

일반화 선형혼합모형(GLMM)을 이용한 청성안정유발반응(ASSR) 연구

박상현¹ and 장원철*,¹

¹대한민국 서울특별시 08826, 서울대학교 자연과학대학 통계학과, *지도교수

초록 본 연구에서는 청성안정유발반응(auditory steady state evoked responses; ASSR)을 통해 측정된 뇌전도로 행동을 예측하는 모형을 통해 청각 피질(auditory cortex; AC), 전두엽 피질(prefrontal cortex; PFC) 및 두정엽 피질(posterior parietal cortex; PPC)의 뇌전도가 생쥐의 행동 결정에 어떤 방식으로 연관되는지를 밝히려고 시도하였다. 이를 위해 기존 예측 모형의 한계를 지적하고 개선된 예측 모형인 일반화 선형혼합모형(generalized mixed effect model; GLMM)을 제안한다. GLMM은 고정효과(fixed effect)와 랜덤효과(random effect)를 모두 고려함으로써 반복측정으로 인한 상관관계 구조를 반영할 수 있고, 결과적으로 약 .925의 F score를 보였다.

KEYWORDS

GLMM

ASSR

Prediction model

CONTENTS

서론	2
모형 적합	2
GLMM 모형	2
자료 전처리	6
탐색적 자료분석	8
GLMM 모형 적합	10
CV 및 시뮬레이션	11
결과	15
코드 및 분석도구	18
Acknowledgement	18

Manuscript compiled: Wednesday 3rd June, 2020

¹교신저자: 박상현, 대한민국 서울특별시 08826, 서울대학교 자연과학대학 통계학과, Email: lkd1962@naver.com

Technicality	18
라플라스 근사를 이용한 GLMM의 로그가능도 근사	18
GLMM의 고정효과에 대한 검정통계량의 극한분포	20
GLMM의 랜덤효과에 대한 예측값	22
이진 분류 문제와 CV	23
베이즈 분류기	25
참고문헌	29

서론

ASSR(Auditory Steady State Evoked Responses; 청성안정유발반응) 실험은 주기적으로 빠르게 반복되는 청각 자극에 대한 EEG(Electroencephalography)나 LFP(Local Field Potential)를 관측하는 실험이다 (Bohórquez & Özdamar 2008). 1981년 Galambos와 동료들이 40Hz대의 청각 자극에서 안정파 유발반응 (steady state evoked potential)을 보고한 것을 시작으로 (Galambos *et al.* 1981), Rickards, Piocton, Choen 등이 다양한 주파수에서의 ASSR 반응을 광범위하게 연구하며 신경과학 및 청각 임상에서의 표준적인 실험절차로 자리잡게 되었다 (Rickards & Clark 1982; Picton *et al.* 1987; Kuwada *et al.* 1986; Cohen *et al.* 1991). 특히 이는 다른 자극과는 달리 주어진 주파수에 해당하는 와우(cochlea)의 좁은 부위를 자극하여 주파수 특이성이 높고 (John & Picton 2000), 반응의 자동탐지 알고리즘이 잘 개발되어 객관적인 반응 탐지가 가능하다는 장점이 있다 (Cone-Wesson *et al.* 2002).

본 연구에서는 독립성 가정이 위배된다는 점을 해결할 수 있는 방안으로 GLMM(Generalized Linear Mixed Effect Model; 일반화 선형혼합모형)을 이용한 예측을 제안하고자 한다. GLMM은 오차항이 하나인 t 검정이나 Wilcoxon 부호순위합 검정의 모형과는 달리 여러 개의 오차항을 동시에 고려할 수 있는 일반적인 통계 모형으로 반복측정으로 인해 생기는 종속성을 서로 독립으로 간주할 수 있는 여러 개의 오차 원인으로 분해하여 모형에 반영한다 (Pinheiro & Bates 2006; Agresti *et al.* 2000). 나아가, GLMM은 예측 모형으로도 활용될 수 있다는 장점을 가진다. 물론 예측만을 위한 ANN(Artificial Neural Network) 등의 비모수적 방법이 계산 신경과학 분야에서 널리 이용되고 있지만, 이가 가지는 본질적인 해석의 어려움 때문에 통계적 추론이 필요한 분야에까지 적용되기는 쉽지 않다. GLMM은 모수적인 모형으로서 해석의 용이함을 유지하면서 예측 모형으로도 나쁘지 않은 성능을 발휘한다.

모형 적합

GLMM 모형

다중선형회귀모형은 설계행렬 X , 반응변수 벡터 \mathbf{Y} , 회귀계수 벡터 β , 오차 $\varepsilon \sim \mathbf{N}_n(\mathbf{0}, \sigma^2 I_n)$ 에 대해 다음과 같은 모형으로 정의된다.

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

여기서 β 는 비록 알 수는 없지만 고정된 상수로서, 위의 모형은 설명변수의 가중합에 약간의 오차가 더하여져 반응 Y 가 결정된다고 보아 이때 가중치의 역할을 하는 β 를 추정하고자 한다. 곧 Y 의 랜덤성은 오로지 오차 ε 에 기인한다. 이러한 다중선형회귀모형에 더 이상 상수가 아닌 확률변수를 회귀계수로 추가한 모형을 생각할 수 있는데, 이를 LME(Linear Mixed Effect Model; 선형 혼합모형)라 하며, 다음과 같이 정의된다 ([Demidenko 2013](#)).

$$\mathbf{Y} = \underset{n \times 1}{X} \underset{n \times p}{\beta} + \underset{n \times q}{Z} \underset{q \times 1}{\gamma} + \underset{n \times 1}{\varepsilon}$$

여기서 Z 는 랜덤효과 γ 의 설계행렬이며 $\gamma \sim \mathbf{N}_q(\mathbf{0}, \Psi)$, $\varepsilon \sim \mathbf{N}_n(\mathbf{0}, \Sigma)$ 이고 γ 와 ε 은 서로 독립으로 가정한다. 새롭게 추가된 랜덤효과 γ 와 구별하기 위해 고정된 상수인 β 를 고정효과라 한다. 랜덤효과의 도입에 더하여 일반적으로 LME에서는 오차항의 분산 또한 더 이상 $\sigma^2 I$ 으로 제한하지 않는다. 따라서 LME에서는 반응 Y 가 결정되는 데 있어 더 이상 오차 ε 만이 유일한 랜덤성으로 적용하지 않는다. 설명변수의 가중합에서 가중치의 일부에도 랜덤성이 있다고 보는 것이다. 일견 이렇게 랜덤효과를 도입하는 것이 어떠한 차이를 만드는지 명확하지 않을 수 있으나, 다중선형회귀모형에서 indicator variable을 더미(dummy) 변수로 사용하여 분산분석을 온전히 기술할 수 있듯이 위의 LME는 아주 폭넓은 구조의 자료를 다룰 수 있도록 해 준다 ([Searle & McCulloch 2001](#)).

이제 LME에서 γ 가 조건부로 주어지면 이때 $\mathbf{Y}|\gamma$ 는 다중선형모형으로 축소된다는 점에 주목하고 다중선형모형을 GLM으로 일반화할 때 사용한 아이디어를 그대로 차용하여 오차가 정규분포가 아닌 경우로 LME를 확장하면 GLMM을 얻는다. 즉, overdispersion이 없는 지수족 분포의 PDF(Probability Density Function; 확률밀도함수) $f(x; \eta) = h(x) \exp(\eta x - A(\eta))$ 와 적당한 link function $g : \mathbb{R} \rightarrow \mathbb{R}$ 에 대해 GLMM은

$$f_{Y_i|\gamma}(y_i|\gamma) = h(y_i) \exp(\eta_i y_i - A(\eta_i)) \quad (1)$$

$$g(\mathbf{E}(Y_i|\gamma)) = \eta_i = \mathbf{x}'_i \beta + \mathbf{z}'_i \gamma$$

으로 정의된다. 여기서 $\gamma \sim \mathbf{N}_q(\mathbf{0}, \Psi)$ 이고 $i = 1, \dots, n$ 이며 \mathbf{x}'_i 와 \mathbf{z}'_i 는 각각 LME에서의 고정효과의 설계행렬 X 와 랜덤효과의 설계행렬 Z 의 i 번째 행에 대응된다.

전술한 바와 같이 GLMM은 단순히 랜덤효과를 더한다는 표면적인 의미를 넘어 다양한 구조를 갖는 자료를 다룰 수 있도록 해 주는데, 본 연구의 주제인 반복측정된 자료에 이가 어떻게 적용될 수 있는지 살펴보자. 피실험체에 m 마리에 대해 각각 n 번 반복측정하여 생성된 자료의 경우, 각 자료간의 독립성은 보장할 수 없겠지만 이때 측정에 더해지는 오차를 피실험체 수준의 오차와 반복측정 수준의 오차로 분해하면 이러한 오차들 간의 독립성은 보장할 수 있을 것이다. 따라서 다음의 로지스틱 선형혼합모형을 생각할 수 있다.

$$Y_{ij} | \varepsilon_i \stackrel{\text{iid}}{\sim} \text{Bern}(p_{ij}) \quad (2)$$

$$g(p_{ij}) = \underset{p \times 1}{\mathbf{x}_{ij}'} \underset{p \times 1}{\beta} + \varepsilon_i$$

여기서 $g(x) = \log(x/(1-x))$, $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathbf{N}(0, \sigma^2)$ 이고 $i = 1, \dots, m$ 는 피실험체를 나타내는 첨자이며 $j = 1, \dots, n$ 는 특정 피실험체에서 측정된 순서를 나타내는 첨자이다. 위의 모형에서 반복측정 수준의

모형만 보면 Y_{ij} 는 i 번째 피실험체의 j 번째 반복에서의 반응, p_{ij} 는 설명변수의 선형결합으로 표현되는 모수(베르누이 분포에서의 성공확률)로 이는 고전적인 로지스틱 회귀모형이다. 그러나 위의 모형은 모수 p_{ij} 에 피실험체 수준의 오차 ε_i 를 더하여 Y_{ij} 가 서로 독립이 아니라는 문제점을 해결한다. 즉, 실험 과정에 있어 오차가 매번의 반복에서 발생할 뿐만 아니라, 피실험체를 선택하는 과정에서도 발생하여 더해진다는 것이다. 이때 작용하는 반복측정 수준의 오차와 피실험체 수준의 오차에 대해서는 어느정도 그 독립성을 보장할 수 있기에 이는 신경과학계의 반복측정 실험에 적합한 모형이라 할 수 있다.

이제 $\text{Bern}(p)$ 의 확률밀도함수가 $h(x) = \mathbf{1}_{\{0,1\}}(x)$, $\eta = \log(p/(1-p))$, $A(\eta) = -\log(1-p)$ 에 대해 (여기서 **1**는 지시함수)

$$f(x) = h(x) \exp(\eta x - A(\eta)) \quad (3)$$

와 같이 지수족 분포로 표현될 수 있다는 점을 떠올리면

$$\underset{mn \times p}{X} = \begin{bmatrix} \mathbf{x}'_{11} \\ \mathbf{x}'_{12} \\ \vdots \\ \mathbf{x}'_{mn} \end{bmatrix}, \underset{mn \times m}{Z} = \begin{bmatrix} \mathbf{1}_n & \mathbf{0}_n & \cdots & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{1}_n & \cdots & \mathbf{0}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_n & \mathbf{0}_n & \cdots & \mathbf{1}_n \end{bmatrix}, \underset{m \times 1}{\gamma} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix} \quad (4)$$

와 $\Psi = \sigma^2 I_m$, logistic link $g(x) = \log(x/(1-x))$ 에 대해

$$\begin{aligned} f_{Y_{ij}|\gamma}(y_{ij}|\gamma) &= h(y_{ij}) \exp(\eta_{ij} y_{ij} - A(\eta_{ij})) \\ g(\mathbf{E}(Y_{ij}|\gamma)) &= \eta_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\gamma \end{aligned} \quad (5)$$

로 쓸 수 있으며 이때 $\gamma \sim \mathbf{N}_m(\mathbf{0}, \Psi)$ 가 되어 방금 살펴보았던 GLMM이 된다. (여기서 \mathbf{z}'_{ij} 는 행렬 Z 의 $n(i-1) + j$ 번째 행을 의미한다.) 이로부터 GLMM이 반복측정이 빈번한 신경과학계의 자료에 적합한 모형임을 알 수 있다.

이제 GLMM의 모수 추정과 검정에 대해 생각해보자. GLMM 모형(식 1)에서 추정해야 할 모수로는 고정효과 β 와 분산구조 Ψ 가 있다. 여기서 γ 는 랜덤효과로 고정된 미지의 상수가 아니므로 추정의 대상이 아니다. GLM에서와 마찬가지로 추정은 MLE(Maximum Likelihood Estimation; 최대가능도 추정)을 이용하는데, \mathbf{Y} 의 로그가능도함수를 구하면 분산구조 Ψ 를 알고 있다는 가정 하에 베이즈 정리로부터 다음과 같이 구해진다.

$$\begin{aligned} l(\beta) &= \log \int_{\mathbb{R}^q} f_{\mathbf{Y}|\gamma}(\mathbf{Y}|\gamma) f_\gamma(\gamma) d\gamma \\ &= \log \int_{\mathbb{R}^q} \exp \left(\sum_{i=1}^n [\mathbf{z}'_i \gamma Y_i - A(\mathbf{x}'_i \beta + \mathbf{z}'_i \gamma)] - \frac{\gamma' \Psi^{-1} \gamma}{2} \right) d\gamma - \frac{1}{2} \log \det \Psi + \sum_{i=1}^n \mathbf{x}'_i \beta Y_i + \text{const.} \end{aligned} \quad (6)$$

대부분의 경우 위 식의 적분이 닫힌 형태로 구해지지 않으므로 adaptive GHQ(Gauss-Hermite Quadrature)나 라플라스 근사, MCMC(Markov Chain Monte Carlo Simulation), MCEM(Monte Carlo Expectation

Maximization) 등을 통한 수치적분 기법으로 적분을 근사하여 계산한다 (Capanu *et al.* 2013). 실제로는 대부분의 경우 분산구조 Ψ 또한 알지 못하므로 이를 동시에 추정해야 하고, 이를 위해 적당한 초기값에서 시작하여 조금씩 Ψ 와 β 를 update하는 반복 알고리즘이 사용된다.

