

# $L_{1/2}$ Norm Regularized Echo State Network for Chaotic Time Series Prediction

Meiling Xu<sup>1</sup>, Min Han<sup>1(✉)</sup>, and Shunshoku Kanae<sup>2</sup>

<sup>1</sup> Faculty of Electronic Information and Electrical Engineering,  
Dalian University of Technology, Dalian, China  
minhan@dlut.edu.cn

<sup>2</sup> Department of Electrical, Electronic and Computer Engineering,  
Fukui University of Technology, Fukui, Japan

**Abstract.** Echo state network contains a randomly connected hidden layer and an adaptable output layer. It can overcome the problems associated with the complex computation and local optima. But there may be ill-posed problem when large reservoir state matrix is used to calculate the output weights by least square estimation. In this study, we use  $L_{1/2}$  regularization to calculate the output weights to get a sparse solution in order to solve the ill-posed problem and improve the generalized performance. In addition, an operation of iterated prediction is conducted to test the effectiveness of the proposed  $L_{1/2}$ ESN for capturing the dynamics of the chaotic time series. Experimental results illustrate that the predictor has been designed properly. It outperforms other modified ESN models in both sparsity and accuracy.

**Keywords:** Echo state networks ·  $L_{1/2}$  norm regularization · Chaotic time series · Prediction

## 1 Introduction

The echo state network (ESN) is a novel kind of recurrent neural networks. Only the output weights are modified by a simple and efficient linear regression algorithm [1]. It can overcome the local minima and vanishing gradient problems associated with traditional neural networks training algorithms. Owing to the above merits, ESNs have been extensively studied in time series prediction [2–4].

However, when least square method is used to compute the readout weights, the large reservoir state matrix may be ill-posed [5], which adversely affect the generalization of the model. To solve this problem, a series of regularization techniques have been applied in the training process of echo state networks [2, 6–8]. They are computational efficient and not prone to over fitting. For example, J.J. Steil proposed a modified  $L_2$  norm regularized echo state network to reduce the risk of error amplification and boost model's generalization [6]. Han added an  $L_1$  norm penalty term in the objective function to control the model's complexity [7]. But the  $L_2$  norm regularization is a biased estimation and the  $L_1$  norm does not satisfy oracle property [8]. Recently,  $L_{1/2}$  penalty which has many promising properties, such as unbiasedness, sparsity and oracle property, has been proposed and attracted growing attention [9, 10].

In this paper, we combine the L<sub>1/2</sub> norm regularization method with ESNs, termed L<sub>1/2</sub>ESN, to improve the model's generalization ability.

For a chaotic time series, two trajectories in the same attractor will diverge significantly after a period of sample-by-sample iteration. Hence, minimizing the root mean square value of the prediction error is necessary, but not sufficient, for a successful mapping. The short-term predictability of the proposed model is considered herein. It is realized by iterated prediction which is to feed the output back to the input and form an autonomous system [11].

This paper is organized as follows. In Sect. 2, we give a brief introduction to general ESNs. Then in Sect. 3, we propose the L<sub>1/2</sub> ESN model and the iterated prediction. Afterwards, experimental results are given in Sect. 4. Finally, in Sect. 5, we draw the conclusions.

## 2 Echo State Networks

An echo state network consists in an input layer, a hidden layer and an output layer. The hidden layer, called dynamic reservoir, contains a large number of neurons and is regarded as a supplier of interesting dynamics [1]. The input-to-reservoir weight matrix  $\mathbf{W}_{in}$  and the recurrent reservoir weight matrix  $\mathbf{W}_x$  are generated randomly, whereas the reservoir-to-output weight matrix  $\mathbf{W}_{out}$  is adapted via supervised learning [2].

