The number of HCV-infected person modeling via Gaussian-HMMs

김수현[1)]

Abstract

본 논문에서는 Gaussian Hidden Markov Model(GHMM)을 이용하여 2001년부터 현재(2022년 12월 1주차)까지 일주일 단위로 C형간염 바이러스(HCV)에 감염된 사람들의 수를 모델링하였다. GHMM의 상태(state) 수를 조절하여 state 수가 2개(2-state GHMM)인 경우부터 7개(7-state GHMM)까지 모델을 적합시켰다. 모델 성능 비교 지표로써 Akaike Information Criterion(AIC) 기준으로는 6-state GHMM이 가장 좋은 성능을 보였고 Bayesian Information Criterion(BIC) 기준으로는 4-state GHMM이 가장 좋은 성능을 보였다. 이 논문에서는 BIC를 중점적으로 보았으며 최종적으로는 4-state GHMM을 선택하였고, 감염자 수가 최근 3년 동안 감소 추세에 있음을 확인할 수 있었다.

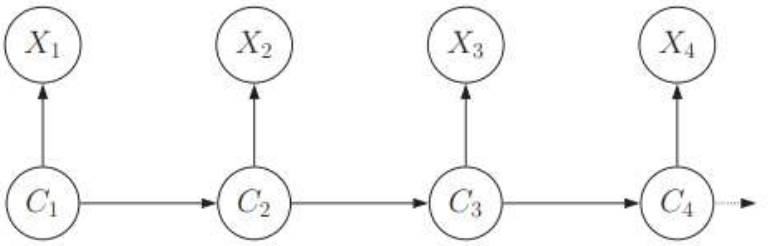
1. Introduction

HCV는 C형간염 바이러스(Hepatitis C Virus) 감염에 의한 급, 만성 간질환이다. 주된 감염원은 HCV에 오염된 혈액이나 기구이며 전파경로는 주사기 공동 사용, 수혈, 혈액투석, 성접촉 등 혈액매개 전파이다. 따라서 일상생활에서 사람 간 전파 가능성은 극히 낮다. 잠복기는 2주에서 6개월이며 평균적으로 6주에서 10주이다. 급성 C형간염과 만성 C형간염으로 나뉘는데 급성 C형간염은 초기 감염 후 70~80%의 환자에서 무증상이며 서서히 감기 몸살과 구역질, 식욕부진 등의 증상이 나타난다. 만성 C형간염은 60~80%의 환자에서 무증상이며 만성 피로감, 간부전 등의 간경변증이 발생한다. 진단은 HCV 특이 유전자(RNA)를 검출해서 양성 반응이 나온 경우 감염되었다고 판정한다.

Hidden Markov Model(HMM)은 관찰 가능한 관측치와 관측 불가능한 어떤 상태의 두 가지 요소로 구성되어 있는 모델이다. 관측 가능한 관측치는 오로지 관측 불가능한 상태에만 영향을 받게 되고 관측 불가능한 상태들은 마르코프 과정(Markov Process)을 따르는 것을 가정한다. 따라서 상태를 직접적으로 관측할 수는 없고, 상태들로부터 야기된 결과만이 관측할 수 있다. GHMM은 관측 불가능한 상태의 분포를 가우시안 분포(Gaussian distribution)로 가정한 것이다.

HMM은 시간의 흐름에 따라 변화하는 패턴을 인식하는 분석에 유용한 모델이다. 음성 인식, 동작 인식(Gesture Recognition), 부분 방전(Partial discharge), 생물정보학 분야에서 이용된다.

