Supplementary Materials for
Do RNN and LSTM have Long Memory?

Jingyu Zhao [1] **Feiqing Huang** [1] **Jia Lv** [2] **Yanjie Duan** [2] **Zhen Qin** [2] **Guodong Li** [1] **Guangjian Tian** [2]

Abstract

This supplementary material contains proofs, detailed remarks and additional theoretic and numerical results to support the theory and claims in the main paper. We also repeat some necessary contents here for easy reference.

A. Detailed Theoretical Results

A.1. Proof of Theorem 1

Assumption 1. *(i) The joint density function of $\varepsilon^{(t)}$ is continuous and positive everywhere; (ii) For some $\kappa \geq 2$, $E\|\varepsilon^{(t)}\|^{\kappa} < \infty$.*

Theorem 1. *Under Assumptions 1, if there exist real numbers $0 < a < 1$ and b such that $\|\mathcal{M}(x)\| \leq a\|x\| + b$, then recurrent network process (7) is geometrically ergodic, and hence has short memory.*

Proof. Let $Y^{(t)} = (y^{(t)\prime}, s^{(t)\prime})'$ and $r = p + q$. Rewrite model (7) as

$$Y^{(t)} = \mathcal{M}(Y^{(t-1)}) + e^{(t)}, \tag{1}$$

where $Y^{(t)}, e^{(t)} \in \mathbb{R}^r$ and $\mathcal{M} : \mathbb{R}^r \to \mathbb{R}^r$ is a general nonlinear function.

Let $\mathcal{B}^r$ be the class of Borel sets of $\mathbb{R}^r$ and ν_r be the Lebesgue measure on $(\mathbb{R}^r, \mathcal{B}^r)$. Then, $\{Y^{(t)}\}$ is a homogeneous Markov chain on the state space $(\mathbb{R}^r, \mathcal{B}^r, \nu_r)$ with the transition probability

$$P(x, A) = \int_A f(z - \mathcal{M}(x))dz, \quad x \in \mathbb{R}^r \text{ and } A \in \mathcal{B}^r, \tag{2}$$

where $f(\cdot)$ is the density of $e^{(t)}$. Observe that, from Assumption 1, the transition density kernel in (2) is positive everywhere, and thus $\{Y^{(t)}\}$ is ν_r-irreducible.

We prove by showing that Tweedie's drift criterion (Tweedie, 1983) holds, i.e. there exists a small set G with $\nu_r(G) > 0$ and a non-negative continuous function $\psi(x)$ such that

$$E\{\psi(Y^{(t)})|Y^{(t-1)} = x\} \leq (1 - \epsilon)\psi(x), \quad x \notin G, \tag{3}$$

and

$$E\{\psi(Y^{(t)})|Y^{(t-1)} = x\} \leq M, \quad x \in G, \tag{4}$$

for some $0 < \epsilon < 1$ and $0 < M < \infty$.

Given that $\|\mathcal{M}(x)\| \leq a\|x\| + b$, where $a < 1$, we have

$$\begin{aligned} & E\left(\|Y^{(t)}\|^{\kappa} \middle| Y^{(t-1)} = x\right) \\ \leq\; & \|\mathcal{M}(x)\|^{\kappa} + E\|e^{(t)}\|^{\kappa} \\ \leq\; & |a|^{\kappa}\|x\|^{\kappa} + |b|^{\kappa} + E\|e^{(t)}\|^{\kappa}. \end{aligned}$$

Define test function $\psi(x) = 1 + \|x\|^{\kappa} > 0$. Then,

$$\begin{aligned} & E\left(\psi(Y^{(t)}) \middle| Y^{(t-1)} = x\right) \\ \leq\; & 1 + |a|^{\kappa}\|x\|^{\kappa} + |b|^{\kappa} + E\|e^{(t)}\|^{\kappa} \\ \leq\; & \rho\psi(x) + 1 - \rho + |b|^{\kappa} + E\|e^{(t)}\|^{\kappa}, \end{aligned}$$

where $\rho = |a|^{\kappa} < 1$.

Denote $\epsilon = 1 - \rho - \frac{\left(1-\rho+|b|^{\kappa}+E\left(\|e^{(t)}\|\right)^{\kappa}\right)}{\psi(x)}$ and $G = \{x : \|x\| \leq L\}$ such that $\psi(x) > 1 + \frac{|b|^{\kappa}+E\|e^{(t)}\|^{\kappa}}{1-\rho}$ for all $\|x\| > L$. We obtain that conditions (3) and (4) hold.

Moreover, $E\left(\phi(Y^{(t)}) \middle| Y^{(t-1)} = x\right)$ is continuous with respect to x for any bounded continuous function $\phi(\cdot)$, then $\{Y^{(t)}\}$ is a Feller chain. By Feigin & Tweedie (1985), G is a small set. By referring to Theorem 4(ii) in Tweedie (1983) and Theorem 1 in Feigin & Tweedie (1985), $\{Y^{(t)}\}$ is geometrically ergodic with a unique strictly stationary solution.

□

A.2. Proof of Theorem 2

Theorem 2. *Under Assumption 1, linear recurrent network process (8) is geometrically ergodic if and only if spectral radius $\rho(W) < 1$. Model (8) hence has short memory.*

Proof. Proof of a similar result might exist in the literature, but we are unaware of the specific paper(s). For the convenience of the readers, we outline the proof here.

Let the Markov chain $\{Y^{(t)}\}$ and its state space be defined as in (i). Under the linear setting, model (7) can be written as

$$Y^{(t)} = WY^{(t-1)} + e^{(t)}, \tag{5}$$

where $Y^{(t)} \in \mathbb{R}^r$ and $W \in \mathbb{R}^{r\times r}$, and the transition probability can be written as

$$P(x, A) = \int_A f(z - Wx)dx, \quad x \in \mathbb{R}^r \text{ and } A \in \mathcal{B}^r. \tag{6}$$

Under Assumption 1, $\{Y^{(t)}\}$ is ν_r-irreducible.

First, suppose $\rho(W) < 1$. Then, there exists an integer s such that $\|W^s\| < 1$. In the following, we prove that s-step Markov chain $\{Y^{(ts)}\}$ satisfies Tweedie's drift criterion (Tweedie, 1983), i.e., there exists a small set G with $\nu_r(G) > 0$ and a non-negative continuous function $\psi(x)$ such that

$$E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} \leq (1-\epsilon)\psi(x), \quad x \notin G, \tag{7}$$

and

$$E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} \leq M, \quad x \in G, \tag{8}$$

for some constant $0 < \epsilon < 1$ and $0 < M < \infty$.