GLMM을 적합한 이후에는 특정 고정효과 β_i 대해 가설 $H_0 : \beta_i = 0$ 을 검정하거나 특정 랜덤효과 γ 에 대응되는 분산 σ_γ^2 에 대해 가설 $H_0 : \sigma_\gamma^2 = 0$ 을 검정함으로써 관심 있는 뇌 활동이 반응의 유무에 유의미한 영향을 주는지에 대한 통계적 추론을 진행할 수 있다. 먼저 고정효과 β 에 대해 보다 일반적인 선형가설 $H_0 : T\beta = \xi \in \mathbb{R}^r$ (단, $\text{rk } T = r < p$)을 검정하는 경우, 이를 검정하는 방법으로 크게 LRT(Likelihood Ratio test; 가능도비 검정), Wald 검정, Rao 검정을 생각해볼 수 있다. 각 검정은 다음 검정통계량의 계산을 필요로 한다.

$$\Lambda = -2[l(\hat{\beta}, \hat{\Psi}) - l(\tilde{\beta}, \tilde{\Psi})] \quad (7)$$

$$W = (T\hat{\beta} - \xi)' \{ T[\nabla_{\beta}^2 l(\hat{\beta}, \hat{\Psi})]^{-1} T' \}^{-1} (T\hat{\beta} - \xi) \quad (8)$$

$$R = \nabla_{\beta} l(\tilde{\beta}, \tilde{\Psi})' [\nabla_{\beta}^2 l(\tilde{\beta}, \tilde{\Psi})]^{-1} \nabla_{\beta} l(\tilde{\beta}, \tilde{\Psi}) \quad (9)$$

여기서 $\hat{\beta}$ 와 $\hat{\Psi}$ 는 각각 full model의 β 와 Ψ 의 MLE이며 $\tilde{\beta}$ 와 $\tilde{\Psi}$ 는 각각 귀무가설 하에서 reduced model의 β 와 Ψ 의 MLE이다. 그렇다면 $\Lambda, W, R \Rightarrow \chi^2(p-r)$ 임을 보일 수 있고, 이로부터 신뢰수준 α 의 (근사적인) 기각역을 정할 수 있다. 각 검정에는 나름의 장단점이 있는데, LRT의 경우 full model에서의 MLE와 reduced model에서의 MLE를 모두 계산하여야 하므로 비교적 많은 계산을 필요로 하지만 다른 두 검정에 비해 정확하며, Wald 검정과 Rao 검정의 경우 full model에서의 MLE와 reduced model에서의 MLE 중 하나만 계산하면 되지만 그 검정통계량 자체가 LRT의 근사이며 비교적 덜 정확하다. 다만, 근사적으로 세 검정은 모두 동치임이 알려져 있다 (Tuerlinckx *et al.* 2006). 이와 비슷하게 랜덤효과의 분산 성분에 대하여서도 LRT를 수행할 수 있지만, 검정통계량의 극한분포가 카이제곱분포가 아니므로 특별한 경우가 아닌 이상 MCMC와 같은 시뮬레이션 기법을 통해 그 분포를 수치적으로 근사해야 하는 어려움이 있다 (Zhang & Lin 2008).

랜덤효과 γ 의 경우, 전술하였다시피 이는 추정의 대상이 아니다. 대신 그 realization에 대한 예측은 가능한데, 관측값 \mathbf{Y} 가 주어지고 고정효과 β 와 분산구조 Ψ 를 알고 있다면 랜덤효과에 대한 예측값으로 MSPE(Mean Squared Prediction Error)를 최소화하는 조건부기댓값을 생각해볼 수 있다.

$$\tilde{\gamma} = \mathbf{E}_{\beta, \Psi}(\gamma | \mathbf{Y}) = \underset{\gamma_* \in \mathbb{R}^q}{\operatorname{argmin}} \mathbf{E}_{\beta, \Psi} \|\gamma - \gamma_*\|^2 \quad (10)$$

실제로는 β 와 Ψ 를 알지 못하므로 이를 앞서 소개한 방법들을 이용하여 구한 추정량 $\hat{\beta}$ 와 $\hat{\Psi}$ 으로 대신하여 γ 를 다음과 같이 예측할 수 있다 (McCulloch & Neuhaus 2011).

$$\begin{aligned} \tilde{\gamma} &= \mathbf{E}_{\hat{\beta}, \hat{\Psi}}(\gamma | \mathbf{Y}) = \frac{\int_{\mathbb{R}^q} \gamma f_{\mathbf{Y}|\gamma}(\mathbf{Y}|\gamma) f_{\gamma}(\gamma) d\gamma}{\int_{\mathbb{R}^q} f_{\mathbf{Y}|\gamma}(\mathbf{Y}|\gamma) f_{\gamma}(\gamma) d\gamma} \\ &= \frac{\int_{\mathbb{R}^q} \gamma \exp(\sum_{i=1}^n [\mathbf{z}'_i \gamma y_i - A(\mathbf{x}'_i \hat{\beta} y_i + \mathbf{z}'_i \gamma y_i)]) - \gamma \hat{\Psi}^{-1} \gamma / 2 d\gamma}{\int_{\mathbb{R}^q} \exp(\sum_{i=1}^n [\mathbf{z}'_i \gamma y_i - A(\mathbf{x}'_i \hat{\beta} y_i + \mathbf{z}'_i \gamma y_i)]) - \gamma \hat{\Psi}^{-1} \gamma / 2 d\gamma} \end{aligned} \quad (11)$$

이를 본 연구에서 사용할 모형(식 2)에 적용하면 i 번째 피실험체의 새로운 뇌 활동 정보 $\mathbf{x}_* \in \mathbb{R}^p$ 에

■ 표 1 이상이 있는 것으로 판단되어 분석에서 제외된 채널

피실험체	1	2	3	4	5	6	7	8
이상채널	AC _L	PFC _L	BF	BF	BF	-	AC _L , BF	-

대해 반응 유무의 확률 $p_* \in [0, 1]$ 를 다음과 같이 예측할 수 있다.

$$\tilde{p}_* = g^{-1}(\mathbf{x}'_* \hat{\beta} + \tilde{\epsilon}_l) \quad (12)$$

곧, GLMM은 반응 유무를 예측하는 예측모형으로도 활용할 수 있어 해석의 용이함과 활용성을 두루 갖춘 모형이다.

자료 전처리

이상 채널 제거 전기적인 신호로 측정되는 뇌파에는 뇌의 활동 이외에도 각종 잡음이 뒤섞이어 기록된다. 일반적으로 이러한 잡음을 뇌파에서 분리해 내는 것은 쉬운 일이 아니며, 이에 대한 다양한 시도와 제안이 이루어져 왔다 ([Shoker et al. 2005](#); [Nolan et al. 2010](#); [Lawhern et al. 2013](#); [Mognon et al. 2011](#)). 하지만 이들은 대부분 사람의 두피에서 측정된 EEG에서의 잡음 제거를 목적으로 고안된 방법들이어서 본 실험에서 사용된 쥐의 LFP에 곧바로 적용하기에는 어려움이 있다. 또한, 사람의 EEG에서의 잡음의 주된 요인이 피험자의 눈깜빡임과 같은 필수불가결한 생리적 활동이나 두피에서의 전기 신호의 간섭에 기인하는 반면, 쥐의 LFP에서는 이러한 요인들로 인한 잡음이 거의 없어 비교적 잡음이 영향이 적다. 그럼에도 불구하고 측정 전극 자체의 이상과 같은 이유로 극히 비정상적인 LFP가 기록되는 경우가 있는데, 다행히 이런 경우는 그림 1과 같이 LFP의 시계열도를 직접 그려봄으로써 육안으로 그 이상유무를 쉽게 판단할 수 있다. 본 연구에서는 보수적인 방식을 택하여 이상이 있는 채널의 자료를 제거하고 분석을 수행하였다. 이상이 있는 것으로 판단된 채널은 표 1과 같다.

LFP 정규화 여러 피실험체 간의 LFP를 비교하기 위해 이들을 적당히 정규화하는 작업이 필요하다. 우선 총 5초의 LFP 중에서 분석의 대상인 prestimulus 기간인 -2초부터 0초까지의 자료만을 잘라내었다.³ 다음으로, 한 시행에서 관측한 LFP의 표본평균이 0이 되도록 조정해 주었다. 이 이상의 정규화는 이후 분석에 영향을 줄 수 있다고 판단하여 진행하지 않았다.

Bandpower 계산 LFP를 시계열로 생각하여 이를 그대로 분석하는 것도 하나의 방법이겠으나, 신경 과학계에서는 LFP의 bandpower를 분석하는 경우가 많다. 이는 특정 주파수대의 bandpower가 나름의 의미를 가지는 것으로 밝혀져 있어 이후 분석 결과의 해석이 용이하기 때문이다 ([Subha et al. 2010](#)). 이러한 관례를 따라 본 연구에서도 LFP의 bandpower를 계산하여 사용하였다.

³ 이때, 자료의 양 끝에서 왜곡이 발생하는 것을 막기 위해 실제로는 0.2초의 epoch을 두고 -1.8초부터 -0.2초까지를 잘랐다. 실제로, 그림 1을 보면 자극이 시작된 0초 부근에서 펠스 형태가 일관되게 관측됨을 알 수 있다. 이러한 뇌파는 청각 자극의 제시에 따른 것으로, 자극이 주어지기 이전의 자료만을 바탕으로 자극에 대한 반응의 상관관계를 분석하는 본 분석의 목표에 비추어 보았을 때 고려되지 않아야 함이 타당하다.

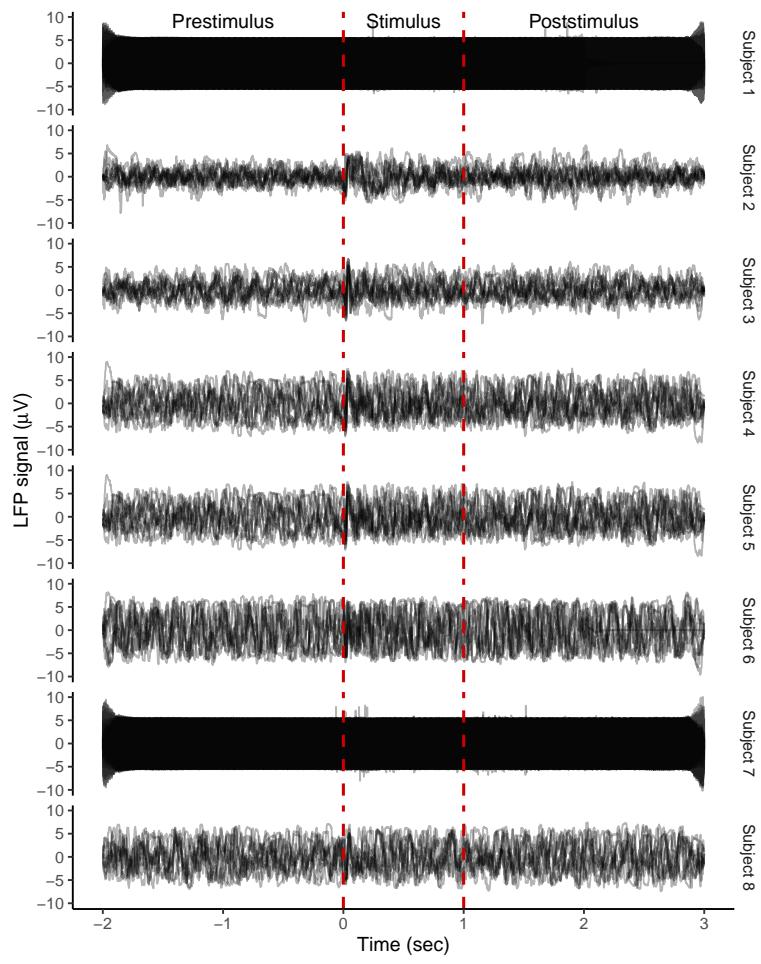


그림 1 AC_L 채널에서 측정된 LFP의 시계열도. 각 피실험체 별로 자극이 시작된 순간을 0초로 하여 세션 1의 첫 10 번의 시행에서 관측한 자료를 겹쳐서 나타내었다. 빨간 점선은 prestimulus, stimulus, poststimulus 구간을 나타낸다. 육안으로도 쉽게 1번과 7번 피실험체의 AC_L 채널 전극에 이상이 있음을 명백히 알 수 있다. 한편, 자극이 시작되는 0초 부근에서 펄스 형태의 뇌파가 지속적으로 관측된다는 점도 확인할 수 있다.

표 2 신경과학계에서 흔히 분석하는 주파수대

명칭	Delta	Theta	Beta	Gamma	High gamma
주파수대(Hz)	[1, 4]	[4, 12]	[12, 30]	[30, 50]	[50, 100]

시계열 $\{X_t\}$ 에 대해 이의 푸리에 변환을 $\{\tilde{X}_\omega\}$ 라 하면 $\{X_t\}$ 의 PSD(Power Spectral Density; 스펙트럼 밀도) $S_X(\omega)$ 는 다음과 같이 정의된다.

$$S_X(\omega) = \mathbf{E}|\tilde{X}_\omega|^2 \quad (13)$$

하지만, 일반적으로 임의의 시계열에 대해 그 푸리에 변환이 존재함을 보장할 수 없으므로 적분의 순서를 바꾸어 $\{X_t\}$ 의 PSD를

$$S_X(\omega) = \lim_{T \rightarrow \infty} \mathbf{E} \left| \frac{1}{\sqrt{T}} \int_0^T X_t e^{-i\omega t} dt \right|^2 \quad (14)$$

로 정의하는 경우가 많다 (Miller & Childers 2012). 여기서 i 는 허수단위이다. 이제 주어진 주파수대(band) $I \subseteq \mathbb{R}$ 에 대해 bandpower BP_I 는 I 에서의 S_X 의 적분, 즉

$$BP_I = \int_I S_X(\omega) d\omega \quad (15)$$

로 정의된다.

실제로는 모든 $t \in \mathbb{R}$ 에 대해 X_t 를 관측하지 못하므로 PSD를 추정해야 한다. 그 방법에는 Bartlett의 방법, LSSA(Least Squares Spectral Analysis) 등과 같은 비모수적인 방법부터 ARMA(AutoRegressive Moving Average) 추정, MUSIC(MULTiple SIgnal Classification) 방법 등 모수적인 방법에 이르기까지 다양한 방법이 제시된 바 있으며 현재도 SDE(Spectral Density Estimation)이라는 분야로 활발히 연구되고 있다 (Stoica et al. 2005). 여기에서는 비모수적인 방법 중 비교적 단순하면서도 널리 사용되는 방법으로, 관측된 자료를 DFT(Discrete Fourier Transform) 하여 periodogram을 구한 다음 이를 PSD의 추정량으로 사용하였다. 이후 추정된 PSD의 I 에서의 적분을 사각형법으로 근사하여 최종적인 bandpower의 추정량을 얻었다.

