lower and the densities of the scores for whatever change in price direction generally overlap.

$\rightarrow$ Existence of patterns that are truly indicative of the direction of the price movement, driving predictive probabilities close to one.

Distribution of the scores III

False Positives:

- Bias towards the majority class for correct labels 2 and 3
- The same distribution on class 1 TPs almost identically replicates on class 2 and 3 FPs: the model interprets certain patterns in the features as remarkably non-indicative of the true class 2 and 3 labels causing an over- flow of low scores for both of them.
- By excluding a relevant probability mass on classes 2 and 3, this is reversed in class 1, following a distribution being very close to that observed on class 1 TPs.

Distribution of the scores III

False Positives:

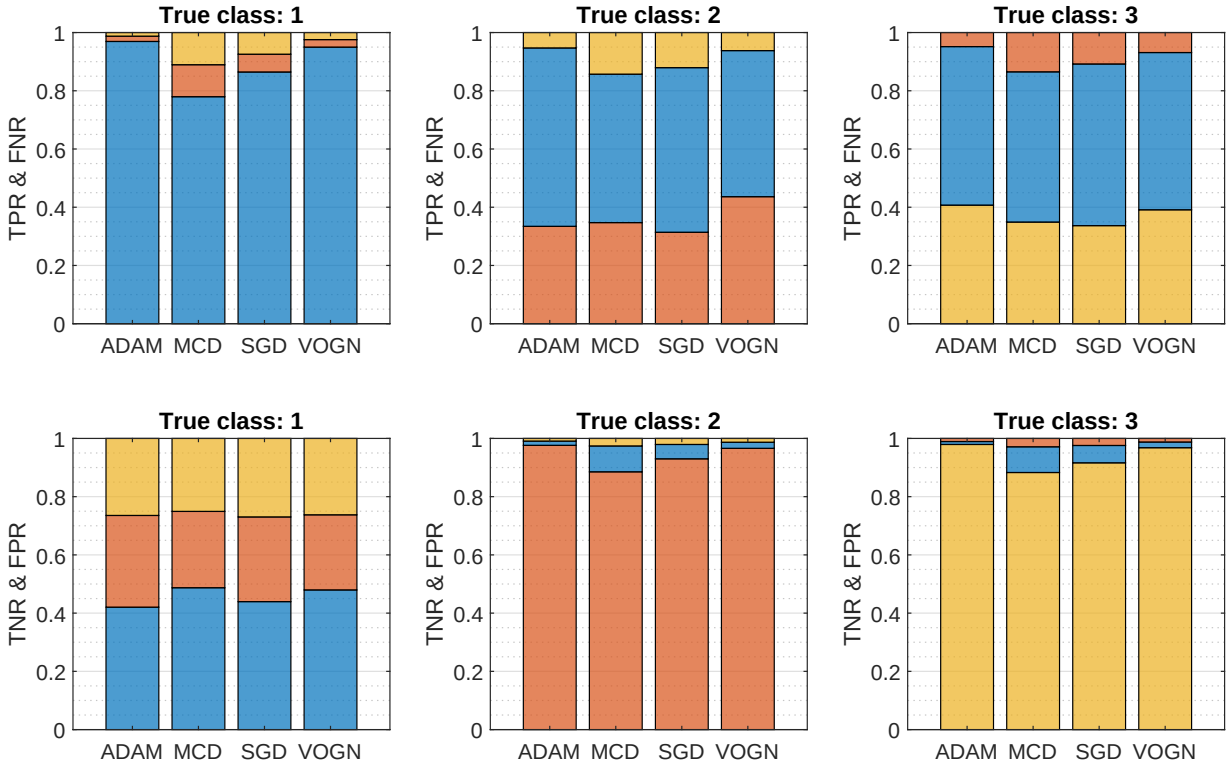
- Bias towards the majority class for correct labels 2 and 3
- The same distribution on class 1 TPs almost identically replicates on class 2 and 3 FPs: the model interprets certain patterns in the features as remarkably non-indicative of the true class 2 and 3 labels causing an over- flow of low scores for both of them.
- By excluding a relevant probability mass on classes 2 and 3, this is reversed in class 1, following a distribution being very close to that observed on class 1 TPs.

$\rightarrow$ This suggests that the model well-distinguishes patterns indicative of classes 2 and 3 and, when these are absent, class 1 classification is enforced.

- Patterns indicative of classes 3 and 2 are causing false positives in classes 2 and 3: typical features for classes 3 and 2 are observed for mid-prices eventually moving in the opposite direction.

(Distribution of the scores IV)

Similar conclusions also supported by further joint analyses on TP, TN, FP, FN on a class-by-class basis.

