JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 21, No. 2 : 259~264 Feb. 2017

한국 신생아의 출생체중 데이터 보정

신형식*

Adjustment of Korean Birth Weight Data

Hyungsik Shin*

School of Electronic and Electrical Engineering, Hongik University, Seoul 04066, Korea

요 약

신생아의 출생체중은 자궁내발육부전이나 과체중출생아를 진단하는 데 사용되는 등, 의학적으로 여러 가지 중요한 정보를 제공한다. 본 논문에서는 2011년부터 2013년까지 한국에서 태어난 신생아의 출생체중 데이터를 분석하고, 생물학적으로 부자연스러운 체중 분포를 관찰할 수 있음을 보인다. 이러한 비상식적인 체중 분포는 데이터 수집 과정 등에서 오류가 존재함을 의미하는데, 특히 임신주수가 28주에서 32주인 신생아들의 체중 데이터에서 현저한 오류 데이터를 관찰할 수 있다. 이를 보정하기 위해, 본 논문은 가우시안 혼합 모델을 사용하여 오류 데이터와 정상 데이터를 예측하고, 오류 데이터로 예측된 자료들을 삭제하는 과정을 제안한다. 제안된 보정 과정을 통하여 보다 자연스럽고 의학적으로 의미 있는 출생체중 백분율을 구할 수 있음을 보인다.

ABSTRACT

Birth weight of a new born baby provides very important information in evaluating many clinical issues such as fetal growth restriction. This paper analyzes birth weight data of babies born in Korea from 2011 to 2013, and it shows that there is a biologically implausible distribution of birth weights in the data. This implies that some errors may be generated in the data collection process. In particular, this paper analyzes the relationship between gestational period and birth weight, and it is shown that the birth weight data mostly of gestational periods from 28 to 32 weeks have noticeable errors. Therefore, this paper employs the finite Gaussian mixture model to classify the collected data points into two classes: non-corrupted and corrupted. After the classification the paper removes data points that have been predicted to be corrupted. This adjustment scheme provides more natural and medically plausible percentile values of birth weights for all the gestational periods.

키워드 : 출생체중, 임신주수, 데이터분석, 가우시안 혼합 모델

Key word : birth weight, gestational period, data analysis, Gaussian mixture model

Received 16 January 2017, **Revised** 16 January 2017, **Accepted** 19 January 2017

* **Corresponding Author** Hyungsik Shin (E-mail:hyungsik.shin@hongik.ac.kr, Tel:+82-2-320-1123)
School of Electronic and Electrical Engineering, Hongik University, Seoul 04066, Korea

Open Access http://doi.org/10.6109/jkiice.2017.21.2.259 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

신생아의 출생체중은 자궁내발육부전이나 과체중출생아를 진단하는 데 사용되는 등, 의학적으로 여러 가지 중요한 정보를 제공한다[1, 2]. 따라서 한국에서 태어나는 신생아의 임신 기간(Gestational period)과 출생체중(Birth weight)의 관계를 파악하여 일선 의료 현장에서 활용하는 것은 매우 중요한 일이다.

신생아의 출생체중은 임신 기간에 따라 달라지는데, 일반적으로 임신 기간이 짧을수록 출생체중도 감소하는 경향이 있다. 이는 수태 기간이 짧을수록 모체 안에서 성장할 기간이 줄어듦으로 인하여 발생하는 자연적인 결과라고 예상할 수 있다.

하지만, 실제 한국의 통계청에서 수집한 자료를 분석하면 이러한 기대와는 다른 모습을 관찰할 수 있으며, 통계청에서 수집한 자료에 오류가 섞여있음이 현저하게 드러난다[3]. 이러한 오류 데이터가 발생하는 이유에 대해서는 아직 확실하게 밝혀진 바는 없으며, 다만 여러 가지 이유가 있을 것으로 학자들이 예측하고 있다.

본 논문에서는 한국 통계청에서 수집한 데이터의 오류를 제거하여 한국 신생아의 임신주수와 출생체중의 관계를 보다 현실화하는 과정을 소개한다. 이와 비슷한 연구가 논문[3-6]에 의하여 소개되었으나, 사용된 데이터가 비교적 오래된 데이터이므로 최근의 관계를 나타낸다고 보기는 어렵다. 또한, 해외에서도 비슷한 연구가 진행되었으나[7-9], 한국인의 특성에 잘 맞지 않을 것으로 예상된다. 이에 본 논문은 2011년부터 2013년까지의 최근 데이터를 활용하여 오류 데이터를 제거하고, 이를 통하여 한국인의 임신 기간과 출생체중 사이의 최근의 관계를 정립한다.