We iterate (5) s times and obtain

$$Y^{(ts)} = W^s Y^{((t-1)s)} + \left(e^{(ts)} + \sum_{j=1}^{s-1} W^j e^{(ts-j)}\right).$$

Let $g(x) = 1 + \|x\|^\kappa$, and it can be verified that

$$\begin{aligned} &E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} \\ &\leq 1 + \|W^s\|^\kappa \|x\|^\kappa + E\left(e^{(ts)} + \sum_{j=1}^{s-1} W^j e^{(ts-j)}\right) \\ &\leq \psi(x)\|W^s\|^\kappa + C, \end{aligned}$$

where $C = 1 + E(e^{(ts)} + \sum_{j=1}^{s-1} W^j e^{(ts-j)}) - \|W^s\|^\kappa < \infty$. Note that $\|W^s\|^\kappa < 1$. Then there exists $L > 0$, such that

$$E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} \leq (1-\epsilon)\psi(x), \quad \forall \|x\| > L,$$

and

$$E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} \leq M < \infty, \quad \forall \|x\| \leq L.$$

and $G = \{x : \|x\| \leq L\}$ with $\nu_r(G) > 0$.

Moreover, because for each bounded continuous function $\phi(\cdot)$, $E\{\phi(Y^{(ts)})|Y^{((t-1)s)} = x\}$ is continuous with respect to x, $\{Y^{(ts)}\}$ is a Feller chain. And $\{Y^{(ts)}\}$ is ν_r-irreducible. This implies that G is a small set (Feigin & Tweedie, 1985). By referring to Theorem 4(ii) in Tweedie (1983), we can show that $\{Y^{(ts)}\}$ is geometrically ergodic with a unique strictly stationary solution. By Lemma 3.1 of Tjøstheim (1990), $\{Y^{(t)}\}$ is geometrically ergodic.

Then, we prove the necessity. Suppose that model (5) is geometrically ergodic, then there exists a strictly stationary solution $\{Y^{(t)}\}$ to model (5) (Feigin & Tweedie, 1985). And then the Markov chain $Y^{(t)}$ have a stationary distribution $\pi(\cdot)$, from which we can generate $Y^{(0)}$, and iteratively obtain the sequence $\{Y^{(t)}\}$. It is nonanticipative and equation (5) holds.

From (6), it holds that

$$P(Y^{(t)} \in A|Y^{(t-1)} = x) = P(x, A) > 0$$

as $\nu_r(A) > 0$. Let H be any affine invariant subspace of $\mathbb{R}^r$ under model (5), i.e. $\{Wx + e^{(t)} : x \in H\} \subseteq H)$ with probability one . If $\nu_r(\mathbb{R}^r - H) \neq 0$, then for any $x \in H$, $P(Wx + e^{(t)} \in H) < 1$. As a result, $\mathbb{R}^r$ is the unique affine invariant subspace, and hence model (5) is irreducible. Thus, by Theorem 2.5 in Bougerol & Picard (1992), we have that the the top Lyapounov exponent is strictly negative, and thus spectral radius $\rho(W) = \|W^s\|^{1/s} < 1$. This completes the proof of (ii).

□

A.3. Proof of Corollary 1

Corollary 1. *Suppose that the output and activation functions, $g(\cdot)$ and $\sigma(\cdot)$, at (10) are continuous and bounded. If Assumption 1 holds, then the RNN process is geometrically ergodic and has short memory.*

Proof. Need to show that there always exist real numbers $a < 1$ and b such that $\|\mathcal{M}_{\text{RNN}}(u, v)\| \leq a\|(u', v')'\| + b$.

Since $g(\cdot)$ and $\sigma(\cdot)$ are bounded, there exist positive constants M_1 and M_2 such that $\|g(W_{zh}\sigma(W_{hh}v + W_{hy}u + b_h) + b_z)\|_{l_1} \leq M_1$, $\|\sigma(W_{hh}v + W_{hy}u + b_h)\|_{l_1} \leq M_2$ for any $u \in \mathbb{R}^p, v \in \mathbb{R}^q$.

Let $a = a_0 \in (0, 1)$ and $b = M_1 + M_2$, we have $\|\mathcal{M}_{\text{RNN}}(u, v)\|_{l_1} - a_0\|(u', v')'\|_{l_1} \leq M_1 + M_2 - a_0\|u\|_{l_1} - a_0\|v\|_{l_1} \leq b = M_1 + M_2$. By Theorem 1, model (10) with bounded and continuous output and activation function is geometrically ergodic and has short memory. □

A.4. Apply Theorem 1 to LSTM networks with $p = q = 1$

We use an LSTM process with $p = q = 1$ as an example to illustrate the application of Theorem 1 to LSTM networks, and prepare readers for Corollary 2. Assume that the norm $\|\cdot\|$ in Theorem 1 is the l_1 norm. Although sigmoid is used by default as the activation functions for the gates, we also consider $\sigma(\cdot)$ as ReLU or tanh for theoretical interests here. For output function $g(\cdot)$, we consider commonly used linear, sigmoid and softmax functions. We summarize our results in Table A1.

A.5. Proof of Corollary 2

Corollary 2. *The input series features $\{y^{(t-1)}\}$ are scaled to the range of $[-1, 1]$. Suppose that $M := \sup_{x \in B_\infty^q} \|g(W_{zh}x + b_z)\|_{l_1} < \infty$ and $\sigma(\|W_{fh}\|_{l_\infty} + \|W_{fy}\|_{l_\infty} + \|b_f\|_{l_\infty}) \leq a$ for some $a < 1$, where B_∞^q is the q-dimensional l_∞-ball and $\|W\|_{l_\infty} = \max_{1 \leq i \leq m} \sum_{j=1}^n |w_{ij}|$ is the matrix l_∞-norm. If Assumption 1 holds, then the LSTM process at (14) is geometrically ergodic and has short memory.*