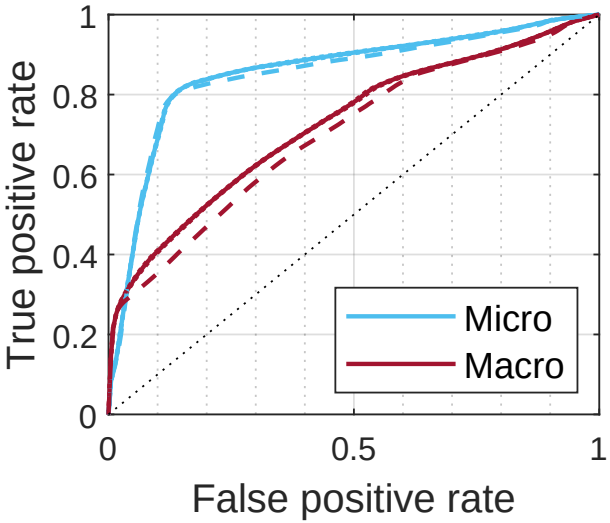


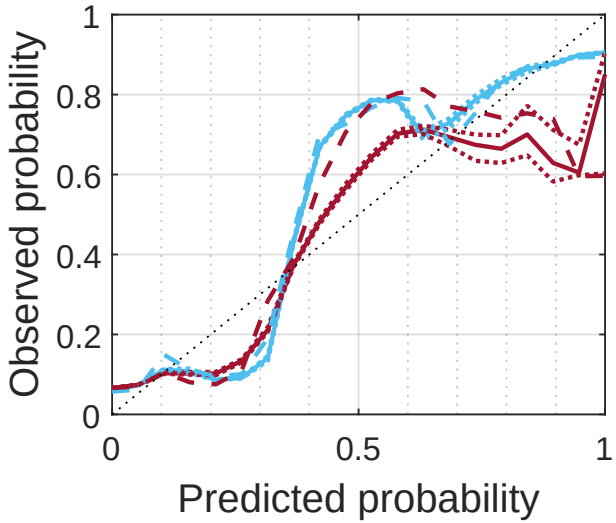
$\rightarrow$ check the paper for an extensive discussion