실제 계산에는 MATLAB 함수 bandpower가 사용되었으며 이때 사용한 주파수대 I 는 표 2와 같다. 각각의 주파수대는 전술한 바와 같이 신경과학계에서 이미 광범위하게 연구되어 나름의 의미가 많이 밝혀져 있다 (Ehlers & Kupfer 1997; Vertes 2005; Baker 2007; Crick & Koch 1990).

탐색적 자료분석

GLMM 적합에 앞서, 간단한 탐색적 자료분석을 수행하였다. 먼저, 각 피실험체별로 Hit 그룹과 Miss 그룹의 bandpower 분포를 살펴보았다. 그림 2의 A, C는 AC_L 채널 beta band의 bandpower 분포를 나타낸 것으로서, 이로부터 bandpower의 분포가 심하게 right-skew되어 있음을 알 수 있다. 이는 다른 채널이나 다른 주파수대에 대해서도 마찬가지였다. 이후 분석의 편의를 위해 bandpower에 로그 변환을 취하였다.

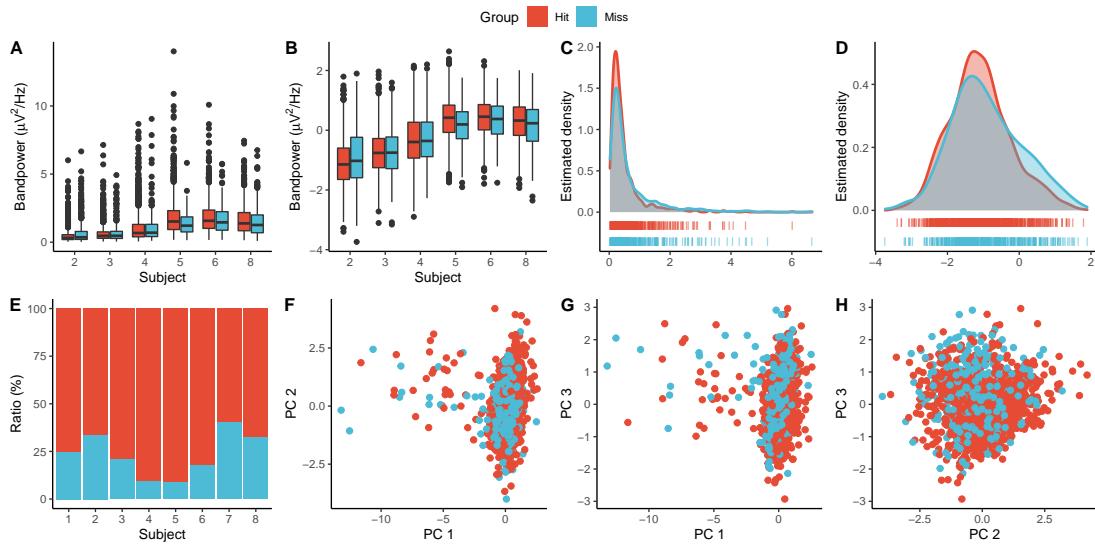


그림 2 로그 변환 **A** 전과 **B** 후 AC_L 채널 delta band의 bandpower로 그린 box plot. 그 중에서 2번 피실험체의 로그 변환 **C** 전과 **D** 후의 추정된 분포. 추정된 분포는 Gaussian 커널을 이용한 커널분포추정의 결과이며 이때의 bandwidth는 $0.9n^{-1/5} \min\{\hat{\sigma}, IQR/1.34\}$ 로 추정하였다 (로그 변환 전 Hit: 0.0609, Miss: 0.1143, 로그 변환 후 Hit: 0.1789, Miss: 0.2590). 여기서 n 은 표본의 수이며 $\hat{\sigma}$ 는 표본표준편차이다. **E** 피실험체별 Hit 그룹과 Miss 그룹 간의 비율. 피실험체에 따라 Hit의 비율이 Miss 비율보다 작게는 1.5배(피실험체 7)에서 많개는 10.4배(피실험체 5) 가량 더 많았다. **F~H** 1번 피실험체의 3개의 주성분을 추출하여 그린 산점도. 어느 주성분 방향에서 보나 두 그룹이 서로 거의 구분이 되지 않고 잘 섞여 있음을 알 수 있다. 주성분 분석에는 R 함수 prcomp이 사용되었으며 피실험체별로 이상 채널을 제외한 모든 채널에서의 bandpower를 하나의 변량으로 하여 주성분분석을 수행하였다. 이때, 각 변량의 단위가 모두 같은 것을 고려하여 scaling은 하지 않았다.

다음으로, 각 피실험체별로 Hit 그룹과 Miss 그룹 간의 비율을 살펴보았다. 그림 2의 E는 그 비율을 그래프로 나타낸 것인데, 이로부터 Hit 그룹이 Miss 그룹보다 훨씬 더 많은, 두 그룹간의 불균형이 뚜렷한 자료임을 확인할 수 있다.

이어서 두 그룹이 서로 어떻게 섞여 있는지를 보기 위해 각 피실험체별로 PCA(Principal Component Analysis; 주성분분석)를 수행하였다. 그림 2의 F~H는 그 중 1번 피실험체의 3개의 주성분으로 그린 산점도로, Hit과 Miss 두 그룹이 서로 거의 구분이 되지 않고 잘 섞여 있음을 알 수 있다. 다른 피실험체의 경우에도 이와 비슷하였다.

GLMM 모형 적합

모형 기술의 편의를 위해 i 번째 피실험체의 j 번째 세션의 i 번째 시행에서 피실험체가 GO 자극에 튜브를 향았으면 (Hit이면) $Y_{ijk} = 1$ 그렇지 않았으면 (Miss이면) $Y_{ijk} = 0$ 이라 하자. 또한, 해당 시행에서 A 부위에서 관측된 I 주파수대의 bandpower를 $\text{BP}_{ijk}^I(\text{A})$ 로 쓰자. 예컨대 해당 시행에서 AC_L 의 delta 주파수대에 해당하는 bandpower는 $\text{BP}_{ijk}^\delta(\text{AC}_L)$ 이다. 그렇다면 GLMM에서의 설명변수는 벡터

$$\mathbf{x}_{ijk} = \begin{bmatrix} \text{BP}_{ijk}^\delta(\text{AC}_L) \\ \text{BP}_{ijk}^\theta(\text{AC}_L) \\ \vdots \\ \text{BP}_{ijk}^{\gamma^+}(\text{PPC}_L) \end{bmatrix} \quad (16)$$

로 쓸 수 있다. 다만, 앞서 데이터 전처리 과정에서 본 바와 같이 피실험체마다 이상 채널이 있으므로 25개의 설명변수를 다 쓰지는 못하고, BF 채널의 자료와 1, 2, 7번 피실험체의 자료를 제외한 차원이 20인 도합 3026개의 자료를 분석하였다.

앞서 GLMM을 소개하며 생각한 모형과는 달리, 이 실험에서는 반복측정이 시행 수준과 세션 수준의 두 가지 단계로 중첩되어 이루어졌다. 따라서 시행 수준, 세션 수준, 피실험체 수준에서 각각 서로 독립인 오차가 발생함을 기술하는 다음과 같은 GLMM을 생각하자.

$$Y_{ijk} | \epsilon_{ij}, \varphi_i \stackrel{\text{iid}}{\sim} \text{Bern}(p_{ijk}) \quad (17)$$

$$g(p_{ijk}) = \mathbf{x}'_{ijk}\beta + \epsilon_{ij} + \varphi_i$$

여기서 $g(x) = \log(x/(1-x))$ 이며 $\epsilon_{ij} \stackrel{\text{iid}}{\sim} \mathbf{N}(0, \sigma^2)$, $\varphi_i \stackrel{\text{iid}}{\sim} \mathbf{N}(0, \tau^2)$ 고 서로 독립이다. GLM의 적합과정을 비슷하게 따라간다고 한다면, 20개의 설명변수를 모두 사용하여 적합을 한 후에 AIC(Akaike information criterion; Akaike 정보기준)나 BIC(Bayesian information criterion; 베이지안 정보기준) 등을 사용하여 모형 선택의 과정을 거쳐야 할 것이다. 충분한 시간과 컴퓨팅 자원이 있다면 이렇게 할 수 있겠지만, 한 번의 적합에 비교적 오랜 시간이 걸리는 GLMM의 특성상 20개나 되는 변수를 모두 사용하여 모형 선택을 진행하는 것은 현실적으로 쉽지 않다. 이에 본 연구에서는 elastic net을 이용하여 설명변수 중 일부를 선별(prescreening)하는 방식을택하였다.

■ 표 3 Elastic net으로 선별된 설명변수

채널	주파수대
AC _L	delta, gamma, high gamma
AC _R	delta, theta, beta, gamma, high gamma
PFC _L	delta, theta, beta, gamma, high gamma
PPC _L	delta, theta, beta, high gamma

Elastic net은 기존의 ridge regression과 LASSO(Least Absolute Shrinkage and Selection Operator)의 아이디어를 합한 것으로, 편차와 분산의 균형을 잡아 평균제곱오차를 줄이는 동시에 모형의 sparsity를 추구하는 것을 목표로 한다. 이를 위해 elastic net은 계수 추정을 위한 목적함수에 L_1 패널티와 L_2 패널티를 추가하여 다음과 같이 계수를 추정한다 (Friedman *et al.* 2010).

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ -l(\beta) + \lambda \left[\alpha \frac{\|\beta\|_2^2}{2} + (1-\alpha) \|\beta\|_1 \right] \right\} \quad (18)$$

여기서 l 은 모형의 로그가능도이며, $\lambda \geq 0$, $\alpha \in [0, 1]$ 는 각각 패널티의 정도와 두 패널티간의 균형을 조절하는 hyperparameter로 사용자가 적절히 설정해주어야 하는 값이다.

본 연구에서 최종적으로 적합하고자 하는 모형이 로지스틱 GLMM이므로 로지스틱 GLM에 elastic net을 사용하여 변수 선별을 진행하였다. Hyperparameter tuning은 10-fold CV를 통해 구한 AUC(Area Under the Curve)를 기준으로 1sd rule을 적용하였다. Tuning 결과 $\alpha = 0.15$, $\lambda = 0.0118$ 가 선택되었으며 해당 α 와 β 로 elastic net을 적합한 결과 20개의 변수 중 17개의 변수가 선별되었다. 그림 3과 표 3는 각각 hyperparameter tuning 결과와 변수 선별 결과를 나타낸 것이다.

CV 및 시뮬레이션

전술하였던 바와 같이 GLMM은 예측 모형으로도 활용할 수 있다. 만약 적합한 GLMM이 다른 방법들에 비해 높은 성능을 발휘한다면 이는 GLMM이 true model이라는 강한 증거로 사용될 수 있을 것이고, 더 높은 성능의 예측모형을 개발의 좋은 출발점으로 삼을 수도 있을 것이다. 이러한 GLMM의 장점을 앞서 적합한 GLMM으로써 보이기 위해 주어진 bandpower로부터 피실험체의 반응을 예측하는 다양한 모형을 적합하고, 이를 서로 비교해보았다.

비교의 기준으로 삼을 통계량은 AUC, F score, MCC(Matthews correlation coefficient)이며 각각 다음과 같이 정의된다. (여기서는 Hit을 True로 생각한다.)

$$\text{AUC} = \int \text{ROC} \quad (19)$$

$$\text{F score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

AUC와 F score는 분류기의 성능 평가에 널리 이용되는 범용적인 척도이다. MCC는 본질적으로 예측 결과와 실제 결과에 대한 상관계수로 -1 에서 1 까지의 값을 가지며 1 은 완벽한 예측을, 0 은

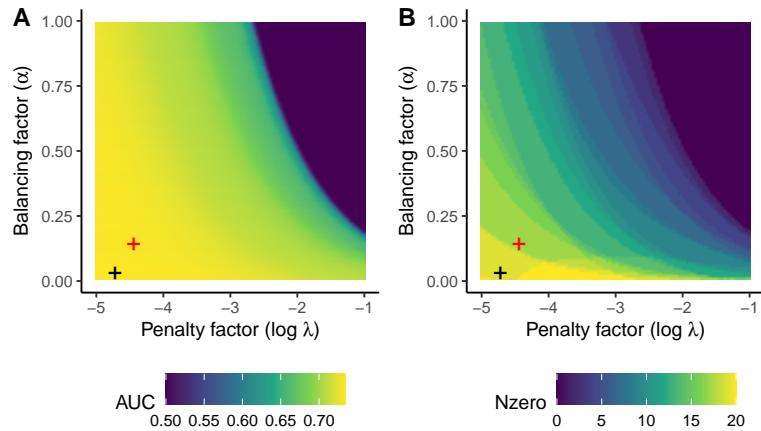


그림 3 Elastic net의 hyperparameter tuning 결과. $\alpha = 0.01i$ ($i = 1, \dots, 99$)와 $\lambda = e^{-1-0.04i}$ ($i = 0, \dots, 100$)에 대해 grid search를 수행하였다. 각 조합에 대해 5번의 10-fold CV를 통해 계산된 **A** AUC와 **B** 선별된 변수의 수. AUC의 최댓값은 $\alpha = 0.04$, $\lambda = 0.0089$ (AUC = 0.7340, 19개의 변수 모두 사용)에서 얻어졌으며 (검은색 +), 이에 1sd rule을 적용하여 얻은 sparse한 모형은 $\alpha = 0.15$, $\lambda = 0.0118$ (AUC = 0.7332, 17개의 변수 사용)에서 얻어졌다 (빨간색 +). 피실험체를 indicator variable로 하여 elastic net을 적합하였으며 그 적합에는 R 함수 glmnet(glmnet 패키지)을 사용하였다.

무작위 예측을, -1 은 완벽히 반대로 예측함을 의미한다. 이러한 MCC는 이진 분류에서 두 그룹간의 불균형이 심할 때 유용한 척도이다 (Chicco & Jurman 2020). 앞서 EDA에서 확인하였듯이 Hit 그룹과 Miss 그룹의 비율 차이가 많게는 8배가 될 정도로 불균형이 심하므로 MCC는 분류기를 비교하는 좋은 기준이 될 것이다.

비교의 대상으로 적합한 분류기들은 다음과 같다.