Proof. Let $a = a_0 \in (0, 1)$ and $b = M + 2q$, we have

$$\begin{aligned}
&\|\mathcal{M}_{\mathrm{LSTM}}(u, v, w)\|_{l_1} - a_0 \|(u', v', w')'\|_{l_1} \\
&\leq \|g(W_{zh}x + b_z)\|_{l_1} + \|x\|_{l_1} \\
&\qquad + \|\mathbf{1}_q + f(u, v) \odot w\|_{l_1} - a_0 \|(u', v', w')'\|_{l_1} \\
&\leq M + q + q + \|f(u, v)\|_{l_\infty} \|w\|_{l_1} \\
&\qquad - a_0\|u\|_{l_1} - a_0\|v\|_{l_1} - a_0\|w\|_{l_1} \\
&\leq M + 2q - a_0\|u\|_{l_1} - a_0\|v\|_{l_1} \\
&\qquad + (\|f(u, v)\|_{l_\infty} - a_0)\|w\|_{l_1} \\
&\leq b = M + 2q,
\end{aligned}$$

where $x = o(u, v) \odot \tanh(i(u, v) \odot \tanh(W_{ch}v + W_{cy}u + b_c) + f(u, v) \odot w) \in B_\infty^q$, $v = h^{(t)} \in B_\infty^q$, and $u = y^{(t-1)} \in B_\infty^p$. The second inequality holds due to the definition of M and $x \in B_\infty^q$. The forth inequality holds due to $\|f(u, v)\|_{l_\infty} = \|\sigma(W_{fh}v + W_{fy}u + b_f)\|_{l_\infty} \leq \sigma(\|W_{fh}\|_{l_\infty} + \|W_{fy}\|_{l_\infty} + \|b_f\|_{l_\infty}) \leq a_0$.

By Theorem 1, model (14) is geometrically ergodic and has short memory. □

A.6. Proof of Theorem 3

Theorem 3. *In terms of Definition 3, the MRNNF has the capability of handling long-range dependence data, while the RNN cannot.*

Proof. Without loss of generality, assume that the linear activation and output functions are identity.

(1) The RNN process can be written as

$$\begin{cases} y^{(t)} = W_{zh}h^{(t)} + \varepsilon^{(t)} \\ h^{(t)} = W_{hh}h^{(t-1)} + W_{hx}x^{(t)} \end{cases}.$$

Then, $h^{(t)} = (I - W_{hh}\mathcal{B})^{-1} W_{hx}x^{(t)}$, and we have

$$y^{(t)} = W_{zh}(I - W_{hh}\mathcal{B})^{-1} W_{hx}x^{(t)} + \varepsilon^{(t)}.$$

Let $y^{(t)} = \sum_{k=0}^\infty A_k x^{(t-k)} + \varepsilon^{(t)}$. Since $(I - W_{hh}\mathcal{B})^{-1} = \sum_{k=0}^\infty W_{hh}^k \mathcal{B}^k$, we have $A_k = W_{zh}W_{hh}^k W_{hx}$, and $(A_k)_{ij}$ decays exponentially for all i, j.

(2) The MRNNF process can be written as

$$\begin{cases} y^{(t)} = W_{zh}h^{(t)} + W_{zm}m^{(t)} + \varepsilon^{(t)} \\ h^{(t)} = W_{hh}h^{(t-1)} + W_{hx}x^{(t)} \\ m^{(t)} = W_{mm}m^{(t-1)} + W_{mf}((I - \mathcal{B})^d - I)x^{(t)} \end{cases}.$$

Then,

$$\begin{cases} h^{(t)} = (I - W_{hh}\mathcal{B})^{-1} W_{hx}x^{(t)} \\ m^{(t)} = (I - W_{mm}\mathcal{B})^{-1} W_{mf}((I - \mathcal{B})^d - I)x^{(t)} \end{cases}.$$

Let $y^{(t)} = \sum_{k=0}^\infty A_k x^{(t-k)} + \varepsilon^{(t)}$, then $A_k = C_k + D_k$, where

$$\begin{cases} \sum_{k=0}^\infty C_k x^{(t-k)} = W_{zh}(I - W_{hh}\mathcal{B})^{-1} W_{hx}x^{(t)} \\ \sum_{k=0}^\infty D_k x^{(t-k)} = \\ \qquad W_{zm}(I - W_{mm}\mathcal{B})^{-1} W_{mf}((I - \mathcal{B})^d - I)x^{(t)} \end{cases}.$$

From part (1) we know that the entries in C_k decay exponentially as well as the entries in the first part $W_{zm}(I - W_{mm}\mathcal{B})^{-1}$ in D_k. Since $(I - \mathcal{B})^d - I = \sum_{k=1}^\infty W_k \mathcal{B}^k$ and W_k's are diagonal matrices with $(W_k)_{ii} \sim k^{-d_i - 1}$, the decay of D_k is dominated by $(I - \mathcal{B})^d - I$, and entries of A_k decay at some polynomial rate $k^{-d_j - 1}$. □

A.7. Memory Property of Constant-gates-LSTM and Constant-gates-MLSTM

Theorem 4. *In terms of Definition 3, the constant-gates-MLSTM has the capability of handling long-range dependence data, while the constant-gates-LSTM cannot.*

Proof. Without loss of generality, assume that the linear activation and output functions are identity.

(1) The constant-gates-LSTM process can be written as

$$\begin{cases} y^{(t)} = W_{zh}h^{(t)} + \varepsilon^{(t)} \\ \tilde{c}^{(t)} = W_{ch}h^{(t-1)} + W_{cx}x^{(t)} \\ c^{(t)} = D_i\, \tilde{c}^{(t)} + D_f\, c^{(t-1)} \\ h^{(t)} = D_o\, c^{(t)} \end{cases},$$

where D_i, D_f and D_o are matrices obtained by diagonalize the constant gates.

Then, $(I - D_f\mathcal{B})c^{(t)} = D_i\tilde{c}^{(t)} = D_i(W_{ch}h^{(t-1)} + W_{cx}x^{(t)}) = D_iW_{ch}D_o\, c^{(t-1)} + D_iW_{cx}x^{(t)}$, and we have $(I - (D_f + D_iW_{ch}D_o)\mathcal{B})c^{(t)} = D_iW_{cx}x^{(t)}$. Thus, $c^{(t)} = (I - (D_f + D_iW_{ch}D_o)\mathcal{B})^{-1}D_iW_{cx}x^{(t-1)}$, then $y^{(t)} = W_{zh}D_o(I - (D_f + D_iW_{ch}D_o)\mathcal{B})^{-1}D_iW_{cx}x^{(t)} + \varepsilon^{(t)}$. From the proof of Theorem 3 (1) we know that writing

Table 1. Application of Theorem 1 to specific LSTMs.

