Appendices to Memory-Gated Recurrent Networks

Yaquan Zhang[3], Qi Wu[2], Nanbo Peng[1], Min Dai[4], Jing Zhang[2], and Hu Wang[1]

[1]JD Digits, {pengnanbo, wanghu5}@jd.com
[2]School of Data Science, City University of Hong Kong, qiwu55@cityu.edu.hk, jingzha28-c@my.cityu.edu.hk
[3]Department of Mathematics and Risk Management Institute, National University of Singapore, rmizhya@nus.edu.sg
[4]Department of Mathematics, Risk Management Institute, and Chong-Qing & Suzhou Research Institutes, National University of Singapore, mindai@nus.edu.sg

A Simulation experiments

To fix the constant terms in the data generation processes, 10 stock pairs are randomly selected from Table 2 of Yan et al. (2019). The stock pairs are (IBM, KO), (BA, CAT), (DWDP, JNJ), (CVX, PG), (IBM, JNJ), (NKE, WMT), (BA, PG), (INTC, KO), (AAPL, NKE), (MMM, DIS). The constant terms corresponding to each stock are listed in Table 1.

Component dimensions of neural networks in the simulation experiment are given in table 2.

A.1 Theoretical minimum MSE

When MSE is chosen to be the loss function of the prediction problem, the best predictor is the conditional expectation (Shumway and Stoffer, 2017).

$$\begin{aligned}E[100y_1(t)y_2(t)|\mathcal{F}_{t-1}] =&100\left((E[\alpha_1(t)|\mathcal{F}_{t-1}]+E[\gamma_1(t)|\mathcal{F}_{t-1}]E[g(\omega_1(t);u_1(t),v_1(t))|\mathcal{F}_{t-1}]\right)E[y_2(t)|\mathcal{F}_{t-1}]+\\&E[\beta_1(t)|\mathcal{F}_{t-1}]E[g(\omega_M(t);u_{M,1}(t),v_{M,1}(t))|\mathcal{F}_{t-1}]\times\\&(E[\alpha_2(t)|\mathcal{F}_{t-1}]+E[\gamma_2(t)|\mathcal{F}_{t-1}]E[g(\omega_2(t);u_1(t),v_2(t))|\mathcal{F}_{t-1}])+\\&E[\beta_1(t)|\mathcal{F}_{t-1}]E[\beta_2(t)|\mathcal{F}_{t-1}]\times\\&\left.E[g(\omega_M(t);u_{M,1}(t),v_{M,1}(t))g(\omega_M(t);u_{M,2}(t),v_{M,2}(t))|\mathcal{F}_{t-1}]\right).\end{aligned}\tag{1}$$

To evaluate equation (1), a primary observation is that the conditional distribution of an AR process is Gaussian. To be specific, we consider the following AR(5) process

$$p(t)=\mu+\sum_{j=1}^{5}\zeta_j p(t-j)+\epsilon(t),$$

with Gaussian random noise $\epsilon(t)\sim N(0,\sigma^2)$. Given observations up to step $t-1$, the conditional distribution of $p(t)$ is Gaussian with mean $\phi_p(t)=\mu+\sum_{j=1}^{5}\zeta_j p(t-j)$ and variance σ^2. Moreover,

$$E[\exp(p(t))|\mathcal{F}_{t-1}]=\exp\left(\phi_p(t)+\frac{\sigma^2}{2}\right).$$

	μ_{α}	$\mu_{\log\beta}$	$\mu_{\log u_M}$	$\mu_{\log v_M}$	$\mu_{\log\gamma}$	$\mu_{\log u}$	$\mu_{\log v}$
AAPL	0.008	-1.024	0.000	0.175	-0.840	0.215	0.159
BA	-0.007	-1.026	0.183	0.182	-0.842	0.164	0.120
CAT	0.020	-0.975	0.000	0.202	-0.847	0.199	0.153
CVX	0.011	-1.021	0.000	0.193	-0.849	0.172	0.138
DIS	0.002	-1.001	0.156	0.214	-0.862	0.196	0.151
DWDP	-0.007	-0.994	0.176	0.186	-0.866	0.198	0.141
IBM	0.021	-0.942	0.000	0.198	-0.886	0.218	0.178
INTC	0.012	-0.948	0.000	0.149	-0.873	0.168	0.141
JNJ	-0.003	-1.012	0.189	0.210	-0.858	0.227	0.160
KO	0.007	-0.979	0.117	0.198	-0.856	0.208	0.153
MMM	0.001	-0.964	0.186	0.198	-0.862	0.199	0.161
NKE	-0.002	-0.995	0.267	0.200	-0.793	0.347	0.297
PG	0.010	-0.979	0.096	0.201	-0.844	0.210	0.161
WMT	-0.007	-0.984	0.183	0.142	-0.871	0.181	0.146

Table 1: Constant terms of parameter processes matching with randomly selected stocks from table 2 of Yan et al. (2019).