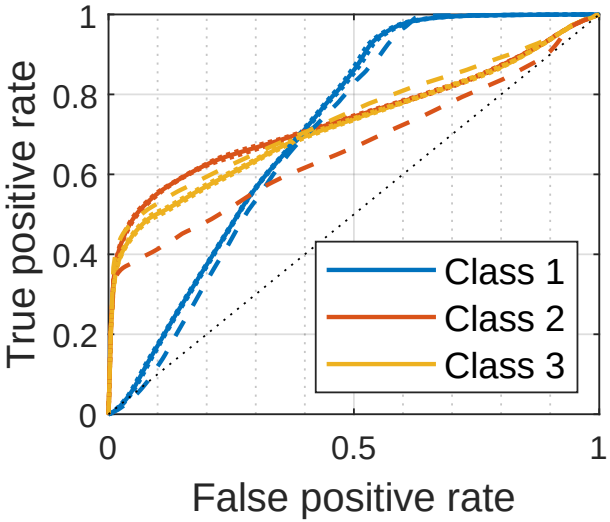
Performance measures

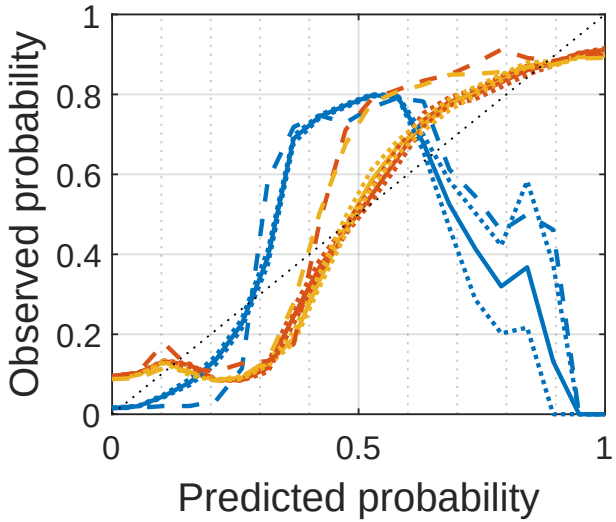
	Any	Precision		Recall		f1-score	
	Micro	Macro	Weighted	Macro	Weighted	Macro	Weighted
VOGN sample-by-sample							
Mean	0.774	0.736	0.763	0.592	0.774	0.636	0.751
Median	0.774	0.736	0.763	0.592	0.774	0.636	0.751
Min	0.772	0.730	0.761	0.589	0.772	0.633	0.749
Max	0.776	0.743	0.766	0.596	0.776	0.638	0.752
VOGN based on forecasts' function							
Mean($\hat{Y}_i$)	0.772	0.731	0.761	0.591	0.772	0.633	0.749
Median($\hat{Y}_i$)	0.774	0.736	0.763	0.592	0.774	0.636	0.751
Mode($\hat{Y}_i$)	0.774	0.737	0.763	0.592	0.774	0.636	0.751
VOGN predictive distribution							
$\hat{Y}_{pred}$	**0.774**	0.737	0.763	**0.592**	**0.774**	**0.636**	**0.751**
$\hat{Y}_{pred}$ (med.)	0.774	0.737	0.763	0.592	0.774	0.636	0.751
Other optimizers							
ADAM	0.772	**0.767**	**0.770**	0.570	0.772	0.619	0.741
MCD (mea.)	0.581	0.450	0.598	0.460	0.581	0.454	0.588
MCD (pred.)	0.638	0.500	0.630	0.492	0.638	0.495	0.634
SGD	0.687	0.556	0.660	0.505	0.687	0.522	0.667
Differences							
Min - ADAM	0.0%	-3.8%	-0.9%	1.8%	0.0%	1.3%	0.7%
$\hat{Y}_{pred}$ - ADAM	0.2%	-3.1%	-0.7%	2.2%	0.2%	1.6%	0.9%
$\hat{Y}_{pred}$ - MCD (pred.)	13.6%	23.7%	13.3%	10.0%	13.6%	14.0%	11.7%
$\hat{Y}_{pred}$ - SGD	8.7%	18.1%	10.4%	8.7%	8.7%	11.4%	8.3%