		Activation function σ	
		ReLU or identity	sigmoid or tanh
Output function g	identity	$\lvert w_{oh}\rvert + \lvert w_{ih}\rvert + \lvert w_{zh}w_{oh}\rvert \le a$, $\lvert w_{oy}\rvert + \lvert w_{iy}\rvert + \lvert w_{zh}w_{oy}\rvert \le a$, $\lvert w_{fh}v + w_{fy}u + b_f\rvert \le a$	No
	sigmoid	$\lvert w_{oh}\rvert + \lvert w_{ih}\rvert \le a$, $\lvert w_{oy}\rvert + \lvert w_{iy}\rvert \le a$, $\lvert w_{fh}v + w_{fy}u + b_f\rvert \le a$	$\lvert\sigma(w_{fh} + w_{fy} + b_f)\rvert \le a$
	softmax	$\lvert w_{oh}\rvert + \lvert w_{ih}\rvert \le a$, $\lvert w_{oy}\rvert + \lvert w_{iy}\rvert \le a$, $\lvert w_{fh}v + w_{fy}u + b_f\rvert \le a$	$\lvert\sigma(w_{fh} + w_{fy} + b_f)\rvert \le a$

$y^{(t)} = \sum_{k=0}^{\infty} A_k x^{(t-k)} + \varepsilon^{(t)}$, we have all the entries of A_k decay exponentially.

(2) The constant-gates-MLSTM process can be written as

$$\begin{cases} y^{(t)} = W_{zh}h^{(t)} + \varepsilon^{(t)} \\ \tilde{c}^{(t)} = W_{ch}h^{(t-1)} + W_{cx}x^{(t)} \\ (I-\mathcal{B})^d\, c^{(t)} = D_i \tilde{c}^{(t)} \\ h^{(t)} = D_o c^{(t)} \end{cases}.$$

Then, $(I-\mathcal{B})^d\, c^{(t)} = D_i(W_{ch}h^{(t-1)} + W_{cx}x^{(t)}) = D_iW_{ch}D_oc^{(t-1)} + D_iW_{cx}x^{(t)}$, and we have $((I-\mathcal{B})^d - D_iW_{ch}D_o\mathcal{B})\, c^{(t)} = D_iW_{cx}x^{(t)}$. Thus, $c^{(t)} = ((I-\mathcal{B})^d - D_iW_{ch}D_o\mathcal{B})^{-1}D_iW_{cx}x^{(t)}$, then $y^{(t)} = W_{zh}D_o((I-\mathcal{B})^d - D_iW_{ch}D_o\mathcal{B})^{-1}D_iW_{cx}x^{(t)} + \varepsilon^{(t)}$.

Now we need to obtain the rate of polynomial $((I-\mathcal{B})^d - C\mathcal{B})^{-1}$ for some matrix $C = D_iW_{ch}D_o$. Let $((I-\mathcal{B})^d - C\mathcal{B})^{-1} = \sum_{j=0}^{\infty}\Theta_j\mathcal{B}^j$, then $(\sum_{j=0}^{\infty}\Theta_j\mathcal{B}^j)((I-\mathcal{B})^d - C\mathcal{B}) = I$. Thus,

$$(\sum_{j=0}^{\infty}\Theta_j\mathcal{B}^j)(I-\mathcal{B})^d = I + C\sum_{j=0}^{\infty}\Theta_j\mathcal{B}^{j+1}$$
$$(\sum_{j=0}^{\infty}\Theta_j\mathcal{B}^j)(\sum_{k=0}^{\infty}W_k\mathcal{B}^k) = I + C\sum_{j=0}^{\infty}\Theta_j\mathcal{B}^{j+1}$$
$$\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\Theta_j\mathcal{B}^jW_k\mathcal{B}^k = I + C\sum_{j=0}^{\infty}\Theta_j\mathcal{B}^{j+1}.$$

Equate the coefficients for each B^j term for $j = 0, 1, 2...$, we have

$$\begin{cases} \Theta_0 & = I \\ \Theta_1 & = C - W_1 \\ \Theta_2 & = C\Theta_1 - W_1\Theta_1 - W_2 \\ \Theta_3 & = C\Theta_2 - W_1\Theta_2 - W_2\Theta_1 - W_3 \\ & \vdots \\ \Theta_k & = C\Theta_{k-1} - \sum_{j=1}^{k} W_j\Theta_{k-j} \end{cases}.$$

The Θ_k's are dominated by the W_k term and the elements decay at the same rate as W_k, which is k^{-d_j-1}. □

Table 2. Overall performance in terms of MAE. Average MAE and the standard deviation (in brackets) are reported.

	ARFIMA	DJI (x100)	Traffic	Tree
RNN	0.9310 (0.1550)	0.1977 (0.0242)	233.442 (12.391)	0.2240 (0.0064)
RNN2	0.9310 (0.1430)	0.1861 (0.0164)	233.419 (12.378)	0.2229 (0.0057)
RWA	1.3330 (0.0030)	0.2052 (0.0164)	233.137 (7.425)	0.2379 (0.0001)
MRNNF	0.8800 (0.0790)	**0.1809** (0.0168)	**232.554** (11.954)	0.2206 (0.0034)
MRNN	**0.8710** (0.0900)	0.1835 (0.0165)	232.794 (12.149)	**0.2202** (0.0037)
LSTM	0.9070 (0.0940)	0.1841 (0.0182)	234.055 (11.149)	0.2215 (0.0051)
MLSTMF	0.9240 (0.1320)	0.1895 (0.0203)	233.142 (11.551)	0.2235 (0.0060)
MLSTM	0.9170 (0.1320)	0.1881 (0.0187)	233.035 (10.793)	0.2234 (0.0061)

B. More Numerical Results

B.1. Autocorrelation Plots of All Datasets

Autocorrelation plots of all 4 datasets, ARFIMA, DJI, traffic and tree, are shown in Figure 1.