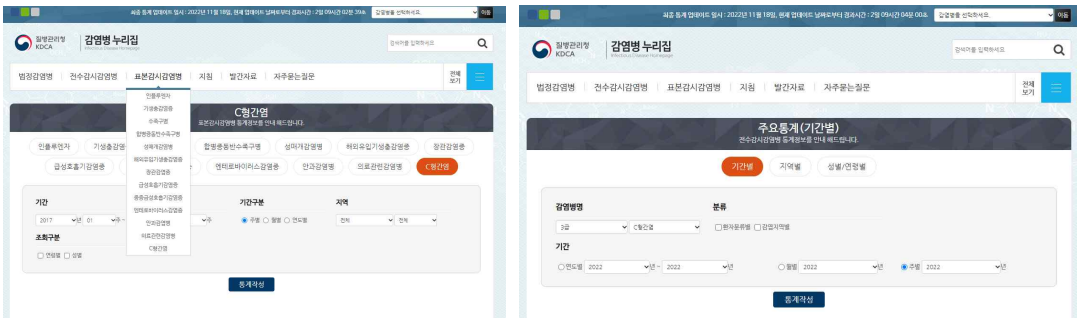
[1)]30019 세종시 세종로 2511, 고려대학교 공공정책대학 경제통계학부 국가통계전공 학사과정, E-mail : i99ksh@korea.ac.kr, 학번: 2018380502



< Directed graph of basic HMM >

2. Data Description and Preprocessing

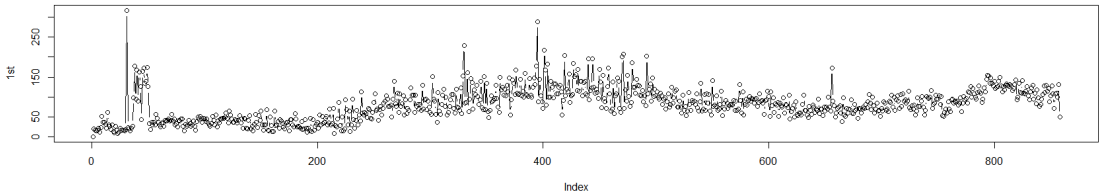
HCV 감염자 수에 대한 데이터는 질병관리청(KDCA) 감염병 누리집에서 제공하고 있다.



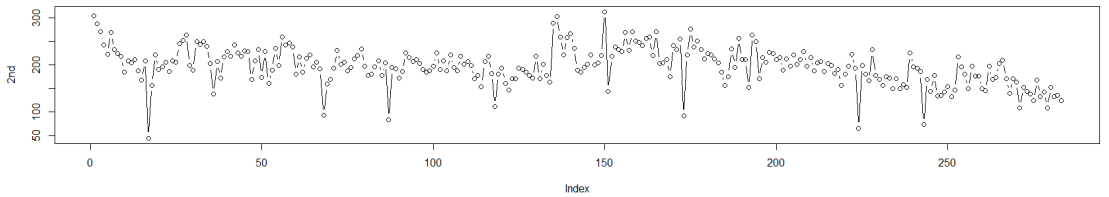
< 질병관리청 감염병 누리집 홈페이지 >

데이터의 특징으로는 2001년 1주차부터 2017년 23주차까지는 표본감시체계로 조사되었고 2017년 24주차부터 현재까지는 전수감시체계로 데이터 수집 기준이 다르다. 표본감시체계는 국가 관리가 필요한 감염병 중 감염병 환자 발생의 전수 보고가 어렵거나 중증도가 비교적 낮고 발생률이 높은 감염병에 대해 일부 표본기관을 지정하여 자료를 지속적, 정기적으로 수집, 분석, 배포하여 이를 감염병의 예방, 관리에 활용하는 감시체계이다. 전수감시체계는 모든 의료기관에서 감염병 환자 등을 진단했을 때 환자 발생을 사례별로 보건 당국에 신고하여, 감염병 환자 등의 관리와 유행 확산 방지 대응을 가능토록 하는 감시체계이다. 따라서 2001년 1주차부터 2017년 23주차까지의 데이터는 표본자료(sample)이며, 2017년 24주차부터 현재까지의 데이터는 모집단(population)이라고 볼 수 있다.