Denote  $\mathbf{m}(i) = [m_1(i), m_2(i), \dots, m_M(i)]^T \in \mathbf{R}^{M \times 1}$  as the collected time series,  $\mathbf{u}(i) = \mathbf{m}(i)$  and  $\mathbf{y}(i) = \mathbf{m}(i+1)$  as the input and output signals at time step  $i$ . The basic state equation is defined by

$$\mathbf{s}(i) = \tanh[\mathbf{W}_{in}\mathbf{u}(i) + \mathbf{W}_x\mathbf{s}(i-1)] \quad (1)$$

where  $\mathbf{s}(i) \in \mathbf{R}^{N \times 1}$  is the state of the network at time step  $i$ .  $\tanh(\cdot)$  is applied element-wise, and the initial state of the reservoir is a zero vector. The dependence on the initial state gradually loses as  $i$  goes to infinity [2]. The linear output layer is defined by

$$\mathbf{y}(i) = [\mathbf{s}(i) : \mathbf{u}(i)]^T \boldsymbol{\beta}_1 + \boldsymbol{\beta}_0 = \mathbf{x}^T(i) \boldsymbol{\beta} \quad (2)$$

where  $[\cdot : \cdot]$  stands for a vertical vector concatenation,  $\mathbf{x}(i) = [\mathbf{s}(i) : \mathbf{u}(i) : 1]$ ,  $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 : \boldsymbol{\beta}_0]$  are the coefficients that to be estimated. Collect  $\mathbf{x}(i)$  column-wise into a matrix  $\mathbf{X} \in \mathbf{R}^{N \times S}$ , and the corresponding  $\mathbf{y}(i)$  row-wise into a matrix  $\mathbf{Y} \in \mathbf{R}^{S \times M}$ , where  $S$  is the number of the samples. Then the readout weights  $\boldsymbol{\beta}$  are computed by a linear regression method which minimizes the mean square error between the network output and the training target output [4]:

$$\min \|\mathbf{Y} - \mathbf{X}^T \boldsymbol{\beta}\|_2^2 \quad (3)$$

where  $\|\cdot\|_2$  stands for L<sub>2</sub> norm. The least square solution of (3) is

$$\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{Y} \quad (4)$$

Sometimes, the above solution (4) is ill-posed because of the approximate collinear components in the high dimensional reservoir state matrix  $\mathbf{X}$ . This implies a bad generalization performance [7]. A solution to this problem is to use regularization, which has the general form:

$$\min \|\mathbf{Y} - \mathbf{X}^T \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_k^k \quad (5)$$

where  $\lambda$  is a regularization parameter that balances the two objective terms, and it is chosen by five-fold cross-validation.  $\|\boldsymbol{\beta}\|_k$  is taken as the  $k$  norm of  $\boldsymbol{\beta}$ .

### 3 $L_{1/2}$ Regularized Echo State Network

In this part, we introduce the  $L_{1/2}$  penalty into echo state networks to improve the prediction performance of echo state networks, i.e.,  $k$  in (5) is equal to  $1/2$ . For a fixed non-negative  $\lambda$ , the cost function (5) thus can be rewritten as

$$\begin{aligned} f(\boldsymbol{\beta}) &= \|\mathbf{Y} - \mathbf{X}^T \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_{\frac{1}{2}}^{\frac{1}{2}} = \left\{ \sum_{i=1}^M \|y_m - \mathbf{x}^T \boldsymbol{\beta}_m\|_2^2 + \lambda \sum_{m=1}^M \|\boldsymbol{\beta}_m\|_{\frac{1}{2}}^{\frac{1}{2}} \right\} \\ \|\boldsymbol{\beta}_m\|_{\frac{1}{2}}^{\frac{1}{2}} &= \sum_{j=1}^p \sqrt{\boldsymbol{\beta}_{mj}}, p = N + M + 1 \end{aligned} \quad (6)$$

$\boldsymbol{\beta}$  is estimated by minimizing the following target function:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \sum_{m=1}^M \left( \sum_{i=1}^S (y_m(i) - \mathbf{x}^T(i) \boldsymbol{\beta}_m)^2 + \lambda \|\boldsymbol{\beta}_m\|_{\frac{1}{2}}^{\frac{1}{2}} \right) \\ = \sum_{i=1}^S (y_1(i) - \mathbf{x}^T(i) \boldsymbol{\beta}_1)^2 + \lambda \|\boldsymbol{\beta}_1\|_{\frac{1}{2}}^{\frac{1}{2}} + \sum_{i=1}^S (y_2(i) - \mathbf{x}^T(i) \boldsymbol{\beta}_2)^2 + \lambda \|\boldsymbol{\beta}_2\|_{\frac{1}{2}}^{\frac{1}{2}} \\ + \cdots + \sum_{i=1}^S (y_M(i) - \mathbf{x}^T(i) \boldsymbol{\beta}_M)^2 + \lambda \|\boldsymbol{\beta}_M\|_{\frac{1}{2}}^{\frac{1}{2}} \end{aligned} \quad (7)$$