Naive Bayes Naive Bayes 분류기는 각 변량이 서로 독립적이라는 다소 극단적인 가정 하에 베이즈 분류기를 계산하여 얻는 분류기이다. 그 이름에서 알 수 있듯이 관측한 표본으로부터 조건부확률을 계산하여 이를 분류에 이용하며, 고전적인 분류기 중에서도 단순한 분류기에 속한다. 본 연구에서는 R 함수 naiveBayes(e1071 패키지)를 이용하여 적합하였다.

LDA LDA(Linear Discriminant Analysis; 선형판별분석)은 두 그룹이 각각 다변량 정규분포를 따른다고 그룹간의 분산이 같다는 가정 하에 베이즈 분류기를 계산하여 얻는 분류기이다. 그 계산과정에서 두 그룹간의 분산은 최대화되고 그룹 내의 분산은 최소화하는 방향을 찾아 자료를 해당 방향으로 사영한다는 점에서 PCA와 그 아이디어가 유사하다. 본 연구에서는 R 함수 lda(MASS 패키지)를 이용하여 적합하였다.

KNN KNN(k -Nearest Neighbor; k 최근접 이웃 알고리즘)은 분류 알고리즘으로, 새로운 관측값 $\mathbf{x} \in \mathbb{R}^p$ 에 대해 이와 가장 가까운 k 개의 관측값(이웃)이 어느 그룹에 속하는지 살펴 다수결에 따라 \mathbf{x} 를 분류한다. 이론적으로 이는 관측값이 충분히 조밀하다는 가정 하에 베이즈 분류기의 근사가 된다. 본 연구에서는 R 함수 knn(class 패키지)을 이용하였다.

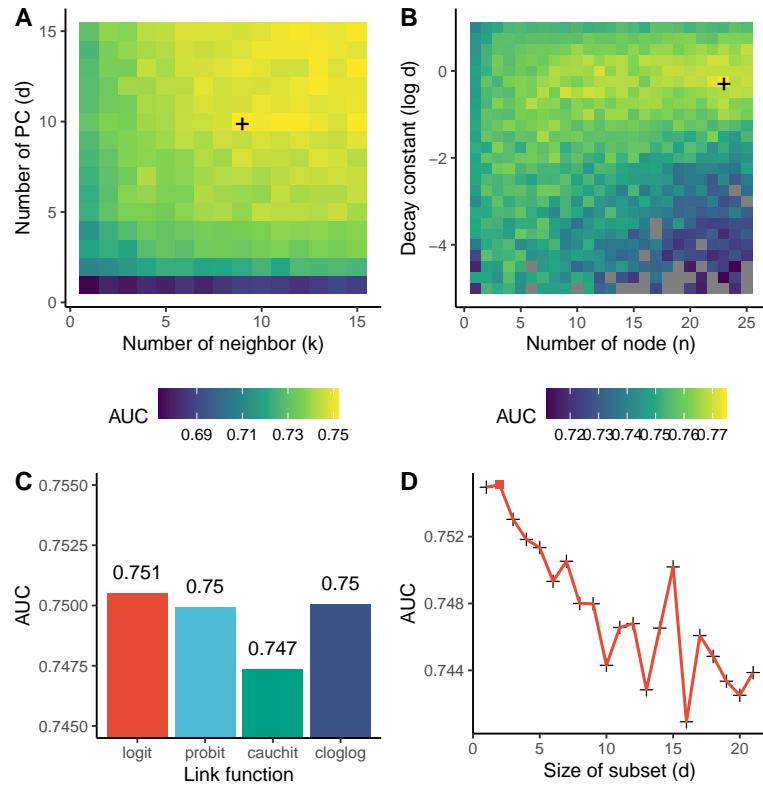


그림 4 A KNN의 hyperparameter tuning 결과. 이웃의 수 $k = 1, \dots, 15$ 과 주성분의 개수 $d = 1, \dots, 15$ 에 대해 grid search를 수행하였으며 KNN은 각각의 피실험체에 대해 적합하였다. 5번의 10-fold CV를 통해 계산된 AUC의 최댓값은 $k = 9$ 과 $d = 10$ 인 경우($AUC = 0.752$)에 얻어졌다 (검은색 +). B ANN의 hyperparameter tuning 결과. 노드의 개수 $n = 1, \dots, 25$ 과 decay constant $d = e^{1-0.25i}$ ($i = 0, \dots, 24$)에 대해 grid search를 수행하여 피실험체를 indicator variable로 하여 ANN을 적합하였다. 5번의 10-fold CV를 통해 계산된 AUC의 최댓값은 $n = 23$ 과 $d = e^{-0.25}$ 인 경우 ($AUC = 0.775$)에 얻어졌다 (검은색 +). C 로지스틱 회귀의 link tuning 결과. Logit, probit, cauchit, complementary log-log(cloglog)에 대해 피실험체를 indicator variable로 하여 로지스틱 회귀모형을 적합하였다. 최초의 full model은 AIC를 기준으로 stepwise 모형 선택을 거쳤으며, 그 결과 선택된 모형으로 5번의 10-fold CV를 수행하고 AUC를 구하였다. AUC의 최댓값은 logistic link를 사용한 경우($AUC = 0.751$)에 얻어졌다. D Random forest의 hyperparameter tuning 결과. 개별 결정트리 적합시 사용할 변수의 수 $d = 1, \dots, 21$ 에 대해 grid search를 수행하였다. 5번의 10-fold CV를 통해 계산된 AUC의 최댓값은 $d = 2$ 인 경우($AUC = 0.755$)에 얻어졌다 (빨간색 네모).

자료가 고차원인 경우에는 차원의 저주에 의해 거리의 의미가 약해지므로 PCA와 같은 차원축소를 거친 자료를 이용하는 것이 바람직하다 (Beyer *et al.* 1999). 본 연구에서는 앞서 수행한 PCA 결과를 사용하였고, 10-fold CV를 통해 계산한 AUC를 기준으로 이웃의 수 k 와 사용할 주성분의 개수 d 를 결정하였다. 그림 4의 A는 hyperparameter tuning 결과를 나타낸 것으로, 9개의 이웃과 10개의 주성분을 사용할 때 AUC를 기준으로 가장 좋은 예측력을 보였다.

ANN ANN(Artificial Neural Network; 인공신경망)은 근래 큰 인기를 끌고 있는 비모수적인 분류기로 인간의 뇌를 모방하여 개개의 뉴런들과 각 뉴런의 활성함수가 합쳐져 동작한다. ANN은 사용되는 분야에 따라 다양한 형태와 변형들이 존재하는데, 본 연구에서는 가장 단순한 형태인 single layer FFNN(FeedForward Neural Network)를 사용하였다. 이론적으로 ANN은 비선형회귀의 일종으로, 아핀 변환의 합성으로 표현된다. 특히 single layer FFNN의 경우 universal approximation theorem에 의해 적절한 조건을 만족하는 어느 함수든 원하는 만큼 근사할 수 있음이 알려져 있다 (Leshno *et al.* 1993).

본 연구에서는 R 함수 nnet(nnet 패키지)을 이용하여 적합하였는데, 각 노드의 가중치 최적화를 위해 CE(Cross-Entropy) 손실함수를 사용하였다. 한편, ANN은 자료에 과적합되는 경향이 있다. 복잡한 구조 속에서 네트워크가 원래의 자료를 그냥 외워버리는 것이다. 이를 방지하기 위해 가중치 최적화를 위한 반복 중에 의도적으로 적합을 방해하는 decay constant d 를 도입한다. 본 연구에서는 10-fold CV를 통해 계산한 AUC를 기준으로 decay constant d 와 은닉층의 노드 개수 n 을 결정하였다. 그림 4의 B는 hyperparameter tuning 결과를 나타낸 것으로, 23개의 노드와 decay constant $e^{-0.25}$ 를 사용할 때 AUC를 기준으로 가장 좋은 예측력을 보였다.

로지스틱 회귀 로지스틱 회귀는 기존 신경과학계의 접근방식대로 반복측정으로 인한 자료의 종속성을 고려하지 않고, 이가 모두 독립하는 가정 하에 사용할 수 있는 분류기이다. 이러한 독립성의 고리를 제외하면 베르누이 분포의 모수를 link 함수로 설명변수의 선형결합과 연관짓는다는 점은 GLMM과 그 아이디어가 동일하다. 다만, 로지스틱 회귀를 이용한 분류는 두 그룹이 이미 잘 분류되어 있는 경우에 이를 적합하는 IRLS (Iteratively Reweighted Least Squares) 알고리즘이 수치적으로 불안정해진다는 단점이 있는데, 앞서 PCA를 통해 확인한 바와 같이 본 연구에서 다루는 자료에서는 Hit과 Miss 두 그룹이 잘 섞여 있으므로 이는 크게 문제되지 않을 것이다.

이때 link 함수로는 앞서 사용한 logit link 외에 probit, cauchit, complementary log-log 등의 다양한 함수를 사용할 수 있다. (표 4 참조) 본 연구에서는 각각의 link 함수에 대해 AIC를 기준으로 stepwise 모형 선택을 진행하여 최적의 모형을 탐색한 후, 10-fold CV를 통해 계산한 AUC를 기준으로 최종적으로 사용할 link 함수를 결정하였다. 본 연구에서는 R 함수 glm을 이용하여 모형을 적합하고, R 함수 stepAIC(MASS 패키지)를 이용하여 stepwise 모형 선택을 수행하였다. 그림 4의 C는 이상의 방식으로 구한 link 함수 별 AUC를 나타낸 것으로 logit link를 사용할 때 AUC를 기준으로 가장 좋은 예측력을 보였다.

■ 표 4 로지스틱 회귀에서의 link 함수

이름	정의
Logit link	$g(x) = \log(x / (1 - x))$
Probit link	$g(x) = \Phi^{-1}(x)^a$
Cauchit link	$g(x) = \tan(\pi(x - 1/2))^b$
Complementary log-log	$g(x) = \log(-\log(1 - x))$

^a Φ : 표준정규분포함수의 누적분포함수

^b 이는 $t(1)$ 분포(Cauchy 분포)의 누적분포함수의 역함수이다.

Random forest Random forest는 여러 개의 결정트리를 적합한 뒤, 각 트리의 예측 결과를 종합하여 최종 예측치를 결정하는 양상을 기법의 일종이다. 이때, 각각의 결정트리는 전체 변량 중의 몇개의 부분집합만을 이용하여 적합되며, 같은 부분집합을 사용하더라도 각 트리의 적합 과정에서 최대 허용 깊이 등의 hyperparameter를 다르게 설정하여 전체 forest가 랜덤성을 띠도록 한다. 이는 과적합의 문제를 막기 위한 것이다.

본 연구에서는 R 함수 `randomForest`(`randomForest` 패키지)를 이용하여 적합하였다. 이때, forest를 구성하는 결정트리의 수는 충분히 많아야 하므로 `randomForest` 함수의 기본값인 500개로 설정하였다. 한편, 각 트리의 적합에 있어 사용할 변량의 부분집합의 크기 d 는 random forest의 성능에 영향을 주는 중요한 hyperparameter이므로 10-fold CV를 통해 계산한 AUC를 기준으로 결정하였다. 그림 4의 D는 hyperparameter tuning 결과를 나타낸 것으로, 변량의 부분집합의 크기가 2일 때 AUC를 기준으로 가장 좋은 예측력을 보였다.

이상의 분류기들을 적합하여 각각 100번의 10-fold CV를 수행하여 AUC, F score, MCC를 계산하였다. 이와 더불어 다음과 같이 정의되는 TPR(True Positive Rate), TNR(True Negative Rate), ACC(Accuracy) 또한 계산하였다.

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{TNR} &= \frac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{ACC} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \end{aligned} \quad (20)$$

결과

그림 5는 CV 결과를 나타낸 것으로 AUC, MCC, F score, ACC 모두 GLMM이 다른 분류기에 비해 우수함을 확인할 수 있다. 특히, TPR과 TNR을 보면, 다른 분류기가 그 비율이 많은 Hit 그룹으로 다소 편향되게 예측하는 것에 비하여 GLMM은 두 그룹간의 불균형에도 불구하고 그 경계를 비교적 잘 구분하고 있음을 알 수 있다.

GLMM 모형을 적합한 결과는 표 5, 6과 같다. 적합 결과를 보면, 우선 랜덤효과에서 피실험체 수준의 오차의 분산이 거의 0임을 알 수 있다. 이는 피실험체 수준의 오차보다 반복측정에 있어서의

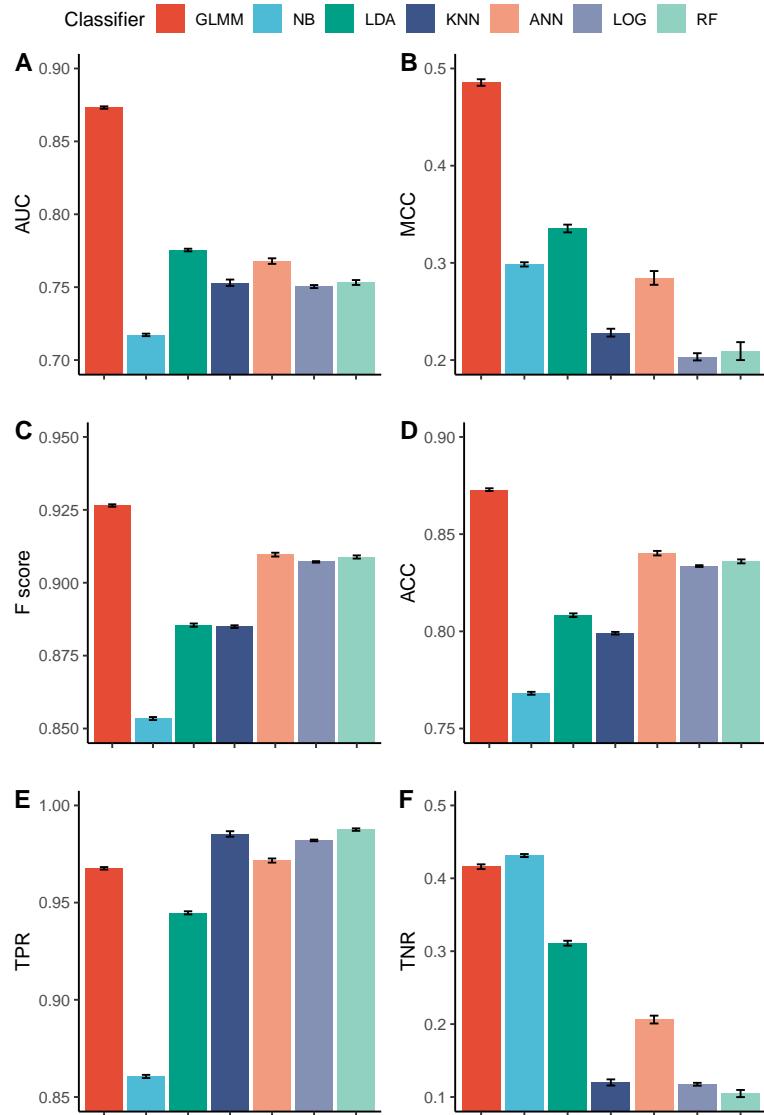


그림 5 GLMM, naive Bayes, LDA, KNN, ANN, 로지스틱 회귀, random forest의 CV 결과. 100번의 10-fold CV를 통해 계산한 **A** AUC, **B** MCC, **C** F score, **D** ACC, **E** TPR, **F** TNR을 나타냈다. 에러바는 1sd를 의미한다. GLMM, ANN, 로지스틱 회귀, random forest는 피실험체를 indicator variable로 하여 하나의 모형을 적합했고, naive Bayes, LDA, KNN는 피실험체별로 하나씩의 모형을 적합했다.