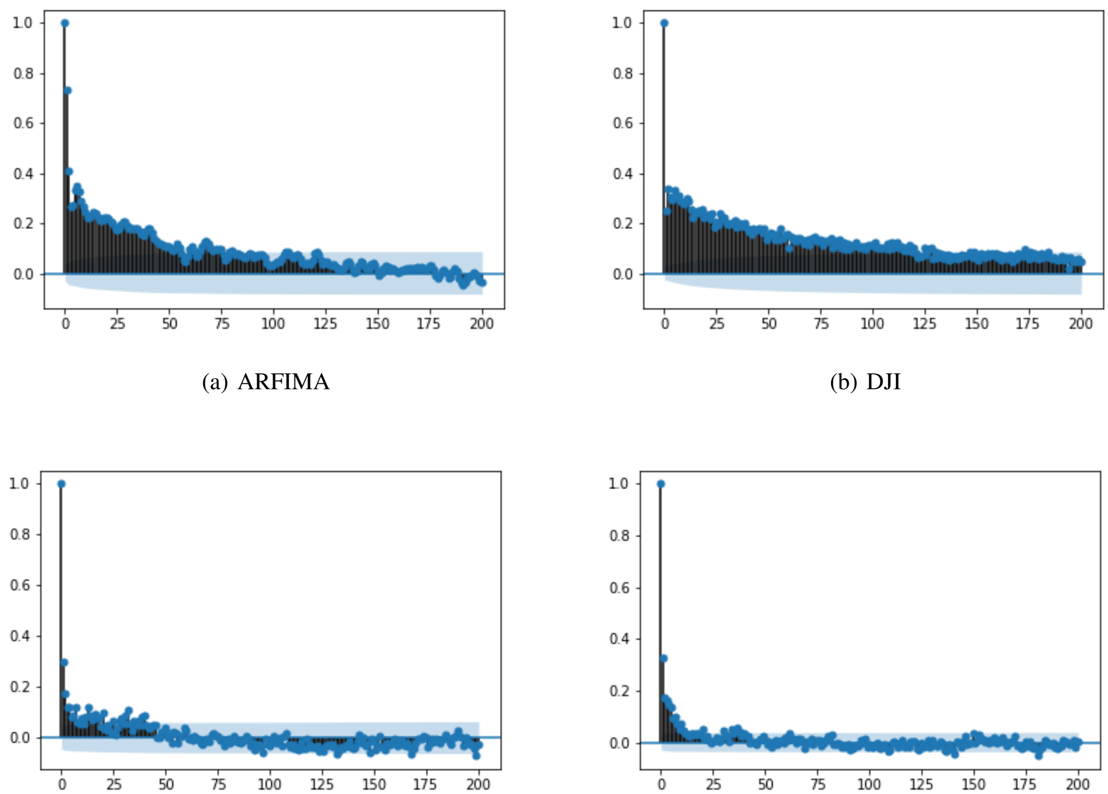
B.2. Overall Performance of the Models

Average RMSE and standard deviation of one-step forecasting is reported in the main paper. We provide results in terms of MAE and MAPE, as well as figures, in this section.

RMSE Boxplot of RMSE for 100 different initializations are shown in Figure 2 for datasets ARFIMA, DJI, traffic and tree.

MAE Average MAE and standard deviation of one-step forecasting is shown in Table 2.

Boxplot of MAE for 100 different initializations are shown



(a) ARFIMA

(b) DJI

(c) traffic

(d) tree

Figure 1. Autocorrelation plots of all 4 datasets.

in Figure 3 for datasets ARFIMA, DJI, traffic and tree.

MAPE Average MAPE and standard deviation of one-step forecasting is shown in Table 3.

Boxplot of MAPE for 100 different initializations are shown in Figure 4 for datasets ARFIMA, DJI, traffic and tree.

B.3. Best Performance of the Models

Best performance of the models, in terms of MAE and MAPE, are shown in Table 4 & 5.

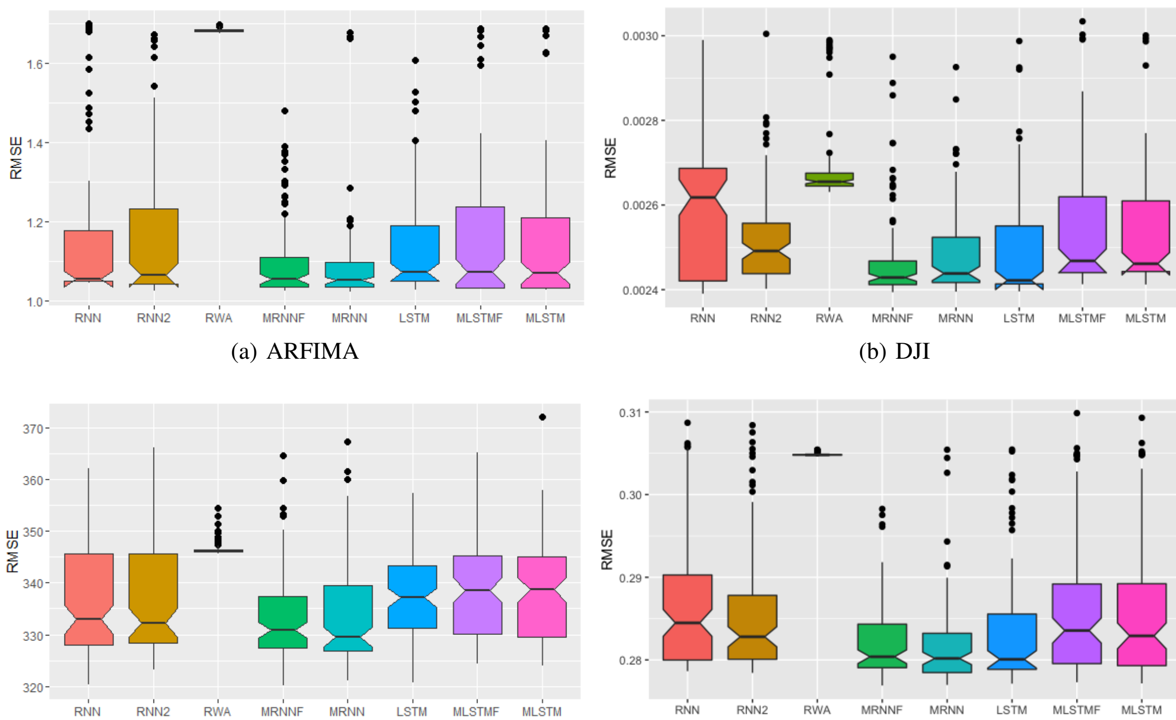
B.4. Performance on a Dataset without Long Memory

We generated a sequence of length 4001 $(2000 + 1200 + 800)$ using model (10), which does not have long memory according to Corollary 1. We refer to this synthetic dataset as the RNN dataset. The boxplots of error measures are presented in Figure 5. From the boxplots we can see that the performance of our proposed models is comparable with that of the true model RNN, except that the variation of the error measures is a bit larger.