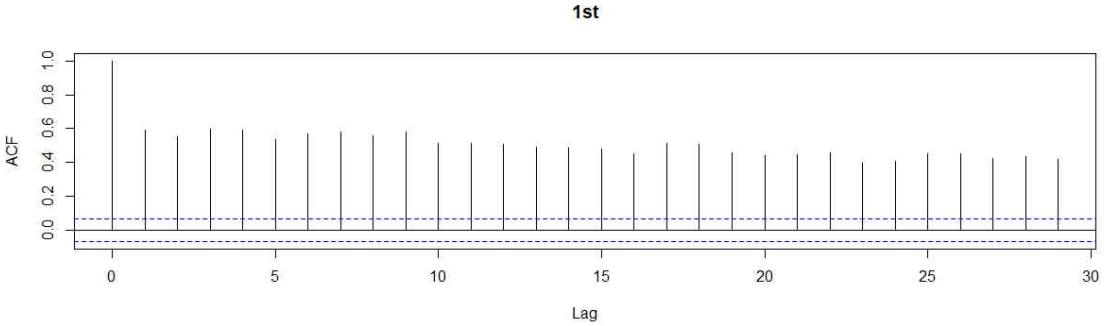
데이터의 index plot과 ACF(Autocorrelation Function), 히스토그램은 아래와 같다.



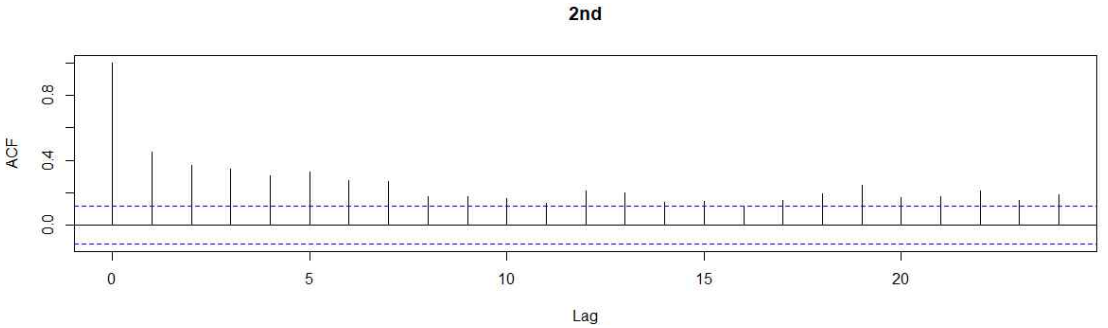
< Index plot (Week1, 2001 - Week23, 2017) >