This can be decomposed as  $M$  independent optimization problems:

$$\min_{\boldsymbol{\beta}_m} \left\{ \|\mathbf{Y}_m - \mathbf{X}^T \boldsymbol{\beta}_m\|_2^2 + \lambda \|\boldsymbol{\beta}_m\|_{\frac{1}{2}}^{\frac{1}{2}} \right\}, \quad m = 1, 2, \dots, M \quad (8)$$

We use coordinate descent algorithm, a “one-at-a-time” approach, to obtain the optimal  $\boldsymbol{\beta}_m$  [10]. For each coefficient, the target function is partially optimized with respect to  $\beta_{mk}$ ,  $k = 1, 2, \dots, p$  while other elements of  $\boldsymbol{\beta}_m$  are fixed at their recently update values.

Equation (9) can be rewritten as

$$f_k(\beta_{mk}) = \sum_{i=1}^S \left[ y_m(i) - x_k(i)\beta_{mk} - \sum_{j \neq k}^p x_j(i)\bar{\beta}_{mj} \right]^2 + \lambda \left( |\beta_{mk}|^{\frac{1}{2}} + \sum_{j \neq k}^p |\bar{\beta}_{mj}|^{\frac{1}{2}} \right) \quad (9)$$

where  $\bar{\beta}_{mj}$  represent the fixed parameters. The details of the coordinate descent algorithm [10] for L<sub>1/2</sub> regularized echo state network are described as follows:

Input: Set  $\beta_m^{\text{int}} = \mathbf{0}$ ,  $m = 1, 2, \dots, M$ , and give a nonnegative constant  $\lambda$ .

Output:  $\beta_m = [\beta_{m1}, \beta_{m2}, \dots, \beta_{mp}]^T$ .

Step 1:  $\beta_m = \beta_m^{\text{int}}$ ;

Step 2: Calculate  $\beta_{mk}$ ,  $k = 1, 2, \dots, p$ ,

$$\text{if } |C_{mk}| \geq \frac{3}{4} \lambda_{mk}^{2/3}, \beta_{mk} = \frac{2}{3} C_{mk} (1 + \cos(2/3\pi - 2/3\varphi)), \text{ else } \beta_{mk} = 0,$$

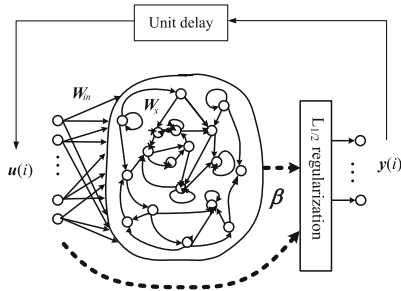
$$\text{where } C_{mk} = \sum_{i=1}^S \left( [y_m(i) - \sum_{j \neq k}^p x_j(i)\bar{\beta}_{mj}] x_k(i) \right) / \sum_{i=1}^S [x_k(i)]^2,$$

$$\lambda_{mk} = \lambda / \sum_{i=1}^S [x_k(i)]^2, \text{ and } \varphi = \arccos((\lambda_{mk}/8)|C_{mk}/3|^{-\frac{3}{2}}).$$

Step 3: if  $\sum_{k=1}^p |\beta_{mk} - \beta_{mk}^{\text{int}}| < 10^{-4}$ , the algorithm stops, else set  $\beta_m^{\text{int}} = \beta_m$ , go back to Step 1.

The coordinate descent algorithm for the L<sub>1/2</sub> regularized echo state network works well for sparsity problems, as it is not necessary to change many irrelevant parameters and recompute partial residuals for each update step.

Chaotic systems are highly sensitive to initial conditions. Small difference may yields significant diverging, rendering long-term prediction impossible [11]. Hence, a pragmatic approach for testing the short-term predictability of the L<sub>1/2</sub> regularized echo



**Fig. 1.** Iterated prediction by L<sub>1/2</sub> regularized echo state network

state network model is to feed the output back to its input, forming an autonomous system, which is a realization of iterated prediction as illustrated in Fig. 1.