B.5. Experiment on Parameter K

Boxplot of RMSE for 100 different initializations are shown in Figure 6, 7, 8 and 9 for datasets ARFIMA, DJI, traffic and tree, respectively. Values of K are appended to the abbreviations of the proposed models to distinguish the settings. For example, model "MRNN25" means the MRNN model with $K = 25$. There are 20 models with different settings in total, and they are sorted by the average RMSE in ascending order from left to right.

For MRNN and MRNNF, the prediction is generally better for a larger K, and they have smaller average RMSE than all the baseline models regardless of the choice of K. Interestingly, the performance of MLSTM and MLSTMF gets better when K becomes smaller, and with $K = 25$, they can outperform LSTM on ARFIMA and traffic datasets. Thus, we recommend a large K for MRNN and MRNNF models, while for the more complicated MLSTM models, K deserves more investigation to balance expressiveness and optimization.



(a) ARFIMA (b) DJI

(c) traffic (d) tree

Figure 2. Boxplot of RMSE for 100 different initializations.

Table 3. Overall performance in terms of MAPE. Average MAPE and the standard deviation (in brackets) are reported.

	ARFIMA	DJI (x100)	Traffic	Tree
RNN	2.5760	1.4371	1.3943	0.2747
	(0.4030)	(0.2566)	(0.1998)	(0.0079)
RNN2	2.5570	1.4407	1.4092	0.2739
	(0.4420)	(0.2106)	(0.1789)	(0.0071)
RWA	**2.2370**	**1.2733**	1.3745	0.2939
	(0.1950)	(0.1702)	(0.1457)	(0.0005)
MRNNF	2.6430	1.5561	1.4270	0.2714
	(0.3380)	(0.2243)	(0.1834)	(0.0042)
MRNN	2.7010	1.5031	1.4253	**0.2706**
	(0.2680)	(0.2045)	(0.1586)	(0.0044)
LSTM	2.5660	1.5725	1.3632	0.2727
	(0.3750)	(0.2283)	(0.1807)	(0.0060)
MLSTMF	2.5100	1.3141	**1.3462**	0.2750
	(0.4690)	(0.1369)	(0.1769)	(0.0074)
MLSTM	2.5500	1.3123	1.3353	0.2748
	(0.4370)	(0.1281)	(0.1926)	(0.0075)

Table 4. Best performance in terms of MAE.

	ARFIMA	DJI (x100)	Traffic	Tree
ARFIMA	0.8190	0.1800	230.99	0.2174
RNN	0.8378	0.1667	213.96	0.2185
RNN2	0.8196	0.1667	**212.69**	0.2186
RWA	1.3307	0.1862	227.01	0.2378
MRNNF	0.8179	**0.1649**	214.88	0.2172
MRNN	**0.8171**	0.1654	213.79	**0.2170**
LSTM	0.8197	0.1671	214.22	0.2170
MLSTMF	0.8191	0.1716	216.15	0.2177
MLSTM	0.8193	0.1702	216.12	0.2175

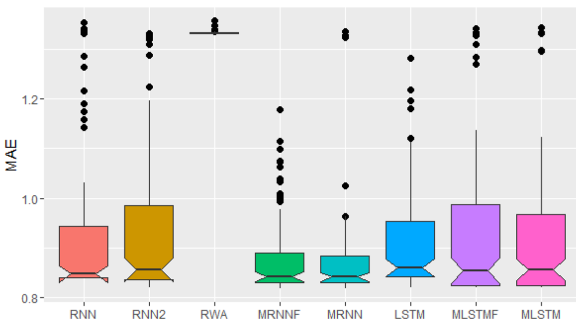
References

Bougerol, P. and Picard, N. Strict stationarity of generalized autoregressive processes. *Annals of Probability*, pp. 1714–1730, 1992.

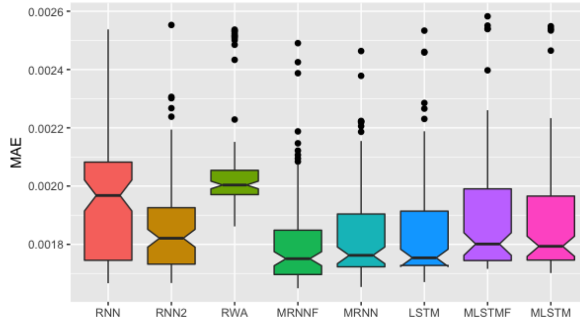
Feigin, P. D. and Tweedie, R. L. Random coefficient autoregressive processes: a markov chain analysis of stationarity and finiteness of moments. *Journal of Time Series Analysis*, 6:1–14, 1985.

Tjøstheim, D. Non-linear time series and markov chains. *Advances in Applied Probability*, 22:587–611, 1990.

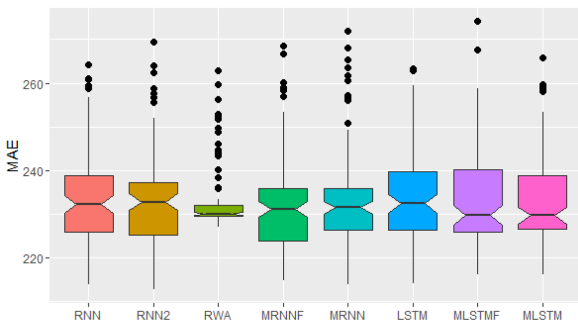
Tweedie, R. *Criteria for rates of convergence of Markov chains, with application to queueing and storage theory*, pp. 260–276. London Mathematical Society Lecture Note Series. Cambridge University Press, 1983. doi: 10.1017/CBO9780511662430.016.