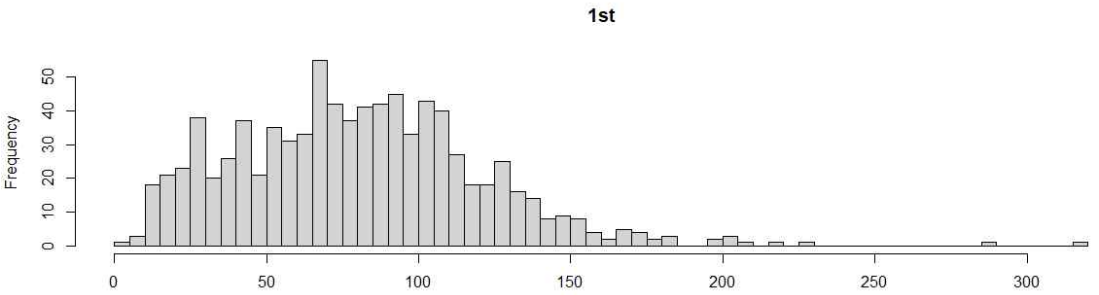
< Index plot (Week24, 2017 - Present) >



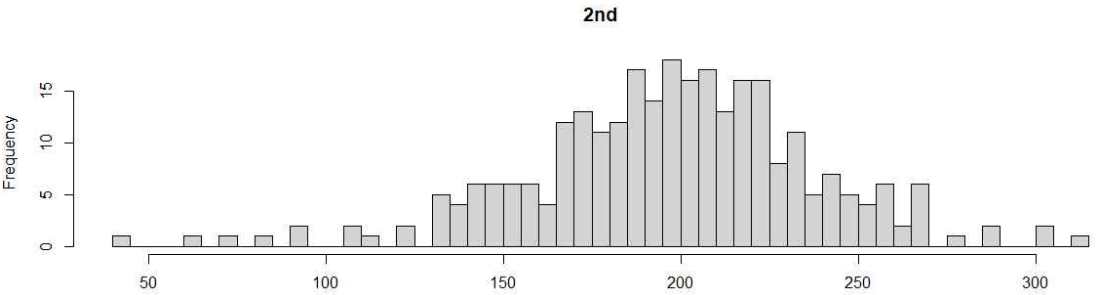
< ACF (Week1, 2001 - Week23, 2017) >



< ACF (Week24, 2017 - Present) >



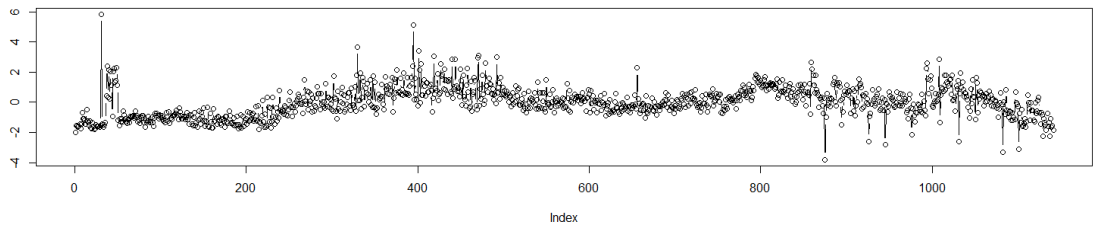
< Histogram (Week1, 2001 - Week23, 2017) >



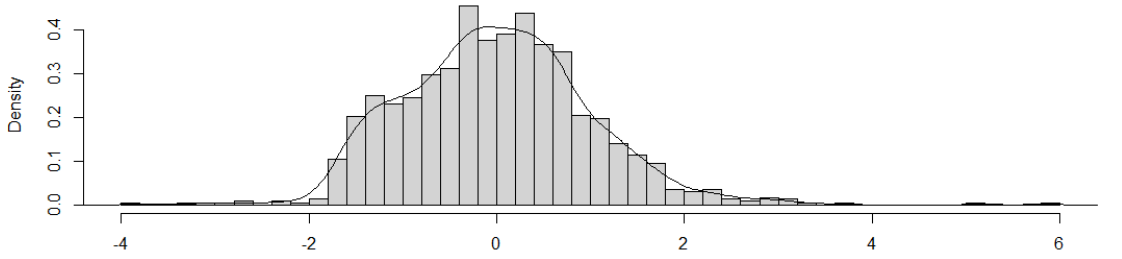
< Histogram (Week24, 2017 - Present) >

이상치(outlier)가 몇 개 관측되긴 하지만 값을 조정하거나 제거하지는 않고 그대로 사용하였다. 또한 ACF를 보면 자기상관이 보인다. 따라서 마르코프 연쇄(Markov chain)를 가정하는 HMM 모델을 사용하는 것은 적절해 보인다.

다른 방식으로 수집된 두 데이터 셋을 하나로 합쳐서 하나의 모델을 만들어 추세를 확인하는 것이 목표이므로 두 데이터 셋을 하나로 합치기 위해서 표준화(standardization)를 하였다. 표준화 후 두 데이터 셋을 합치면 Index plot과 히스토그램은 아래와 같다.