■ 표 5 GLMM 적합 결과 (고정효과)

설명변수	계수	표준편차	LRT Λ	p-value
절편	-4.766	0.608	-	-
AC_Ldelta	-0.030	0.082	4.304	1.68e-5 ***
AC_Lgamma	-0.191	0.136	-0.393	0.694
AC_Lhigh gamma	-0.686	0.179	-3.233	1.22e-3 **
AC_Rdelta	0.061	0.094	4.304	1.68e-5 ***
AC_Rtheta	-0.038	0.138	4.304	1.68e-5 ***
AC_Rbeta	0.013	0.138	1.434	0.152
AC_Rgamma	-0.162	0.137	-2.171	0.030 *
AC_Rhigh gamma	-0.645	0.199	-0.478	0.633
PFC_Ldelta	-0.255	0.075	0.075	1.77e-5 ***
PFC_Ltheta	0.065	0.115	0.115	1.77e-5 ***
PFC_Lbeta	0.556	0.119	0.119	8.12e-4 ***
PFC_Lgamma	-0.317	0.097	-2.335	0.020 *
PFC_Lhigh gamma	0.620	0.816	-2.335	0.020 *
PPC_Ldelta	0.269	0.079	4.144	3.42e-5 ***
PPC_Ltheta	0.184	0.143	-1.967	0.049 *
PPC_Lbeta	0.556	0.142	4.055	5.02e-5 ***
PPC_Lhigh gamma	-0.980	0.184	-5.420	5.96e-8 ***

■ 표 6 GLMM 적합 결과 (랜덤효과)

그룹	설명변수	분산	표준편차
피실험체별 세션	절편	2.72050	1.6494
피실험체	절편	0.01301	0.1141

개개의 세션에서의 오차가 지배적임을 합의한다. 다만, 사실은 피실험체 수준의 변동성이 세션 수준의 변동성으로 고려되어 적합되었을 가능성이 있으므로 선불리 피실험체가 서로 동일하다고 결론지어서는 안될 것이다. 고정효과를 보면, PFC_L과 PPC_L 부위의 다양한 주파수대에 걸쳐 해당 bandpower가 피실험체의 행동 결과에 유의한 영향을 주고 있음을 알 수 있다. 그러나 그 방향성은 제각각인데, 예컨대 PPC_L의 delta bandpower와 beta bandpower는 Hit의 확률과 양의 상관을 가지는 방향으로 유의하고, 동일한 부위의 high gamma bandpower는 Hit의 확률과 음의 상관을 가지는 방향으로 유의하다.

코드 및 분석도구

이상 채널 제거와 LFP 정규화, bandpower 계산은 MATLAB으로 하였다. 나머지 분석은 모두 R로 하였다. 분석에 사용한 코드는 [github https://github.com/eik4862/GLMM](https://github.com/eik4862/GLMM)에서 찾아볼 수 있다.

ACKNOWLEDGEMENT

본 연구에서 사용한 자료는 한국과학기술연구원 신경과학연구단의 한효빈, 이가은, 최지현이 실험을 통해 얻은 것으로 3/2/2020 허락을 받고 사용하였다. 동물 실험에 대한 규정 준수 및 관련된 자세한 사항은 [Han et al. \(2019\)](#) 참조.

TECHNICALITY

라플라스 근사를 이용한 GLMM의 로그가능도 근사

본 절에서는 GLMM의 로그가능도인

$$\begin{aligned} l(\beta, \Psi) = & \log \int_{\mathbb{R}^q} \exp \left(\sum_{i=1}^n [\mathbf{z}'_i \gamma Y_i - A(\mathbf{x}'_i \beta + \mathbf{z}'_i \gamma)] - \frac{\gamma' \Psi^{-1} \gamma}{2} \right) d\gamma \\ & - \frac{1}{2} \log \det \Psi + \sum_{i=1}^n \mathbf{x}'_i \beta Y_i + \text{const.} \end{aligned} \quad (21)$$

를 라플라스 근사를 통해 수치적으로 계산하는 구체적인 방법에 대해 알아본다. 이는 본 연구에서 GLMM의 적합에 사용된 R 함수 `glmer`(lme4 패키지)에서 모형을 적합할 때 적용되는 수치적분법이다. 먼저 적분에 대한 라플라스 근사의 justification을 위해 다음 정리를 증명한다.

정리 원점 $\mathbf{0}$ 근방에서 각각 \mathcal{C}^1 급이고 \mathcal{C}^2 급인 함수 $f, g : \mathbb{R}^d \rightarrow \mathbb{R}$ 에 대해 $f(\mathbf{0}) > 0$ 이고 $\nabla^2 g(\mathbf{0}) > 0$ 이라 하자. 나아가, 임의의 $\delta > 0$ 에 대해 적당한 $\rho > 0$ 가 존재하여 $\|\mathbf{x}\| \geq \delta$ 이면 $g(\mathbf{x}) - g(\mathbf{0}) \geq \rho$ 라 하자. 만약 적당한 $a_0 \in \mathbb{R}$ 가 존재하여 임의의 $a \geq a_0$ 에 대해 적분 $\int_{\mathbb{R}^d} f e^{-ag} d\mu$ 가 절대수렴한다면, $a \rightarrow \infty$ 일 때

$$\int_{\mathbb{R}^d} f e^{-ag} d\mu / \left[f(\mathbf{0}) e^{-ag(\mathbf{0})} \left(\frac{2\pi}{a} \right)^{d/2} \frac{1}{\sqrt{\det \nabla^2 g(\mathbf{0})}} \right] \rightarrow 1 \quad (22)$$

이 성립한다. 여기서 μ 는 르벡측도이다.

PROOF 필요하다면 분자와 분모에 $e^{ag(\mathbf{0})}$ 을 곱하고 g 대신 $g - g(\mathbf{0})$ 을 생각하여 WLOG, $g(\mathbf{0}) = 0$ 이라 해도 된다. 이제 테일러 근사로부터 $\mathbf{x} \rightarrow \mathbf{0}$ 이면 $\nabla g(\mathbf{x}) = \nabla^2 g(\mathbf{0})\mathbf{x} + o(\|\mathbf{x}\|)$ 이고 로피탈의 법칙

으로부터 $\mathbf{x} \rightarrow \mathbf{0}$ 이면 $g(\mathbf{x}) - \mathbf{x}'\nabla^2 g(\mathbf{0})\mathbf{x}/2 = o(||\mathbf{x}||^2)$ 임을 안다. 한편, f 가 $\mathbf{0}$ 에서 \mathcal{C}^1 급이고 $f(\mathbf{0}) > 0$ 임을 생각하면 충분히 작은 임의의 $\varepsilon > 0$ 에 대해 적당한 $\delta > 0$ 가 존재하여 $||\mathbf{x}|| < \delta$ 이면

$$(1 - \varepsilon)f(\mathbf{0}) \leq f(\mathbf{x}) \leq (1 + \varepsilon)f(\mathbf{0}) \quad (23)$$

$$-\varepsilon||\mathbf{x}||^2 \leq g(\mathbf{x}) - \frac{1}{2}\mathbf{x}'\nabla^2 g(\mathbf{0})\mathbf{x} \leq \varepsilon||\mathbf{x}||^2 \quad (24)$$

임을 안다.

이제 적분의 범위를 나누어 $\mathbb{R}^d \setminus B(\mathbf{0}, \delta)$ 에서 적분의 기여를 생각하면, 주어진 g 에 대한 조건으로 부터 $a \geq a_0$ 이면

$$\begin{aligned} \left| \int_{\mathbb{R}^d \setminus B(\mathbf{0}, \delta)} f e^{-ag} d\mu \right| &\leq \int_{\mathbb{R}^d \setminus B(\mathbf{0}, \delta)} |f| e^{-(a-a_0)g} e^{-a_0 g} d\mu \\ &\leq e^{-(a-a_0)\rho} \int_{\mathbb{R}^d} |f| e^{-a_0 g} d\mu \end{aligned} \quad (25)$$

에서 $\int_{\mathbb{R}^d \setminus B(\mathbf{0}, \delta)} f e^{-ag} d\mu = O(e^{-a\rho})$ 이다. 한편, $B(\mathbf{0}, \delta)$ 에서는

$$\begin{aligned} \int_{B(\mathbf{0}, \delta)} f e^{-ag} d\mu &\leq (1 + \varepsilon)f(\mathbf{0}) \int_{\mathbb{R}} \exp\left(-\frac{a}{2}\mathbf{x}'[\nabla^2 g(\mathbf{0}) - 2\varepsilon I_d]\mathbf{x}\right) d\mu(\mathbf{x}) \\ &= (1 + \varepsilon)f(\mathbf{0}) \left(\frac{2\pi}{a}\right)^{d/2} \frac{1}{\sqrt{\det(\nabla^2 g(\mathbf{0}) - 2\varepsilon I_d)}} \end{aligned}$$

으로 위로 유계이다. 이로부터

$$\limsup_{a \rightarrow \infty} a^{d/2} \int_{\mathbb{R}} f e^{-ag} d\mu \leq (1 + \varepsilon)f(\mathbf{0}) \frac{(2\pi)^{d/2}}{\sqrt{\det(\nabla^2 g(\mathbf{0}) - 2\varepsilon I_d)}}$$

이고 ε 이 임의의 양수라는 점을 떠올리면

$$\limsup_{a \rightarrow \infty} a^{d/2} \int_{\mathbb{R}} f e^{-ag} d\mu \leq f(\mathbf{0}) \frac{(2\pi)^{d/2}}{\sqrt{\det(\nabla^2 g(\mathbf{0}))}} \quad (26)$$

을 얻는다.

한편, 적당한 $\lambda > 0$ 가 존재하여 $||\mathbf{x}|| \geq \delta$ 이면 $\mathbf{x}'[\nabla^2 g(\mathbf{0})/2 + \varepsilon I_d]\mathbf{x} \geq \lambda$ 이고 모든 $a > 0$ 에 대해

$$\int_{\mathbb{R}^d} \exp\left(-\frac{a}{2}\mathbf{x}'[\nabla^2 g(\mathbf{0}) + 2\varepsilon I_d]\mathbf{x}\right) d\mu(\mathbf{x}) < \infty \quad (27)$$

이므로 앞선 논의와 비슷하게 하면 $\int_{\mathbb{R}^d \setminus B(\mathbf{0}, \delta)} \exp(-a\mathbf{x}'[\nabla^2 g(\mathbf{0}) + 2\varepsilon I_d]\mathbf{x}/2) d\mu(\mathbf{x}) = O(e^{-a\lambda})$ 임을 안다. 이로부터

$$\begin{aligned} (1 - \varepsilon)f(\mathbf{0}) \int_{B(\mathbf{0}, \delta)} \exp\left(-\frac{a}{2}\mathbf{x}'[\nabla^2 g(\mathbf{0}) + 2\varepsilon I_d]\mathbf{x}\right) d\mu(\mathbf{x}) &\quad (28) \\ &\geq (1 - \varepsilon)f(\mathbf{0}) \int_{\mathbb{R}^d} \exp\left(-\frac{a}{2}\mathbf{x}'[\nabla^2 g(\mathbf{0}) + 2\varepsilon I_d]\mathbf{x}\right) d\mu(\mathbf{x}) - O(e^{-a\lambda}) \\ &= (1 - \varepsilon)f(\mathbf{0}) \left(\frac{2\pi}{a}\right)^{d/2} \frac{1}{\sqrt{\det(\nabla^2 g(\mathbf{0}) + 2\varepsilon I_d)}} + O(e^{-a\lambda}) \end{aligned}$$

이고, 곧

$$\liminf_{a \rightarrow \infty} a^{d/2} \int_{\mathbb{R}^d} f e^{-ag} d\mu \geq (1 - \varepsilon) f(\mathbf{0}) \frac{(2\pi)^{d/2}}{\sqrt{\det(\nabla^2 g(\mathbf{0}) + 2\varepsilon I_d)}} \quad (29)$$

에서 ε 이 임의의 양수라는 점을 떠올리면

$$\liminf_{a \rightarrow \infty} a^{d/2} \int_{\mathbb{R}^d} f e^{-ag} d\mu \geq f(\mathbf{0}) \frac{(2\pi)^{d/2}}{\sqrt{\det \nabla^2 g(\mathbf{0})}} \quad (30)$$

을 얻는다. 이제 식 26과 30로부터 증명이 끝난다. ■

이로부터 적분 $\int_{\mathbb{R}^d} f e^{-ag} d\mu$ 를

$$f(\mathbf{0}) e^{-ag(\mathbf{0})} \left(\frac{2\pi}{a} \right)^{d/2} \frac{1}{\sqrt{\det \nabla^2 g(\mathbf{0})}} \quad (31)$$

으로 근사할 수 있다. 한편, 라플라스 근사의 오차를 줄이기 위해서는 함수 g 의 mass가 $\mathbf{0}$ 근처에 집중되어 있어야 하므로 평행이동을 통해 g 가 $\mathbf{0}$ 에서 최댓값을 가지도록 조정한 후에 근사하는 것이 일반적이다. 이를 GLMM의 로그가능도(식 21)에 적용하기 위해 함수

$$\gamma \mapsto \sum_{i=1}^n [A(\mathbf{x}'_i \beta + \mathbf{z}'_i \gamma) - \mathbf{z}'_i \gamma Y_i] + \frac{\gamma' \Psi^{-1} \gamma}{2} \quad (32)$$

가 $\gamma_* \in \mathbb{R}^q$ 에서 최댓값을 가진다고 하자. 그렇다면 평행이동을 통해 GLMM의 로그가능도에서의 적분은 다음과 같이 쓸 수 있다.

$$\int_{\mathbb{R}^q} \exp \left(\sum_{i=1}^n [\mathbf{z}'_i (\gamma + \gamma_*) Y_i - A(\mathbf{x}'_i \beta + \mathbf{z}'_i (\gamma + \gamma_*))] - \frac{(\gamma + \gamma_*)' \Psi^{-1} (\gamma + \gamma_*)}{2} \right) d\gamma \quad (33)$$

라플라스 근사로부터 위의 적분은

$$\exp \left(- \sum_{i=1}^n A(\mathbf{x}'_i \beta) \right) \frac{(2\pi)^{d/2}}{\sqrt{\det(\sum_{i=1}^n A''(\mathbf{x}'_i \beta + \mathbf{z}'_i \gamma_*) \mathbf{z}'_i \mathbf{z}'_i + \Psi^{-1})}} \quad (34)$$

로 근사할 수 있으므로 곧 식 21를 최대화하는 것은 근사적으로

$$-\sum_{i=1}^n A(\mathbf{x}'_i \beta) - \frac{1}{2} \log \det \left(\sum_{i=1}^n A''(\mathbf{x}'_i \beta + \mathbf{z}'_i \gamma_*) \mathbf{z}'_i \mathbf{z}'_i + \Psi^{-1} \right) - \frac{1}{2} \log \det \Psi + \sum_{i=1}^n \mathbf{x}'_i \beta Y_i \quad (35)$$

를 β, Ψ 에 대해 최대화하는 것과 같다.