This enables us to evaluate $E[\alpha_i(t)|\mathcal{F}_{t-1}]$, $E[\beta_i(t)|\mathcal{F}_{t-1}]$, $E[u_{M,i}(t)|\mathcal{F}_{t-1}]$, $E[v_{M,i}(t)|\mathcal{F}_{t-1}]$, $E[\gamma_i(t)|\mathcal{F}_{t-1}]$, $E[u_i(t)|\mathcal{F}_{t-1}]$ and $E[v_i(t)|\mathcal{F}_{t-1}]$.

It remains to evaluate the conditional expectations involving function g. Given $\omega \sim N(0,1)$, and $u(t)$ whose conditional distribution is lognormal with parameters $\phi_{\log u}(t)$ and $\sigma^2_{\log u} < 1$, we define

$$\begin{aligned}
V_1(u(t)) &:= E\left[\omega u(t)^{\omega}|\mathcal{F}_{t-1}\right] = \frac{\phi_{\log u}(t)}{\left(1-\sigma^2_{\log u}\right)^{3/2}} \exp\left(\frac{\phi^2_{\log u}(t)}{2-2\sigma^2_{\log u}}\right), \\
V_2(u(t)) &:= E\left[\omega^2 u(t)^{\omega}|\mathcal{F}_{t-1}\right] = \frac{1+\phi^2_{\log u}(t)-\sigma^2_{\log u}}{\left(1-\sigma^2_{\log u}\right)^{5/2}} \exp\left(\frac{\phi^2_{\log u}(t)}{2-2\sigma^2_{\log u}}\right).
\end{aligned}$$

The conditional expectations involving function g can be evaluated with V_1 and V_2.

$$\begin{aligned}
&E[g(\omega_M(t); u_{M,i}(t), v_{M,i}(t)|\mathcal{F}_{t-1}] = \frac{1}{A}\left(V_1(u_{M,i}(t)) + V_1\left(\frac{1}{v_{M,i}(t)}\right)\right), \\
&E[g(\omega_i(t); u_i(t), v_i(t)|\mathcal{F}_{t-1}] = \frac{1}{A}\left(V_1(u_i(t)) + V_1\left(\frac{1}{v_i(t)}\right)\right), \\
&E[g(\omega_M(t); u_{M,1}(t), v_{M,1}(t))g(\omega_M(t); u_{M,2}(t), v_{M,2}(t))|\mathcal{F}_{t-1}] \qquad (2) \\
=&\frac{1}{A^2}\left(V_2(u_{M,1}(t)u_{M,2}(t)) + V_2\left(\frac{u_{M,1}(t)}{v_{M,2}(t)}\right) + V_2\left(\frac{u_{M,2}(t)}{v_{M,1}(t)}\right) + V_2\left(\frac{1}{v_{M,1}(t)v_{M,2}(t)}\right)\right) + \\
&\frac{1}{A}\left(V_2(u_{M,1}(t)) + V_2(u_{M,2}(t)) + V_2\left(\frac{1}{v_{M,1}(t)}\right) + V_2\left(\frac{1}{v_{M,2}(t)}\right)\right) + 1.
\end{aligned}$$

	λ	$\tilde{N}$	N	Number of trainable parameters
LSTM	-	-	14	1736
GRU	-	-	17	1734
Channel-wise LSTM (two groups)	1	5	5	1640
	2	4	8	1632
	4	3	12	1776
	8	2	16	1952
Channel-wise LSTM (total split)	1	3	3	3120
	2	2	4	2128
	4	2	8	3360
	8	2	16	6208
mGRN (two groups)	1	10	10	1620
	2	8	16	1616
	4	6	24	1836
	8	3	24	1368
mGRN (total split)	1	4	4	1496
	2	4	8	1872
	4	3	12	1656
	8	2	16	1440

Table 2: Component dimensions and number of trainable parameters of neural networks in the simulation experiment. We limit the total number of trainable parameters to be around 1.8 thousand without including the parameters in the output dense layer. The number is chosen such that to further increase model sizes does not improve validation results for LSTM or GRU. It is impossible to set the numbers of trainable parameters to be exactly the same for all models. In general, we choose the number of parameters of mGRN to be smaller than those of alternative models. Moreover, to obtain a reasonable performance from channel-wise LSTM, we set a lower limit of 2 for the marginal component dimension. This indeed leads to much greater models when variables are totally split in channel-wise LSTM.