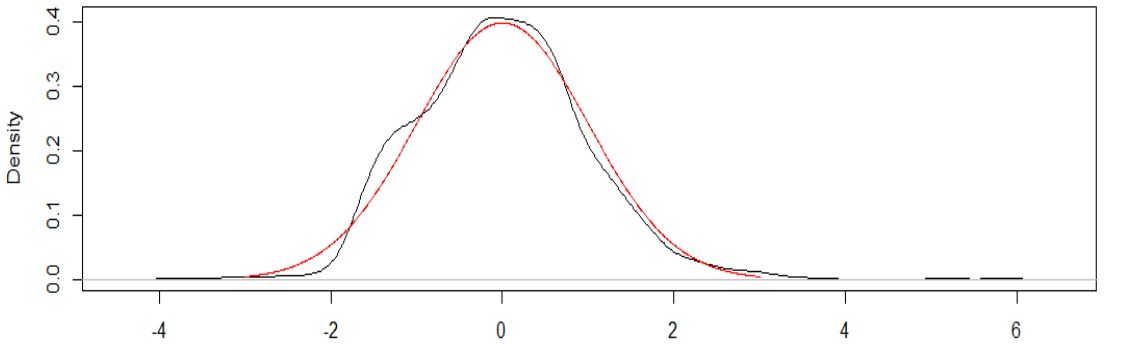


< Index plot of combined data >



< Histogram of combined data >

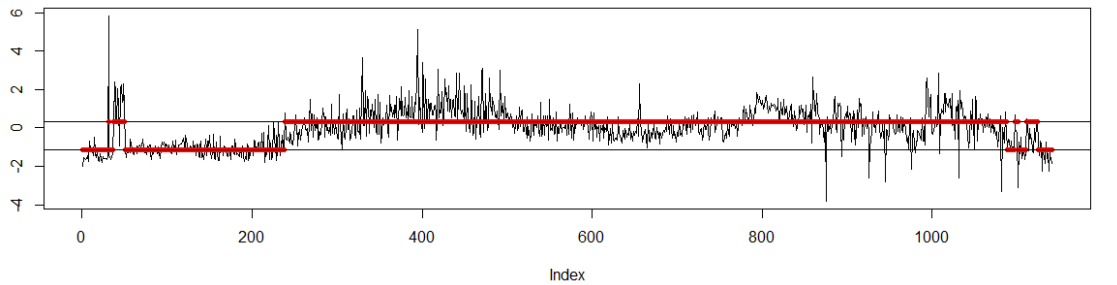
히스토그램의 밀도를 보면 단일분포보다는 혼합분포(mixture distribution)로 모델링하는 것이 필요해 보인다. 따라서 이 역시 상태(state)별 분포가 다른 HMM 모델을 사용하는 것은 적절하다.



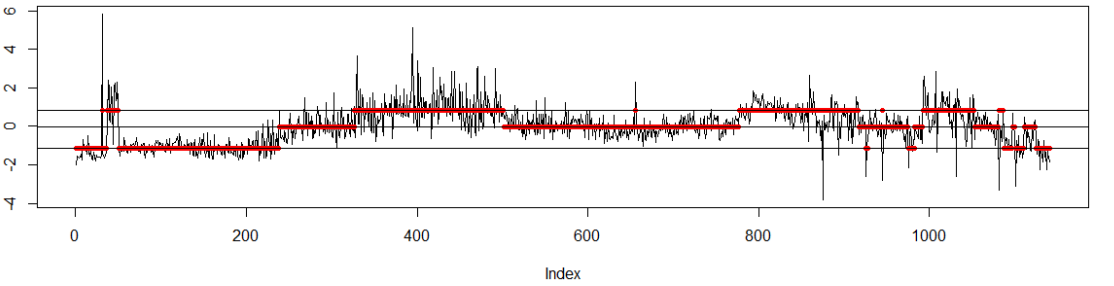
< Density plot of data and density plot of standard normal distribution >

3. Modeling

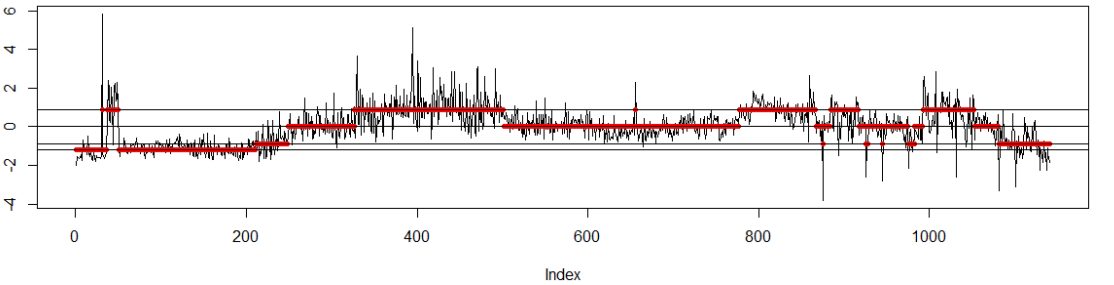
추정해야 할 파라미터의 개수는 $2m + m^2$개다(m은 state의 수). R의 'hmmr' 패키지를 이용해 모델링하였다. 2개의 상태를 가지는 GHMM에서 5개의 상태를 가지는 GHMM의 모델링 결과는 아래와 같다.