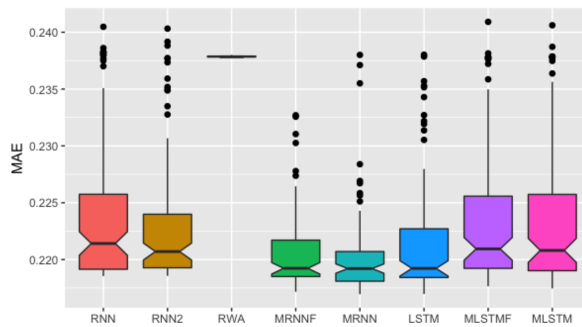
(a) ARFIMA



(b) DJI



(c) traffic

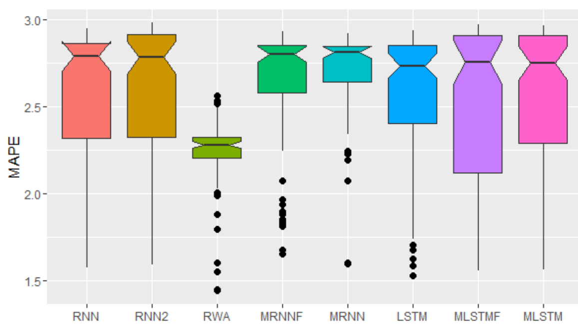


(d) tree

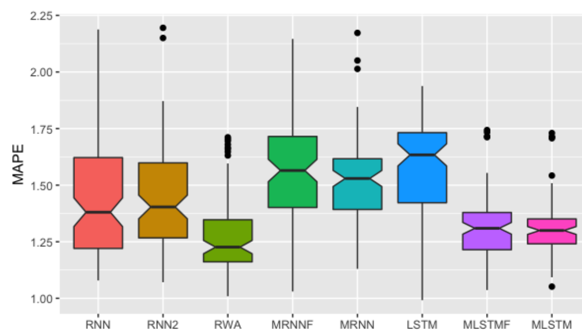
Figure 3. Boxplot of MAE for 100 different initializations.

Table 5. Best performance in terms of MAPE.

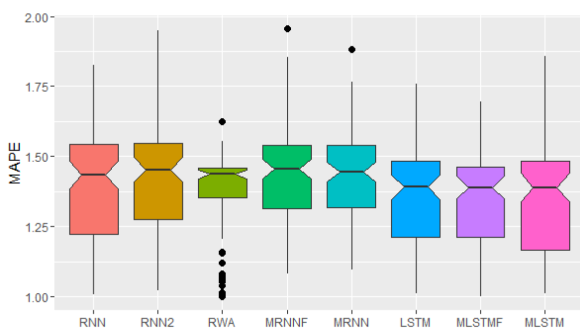
	ARFIMA	DJI	Traffic	Tree
ARFIMA	2.8424	1.8334	1.6942	0.2676
RNN	1.5729	1.0789	1.0075	0.2680
RNN2	1.5905	1.0714	1.0215	0.2680
RWA	**1.4408**	1.0091	**0.9986**	0.2923
MRNNF	1.6508	1.0304	1.0816	0.2670
MRNN	1.5967	1.1303	1.0938	**0.2668**
LSTM	1.5282	**0.9918**	1.0099	0.2675
MLSTMF	1.5565	1.0368	0.9990	0.2673
MLSTM	1.5597	1.0518	1.0098	0.2673



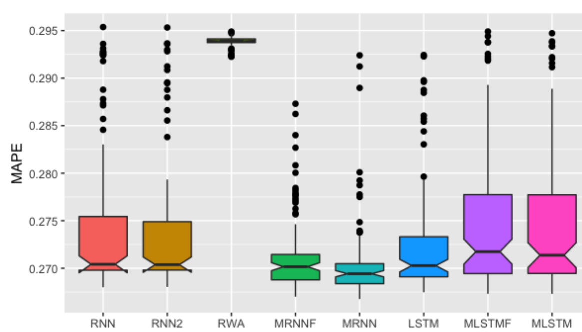
(a) ARFIMA



(b) DJI

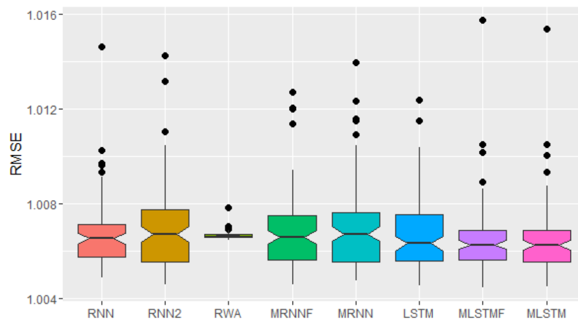


(c) traffic

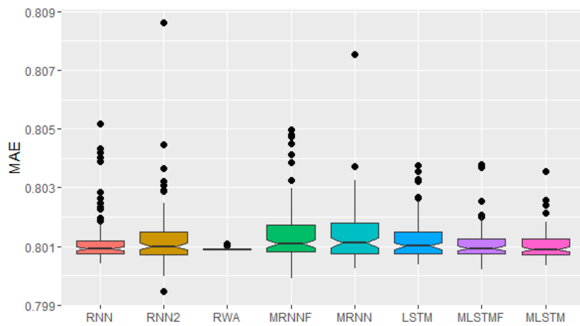


(d) tree

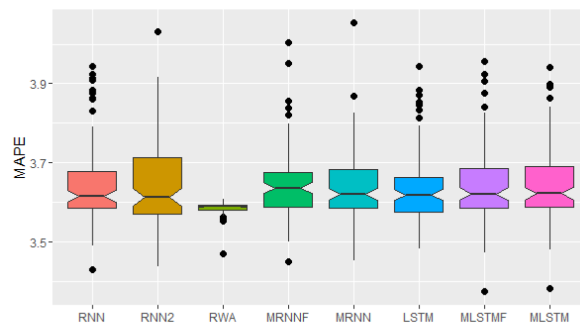
Figure 4. Boxplot of MAPE for 100 different initializations.



(a) RMSE



(b) MAE



(c) MAPE

Figure 5. Boxplot of RMSE, MAE and MAPE for 100 different initializations. Dataset: RNN.

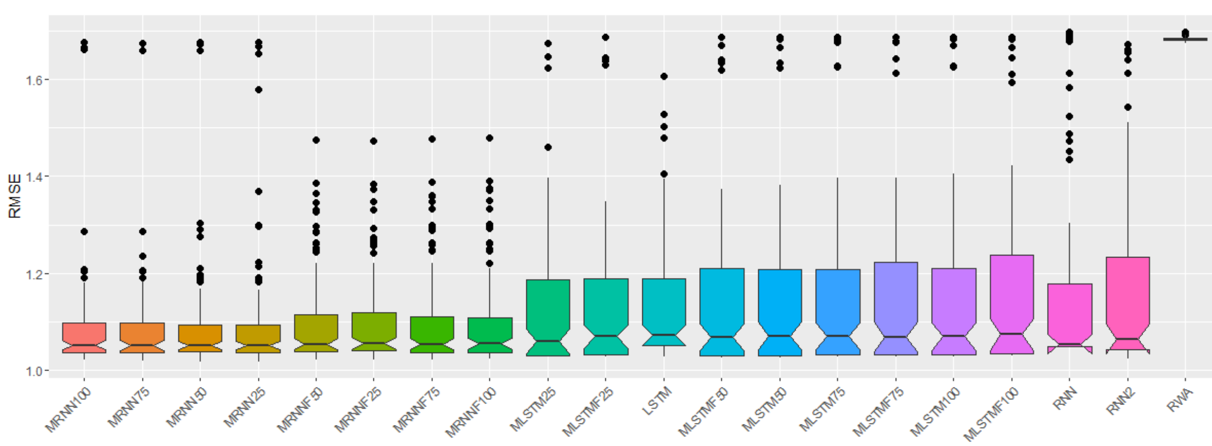


Figure 6. Boxplot of RMSE for 100 different initializations. Dataset: ARFIMA.