Ⅱ. 출생체중 데이터 분석

본 논문에서 분석한 데이터는 한국 통계청에서 수집한 자료로서, 2011년부터 2013년까지의 총 3년간 한국에서 태어난 신생아 중, 임신주수가 22주에서 42주까지의 신생아들의 임신주수와 출생체중 데이터이다[10]. 이러한 조건을 만족하는 신생아 수는 총 1,425,986명이며, 본 논문의 데이터 분석에 모두 포함되었다.

2.1. 출생체중 분포 곡선

우선적으로 분석해 보아야 할 것은, 2011년부터 2013년까지 한국에서 태어난 신생아에 대하여 각각의 임신주수(Gestational period)에 따른 출생체중의 분포도를 알아보는 것이다. 그림 1은 임신주수가 40주인 경우의 신생아 출생체중의 분포도이다. 이 그림에서 볼 수 있듯이, 40주의 임신 기간을 거친 신생아들은 정상분포 곡선으로 예상되는 분포를 따라 출생체중이 분산되어 있음을 알 수 있다[11]. 대략 3.2 kg을 중심으로 좌우 대칭적인 형태로 출생체중이 분포하고 있음을 관찰할 수 있다.

그림 2는 임신주수가 28주인 경우의 신생아 출생체중의 분포도이다. 이 그림을 관찰해보면, 출생체중의 분포가 정상 분포 곡선을 따르는 듯처럼 보이다가, 과체중 구간에서 빈도수가 현저하게 증가하는 또 다른 집단이 위치해 있음을 알 수 있다.

즉, 임신주수가 28주에 해당하는 신생아들의 출생체중 분포는 1.2 kg 부근에서 가장 많은 빈도수를 보이며 대체적으로 정규분포 곡선을 따르는 것처럼 보이지만, 3.0 kg를 중심으로 또 다시 현저하게 증가하는 빈도수를 관찰할 수 있다. 이처럼 과체중 구간에 현저하게 증가하는 빈도수는 생물학적 자연 현상으로부터 비롯되었다기보다는 어떠한 오류 데이터로부터 발생한 것으로 짐작할 수 있다.

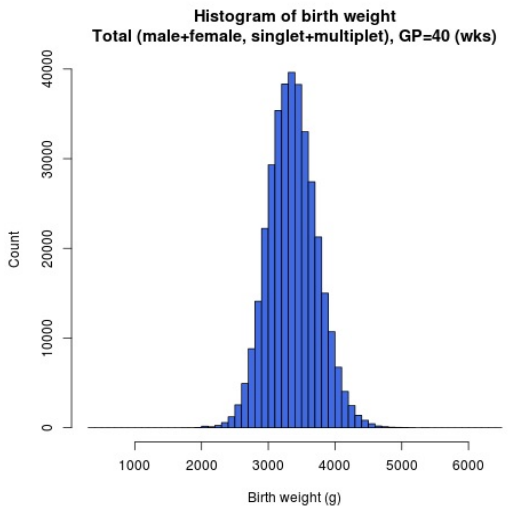

Fig. 1 Histogram of birth weight with gestational period of 40 weeks, 2011-2013, Korea

이러한 현상은 비단 한국에서만 관찰되는 것이 아니며, 논문[7, 9]에서 보고되었듯이 해외에서도 관찰할 수 있다. 많은 학자들이 이러한 데이터가 존재하는 이유에 대하여 연구하고 있으나, 아직 뚜렷한 원인을 밝혀내지 못하였다. 이러한 오류가 발생하는 여러 가지 가능한 원인 중의 하나로는, 산모가 자신의 임신 시기를 정확히 계산하지 못함으로 인해 발생하는 임신주수 계산 착오를 생각해 볼 수 있다.