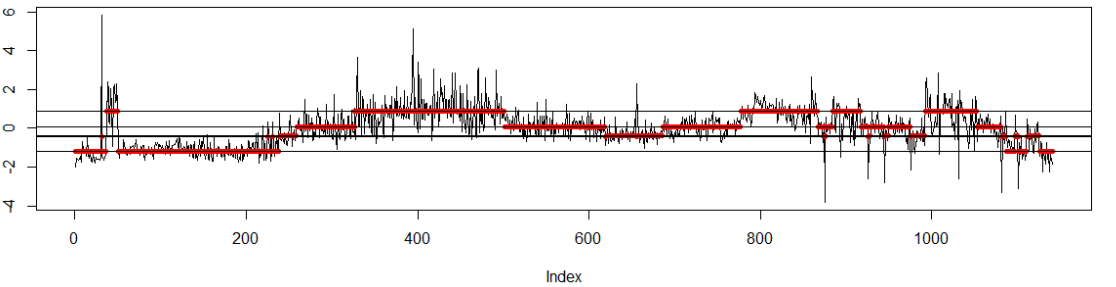
< 2-state GHMM >



< 3-state GHMM >



< 4-state GHMM >



< 5-state GHMM >

각 상태 개수별 모델의 파라미터는 다음과 같이 추정되었다.

$$\hat{\mathbf{\Gamma}} = \begin{bmatrix} 0.993 & 0.007 \\ 0.024 & 0.976 \end{bmatrix} \qquad \begin{aligned} \hat{\mu}_1 &= 0.332 \quad \hat{\sigma}_1^2 = 0.863 \\ \hat{\mu}_2 &= -1.163 \quad \hat{\sigma}_2^2 = 0.381 \end{aligned}$$

< 2-state GHMM >

$$\hat{\Gamma} = \begin{bmatrix} 0.967 & 0.012 & 0.021 \\ 0.023 & 0.969 & 0.008 \\ 0.024 & 0.007 & 0.968 \end{bmatrix} \qquad \begin{aligned} \hat{\mu}_1 &= -0.014 \quad \hat{\sigma}_1^2 = 0.467 \\ \hat{\mu}_2 &= -1.160 \quad \hat{\sigma}_2^2 = 0.422 \\ \hat{\mu}_3 &= 0.812 \quad \hat{\sigma}_3^2 = 0.968 \end{aligned}$$

< 3-state GHMM >

$$\hat{\Gamma} = \begin{bmatrix} 0.971 & 0.000 & 0.016 & 0.013 \\ 0.000 & 0.985 & 0.010 & 0.005 \\ 0.020 & 0.05 & 0.975 & 0.000 \\ 0.051 & 0.000 & 0.000 & 0.949 \end{bmatrix} \qquad \begin{aligned} \hat{\mu}_1 &= 0.009 \quad \hat{\sigma}_1^2 = 0.456 \\ \hat{\mu}_2 &= -1.172 \quad \hat{\sigma}_2^2 = 0.331 \\ \hat{\mu}_3 &= 0.881 \quad \hat{\sigma}_3^2 = 0.892 \\ \hat{\mu}_4 &= -0.886 \quad \hat{\sigma}_4^2 = 0.807 \end{aligned}$$

< 4-state GHMM >

$$\hat{\Gamma} = \begin{bmatrix} 0.000 & 0.000 & 0.743 & 0.000 & 0.257 \\ 0.000 & 0.983 & 0.000 & 0.014 & 0.003 \\ 0.031 & 0.008 & 0.898 & 0.048 & 0.014 \\ 0.021 & 0.011 & 0.001 & 0.966 & 0.000 \\ 0.027 & 0.004 & 0.000 & 0.000 & 0.969 \end{bmatrix} \qquad \begin{aligned} \hat{\mu}_1 &= -0.449 \quad \hat{\sigma}_1^2 = 2.280 \\ \hat{\mu}_2 &= 0.861 \quad \hat{\sigma}_2^2 = 0.854 \\ \hat{\mu}_3 &= -0.359 \quad \hat{\sigma}_3^2 = 0.401 \\ \hat{\mu}_4 &= 0.091 \quad \hat{\sigma}_4^2 = 0.443 \\ \hat{\mu}_5 &= -1.174 \quad \hat{\sigma}_5^2 = 0.367 \end{aligned}$$

