Prediction on Predictions by Ensemble Method

Hideo Hirose and Yuki Koyanagi

I. INTRODUCTION

In observing the increase of the number of patients caused by an infectious disease, it is crucial to predict the final number of infected patients. To determine whether the spread could be an outbreak or not is a great concern to everyone because a possible pandemic may affect the huge economical effect as well as the social damages. To estimate the number of infected patients, the SIR model [1], [10], [13], [4], the ordinary differential equation (ODE) model [3], [7], and the statistical truncated model [2], [5], [6], [8], [9] are considered to be useful to estimate the number of infected patients.

The predicted value for the final number of patients using data until time \( T \) becomes a function (trend) of \( T \). This is called L-plot. We here consider the use of the L-plot to predict the final number of patients, and we defined the decay function using the L-plot. Applying the multiple methodologies to the same data, we could expect the better predicted values. This is called the PoP, the prediction on predictions. As one of the PoP method, we propose to use the ensemble method. By applying these methods to the SARS case, we have found that the ensemble method works well as a PoP method.

II. PRIMARY PREDICTION METHODS

We have made the predictions for disease spread by using three primary prediction models: 1) the SIR model, 2) the ordinary differential equation model, 3) the statistical truncated model.

A. SIR Model

The SIR model is described by simultaneous ordinary differential equations to perform pandemic simulations [1], [10], [13], [4], where \( S \), \( I \), and \( R \) are susceptible, infectious, and removed populations, and the parameters \( \lambda \) and \( \gamma \) the infection rate and the removal rate (recovery rate), respectively.

\[
\begin{align*}
S'(t) &= -\lambda S(t)I(t), \\
I'(t) &= \lambda S(t)I(t) - \gamma I(t), \\
R'(t) &= \gamma I(t). 
\end{align*}
\]

The parameters \( \lambda \) and \( \gamma \) can be computed by using the best-backward solution method, BBS ([3], [7]), when we estimate the parameters \( \lambda \) and \( \gamma \) using the observed data.

B. Ordinary Differential Equation (ODE) Model

The ordinary differential equation (ODE) model [3] uses the generalized logistic distribution such that

\[
G'(t) = \frac{\beta G(t)}{\sigma} \frac{\exp(-(t-\mu)/\sigma)}{1 + \exp(-(t-\mu)/\sigma)},
\]

where, \( G(t) \) corresponds to the number of infected patients at time \( t \).

\[
G(t; \mu, \sigma, \beta) = \frac{N}{1 + \exp(-(t-\mu)/\sigma)}^{\beta},
\]

Here, \( N \) is the final number of infected patients. The parameters are estimated by using the method of least squares, and the optimization is performed by the simplex method [12].

C. Statistical Truncated Model

Although we use the same probability distribution as shown above, the method is different from that. The log-likelihood function

\[
\log L(\theta) = \sum_{i=1}^{r} n_i \log \left\{ \frac{F(t_{i+1}; \theta) - F(t_i; \theta)}{F(t_f; \theta)} \right\},
\]

is used [2], [5], [6], [8], [9], where \( t_f \) denotes the truncation time, \( t_i \) the \( i \)th day from the beginning, and \( n_i \) the number of patients on the \( i \)th day.
III. BEST-BACKWARD SOLUTION METHOD, BBS

In estimating the parameters using the observed data, we use the best-backward solution method, BBS ([7]). This is basically a method of least squares, but some extension is included. The procedure for this method is as follows:

1) We obtain initial estimates for parameters using the simple forward/backward difference method.
2) Using these initial values, we solve differential equations (5) from \( t_T \) to 0 backward, where \( t_T \) is the last time of observation. We, next, compute \( Z_0 \) as shown below,

\[
Z_0 = \sum_{j=1}^{n} (\hat{Y}(t_j) - \tilde{Y}(t_j))^2, \tag{5}
\]

where, \( Y(t_j) = R(t_j) + I(t_j) \) in the SIR model or \( Y(t_j) = NF(t_j) \) in the single distribution model having a cumulative distribution function \( F(t) \); \( \hat{Y}(t_j) \) is the observed value for \( Y(t_j) \); \( \hat{Y}(t_j) \) is the estimated value for \( Y(t_j) \). Here, observed data \( \tilde{Y}(t_j), \ (j = 1, \ldots, n) \) were assumed to be available, where \( t_T = t_n \). We find parameters so that we minimize \( Z_0 \) using the downhill simplex method by [12] by iterating backward-solution until convergence. We have applied this method to the SIR model and the ODE model.