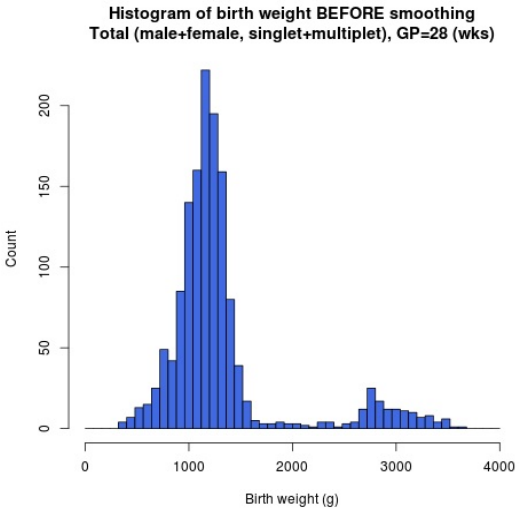

Fig. 2 Histogram of birth weight with gestational period of 28 weeks, 2011-2013, Korea

2.2. 임신주수별 출생체중 백분율

임신주수에 따른 신생아의 출생체중 분포를 표현하는 방법에는 여러 가지가 있을 수 있는데, 본 논문에서는 각 임신주수에 따른 신생아 출생체중의 백분율 분포를 표로 나타내었다.

표 1은 각 임신주수에 해당하는 신생아들의 출생체중에서 각각의 백분율에 해당하는 출생체중을 나타낸 것이다. 표 1의 어두운 배경색으로 표시한 부분에 나타나 있듯이, 임신주수가 28~30 주의 출생체중 데이터에서 백분율이 90% 이상인 구간에 비정상적으로 보이는 과체중 신생아가 많이 분포되어 있음을 알 수 있다.

표 1의 데이터를 그림으로 나타내면 오류데이터로 보이는 자료들이 데이터에 포함되어 있음이 보다 현저하게 나타난다. 그림 3은 표 1의 백분율 계산 결과를 나타낸 것인데, 임신주수가 28~30주에 해당하는 백분율 데이터가 두드러지게 왜곡되어 있는 것을 볼 수 있다. 이 기간 동안의 신생아들은 임신주수가 31~32주의 신생아보다 더 큰 출생체중을 보이는데, 이는 아마도 데이터 수집과정에서의 오류로 인한 것으로 짐작된다.

Table. 1 Birth weight percentiles (in gram) for each gestational period (in week), 2011-2013, Korea

GP	5%	10%	25%	50%	75%	90%	95%
22	404	430	470	520	580	620	640
23	461	500	550	590	650	700	740
24	470	540	620	690	760	820	877
25	510	590	710	790	870	940	980
26	580	680	820	910	1000	1080	1130
27	630	740	910	1020	1130	1239	1300
28	740	850	1030	1190	1340	2251	2900
29	879	990	1163	1340	1480	1680	2970
30	953	1100	1320	1500	1680	2385	3008
31	1100	1250	1478	1670	1830	2010	2201
32	1230	1400	1640	1840	2040	2260	2520
33	1414	1590	1820	2040	2240	2460	2620
34	1600	1780	2020	2250	2475	2710	2890
35	1860	2010	2250	2490	2720	2970	3160
36	2080	2220	2460	2700	2960	3220	3400
37	2320	2490	2720	2980	3240	3500	3680
38	2600	2740	2940	3180	3420	3660	3820
39	2710	2840	3040	3270	3520	3760	3900
40	2800	2930	3130	3370	3610	3850	4000
41	2870	3000	3200	3420	3670	3900	4040
42	2850	2980	3200	3420	3680	3920	4100

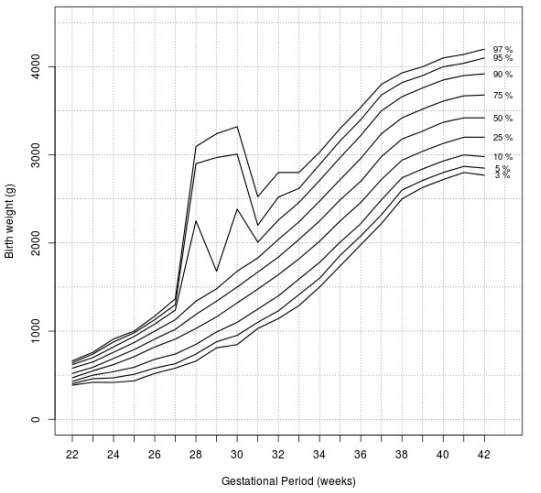

Fig. 3 Birth weight percentiles for gestational periods, 2011-2013, Korea

Ⅲ. 오류 데이터 제거

3.1. 오류 데이터 판별

앞에서 살펴보았듯이, 임신주수가 28~30주까지의 데이터는 오류인 것으로 예측되는 자료들을 포함하고 있다. 따라서 이러한 오류 데이터를 예측하여 제거한 후에 다시 백분율 데이터를 계산하면 좀 더 의학적으로 의미 있는 임신주수와 출생체중 간의 관계를 정립할 수 있을 것으로 기대한다.