ROC and Calibration curves I







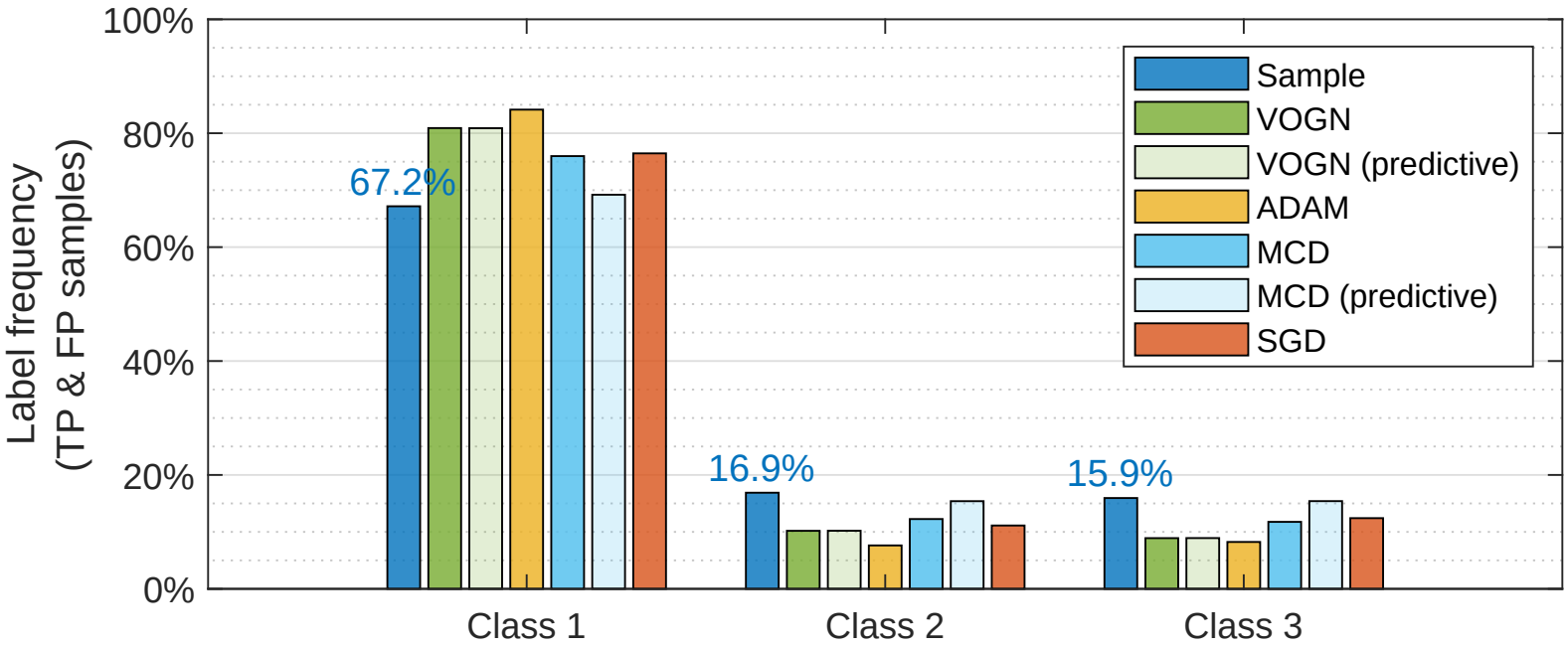


ROC and Calibration curves II

	Single-class task			Multi-class task	
	Class 1	Class 2	Class 3	Micro	Macro
Area under the ROC curve					
VOGN (pred.)	**0.716**	**0.739**	0.722	**0.858**	**0.726**
ADAM	0.697	0.665	**0.742**	0.851	0.702
MCD (pred.)	0.672	0.649	0.657	0.770	0.659
SGD	0.691	0.660	0.656	0.790	0.669
Expected calibration error					
VOGN (pred.)	-0.107	-0.014	**-0.016**	0.035	-0.046
ADAM	-0.104	0.043	0.033	0.040	**-0.009**
MCD (pred.)	**-0.051**	**-0.016**	-0.044	**0.021**	-0.039
SGD	0.153	-0.081	-0.072	-0.021	-0.032
Expected calibration distance					
VOGN (pred.)	0.144	**0.008**	**0.009**	**0.018**	**0.018**
ADAM	0.146	0.021	0.018	0.018	0.030
MCD (pred.)	0.181	0.028	0.012	0.019	0.023
SGD	**0.039**	0.028	0.027	0.024	0.018

Table: Measures related to ROC and Calibration curves

Class frequencies



- Differences are explainable under deeper analyses.
- E.g. MCD alignment is not indicative of a genuine satisfactory performance: for class 1 (classes 2 or 3) this arises from a lower (comparable) TPR and comparable (lower) FPR with respect to the other optimizers.

Conclusion

- A first econometric time-series application with BNN for mid-price movement prediction.
- Promising results showing that Bayesian methods in deep learning are feasible, attractive, and useful for economic applications.
- Discuss how to make use and interpret predictive probabilities, providing insights on their implication in the decision process.
- Analyses on the scores' distribution and TPs, TNs, FPs, FNs allow to grasp important insights into models' learning.
- Optimizer-specific analyses and cross-comparisons (VOGN and ADAM are aligned, VOGN is slightly superior but enables a probabilistic dimension for DL models).

Cite us

Full paper available at:
https://onlinelibrary.wiley.com/doi/full/10.1002/for.2955

Martin Magris, Mostafa Shabani, and Alexandros Iosifidis (2023). "Bayesian bilinear neural network for predicting the mid-price dynamics in limit-order book markets". In: *Journal of Forecasting*

References I

- Ragnar, Frisch (1933). "Editor's Note". In: *Econometrica* 1.1, pp. 1–4.
- Makridakis, Spyros, Robin M Hogarth, and Anil Gaba (2009). "Forecasting and uncertainty in the economic and business world". In: *International Journal of Forecasting* 25.4, pp. 794–812.
- Varian, Hal R (2014). "Big data: New tricks for econometrics". In: *Journal of Economic Perspectives* 28.2, pp. 3–28.
- Khan, Mohammad Emtiyaz and W. Lin (2017). "Conjugate-computation variational inference: converting variational inference in non-conjugate models to inferences in conjugate models". In: *20th International Conference on Artificial Intelligence and Statistics*, pp. 878–887.
- Tsantekidis, Avraam et al. (2017). "Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks". In: *19th IEEE Conference on Business Informatics*, pp. 7–12.

References II

- Ntakaris, Adamantios et al. (2019). "Feature Engineering for Mid-Price Prediction With Deep Learning". In: *IEEE Access* 7, pp. 82390–82412.
- Osawa, Kazuki et al. (2019). "Practical Deep Learning with Bayesian Principles". In: *Advances in Neural Information Processing Systems*. Vol. 32, pp. 1–13.
- Tran, Dat Thanh et al. (2019). "Temporal Attention-Augmented Bilinear Network for Financial Time-Series Data Analysis". In: *IEEE Transactions on Neural Networks and Learning Systems* 30.5, pp. 1407–1418.
- Dixon, Matthew F, Igor Halperin, and Paul Bilokon (2020). *Machine learning in Finance*. Vol. 1170. Springer.
- Salinas, David et al. (2020). "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* 36.3, pp. 1181–1191.