< 5-state GHMM >

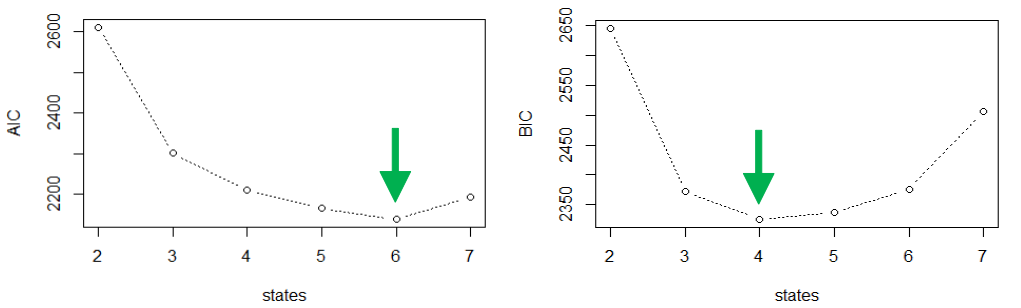
5-state GHMM의 특징은 전이확률행렬(transition probability matrix, TPM)에 나타나는데 state 1에서 state 1로의 전이확률(transition probability)이 0에 수렴한다는 것이다. 또한 state 1의 분포 파라미터를 보면 분산이 2.280으로 다른 분포의 분산에 비해 매우 크게 추정되었다. 이는 state 1로 상태가 변한 경우에는 상태가 지속됨 없이 바로 다른 상태로 넘어간다는 이야기이고 과적합(overfitting)이 발생했다고 볼 수도 있다. 따라서 5-state부터는 이처럼 지속되지 않는 상태 즉, 거쳐 가는 상태의 분포가 계속해서 발생할 것이며 과적합 위험성이 커진다.

4. Model Selection and Checking

각 상태 개수별 모델의 AIC와 BIC는 다음과 같이 추정되었다.

	2-state	3-state	4-state	5-state	6-state	7-state
AIC	2610.072	2301.231	2209.518	2165.289	2139.452	2193.809
BIC	2645.349	2371.787	2325.430	2336.638	2376.316	2506.267