Lastly, by substituting the respective parts into equation (1), we have

$$\begin{aligned} E[100y_1(t)y_2(t)|\mathcal{F}_{t-1}] =& 100\,\Big(\phi_{\alpha_1(t)}E[y_2(t)|\mathcal{F}_{t-1}] + \\ & \exp\left(\phi_{\log\gamma_1}+\frac{\sigma^2_{\log\gamma_1}}{2}\right)E[g(\omega_1(t);u_1(t),v_1(t))|\mathcal{F}_{t-1}]E[y_2(t)|\mathcal{F}_{t-1}]+ \\ & \exp\left(\phi_{\log\beta_1}+\frac{\sigma^2_{\log\beta_1}}{2}\right)E[g(\omega_M(t);u_{M,1}(t),v_{M,1}(t))|\mathcal{F}_{t-1}]\times \\ & \left(\phi_{\alpha_2(t)}+\exp\left(\phi_{\log\gamma_2}+\frac{\sigma^2_{\log\gamma_2}}{2}\right)E[g(\omega_2(t);u_1(t),v_2(t))|\mathcal{F}_{t-1}]\right)+ \\ & \exp\left(\phi_{\log\beta_1}+\frac{\sigma^2_{\log\beta_1}}{2}\right)\exp\left(\phi_{\log\beta_2}+\frac{\sigma^2_{\log\beta_2}}{2}\right)\times \\ & E[g(\omega_M(t);u_{M,1}(t),v_{M,1}(t))g(\omega_M(t);u_{M,2}(t),v_{M,2}(t))|\mathcal{F}_{t-1}]\Big)\,, \end{aligned}$$

where

$$\begin{aligned} E[y_2(t)|\mathcal{F}_{t-1}] =& \phi_{\alpha_2(t)}+\exp\left(\phi_{\log\beta_2}+\frac{\sigma^2_{\log\beta_2}}{2}\right)E[g(\omega_M(t);u_{M,2}(t),v_{M,2}(t))|\mathcal{F}_{t-1}]+ \\ & \exp\left(\phi_{\log\gamma_2}+\frac{\sigma^2_{\log\gamma_2}}{2}\right)E[g(\omega_2(t);u_1(t),v_2(t))|\mathcal{F}_{t-1}], \\ \phi_p(t) =& \mu_p+0.9p(t-1)-0.8p(t-2)+0.7p(t-3)-0.6p(t-4)+0.5p(t-5) \\ & \text{for}\quad p=\alpha_i,\log\beta_i,\log u_{M,i},\log v_{M,i},\log\gamma_i,\log u_i,\log v_i \quad\text{and}\quad i=1,2, \\ \sigma_p =& 0.1 \quad\text{for}\quad p=\alpha_i,\log\beta_i,\log u_{M,i},\log v_{M,i},\log\gamma_i,\log u_i,\log v_i \quad\text{and}\quad i=1,2, \end{aligned}$$

and conditional expectations involving function g are given by equation (2).

B MIMIC-III Data Set

The full experiment results (including the 95% confidence intervals) are given in table 3.

	AUC-ROC	AUC-PR
Logistic regression	0.848 (0.828, 0.868)	0.474 (0.419, 0.529)
LSTM	0.855 (0.835, 0.873)	0.485 (0.431, 0.537)
Channel-wise LSTM	**0.862** (0.844, 0.881)	0.515 (0.464, 0.568)
mGRN	**0.862** (0.843, 0.880)	**0.523** (0.469, 0.575)

(a) In-hospital mortality

	AUC-ROC	AUC-PR
Logistic regression	0.870 (0.867, 0.873)	0.214 (0.205, 0.223)
LSTM	0.892 (0.889, 0.895)	0.324 (0.314, 0.333)
Channel-wise LSTM	0.906 (0.903, 0.909)	0.333 (0.323, 0.344)
mGRN	**0.911** (0.908, 0.913)	**0.347** (0.338, 0.358)

(b) Decompensation

	Kappa	MAD
Logistic regression	0.402 (0.401, 0.404)	162.3 (161.8, 162.8)
LSTM	0.438 (0.436, 0.440)	**123.1** (122.6, 123.5)
Channel-wise LSTM	0.442 (0.440, 0.444)	136.6 (136.1, 137.1)
mGRN	**0.447** (0.445, 0.449)	124.6 (124.1, 125.0)

(c) Length of stay

	Macro AUC-ROC	Micro AUC-ROC
Logistic regression	0.739 (0.734, 0.743)	0.799 (0.796, 0.803)
LSTM	0.770 (0.766, 0.775)	0.821 (0.818, 0.825)
Channel-wise LSTM	0.776 (0.772, 0.781)	0.825 (0.822, 0.828)
mGRN	**0.779** (0.774, 0.783)	**0.826** (0.823, 0.830)

(d) Phenotyping

Table 3: Model performance on the MIMIC-III data set (Johnson et al., 2016). Except for those of mGRN, all results are taken from Harutyunyan et al. (2019). Greater values are better for all metrics except mean absolute difference (MAD). The bold numbers are the best results. Following Harutyunyan et al. (2019), the reported results are the mean values calculated by resampling the test sets Q times with replacement ($Q = 10000$ for in-hospital mortality prediction and phenotype classification, and $Q = 1000$ for decompensation and length-of-stay prediction tasks). 95% confidence intervals are given in parentheses.

References

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18.

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Shumway, R. H. and Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples*. Springer.

Yan, X., Wu, Q., and Zhang, W. (2019). Cross-sectional learning of extremal dependence among financial assets. In *Advances in Neural Information Processing Systems*, pages 3852–3862.