3.2. Gaussian mixture model

본 논문에서는 오류 데이터로 보이는 항목을 예측하기 위하여, 각각의 임신주수에 해당하는 출생체중 데이터에는 두 개의 Gaussian 분포가 함께 혼합되어 있으며, 과체중 쪽의 Gaussian 분포는 오류에 의해 생성된 데이터라고 가정하였다. 이렇게 두 개의 Gaussian 분포가 함께 혼합되어 있는 것은 Gaussian mixture model로 표현할 수 있다. Gaussian mixture model을 가정하여 오류 데이터를 제거하는 과정은 논문[12, 13]에 소개되었다.

이러한 가정 하에, 오류 데이터로 예측되는 항목들을 모두 제거한 후, 해당 임신주수의 출생체중 백분율을 다시 계산하였다. 이러한 과정을 거쳐 얻어진 백분율 데이터는 보정 과정 이전의 백분율 데이터보다 오류가 적을 것으로 예상된다. 또한, 보정된 백분율 데이터는 일선 의료 현장에서 좀 더 정확하게 여러 가지 신생아 관련 질환을 진단하는데 도움을 줄 수 있을 것으로 기대된다.

3.3. Scikit-learn

위에서 언급한 두 개의 Gaussian 분포가 함께 혼합되어 있는 것은 Gaussian mixture model 로 표현할 수 있다. 많은 소프트웨어 패키지들이 Gaussian mixture model을 제공하는데, 본 논문은 오픈소스(Open-source) 소프트웨어인 scikit-learn[14]을 활용하였다.

프로그래밍 언어 Python에 기반을 둔 패키지인 scikit-learn은 다양한 기계학습(machine learning) 및 인공지능(artificial intelligence) 알고리즘들을 구현한다. 이를 통하여 scikit-learn 사용자들로 하여금 비교적 쉽게 여러 가지 데이터 분석 알고리즘을 사용자 데이터에 적용할 수 있도록 하였다. Gaussian mixture model도 scikit-learn에 구현되어 있다.

3.4. 오류데이터 제거

본 논문은 위에서 언급한 Gaussian mixture model을 사용하여 각각의 임신주수에서 오류로 의심되는 데이터들을 예측하고 제거하는 방법을 적용하였다. 즉, 각각의 신생아 데이터에 대하여, Gaussian mixture model을 통해 예측한 결과, 과체중 쪽에 분포한 Gaussian으로부터 생성된 데이터로 예측될 경우, 해당 데이터를 삭제하였다.

임신주수가 28주인 신생아들의 데이터에서 오류 데이터로 의심되는 항목들을 제거한 후의 출생체중 분포도를 그려보면 그림 4와 같다. 그림 4에서 볼 수 있듯이, 과체중 구간에 존재하였던 오류 데이터일 것으로 예측되는 빈도수들이 모두 제거되었음을 알 수 있다. 이와 같은 보정 작업을 임신주수 28~32 주에 걸쳐 총 5주에 해당하는 데이터에 각각 적용한 후, 다시 백분율을 계산한 결과는 표 2에 나타내었다.

위와 같은 보정 과정을 거친 백분율 데이터를 그림으로 표현하면 그림 5와 같다. 그림 3에서 현저하게 나타나 있던 오류 데이터들이 모두 제거되어, 임신 주수가 늘어날수록 출생체중도 자연스럽게 증가하는 백분율 분포를 얻었음을 관찰할 수 있다.