IV. L-PLLOT

Figure 1 shows the observed and predicted number of cumulative patients using the SARS case data in Hong Kong in 2003 by the day of truncation; the prediction method is the statistical truncated model. It is very difficult to grasp the whole prediction trend in the figure.

![Fig. 1. Observed and Predicted Number of Cumulative Patients in the Case of SARS.](image)

The predicted value for the final number of patients using data until time \( T \) becomes a function (trend) of \( T \). We call this trend plot “L-plot” here. The L-plot shows us how early the prediction method predicts the final number of patients; see Figure 2, which demonstrates the SARS case. It would be beneficial if we can consider to use the L-plot in predicting the final number of patients easily. Figure 3 shows an illustration of the L-plots by various methods for the SARS case.

![Fig. 2. Concept of the L-plot.](image)

![Fig. 3. L-plots by various methods for the SARS case.](image)

V. PREDICTION ON PREDICTIONS, PoP

We are apt to select the best model from many models in an accuracy sense. For example, we often explain that the newly proposed method is superior to the conventional methods. If this tendency is always true, then this makes sense. However, we sometimes encounter cases that method A and B produce the similar results, but method C does not; in one case, A is better than B, but in another case, B is better than A. The results vary according to the situations. We cannot simply accept which is better deterministically. Let us take a new look at the prediction method. That is, we consider to use the combination method of these methods. In other words, we will make a prediction using the predicted values already observed. We call this the PoP, the prediction on predictions. One idea for this is to use the trend of the predicted final values (use of the decay function, shown...
later), and the other is to select the better candidates for predicting the final value (use of the ensemble method, shown later).

A. Use of the Decay Function

Looking at a trend itself by each prediction method (SIR, ODE, or truncated) in Figure 3, we may imagine a continuous curve fitted to the trend and its limiting value will converge to a constant value as days go on. Then, we assume the function

$$d_i(t) = c_i - b_i \exp(-a_i t)$$

(6)

where $i$ means the prediction method id; $a, b, c$ are constants to be fitted. The limiting value is $c_i$. Figure 4 shows this conceptual idea to use the decay function. We may fit a curve decaying to each predicted trend using the observed values until the truncation time $T$.

$$P = \sum_{i=n+1}^{2n+1} \left( \sum_{j=1}^{2n+1-i} \right) p^j q^{2n+1-j}.$$ 

Figure 5 shows the relationship between $p$ and $P$. We can see that $P > p$ whenever $p > 0.5$. For example, the values of $P$ are

$$P = 0.844, \quad (n = 1, \quad p = \frac{3}{4})$$

(8)

$$P = 0.790, \quad (n = 2, \quad p = \frac{2}{3})$$

(9)

$$P = 0.896, \quad (n = 2, \quad p = \frac{3}{4})$$

(10)

This shows the effectiveness of the use of ensemble method.

B. Use of the Ensemble Method

When the observed data includes the randomness, a much more accurate estimation method may be applicable; that is, two heads are better than one. The idea is similar to the ensemble methods [14].

If each individual has the same probability $p$ for success, then the value of the majority votes $P$ can be expressed as

$$P = \frac{2n+1}{2n+1} \left( \sum_{j=1}^{2n+1-i} \right) p^j q^{2n+1-j}.$$ 

Figure 6 shows the L-plots as in Figure 3 by adding the L-plot using the ensemble method for the SARS case.

VI. THE CONDITIONS IN THE SIMULATION STUDY

We are dealing with the SARS case here. The literature [11] shows the mean incubation period of the disease is estimated to be 6.4 days (95% confidence interval is [5.2, 7, 7]). If we apply the SIR method to the real data case, we refer to this information. However, in the simulation study mimicked to the SARS case, we assume that the incubation period is just one day because we try to compare the results with those obtained by other methods which may not require the value of the incubation period.