## 4 Simulations

In this section, we evaluate the performance of the proposed  $L_{1/2}$  norm regularized echo state network ( $L_{1/2}$ ESN) on a typical chaotic system-Lorenz system. Some other models are conducted to compare with our proposed model: Elman network, echo state networks with  $L_1$  penalty ( $L_1$ ESN) [7], echo state networks with  $L_2$  penalty ( $L_2$ ESN) [2], and echo state networks with elastic net penalty (EESN) [8]. The accuracy of the prediction models is assessed by the root mean square error (RMSE), normalized root mean square error (NRMSE) and the symmetric mean absolute percentage error (SMAPE). The results provided herein are averages over 20 different random reservoir initializations. The Lorenz equations are defined by

$$dx/dt = \sigma(y - x), \quad dy/dt = x(\rho - z) - y, \quad dz/dt = xy - \beta z \quad (10)$$

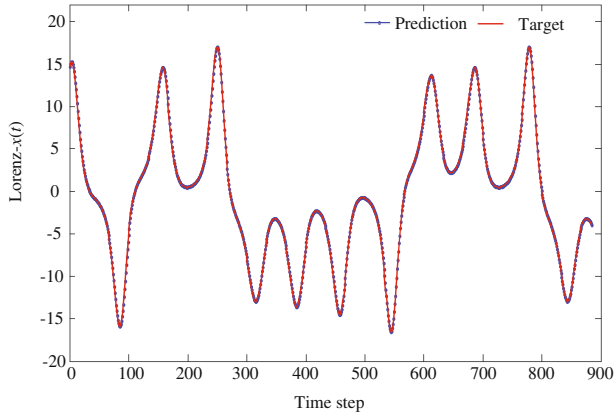
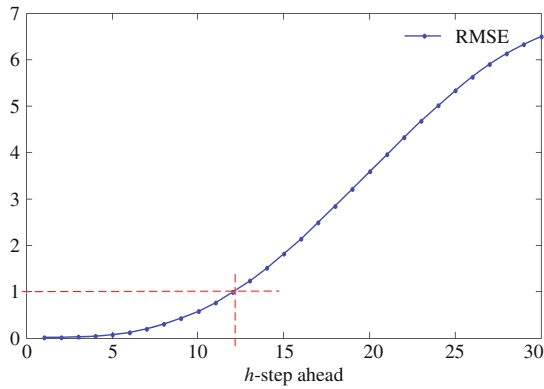
When  $\sigma = 10$ ,  $\beta = 3/8$  and  $\rho = 28$ , the system exhibits chaotic behavior. We use Runge-Kutta method to generate 5000 points with time step 0.01 from the initialized point (1, 1, 1). The Lorenz series is then reconstructed into the phase space with delay times 8, 8, 12, and embedding dimensions 2, 1, 7 for  $x$ ,  $y$ ,  $z$  series respectively. Afterwards, the first 80 % samples are used to train the model with 100 samples warming up the reservoir, and the remaining samples are used to test the model. The spectral radius of  $\mathbf{W}_x$  is set as 0.9, the sparse connectivity of  $\mathbf{W}_x$  is set as 0.05, and the input scaling parameter of  $\mathbf{W}_{in}$  is set as 0.01. Some other parameters are given in Table 1, where the regularization parameter  $\lambda$  is chosen by five-fold cross-validation.

The obtained results of all the evaluated models are provided in Table 1, and the one-step-ahead prediction curves for Lorenz- $x(t)$  produced by  $L_{1/2}$ ESN are plotted in Fig. 2. As can be seen, the  $L_{1/2}$ ESN with  $\lambda$  equal to  $10^{-7.5}$  performs much better than other models, being capable of obtaining much sparser output weights and lower RMSE, NRMSE, and SMAPE values, but when  $\lambda$  is chosen as  $10^{-7.5}$ , the effect of  $L_1$  norm is very little as all the weights are nonzero. When  $\lambda$  increases to  $5 \times 10^{-5}$ , the  $L_1$  norm makes sense since a large proportion of the output weights are zero. We can make the comment that  $L_{1/2}$ ESN is more prone to get a sparse solution than  $L_1$ ESN. One point needs to be emphasized is that  $\lambda$  is not the bigger the better, because a big  $\lambda$  will generate a big deviation in the estimation of  $\mathbf{W}_{out}$ . We also note that there are a large number of unknown weights in Elman network even for a small hidden layer. In this experiment, the numbers of input layer to hidden layer connections, hidden layer recurrent connections, and hidden layer to output layer connections are 350, 125, 350 respectively for 35 hidden nodes. The too many unknown parameters make the Elman network underfitting.