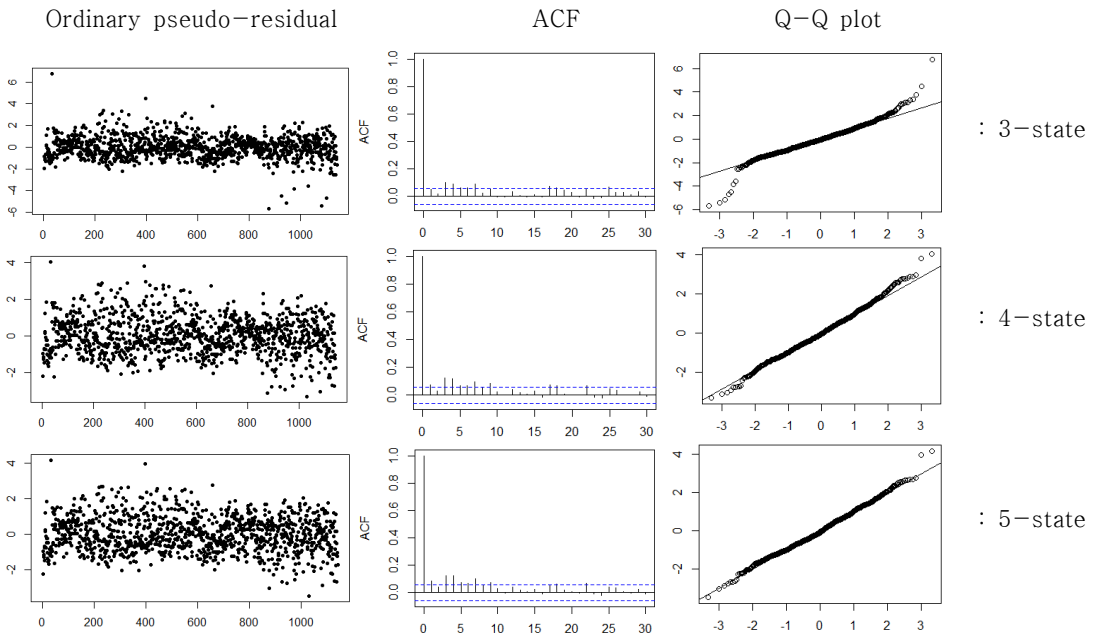
< Table of AIC and BIC >



< Graph of AIC and BIC >

AIC 기준으로 보면 6-state일 때의 값이 2139.452로 가장 작다. 그리고 BIC 기준으로 보면 4-state일 때의 값이 2325.430으로 가장 작다. AIC와 BIC 모두 값이 작을수록 좋은 성능을 보이는데 어떤 지표를 따를 것인지에 따라 선택되는 모델이 다르다. 5-state GHMM의 TPM을 보았을 때 5-state부터는 과적합 될 위험성이 있다고 판단했고 본 연구에서는 BIC 지표를 모델 비교 지표로 선택하였다.

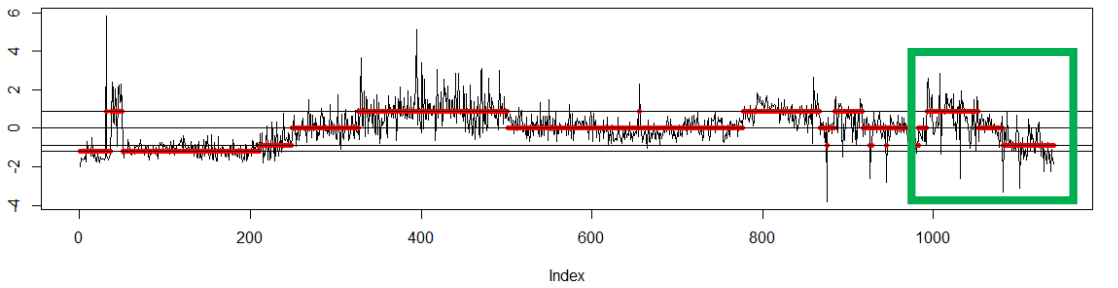
이후 잔차분석을 하였으며 분석 결과는 아래와 같이 요약되었다.



3-state의 Q-Q plot을 보면 양측 꼬리가 퍼지는 경향이 있다. 이는 모델이 모든 관측값에 대해 설명을 잘 하고 있지 못하다고 볼 수 있고 4-state와 5-state의 Q-Q plot을 보면 어느 정도 잘 적합되었다고 판단할 수 있다. 또한 자기상관도 거의 없으며 pseudo-residual plot도 확인을 해보면 4-state와 5-state의 차이가 크게 나지 않는다고 볼 수 있다. 통계학은 더 적은 파라미터를 가지는 모델이 선호되는 것을 고려하면 4-state 모델을 선택하는 것이 적절하다고 할 수 있다.

5. Conclusion

AIC 기준으로는 6-state GHMM이, BIC 기준으로는 4-state GHMM이 best model로 선택되었다. 그러나 5-state GHMM의 추정된 TPM에서 보이듯이 5-state부터는 과적합의 위험성이 있다고 판단하였고 본 연구에서는 BIC 기준을 따르는 4-state GHMM을 best model로 선정하였다.



< Result of 4-state GHMM >

위 그림에서 녹색 박스 부분은 최근 3년 정도의 추세를 보여준다. 이 부분을 보았을 때, 시간이 지날수록 더 낮은 평균값을 가지는 상태가 선택되는 것을 볼 수 있다. 따라서 HCV 감염자 수가 유의미하게 줄어들고 있다고 판단할 수 있다.

Reference

[1] Walter Zucchini, lain L. MacDonald, Roland Langrock. *Hidden Markov Models for Time Series : An Introduction Using R, Second Edition*. Chapman & Hall, 2016.

[2] Sook-Hyang Jeong, EunSun Jang, Hwa Young Choi, Kyung-Ah Kim, WankyoChung, Moran Ki (2017), "**Current Status of Hepatitis C Virus Infection and Countermeasures in South Korea**". Epidemiology and Health, Vol 39; 2017. https://doi.org/10.4178/epih.e2017017