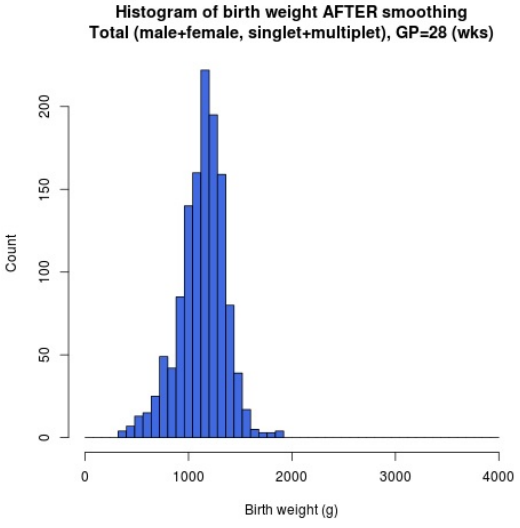

Fig. 4 Adjusted histogram of birth weight with gestational period of 28 weeks, 2011-2013, Korea

Table. 2 Adjusted birth weight percentiles (in gram) for each gestational period (in week), 2011-2013, Korea

GP	5%	10%	25%	50%	75%	90%	95%
22	404	430	470	520	580	620	640
23	460	500	550	590	650	700	740
24	470	540	620	690	760	820	877
25	510	590	710	790	870	940	980
26	580	680	820	910	1000	1080	1130
27	630	740	910	1020	1130	1239	1300
28	720	830	1010	1160	1280	1390	1460
29	870	970	1150	1320	1450	1580	1650
30	940	1077	1300	1470	1620	1760	1840
31	1130	1260	1470	1660	1800	1940	2020
32	1320	1460	1650	1820	2000	2150	2230
33	1414	1590	1820	2040	2240	2460	2620
34	1600	1780	2020	2250	2475	2710	2890
35	1860	2010	2250	2490	2720	2970	3160
36	2080	2220	2460	2700	2960	3220	3400
37	2320	2490	2720	2980	3240	3500	3680
38	2600	2740	2940	3180	3420	3660	3820
39	2710	2840	3040	3270	3520	3760	3900
40	2800	2930	3130	3370	3610	3850	4000
41	2870	3000	3200	3420	3670	3900	4040
42	2850	2980	3200	3420	3680	3920	4100

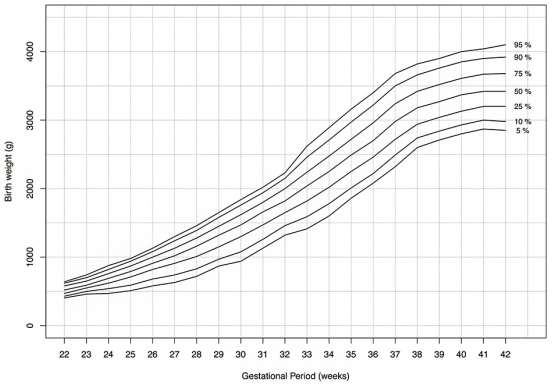

Fig. 5 Adjusted birth weight percentiles for gestational periods, 2011-2013, Korea

Ⅳ. 결 론

본 논문에서는 2011년부터 2013년까지 한국에서 태어난 신생아 약 142만 명의 출생체중과 임신 주수(Gestational period)와의 관계를 분석하였다. 기존의 여러 연구에서도 보고되었듯이, 출생체중 데이터는 오류로 짐작되는 자료들이 포함되어 있다. 본 논문은 이러한 오류로 짐작되는 데이터들을 예측하고 제거하는 방법을 제시하고 최근의 데이터에 적용하였다. 이러한 보정 과정을 거친 후의 데이터는 좀 더 의학적으로 의미있는 임신 주수와 출생체중의 관계를 도출하는데 도움이 될 것으로 기대한다.

각 임신주수 별로 출생체중의 분포를 살펴보면, 정상분포곡선을 따르는 듯 보이지만, 과체중 구간에서 또 다른 높은 빈도수를 가지는 집단을 관찰할 수 있는데, 이는 오류로부터 기인한 것으로 짐작된다. 본 논문에서는 이를 finite Gaussian mixture model을 사용하여, 두 개의 Gaussian이 혼합되어 있는 분포로 가정한 후, 과체중 구간에 위치한 Gaussian분포 구간을 오류 데이터로 판단하고 데이터 집단에서 제거하였다. 이를 구현하기 위하여 데이터 분석 소프트웨어 패키지인 scikit-learn을 활용하였다.

위와 같은 보정 과정을 거친 후의 데이터를 사용하여 다시 임신 주수와 출생체중의 백분율을 계산하면, 임신주수가 길어질수록 출생체중도 자연스럽게 증가하는 백분율 분포곡선을 구할 수 있다. 이렇게 얻어진 백분율 분포곡선은 신생아와 관련된 여러 가지 의료 활동에 있어서 보다 정확한 진단 기준을 마련하는데 도움을 줄 것으로 기대한다.