GLMM의 고정효과에 대한 검정통계량의 극한분포

본 절에서는 GLMM의 고정효과에 대한 선형가설

$$H_0 : T\beta = \xi \quad H_1 : T\beta \neq \xi \quad (\text{rk } T = r < p) \quad (36)$$

을 검정하는 LRT 검정통계량

$$\Lambda = -2[l(\hat{\beta}, \hat{\psi}) - l(\tilde{\beta}, \tilde{\psi})] \quad (37)$$

의 극한분포가 $\chi^2(p - r)$ 임을 보이도록 한다. 이를 위해 보다 일반적인 다음 정리를 증명한다.

정리 모형 $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta \in \Theta}$ 이 $\vartheta \in \Theta^\circ \subseteq \mathbb{R}^d$ 에 대해 differentiable in quadratic mean⁴이고 Fisher 정보행렬 I_ϑ 가 가역이라 하자. 이제 가설

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1 \quad (\Theta_0 \sqcup \Theta_1 = \Theta) \quad (39)$$

를 검정하는 상황을 생각하고, 적당한 $L^2(\mathbf{P}_\vartheta)$ 함수 $\hat{l} : \mathbb{R} \rightarrow \mathbb{R}$ 이 존재하여 임의의 $x \in \mathbb{R}$ 와 ϑ 의 근방 $\mathcal{N} \subseteq \Theta$ 에 속하는 임의의 $\theta_1, \theta_2 \in \mathcal{N}$ 에 대해

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq \hat{l}(x) \|\theta_1 - \theta_2\| \quad (40)$$

라 하자. 여기서 p_θ 는 확률측도 \mathbf{P}_θ 의 확률밀도함수이다. 만약 θ 의 MLE $\hat{\theta}_n$ 과 귀무가설 하에서의 MLE $\tilde{\theta}_n$ 에 대해 $\hat{\theta}_n, \tilde{\theta}_n \xrightarrow{\mathbf{P}_\vartheta} \theta_0$ 이며 집합 $H_n = \sqrt{n}(\Theta - \vartheta), H_n^0 = \sqrt{n}(\Theta_0 - \vartheta) \subseteq \mathbb{R}^d$ 가 각각 집합 $H, H^0 \subseteq \mathbb{R}^d$ 으로 수렴하면⁵ LRT의 검정통계량 $\Lambda_n = -2[l(\hat{\theta}_n) - l(\tilde{\theta}_n)]$ 에 대해

$$\Lambda_n \xrightarrow{\mathbf{P}_\vartheta + \mathbf{h}/\sqrt{n}} \|I_\vartheta^{1/2} \mathbf{Z} - I_\vartheta^{1/2} H^0\|^2 - \|I_\vartheta^{1/2} \mathbf{Z} - I_\vartheta^{1/2} H\|^2 \quad (41)$$

이 성립한다.⁶ 여기서 $\mathbf{Z} \sim \mathbf{N}_d(\mathbf{0}, I_\vartheta^{-1})$ 이다.

PROOF 검정하고자 하는 가설은 가설 $H_0 : \mathbf{h} \in H_n, H_1 : \mathbf{h} \in H_n^0$ 을 검정통계량

$$\Lambda_n = 2 \sup_{\mathbf{h} \in H_n} \log \frac{\prod_{i=1}^n p_{\vartheta+\mathbf{h}/\sqrt{n}}(X_i)}{\prod_{i=1}^n p_\vartheta(X_i)} - 2 \sup_{\mathbf{h} \in H_n^0} \log \frac{\prod_{i=1}^n p_{\vartheta+\mathbf{h}/\sqrt{n}}(X_i)}{\prod_{i=1}^n p_\vartheta(X_i)} \quad (42)$$

로 검정하는 것과 동등하다. 이제 empirical probability measure \mathbb{P}_n 에 대해 empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbf{P}_\vartheta)$ 을 생각하고, 확률과정 \mathbb{Z}_n 을

$$\mathbb{Z}_n(\mathbf{h}) = n \int_{\mathbb{R}^n} \log \frac{p_{\vartheta+\mathbf{h}/\sqrt{n}}}{p_\vartheta} d\mathbb{P}_n - \mathbf{h}' \int_{\mathbb{R}^n} \hat{l}_\vartheta d\mathbb{G}_n + \frac{1}{2} \mathbf{h}' I_\vartheta \mathbf{h} \quad (43)$$

로 정의하자. 여기서 $\hat{l}_\vartheta = \log p_\vartheta$ 이며 \hat{l}_ϑ 는 l_ϑ 의 differentiable in quadratic mean의 의미로의 미분이다. 그렇다면 Van der Vaart (2000)의 정리 7.12의 증명으로부터 임의의 $M \geq 0$ 에 대해

$$\sup_{\|\mathbf{h}\| \leq M} |\mathbb{Z}_n(\mathbf{h})| \xrightarrow{\mathbf{P}_\vartheta} 0 \quad (44)$$

이고 $\sqrt{n}(\hat{\theta}_n - \vartheta), \sqrt{n}(\tilde{\theta}_n - \vartheta) \xrightarrow{\mathbf{P}_\vartheta} 0$ 임을 알 수 있다. 나아가, 위의 식 44가 충분히 천천히 ∞ 로 발산하는 임의의 수열 $\{M_n\}$ 에 대해서도 성립함을 쉽게 알 수 있다. 이러한 수열 $\{M_n\}$ 을 고정하면 앞선 결론으로부터 MLE $\hat{\theta}_n$ 와 $\tilde{\theta}_n$ 가 $B(\vartheta, M_n/\sqrt{n})$ 에 포함될 확률은 각각 1로 수렴하고, 따라서 Λ_n 의

⁴ 모형 $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta \in \Theta}$ 이 differentiable in quadratic mean이라 함은 임의의 $\theta \in \Theta$ 에 대해 확률측도 \mathbf{P}_θ 의 확률밀도함수 p_θ 에 대해 적당한 가측함수 $\hat{l}_\theta : \mathbb{R} \rightarrow \mathbb{R}^d$ 이 존재하여 $\mathbf{h} \rightarrow \mathbf{0}$ 일 때

$$\int_{\mathbb{R}} \left(\sqrt{p_{\theta+\mathbf{h}}} - \sqrt{p_\theta} - \frac{1}{2} \mathbf{h}' \hat{l}_\theta \sqrt{p_\theta} \right)^2 d\mu = o(\|\mathbf{h}\|^2) \quad (38)$$

임을 의미한다.

⁵ 집합 $A_n \subseteq \mathbb{R}^d$ 이 $A \subseteq \mathbb{R}^d$ 로 수렴한다 함은 임의의 수열 $a_n \in A_n$ 에 대해 $a_n \rightarrow a$ 이면 $a \in A$ 이고, 나아가 이의 임의의 부분열 $\{a_{n_i}\}$ 에 대해 $a_{n_i} \rightarrow a_*$ 이면 $a_* \in A$ 임을 의미한다.

⁶ 집합 $A \subseteq \mathbb{R}^d$ 와 점 $\mathbf{x} \in \mathbb{R}^d$ 에 대해 $\|\mathbf{x} - A\|$ 는 $\inf_{\mathbf{y} \in A} \|\mathbf{x} - \mathbf{y}\|$ 로 정의한다.

극한분포는 집합 H_n 과 H_n^0 을 각각 $H_n \cap B(\mathbf{0}, M_n)$ 과 $H_n^0 \cap B(\mathbf{0}, M_n)$ 로 바꾸더라도 변함이 없다. 또한, 집합 $H_n \cap B(\mathbf{0}, M_n)$ 과 $H_n^0 \cap B(\mathbf{0}, M_n)$ 이 각각 H 와 H^0 으로 수렴함도 자명하다. 이제 식 44으로부터 간단한 산수를 통해

$$\begin{aligned}\Lambda_n &= 2 \sup_{\mathbf{h} \in H_n} n \int_{\mathbb{R}^n} \log \frac{p_{\vartheta+\mathbf{h}/\sqrt{n}}}{p_\vartheta} d\mathbb{P}_n - 2 \sup_{\mathbf{h} \in H_n^0} n \int_{\mathbb{R}^n} \log \frac{p_{\vartheta+\mathbf{h}/\sqrt{n}}}{p_\vartheta} d\mathbb{P}_n \\ &= 2 \sup_{\mathbf{h} \in H_n} \left(\mathbf{h}' \int_{\mathbb{R}^n} l_\vartheta d\mathbb{G}_n - \frac{1}{2} \mathbf{h}' I_\vartheta \mathbf{h} \right) - 2 \sup_{\mathbf{h} \in H_n^0} \left(\mathbf{h}' \int_{\mathbb{R}^n} l_\vartheta d\mathbb{G}_n - \frac{1}{2} \mathbf{h}' I_\vartheta \mathbf{h} \right) + o_{\mathbb{P}_\vartheta}(1) \\ &= \left\| I_\vartheta^{-1/2} \int_{\mathbb{R}^n} l_\vartheta d\mathbb{G}_n - I_\vartheta^{1/2} H_0 \right\|^2 - \left\| I_\vartheta^{-1/2} \int_{\mathbb{R}^n} l_\vartheta d\mathbb{G}_n - I_\vartheta^{1/2} H \right\|^2 + o_{\mathbb{P}_\vartheta}(1)\end{aligned}\quad (45)$$

임을 안다. 마지막 등호는 Van der Vaart (2000)의 보조정리 7.13으로부터 성립하며, 여기서 $o_{\mathbb{P}_\vartheta}(1)$ 은 little o in probability이다. 이제 continuous mapping theorem으로부터 증명이 끝난다. ■

보조정리 확률벡터 $\mathbf{Z} \sim \mathbf{N}_d(0, I_d)$ 와 $\dim V = k$ 인 부분공간 $V < \mathbb{R}^d$ 에 대해 $\|\mathbf{Z} - V\|^2 \sim \chi^2(d - k)$ 이다.

PROOF 부분공간 V 의 기저를 β 라 하고 이를 확장하여 \mathbb{R}^d 의 기저를 구성하여 이를 γ 라 하자. WLOG, γ 의 첫 k 개의 기저가 β 라 하자. 그렇다면 임의의 $\mathbf{x} \in \mathbb{R}^d$ 에 대해 $\|\mathbf{x} - V\|^2$ 는 γ 를 기저로 할 때 \mathbf{x} 좌표 (x_1, \dots, x_d) 를 사용하여 $\sum_{i=k+1}^d x_i^2$ 로 쓸 수 있다. 한편, \mathbb{R}^d 의 표준기저 \mathcal{E} 에서 γ 로의 좌표변환을 기술하는 행렬 $I_\mathcal{E}^\gamma$ 는 직교행렬이므로 γ 를 기저로 할 때 확률벡터 \mathbf{Z} 의 좌표 (Z_1, \dots, Z_d) 에 대해 각 Z_i 는 표준정규분포를 따른다. 이상의 사실을 종합하면 $\|\mathbf{Z} - V\|^2 = \sum_{i=k+1}^d Z_i^2 \sim \chi^2(d - k)$ 이다. ■

먼저 $\xi = \mathbf{0}$ 인 경우를 생각하면 검정하고자 하는 가설(식 36)은 $H_0 : \beta \in \ker T$ 와 동등하고 $\dim \ker T = p - r$ 이다. 이 경우, GLMM이 위의 정리가 요구하는 조건을 모두 만족함을 쉽게 보일 수 있다 (Van der Vaart 2000). 그렇다면 이상의 논의로부터 Λ 의 극한분포가 $\chi^2(p - r)$ 임을 알고, 일반적인 경우에는 적당한 $\beta_0 \in T^{-1}(\xi)$ 에 대해 $\delta = \beta_0 + \beta$ 의 reparameterization을 생각하면 같은 결론에 이른다.

동일한 가설을 검정하는 Wald의 검정통계량과 Rao의 검정통계량

$$W = (T\hat{\beta} - \xi)' \{ T[\nabla_\beta^2 l(\hat{\beta}, \hat{\Psi})]^{-1} T' \}^{-1} (T\hat{\beta} - \xi) \quad (46)$$

$$R = \nabla_\beta l(\tilde{\beta}, \tilde{\Psi})' [\nabla_\beta^2 l(\tilde{\beta}, \tilde{\Psi})]^{-1} \nabla_\beta l(\tilde{\beta}, \tilde{\Psi}) \quad (47)$$

의 경우, 로그가능도에 대한 테일러 근사로부터 $\Lambda = W + o_{\beta, \Psi}(1) = R + o_{\beta, \Psi}(1)$ 임을 보일 수 있고 (Bickel & Doksum 2015), 따라서 두 검정통계량의 극한분포도 $\chi^2(p - r)$ 임을 안다.

GLMM의 랜덤효과에 대한 예측값

본 절에서는 GLMM에서 랜덤효과의 예측에 대한 MSPE $\mathbf{E}_{\beta, \Psi} \|\gamma - \gamma_*\|^2$ 가 $\mathbf{E}_{\beta, \Psi}(\gamma | \mathbf{Y})$ 의해 최소화됨을 보이도록 한다.