Prediction on  $h$ -step horizon by all the evaluated models are conducted by iteratively applying the predictor  $h$  times in a generative mode, where on each step it takes its own last prediction as input to do the next prediction. The  $h$ -step-ahead prediction performance produced by  $L_{1/2}$ ESN in terms of RMSE is depicted in Fig. 3. In addition,

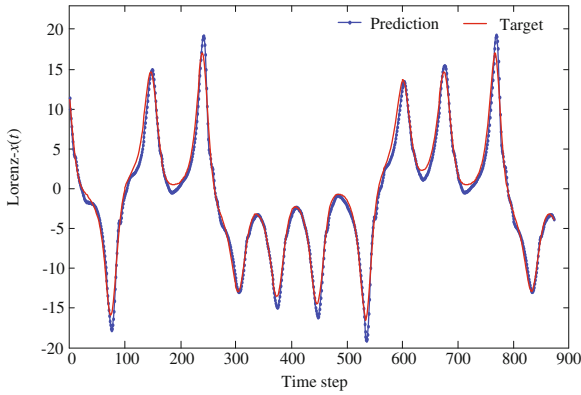
**Table 1.** Main parameters and performance of one-step-ahead prediction of evaluated models

Model	Num. of hidden nodes	$\lambda$	Num. of nonzero output weights	RMSE	NRMSE	SMAPE
Elman	35	—	[350 125 350]	0.2017	0.0256	0.2397
L <sub>1</sub> ESN	100	$5 \times 10^{-5}$	[53 18 56]	0.0153	0.0019	0.0072
L <sub>1</sub> ESN	100	$10^{-7.5}$	[110 110 110]	0.0457	0.0058	0.0214
L <sub>2</sub> ESN	100	$5 \times 10^{-5}$	[110 110 110]	0.0201	0.0025	0.0088
EESN	100	$5 \times 10^{-5}$	[51 35 59]	0.0178	0.0022	0.0245
L <sub>1/2</sub> ESN	100	$10^{-7.5}$	<b>[34 14 39]</b>	<b>0.0057</b>	<b>0.0007</b>	<b>0.0025</b>

**Fig. 2.** One-step-ahead prediction curves produced by L<sub>1/2</sub>ESN**Fig. 3.** RMSE of  $h$ -step-ahead prediction produced by L<sub>1/2</sub>ESN

**Table 2.** RMSEs of  $h$ -step-ahead prediction of all the evaluated models

$h$ -step	Elamn	$L_1$ ESN $\lambda = 5 \times 10^{-5}$	$L_2$ ESN $\lambda = 5 \times 10^{-5}$	EESN $\lambda = 5 \times 10^{-5}$	$L_{1/2}$ ESN $\lambda = 10^{-7.5}$
1	0.2017	0.0153	0.0201	0.0178	<b>0.0057</b>
2	0.2957	0.0170	0.0389	0.0259	<b>0.0103</b>
3	0.4617	0.0512	0.0605	0.0377	<b>0.0192</b>
4	0.6751	0.0767	0.1014	0.0617	<b>0.0378</b>
5	0.9240	0.1458	0.1678	0.1033	<b>0.0709</b>
6	1.2027	0.2159	0.2551	0.1656	<b>0.1224</b>
7	1.5070	0.3363	0.3645	0.2588	<b>0.1960</b>
8	1.8336	0.4260	0.5129	0.3869	<b>0.2946</b>
9	2.1790	0.6374	0.6973	0.5443	<b>0.4213</b>
10	2.5400	0.8288	0.9066	0.7206	<b>0.5777</b>
11	2.9132	1.0611	1.1474	0.9148	<b>0.7652</b>
12	3.2952	1.3095	1.4174	1.1377	<b>0.9839</b>