ACKNOWLEDGMENTS

This work was supported by 2016 Hongik University Research Fund.

REFERENCES

[1] S. Paranjothy, F. Dunstan, W. J. Watkins, M. Hyatt, J. C. Demmler, R. A. Lyons, and D. Fone, “Gestational age, birth weight, and risk of respiratory hospital admission in childhood,” *Pediatrics*, vol. 132, no. 6, pp. e1562-1569, Dec. 2013.

[2] J. L. Richards, C. Hansen, C. Bredfeldt, R. A. Bednarczyk, M. C. Steinhoff, D. Adjaye-Gbewonyo, K. Ault, M.

Gallagher, W. Orenstein, R. L. Davis, and S. B. Omer, "Neonatal outcomes after antenatal influenza immunization during the 2009 H1N1 influenza pandemic: impact on preterm birth, birth weight, and small for gestational age birth," *Clinical Infectious Diseases*, vol. 56, no. 9, pp. 1216-1222, May 2013.

[3] J. J. Lee, "Birth weight for gestational age patterns by sex, plurality, and parity in Korean population," *Korean Journal of Pediatrics*, vol. 50, no. 8, pp. 732-739, Aug. 2007.

[4] G. H. Lee, Y. W. Kim, E. J. Seo, M. S. Son, H. G. Ahn, E. W. Seok, Y. J. Choi, G. J. Kim, S. Y. Kim, B. C. Hwang, Y. D. Choi, S. Y. Kim, and S. J. Sohn, "Change of birth weight-gestational age table," *Korean Journal of Obstetrics and Gynecology*, vol. 44, no. 10, pp. 1851-1856, Oct. 2001.

[5] G. Y. Jung and K. Lee, "Intrauterine growth of Korean infants from 25 weeks to 44 weeks gestation," *Journal of the Korean Pediatric Society*, vol. 33, no. 7, pp. 887-900, Jul. 1990.

[6] J. J. Lee, C. G. Park, and K. S. Lee, "Birth weight distribution by gestational age in Korean population: using finite mixture model," *Korean Journal of Pediatrics*, vol. 48, no. 11, pp. 1179-1186, Nov. 2005.

[7] N. M. Talge, L. M. Mudd, A. Sikorskii, and O. Basso, "United States birth weight reference corrected for implausible gestational age estimates," *Pediatrics*, vol. 133, no. 5, pp. 844-853, May 2015.

[8] J. Villar, L. C. Ismail, C. G. Victora, E. O. Ohuma, E. Bertino, D. G. Altman, A. Lambert, A. T Papageorghiou, M. Carvalho, Y. A. Jaffer, M. G. Gravett, M. Purwar, I. O. Frederick, A. J. Noble, R. Pang, F. C. Barros, C. Chumlea, Z. A. Bhutta, and S. H. Kennedy, "International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21 st Project," *The Lancet*, vol. 384, no. 9946, pp. 857-868, Sep. 2014.

[9] M. S. Kramer, R. W. Platt, S. W. Wen, K. S. Joseph, A. Allen, M. Abrahamowicz, B. Blondel, and G. Breart, "A new and improved population-based Canadian reference for birth weight for gestational age," *Pediatrics*, vol. 108, no. 2, pp. e35-e35, Aug. 2001.

[10] Statistics Korea, Vital Statistics [Internet]. Available: https://mdis.kostat.go.kr/extract/extSurvSearchByDate.do?extcTypeDivCD=E&curMenuNo=UI_POR_P1070.

[11] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065-1076, Sep. 1962.

[12] R. W. Platt, M. Abrahamowicz, M. S. Kramer, K. S. Joseph, L. Mery, B. Blondel, G. Breart, and S. W. Wen, "Detecting and eliminating erroneous gestational ages: a normal mixture model," *Statistics in medicine*, vol. 20, no. 23, pp. 3491-3503, Dec. 2001.

[13] H. Oja, M. Koiranen, and P. Rantakallio, "Fitting mixture models to birth weight data: a case study," *Biometrics*, vol. 47, no. 3, pp. 883-897, Sep. 1991.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, Oct. 2011.

신형식(Hyungsik Shin)

2016년 3월 ~ 현재: 홍익대학교 전자전기공학부 조교수
2011년 3월: Stanford University, Electrical Engineering, 공학박사
※관심분야 : 데이터마이닝, 통계학습, 인공지능, 분산제어