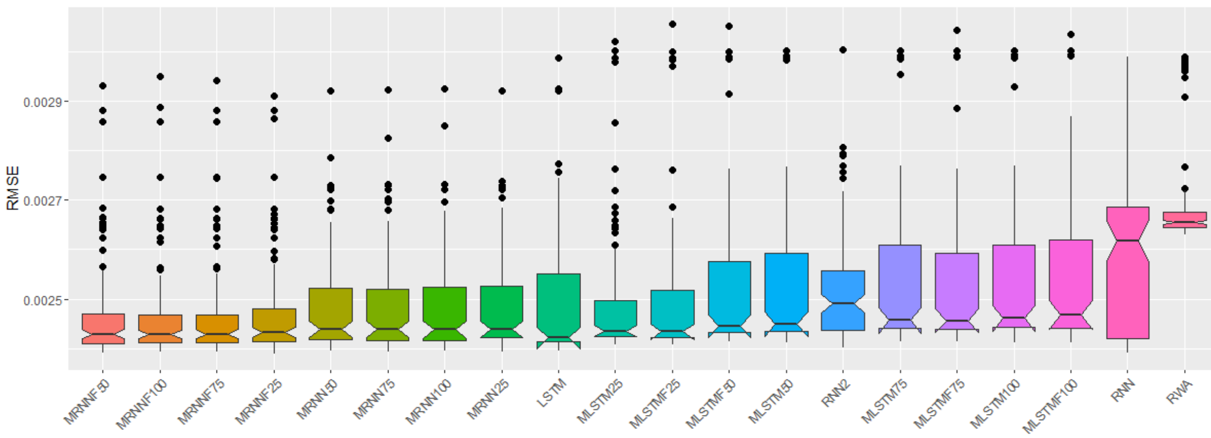


Figure 7. Boxplot of RMSE for 100 different initializations. Dataset: DJI.

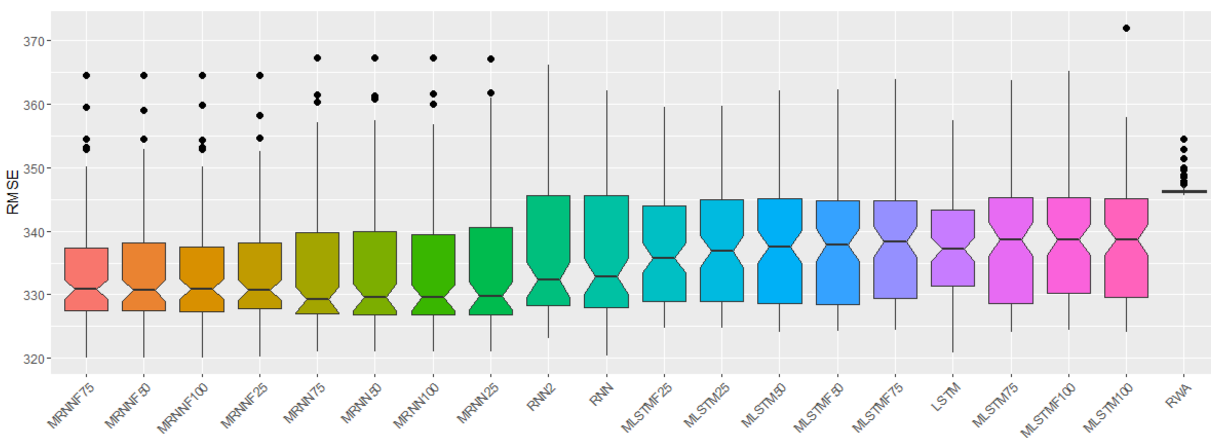


Figure 8. Boxplot of RMSE for 100 different initializations. Dataset: traffic.

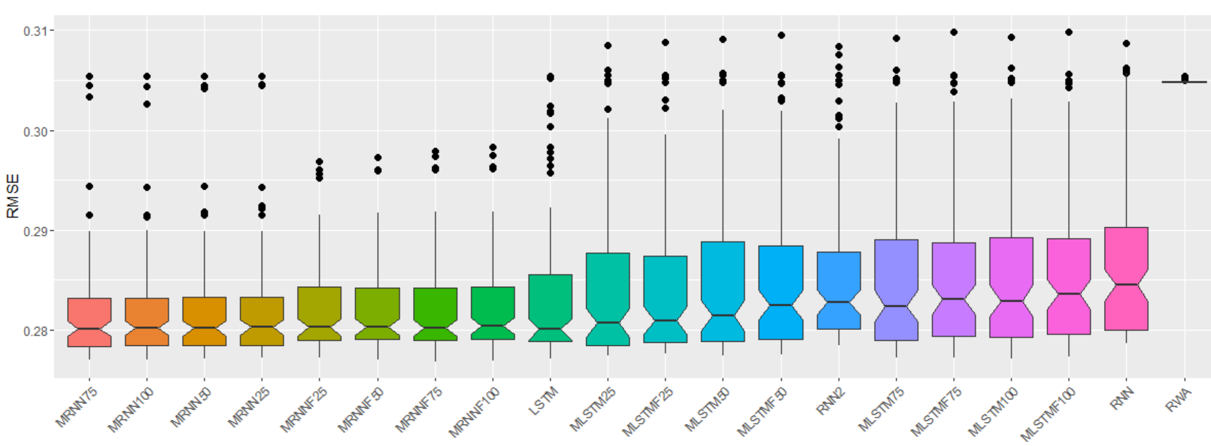


Figure 9. Boxplot of RMSE for 100 different initializations. Dataset: tree.