**Fig. 4.** Twelve-step-ahead prediction curves produced by  $L_{1/2}$ ESN

the  $h$ -step-ahead prediction RMSEs by all the evaluated models with  $h$  ranging from 1 to 12 are shown in Table 2. As can be seen, the prediction of  $L_{1/2}$ ESN is effective and accurate, as all the RMSEs are less than 1. From thirteenth step, the RMSE increases fast, and the prediction becomes inaccurate. The chaotic property of Lorenz series makes it hard for a long period of prediction. The twelve-step-ahead prediction curves produced by  $L_{1/2}$ ESN are shown in Fig. 4. Note that the overall twelve-step-ahead prediction produced by  $L_{1/2}$ ESN is accurate besides some peak points. This illustrates that the  $L_{1/2}$ ESN model can predict the Lorenz series in a short term.

## 5 Conclusions

In this paper, we propose a new model for forecasting multivariate time series, called L<sub>1/2</sub>ESN. It applies L<sub>1/2</sub> norm regularization to calculate the output weights of an ESN to solve the ill-posed problem and improve the prediction performance. The L<sub>1/2</sub> penalty imposed on the output weights outweighs L<sub>1</sub> penalty in terms of sparsity. Short-term prediction is realized by iterated prediction. Experiments results of Lorenz time series show that our proposed model obtain a higher accuracy for both one-step-ahead prediction and multiple-step-ahead prediction. The short-term predictability of the chaotic time series demonstrates the effectiveness of the proposed prediction model.

**Acknowledgement.** This work was supported by National Natural Science Foundation of China under Grant 61374154.

## References

1. Jaeger, H., Haas, H.: Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science* **304**(5667), 78–80 (2004)
2. Lukoševičius, M., Jaeger, H., Schrauwen, B.: Reservoir computing trends. *KI-Künstliche Intelligenz* **26**(4), 365–371 (2012)
3. Soh, H., Demiris, Y.: Spatio-temporal learning with the online finite and infinite echo-state gaussian processes. *IEEE Trans. Neural Netw. Learn. Syst.* **26**(3), 522–536 (2015)
4. Yuenyong, S., Nishihara, A.: Evolutionary pre-training for CRJ-type reservoir of echo state networks. *Neurocomputing* **149**, 1324–1329 (2015)
5. Chatzis, S.P., Demiris, Y.: Echo state gaussian process. *IEEE Trans. Neural Networks* **22**(9), 1435–1445 (2011)
6. Reinhart, R.F., Steil, J.J.: Regularization and stability in reservoir networks with output feedback. *Neurocomputing* **90**, 96–105 (2012)
7. Han, M., Ren, W.J., Xu, M.L.: An improved echo state network via L<sub>1</sub>-norm regularization (in Chinese). *Acta Automatica Sin.* **40**(11), 2428–2435 (2014)
8. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. Royal Stat. Soc. Ser. B (Stat. Methodol.)* **67**(2), 301–320 (2005)
9. Xu, Z.B., Chang, X.Y., Xu, F.M., Zhang, H.: L<sub>1/2</sub> regularization: a thresholding representation theory and a fast solver. *IEEE Trans. Neural Netw. Learn. Syst.* **23**(7), 1013–1027 (2012)
10. Liang, Y., Liu, C., Luan, X.Z., Leung, L.S., Chan, T.M., Xu, Z.B., Zhang, H.: Sparse logistic regression with a L<sub>1/2</sub> penalty for gene selection in cancer classification. *BMC Bioinform.* **14**(1), 198 (2013)
11. Haykin, S.S.: *Neural Networks and Learning Machines*, 3rd edn., pp. 711–722. Pearson Education, Prentice Hall, Upper Saddle River (2009)



Neural Information Processing

23rd International Conference, ICONIP 2016, Kyoto,

Japan, October 16–21, 2016, Proceedings, Part III

Akira, H.; Seiichi, O.; Doya, K.; Kazushi, I.; Minho, L.;

Derong, L. (Eds.)

2016, XVIII, 651 p. 215 illus., Softcover

ISBN: 978-3-319-46674-3