정리 확률공간 $(\Omega, \mathcal{F}, \mathbb{P})$ 와 확률벡터 $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$, sub-sigma field $\mathcal{E} \subseteq \mathcal{F}$ 를 생각하자. 그렇다면 임의의 \mathcal{E} -가측인 확률벡터 $\mathbf{Y} : \Omega \rightarrow \mathbb{R}^d$ 에 대해 $\mathbf{X}, \mathbf{Y} \in L^2$ 라면 $\mathbf{E}([\mathbf{X} - \mathbf{E}(\mathbf{X}|\mathcal{E})]'\mathbf{Y}) = 0$ 이다.

PROOF 이는 $\mathbf{E}([\mathbf{X} - \mathbf{E}(\mathbf{X}|\mathcal{E})]'\mathbf{Y}) = \mathbf{E}(\mathbf{E}([\mathbf{X} - \mathbf{E}(\mathbf{X}|\mathcal{E})]'\mathbf{Y}|\mathcal{E})) = \mathbf{E}(\mathbf{E}(\mathbf{X} - \mathbf{E}(\mathbf{X}|\mathcal{E})|\mathcal{E})'\mathbf{Y}) = \mathbf{0}$ 에서 자명하다. ■

따름정리 확률공간 $(\Omega, \mathcal{F}, \mathbb{P})$ 와 확률벡터 $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$, sub-sigma field $\mathcal{E} \subseteq \mathcal{F}$ 에 대해 $\mathbf{X} \in L^2$ 라 하자. 그렇다면 $\operatorname{argmin}_{\mathbf{Y} \in L^2(\mathcal{E}, \mathbb{P})} \mathbf{E}\|\mathbf{X} - \mathbf{Y}\|^2 = \mathbf{E}(\mathbf{X}|\mathcal{E})$ 이다.

PROOF 임의의 $\mathbf{Y} \in L^2(\mathcal{E}, \mathbb{P})$ 에 대해 $\mathbf{E}\|\mathbf{X} - \mathbf{Y}\|^2 = \mathbf{E}\|\mathbf{X} - \mathbf{E}(\mathbf{X}|\mathcal{E})\|^2 + \mathbf{E}\|\mathbf{E}(\mathbf{X}|\mathcal{E}) - \mathbf{Y}\|^2 + 2\mathbf{E}((\mathbf{X} - \mathbf{E}(\mathbf{X}|\mathcal{E}))(\mathbf{E}(\mathbf{X}|\mathcal{E}) - \mathbf{Y}))$ 인데 여기서 $\mathbf{E}(\mathbf{X}|\mathcal{E}) - \mathbf{Y} \in L^2(\mathcal{E}, \mathbb{P})$ 이므로 위의 정리로부터 $\mathbf{E}\|\mathbf{X} - \mathbf{Y}\|^2 \geq \mathbf{E}\|\mathbf{X} - \mathbf{E}(\mathbf{X}|\mathcal{E})\|^2$ 이다. 증명은 이로써 충분하다. ■

이상의 결과로부터 MSPE를 최소화하는 γ 의 예측값은 $\mathbf{E}_{\beta, \Psi}(\gamma|\mathbf{Y})$ 임이 분명하다.

이진 분류 문제와 CV

본 절에서는 일반적인 이진 분류 문제를 염밀하게 기술하고, 이로부터 본 연구에서 CV를 통해 각종 통계량을 추정한 것에 대한 justification을 진행하고자 한다. 통계적 결정이론의 관점에서 이진 분류 문제는 주어진 확률공간 $(\Omega, \mathcal{F}, \mathbb{P})$ 와 확률벡터 $(\mathbf{X}, G) : \Omega \rightarrow \mathbb{R}^d \times \{0, 1\}$ 에 대해 realize된 (\mathbf{X}, G) 에서 \mathbf{X} 만으로 G 를 예측하는 문제로 생각할 수 있다.⁷ 이때 분류의 방법, 즉 분류기는 확률벡터 \mathbf{X} 의 지지집합 \mathcal{X} 에 대해 가측함수 $\phi : \mathcal{X} \rightarrow \{0, 1\}$ 로 기술되며 대부분의 경우 (\mathbf{X}, G) 의 joint PDF f 는 알려져 있지 않다고 가정한다.

분류기 ϕ 의 성능을 나타내는 방법으로 흔히 사용되는 confusion matrix는 분할표의 일종이다. 만약 N 개의 독립표본 $(\mathbf{X}_i, G_i) \stackrel{\text{iid}}{\sim} f$ 에 대해 분류기 ϕ 로써 분류를 진행했다면 다음과 같은 confusion matrix를 구성할 수 있다.

$$\begin{array}{cc} & \text{Pos} \quad \text{Neg} \\ \text{Pos} & \left[\begin{array}{cc} N_{\text{TP}} & N_{\text{FP}} \\ N_{\text{FN}} & N_{\text{TN}} \end{array} \right] \\ \text{Neg} & \end{array} \quad (48)$$

여기서 왼쪽의 라벨은 분류기 ϕ 가 예측한 G 의 값, 위쪽의 라벨은 실제 G 의 값이며 각 성분은 해당되는 분류 결과의 수이다. (관례를 따라 1을 positive로, 0을 negative로 생각한다.) 그렇다면 confusion matrix의 성분으로 구성된 벡터 $\mathbf{N}_{\text{confu}} = (N_{\text{TP}}, N_{\text{FN}}, N_{\text{FP}}, N_{\text{TN}})'$ 는 적당한 $\mathbf{p} \in \mathbb{R}^4$ 에 대해 다항분포 $\text{Mult}(N, \mathbf{p})$ 를 따르며, 이때의 매개변수 $\mathbf{p} = (p_1, p_2, p_3, p_4)'$ 는 confusion matrix의 의미를 고려했을

⁷ 기계학습의 관점에서 \mathbf{X} 는 관측한 데이터이고 G 는 해당 데이터의 라벨이다.

때 다음과 같이 정의된다.

$$\begin{aligned} p_1 &= \mathbf{P}(\phi(X) = 1 | G = 1) \\ p_2 &= \mathbf{P}(\phi(X) = 0 | G = 1) \\ p_3 &= \mathbf{P}(\phi(X) = 1 | G = 0) \\ p_4 &= \mathbf{P}(\phi(X) = 0 | G = 0) \end{aligned} \quad (49)$$

여기서 \mathbf{P} 는 pushforward measure $\mathbf{X}_*\mathbb{P}$ 이다.

본 연구를 비롯한 여러 상황에서 보편적으로 사용되는 분류기의 성능 지표인 TPR은 다음과 같이 정의된다.

$$TPR = \mathbf{E}\left(\frac{N_{TP}}{N_{TP} + N_{FN}}\right) \quad (50)$$

이는 (\mathbf{X}, G) 의 분포 f 와 분류기 ϕ 에 의존하는 통계량이지만, 대부분의 경우 f 와 ϕ 를 알지 못하므로 직접 구하지 못한다는 한계가 있다. 만약 realize되어 완전히 알고 있는 N 개의 관측값 $(\mathbf{X}_i, G_i) \stackrel{iid}{\sim} f$ 이 주어져 이로부터 ϕ 를 $\hat{\phi}$ 로 적합할 수 있다면

$$TPR \approx \frac{\mathbf{E}(N_{TP})}{\mathbf{E}(N_{TP} + N_{FN})} = \frac{\mathbf{P}(\phi(X) = 1, G = 1)}{\mathbf{P}(G = 1)} \quad (51)$$

에서 TPR의 추정량으로 다음을 생각해볼 수 있다.

$$\widehat{TPR}_{\mathcal{D}} = \frac{\mathbf{P}(\hat{\phi}(X) = 1, G = 1 | \mathcal{D})}{\mathbf{P}(G = 1 | \mathcal{D})} \quad (52)$$

여기서 \mathcal{D} 는 ϕ 를 적합하는데 사용된 주어진 N 개의 관측값의 집합, 즉 $\{(\mathbf{X}_i, G_i)\}_{1 \leq i \leq N}$ 을 의미한다. 아래첨자의 \mathcal{D} 에서 알 수 있듯이 이 추정량은 \mathcal{D} 에 의존하므로 이에 대한 기댓값을 취하여 최종적인 추정량을 얻을 수 있다.

$$\widehat{TPR} = \mathbf{E}_{\mathcal{D}}\left(\frac{\mathbf{P}(\hat{\phi}(X) = 1, G = 1 | \mathcal{D})}{\mathbf{P}(G = 1 | \mathcal{D})}\right) \quad (53)$$

그러나 여전히 (\mathbf{X}, G) 의 joint PDF f 를 알지 못하므로 이는 실제로는 구하지 못하는 값이다. 이런 상황에서, k -fold CV는 식 53를 근사하는 방법으로서 그 justification을 획득한다.

k -fold CV는 기계학습 분야에서 널리 사용되는 검증 내지는 성능 측정 방법으로, 이미 realize되어 완전히 알고 있는 관측값 $(\mathbf{X}_i, G_i) \stackrel{iid}{\sim} f$ 를 랜덤하게 분할하고, 이 중 일부만으로 ϕ 를 적합한 다음, 사용하지 않은 나머지를 통해 추정하고자 하는 통계량을 계산한다.⁸ 보다 구체적으로, k -fold CV를 통해 앞서 소개한 TPR을 구하는 상황을 생각하자. 앞선 논의와 동일하게 N 개의 관측값의 집합을 \mathcal{D} 라고 하고, 적당한 k 를 택하여 (간결한 논의를 위해 N 이 k 의 배수라 하자) \mathcal{D} 를 랜덤하게 k 개의 부분집합 $\mathcal{V}_1, \dots, \mathcal{V}_k$ 로 분할한다. 나아가, 각 $i \leq k$ 에 대해 $\mathcal{D} \setminus \mathcal{V}_i$ 를 통해 적합한 ϕ 를 $\hat{\phi}_{\mathcal{D} \setminus \mathcal{V}_i}$ 라 하자. 그렇다면

⁸ 기계학습의 관점에서는 이때의 관측값 중 ϕ 의 적합에 사용되는 관측값이 train dataset이고, 통계량의 계산에 사용되는 관측값이 test dataset이다. 또한, 이때 ϕ 를 적합하는 과정을 분류기를 훈련시킨다고 이른다.

대수의 법칙으로부터

$$\frac{1}{N/k} \sum_{(\mathbf{X}, G) \in \mathcal{V}_i} \mathbf{1}\{\hat{\phi}_{\mathcal{D} \setminus \mathcal{V}_i}(\mathbf{X}) = 1, G = 1\} \approx \mathbf{P}(\hat{\phi}_{\mathcal{D} \setminus \mathcal{V}_i}(\mathbf{X}) = 1, G = 1 | \mathcal{D} \setminus \mathcal{V}_i) \quad (54)$$

$$\frac{1}{N/k} \sum_{(\mathbf{X}, G) \in \mathcal{V}_i} \mathbf{1}\{G = 1\} \approx \mathbf{P}(G = 1 | \mathcal{D} \setminus \mathcal{V}_i) \quad (55)$$

이고, 곧

$$\begin{aligned} \frac{1}{k} \sum_{i=1}^k \frac{\sum_{(\mathbf{X}, G) \in \mathcal{V}_i} \mathbf{1}\{\hat{\phi}_{\mathcal{D} \setminus \mathcal{V}_i}(\mathbf{X}) = 1, G = 1\}}{\sum_{(\mathbf{X}, G) \in \mathcal{V}_i} \mathbf{1}\{G = 1\}} &\approx \frac{1}{k} \sum_{i=1}^k \frac{\mathbf{P}(\hat{\phi}_{\mathcal{D} \setminus \mathcal{V}_i}(\mathbf{X}) = 1, G = 1 | \mathcal{D} \setminus \mathcal{V}_i)}{\mathbf{P}(G = 1 | \mathcal{D} \setminus \mathcal{V}_i)} \\ &\approx \mathbf{E}_{\mathcal{D}} \left(\frac{\mathbf{P}(\hat{\phi}(X) = 1, G = 1 | \mathcal{D})}{\mathbf{P}(G = 1 | \mathcal{D})} \right) \end{aligned} \quad (56)$$

가 되어 k -fold CV를 통해 계산한 TPR은 식 53에서의 $\widehat{\text{TPR}}$ 의 근사가 된다. 이와 유사한 방법으로 ACC나 MCC와 같은 보다 복잡한 통계량에 대해서도 k -fold CV를 통한 추정을 justify할 수 있다.

한편, k -fold CV를 통해 추정된 ROC와 AUC에 대해서는 조금 다른 justification이 필요하다. 앞선 논의는 기본적으로 confusion matrix에서 출발하였기에 ROC를 정의조차 할 수 없기 때문이다. 이를 위해서 이번에는 분류기를 어떤 가측함수 $\psi : \mathcal{X} \rightarrow [0, 1]$ 와 $c \in [0, 1]$ 에 대해 $\phi : \mathbf{x} \mapsto \mathbf{1}\{\psi(\mathbf{x}) > c\}$ 로 정의되는 함수로 정의한다. 여기서 $\psi(\mathbf{x})$ 를 조건부확률 $\mathbf{P}(G|\mathbf{X} = \mathbf{x})$ 에 대한 근사로 해석한다면 ϕ 는 조건부확률 $\mathbf{P}(G|\mathbf{X} = \mathbf{x})$ 가 c 보다 크면 이를 1로 분류하는 분류기로 생각할 수 있다. 곧 c 는 분류기의 행동을 결정짓는 threshold의 역할을 하여 이때의 TPR과 FPR은 c 에 대해 다음과 같이 정의할 수 있다.

$$\text{TPR}(c) = \mathbf{P}(\phi(X) = 1 | G = 1) = \int_{\{\psi > c\}} f_{\mathbf{X}|G=1} d\mu \quad (57)$$

$$\text{FPR}(c) = \mathbf{P}(\phi(X) = 1 | G = 0) = \int_{\{\psi > c\}} f_{\mathbf{X}|G=0} d\mu \quad (58)$$

이로부터 ROC는 매개곡선 $\text{ROC}(c) = (\text{FPR}(c), \text{TPR}(c))$ 으로 자연스럽게 정의되고, AUC는 곡선 $\text{ROC}(c)$ 와 x 축이 이루는 넓이가 되어 다음의 르벡-스틸체스 적분으로 정의할 수 있다.

$$\text{AUC} = \int_0^1 \text{TPR} d\text{FPR} \quad (59)$$

앞선 논의로부터 k -fold CV를 통해 계산한 TPR과 FPR이 실제 값에 대한 근사가 됨을 알고 있으므로 각각을 $\widehat{\text{TPR}}$ 와 $\widehat{\text{FPR}}$ 로 쓰면 곧 k -fold CV 통해 구한 ROC는 $\widehat{\text{ROC}}(c) = (\widehat{\text{FPR}}(c), \widehat{\text{TPR}}(c))$ 이고 AUC는 $\int_0^1 \widehat{\text{TPR}} d\widehat{\text{FPR}}$ 이 되어 그 justification을 획득한다.