In the simulation, the final number of patients is set to 1,000, and $S_0$ is set to 5,000. The data generation is followed by the generalized logistic distribution function.
with parameter values, $\mu = 3.99$, $\sigma = 12.56$, $\beta = 3.27$, which came from the maximum likelihood estimates in the real SARS case in Hong Kong, 2003 [9]. In this paper, we have performed 100 simulation cases for the purpose of comparison.

VII. PREDICTED RESULTS BY USING THE PoP

To show the trend to each prediction method, we made box-plots using 100 simulation cases as shown in Figure 7. The SIR shows the high bias to the final value in earlier stages. The ODE and the statistical truncated methods show the similar results, revealing the low bias in earlier stages. We may expect that a simple use of the mean value from the three may provide a better value because the ODE and truncated results show the lower bias contrary to the SIR results.

A. Restricted Root Mean Square Error, $rRMSE$

To determine the accuracy for the prediction method, we can use the root mean square error. However, we introduce, here, the restricted root mean square error, $rRMSE$; since the predicted values for the final number of patients sometimes may have very large values or may not converge, we will compress these values to the boundary of the window (see dotted box in Figure 8 on the top). Here, $rRMSE$ is defined as

$$rRMSE(j) \equiv \frac{1}{|\Delta_j|} \sum_{k \in \Delta_j} (\hat{W}_r(\infty|t_T = k) - W(\infty))^2,$$

where $\hat{W}(\infty|t_T = k)$ means the estimate of $W(\infty)$ when using the data from the beginning to the truncation time $t_T$ such as

$$\hat{W}_r(\infty|t_T = k) = \min(\hat{W}(\infty|t_T = k), 2W(\infty)).$$

$\Delta_j$ expresses the days in the target area, and $|\Delta_j|$ denotes the number of days in $\Delta_j$. $W_r(\infty|t_T = k)$ attracts $\hat{W}(\infty|t_T = k)$ at the boundary $2W(\infty)$ if $W(\infty|t_T = k) > 2W(\infty)$. Figure 8 on the bottom shows an illustrative example for the $rRMSE$.

Figure 9 the $rRMSE$ for L-plot of the SIR, ODE, statistical truncated models, ensemble method, and the mean value for the SARS Case. The ensemble method provides a good result.

Selecting the majority votes, we pick up two nearest neighbors out of three, and take a mean value of the two for the new prediction. This is called the ensemble method here. The ensemble method could remove the noisy estimates by the SIR method although the mean value was affected by this noise. Figure 10 shows the $rRMSE$ for L-plot of the SIR, ODE, statistical truncated models, ensemble method, and the mean value after decaying process for the SARS Case. The figure reveals that the decaying process works and that the SIR and the ensemble methods show lower $rRMSE$ values. We may use the ensemble method as a PoP method.

VIII. CONCLUSION

In observing the widely spread of patients caused by infectious diseases or the increase of the number of failures of equipments, it is crucial to predict the final number of infected patients or failures at earlier stages. To estimate the number of infected patients, the SIR model is commonly used even when the size of observed data is small. Other methods, such as the ordinary differential equation model, statistical truncated model are also useful to estimate the number of infected patients. These methods are also applicable to the increase of the number of failures. The predicted value for the final number of patients using data until time $T$ becomes a function (trend) of $T$. We call this L-plot. We here consider the use of the L-plot to predict the final number of patients, and we defined the decay function using the L-plot. Applying the multiple methodologies to the same data, we could expect the better decay function. This is called the PoP, the prediction on predictions. As one of the PoP method, we also proposed to use the ensemble method. The PoP includes to use the simple mean value, the decay function, and the ensemble method. By applying these methods to the SARS case, we have found that the ensemble method works well as one of the PoP methods.

REFERENCES

Fig. 7. Box Plots for the Predicted Trends using the SIR, ODE, Statistical Truncated Models.

Fig. 8. An Illustrative Example for the rRMSE.

Fig. 9. rRMSE for L-plot of the ODE, Statistical Truncated Models, Ensemble Method, Mean Value for the SARS Case.

Fig. 10. rRMSE for L-plot of the SIR, ODE, Statistical Truncated Models, Ensemble Method.