베이즈 분류기

본 절에서는 베이즈 분류기의 optimality를 증명하고 이의 근사라는 관점에서 본 연구에서 적합한 분류기들을 살펴본다. 먼저 다음 정리를 증명하자.

정리 0-1 손실함수를 사용할 때, 위험 $R(\phi) = \mathbf{E}(\mathbf{1}\{\phi(\mathbf{X}) \neq G\})$ 를 최소화하는 분류기 $\phi : \mathcal{X} \rightarrow \{0, 1\}$ 는 다음의 베이즈 분류기이다.

$$\phi_{\text{Bayes}} : \mathbf{x} \mapsto \underset{g=1,2}{\operatorname{argmax}} \mathbf{P}(G = g | \mathbf{X} = \mathbf{x}) \quad (60)$$

PROOF 이는 임의의 분류기 ϕ 에 대해

$$\begin{aligned} R(\phi) &= \int_{\mathbb{R}^d} \mathbf{P}(\phi(\mathbf{x}) \neq G | \mathbf{X} = \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^d} [\mathbf{P}(\phi(\mathbf{x}) = 0 | \mathbf{X} = \mathbf{x}, G = 1) \mathbf{P}(G = 1 | \mathbf{X} = \mathbf{x}) \\ &\quad + \mathbf{P}(\phi(\mathbf{x}) = 1 | \mathbf{X} = \mathbf{x}, G = 0) \mathbf{P}(G = 0 | \mathbf{X} = \mathbf{x})] f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (61)$$

에서 $\mathbf{P}(\phi(\mathbf{x}) = 0 | \mathbf{X} = \mathbf{x}, G = 1)$ 과 $\mathbf{P}(\phi(\mathbf{x}) = 1 | \mathbf{X} = \mathbf{x}, G = 0)$ 이 동시에 1이 되지 못한다는 점에서 자명하다. ■

대부분의 경우 (\mathbf{X}, G) 의 joint PDF를 알지 못하므로 베이즈 분류기를 직접 구하는 것은 불가능하다. 따라서 이를 근사하려는 많은 시도가 있었고, 많은 분류기들은 베이즈 분류기에 대한 나름의 근사로 해석할 수 있다.

Naive Bayes 분류기는 G 가 조건부로 주어졌을 때 \mathbf{X} 의 각 성분이 서로 조건부독립임을 가정한다. 그렇다면 베이즈 정리로부터 다음이 성립한다.

$$\begin{aligned} \mathbf{P}(G = g | \mathbf{X} = \mathbf{x}) &= \frac{\mathbf{P}(G = g) f_{X_1, \dots, X_d | G=g}(x_1, \dots, x_d)}{f_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{\mathbf{P}(G = g) f_{X_1 | G=g}(x_1) f_{X_2, \dots, X_d | G=g, X_1=x_1}(x_2, \dots, x_d)}{f_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{\mathbf{P}(G = g) f_{X_1 | G=g}(x_1) f_{X_2, \dots, X_d | G=g}(x_2, \dots, x_d)}{f_{\mathbf{X}}(\mathbf{x})} \\ &= \dots \\ &= \frac{\mathbf{P}(G = g) \prod_{i=1}^d f_{X_i | G=g}(x_i)}{f_{\mathbf{X}}(\mathbf{x})} \end{aligned}$$

이를 실제로 계산하기 위해서는 $X_i | G = g$ 의 conditional PDF를 커널분포추정 등의 비모수적인 방법으로 추정하거나 적당한 분포를 가정하고 그 모수를 추정해야 하는데, 보통 $X_i | G = g$ 가 정규분포 $N(\mu_{ig}, \sigma_{ig}^2)$ 를 따른다고 가정하는 것이 일반적이다. 분류기를 적합하는 데 사용되는 자료 중에서 $G = g$ 인 것만을 택하고, 이들의 표본평균 \bar{X}_{ig} 와 표본분산 S_{ig}^2 로 μ_{ig} 와 σ_{ig}^2 를 추정하자. 한편, $\mathbf{P}(G = g)$ 의 경우 관측치 중 $G = g$ 인 자료의 비율 $\hat{\pi}_g = n_g / N$ 로 추정하는 것이 일반적이므로 (여기서 n_g 는 $G = g$ 인 자료의 개수이다) 조건부확률 $\mathbf{P}(G = g | \mathbf{X} = \mathbf{x})$ 의 추정량은 다음과 같이 쓸 수 있다.

$$\widehat{\mathbf{P}}(G = g | \mathbf{X} = \mathbf{x}) = \frac{\widehat{\pi}_g}{f_{\mathbf{X}}(\mathbf{x})} \prod_{i=1}^d \frac{1}{\sqrt{2\pi S_{ig}^2}} \exp\left(-\frac{(x_i - \bar{X}_{ig})^2}{2S_{ig}^2}\right) \quad (62)$$

여기서 $f_{\mathbf{X}}$ 또한 알지 못하는 분포이지만 이는 $\mathbf{x} \in \mathbb{R}^d$ 가 주어지면 $\sum_{g=0}^1 \widehat{\mathbf{P}}(G = g | \mathbf{X} = \mathbf{x}) = 1$ 이 되도록 하는 정규화 상수로 주어지므로 별도로 추정할 필요가 없다. 이를 베이즈 분류기 ϕ_{Bayes} 의 정의(식 60)에 그대로 대입하여 정리하면 Naive Bayes 분류기 $\widehat{\phi}_{\text{Naive}}$ 를 얻는다.

$$\widehat{\phi}_{\text{Naive}} : \mathbf{x} \mapsto \mathbf{1} \left\{ \frac{\widehat{\pi}_1}{\widehat{\pi}_0} \prod_{i=1}^d \sqrt{\frac{S_{i0}^2}{S_{i1}^2}} \exp\left(-\frac{(x_i - \bar{X}_{i0})^2}{2S_{i0}^2} - \frac{(x_i - \bar{X}_{i1})^2}{2S_{i1}^2}\right) > 1 \right\} \quad (63)$$

LDA는 naive Bayes 분류기와 비슷한 접근으로 시작하지만, \mathbf{X} 의 각 성분이 서로 조건부독립이라 가정하지 않는다. 우선 베이즈 정리로부터 $\mathbf{P}(G = g|\mathbf{X} = \mathbf{x}) = \mathbf{P}(G = g)f_{\mathbf{X}|G=g}(\mathbf{x})$ 이다. Naive Bayes 분류기가 조건부독립성을 근거로 $f_{\mathbf{X}|G=g}$ 를 계속 marginal PDF의 곱으로 분해해나갔다면 LDA는 $f_{\mathbf{X}|G=g}$ 를 곧바로 추정하고자 한다. 이를 위해서 보통 $\mathbf{X}|G = g$ 가 다변량정규분포 $\mathbf{N}_d(\mu_g, \Sigma)$ 를 따른다고 가정하고, 이때 특히 분산 Σ 는 g 에 의존하지 않는다고 가정한다.⁹ 이제 적합에 사용되는 자료 중에서 $G = g$ 인 것만을 택하여 이들의 표본평균벡터 $\bar{\mathbf{X}}_g$ 로 μ_g 를 추정하고, 이들의 표본공분산행렬 S_g 로부터 계산한 pooled covariance matrix $S_p = \sum_{g=0}^1 (n_g - 1)S_g / (N - 1)$ 로 Σ 를 추정하자. 또한, naive Bayes 분류기에서와 같이 $\mathbf{P}(G = g)$ 는 관측치 중 $G = g$ 인 자료의 비율 $\hat{\pi}_g = n_g/N$ 로 추정하면 조건부확률 $\mathbf{P}(G = g|\mathbf{X} = \mathbf{x})$ 의 추정량은 다음과 같이 쓸 수 있다.

$$\hat{\mathbf{P}}(G = g|\mathbf{X} = \mathbf{x}) = \frac{\hat{\pi}_g}{\sqrt{\det(2\pi S_p)}} \exp\left(-\frac{(\mathbf{x} - \bar{\mathbf{X}}_g)' S_p^{-1} (\mathbf{x} - \bar{\mathbf{X}}_g)}{2}\right) \quad (64)$$

이를 베이즈 분류기 ϕ_{Bayes} 의 정의(식 60)에 그대로 대입하여 정리하면 LDA $\hat{\phi}_{\text{LDA}}$ 를 얻는다.¹⁰

$$\hat{\phi}_{\text{LDA}} : \mathbf{x} \mapsto \mathbf{1} \left\{ (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)' S_p^{-1} \mathbf{x} - \frac{1}{2} (\bar{\mathbf{X}}_1' S_p^{-1} \bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0' S_p^{-1} \bar{\mathbf{X}}_0) + \log \frac{\hat{\pi}_1}{\hat{\pi}_0} > 0 \right\} \quad (65)$$

앞서 살펴본 두 분류기가 모두 특정한 분포를 가정한 것과 달리 나머지 분류기는 별도의 분포 가정 없이 조건부확률 $\mathbf{P}(G = g|\mathbf{X} = \mathbf{x})$ 을 직접 추정하고자 한다. 먼저 KNN에 대해 살펴보기 위해 주어진 $\mathbf{x} \in \mathbb{R}^d$ 의 k 최근접 이웃을 $(\mathbf{X}_1, G_1), \dots, (\mathbf{X}_k, G_k)$ 라 하자. 그렇다면 KNN $\hat{\phi}_{\text{KNN}}$ 은 다음과 같이 정의된다.

$$\hat{\phi}_{\text{KNN}} : \mathbf{x} \mapsto \mathbf{1} \left\{ \sum_{i=1}^k \mathbf{1}\{G_i = 1\} > \frac{k}{2} \right\} \quad (66)$$

만약 자료가 충분히 조밀하다면 대수의 법칙으로부터 $\sum_{i=1}^k \mathbf{1}\{G_i = 1\}/k \approx \mathbf{P}(G = 1|\mathbf{X} = \mathbf{x})$ 의 근사가 성립하고, 따라서 KNN도 베이즈 분류기의 근사로 볼 수 있다. 로지스틱 회귀의 경우 link function $g : \mathbb{R} \rightarrow [0, 1]$ 를 통해 $\mathbf{P}(G = 1|\mathbf{X} = \mathbf{x}) \approx g(\mathbf{x}'\beta)$ 의 형태를 가진다고 가정하고, 이때의 β 를 MLE $\hat{\beta}$ 로 추정한다. 이렇게 정의되는 로지스틱 회귀를 이용한 분류기 $\hat{\phi}_{\text{Log}}$ 는

$$\hat{\phi}_{\text{Log}} : \mathbf{x} \mapsto \mathbf{1} \left\{ g(\mathbf{x}'\hat{\beta}) > \frac{1}{2} \right\} \quad (67)$$

로 쓸 수 있고, 곧 베이즈 분류기의 근사이다. ANN 역시 이와 비슷하게 정의되지만 로지스틱 회귀와 달리 $\mathbf{P}(G = 1|\mathbf{X} = \mathbf{x})$ 가 어떤 함수의 형태를 가지는지 가정하지 않는다. 여러 분야에서 다양한 형태의 ANN이 사용되고 있지만, 기본적인 형태인 k 개의 은닉층을 가지는 FFNN의 경우, 이는 비선형 회귀분석의 일종으로 Affine 변환 A_1, \dots, A_k 와 활성함수 σ 의 합성 $F = \sigma \circ A_k \circ \dots \circ \sigma \circ A_1$ 을 통해 $\mathbf{P}(G = 1|\mathbf{X} = \mathbf{x})$ 를 근사하는 것으로 파악할 수 있다. 곧 이러한 근사 F 를 이용하여 ANN $\hat{\phi}_{\text{ANN}}$ 은

$$\hat{\phi}_{\text{ANN}} : \mathbf{x} \mapsto \mathbf{1} \left\{ F(\mathbf{x}) > \frac{1}{2} \right\} \quad (68)$$

⁹ 만약 여기서 분산 또한 g 에 의존한다고 가정하고 이하의 과정을 따라 베이즈 분류기를 근사하면 QDA(Quadratic Discriminant Analysis; 이차판별분석)를 얻는다.

¹⁰ 한편, LDA는 R. A. Fisher에 의해 1936년 이와는 독립적으로 제안되었다 (Fisher 1936). Fisher가 LDA를 도입할 때 사용한 아이디어는 PCA와 비슷한 것으로, 자료를 LD direction이라 불리는 방향으로 사영함으로써 그룹 간의 분산은 최대화하고 그룹 내의 분산은 최소화하는 것이었다.

로 정의되며, 이로부터 ANN 역시 베이즈 분류기에 대한 근사임을 안다. 한편, 이 근사에 대한 이론적 근거를 마련하는 작업은 현재에도 활발히 진행중인데, 흔히 universal approximation theorem이라 총칭되는 몇몇 결과를 소개하면 다음과 같다.

정리 (Cybenko 1989) 함수 $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ 를 상수가 아닌 유계 연속함수라 하자. 또한, 콤팩트한 $K \subseteq \mathbb{R}^d$ 에 대해 K 에서 연속인 함수의 모임을 $\mathcal{C}(K)$ 라 하자. 그렇다면 1개의 input node와 output node를 가지며 은닉층이 1개이고, σ 를 활성함수로 하는 FFNN의 모임 \mathcal{N} 에 대해 \mathcal{N} 은 $\mathcal{C}(K)$ 에서 uniform norm으로 조밀하다. 즉, 임의의 $\varepsilon > 0$ 과 임의의 $f \in \mathcal{C}(K)$ 에 대해 적당한 $F \in \mathcal{N}$ 가 존재하여 $\sup_{x \in K} |F(x) - f(x)| < \varepsilon$ 이다.

정리 (Hanin & Sellke 2017) 함수 $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ 를 ReLU $\sigma : x \mapsto \max\{0, x\}$ 라 하자. 그렇다면 d 개의 input node와 1개의 output node를 가지며 각 은닉층이 많아야 $d + 4$ 개의 node를 가지고, σ 를 활성함수로 하는 FFNN의 모임 \mathcal{N} 에 대해 \mathcal{N} 은 $L^1(\mathbb{R}^d)$ 에서 L^1 norm으로 조밀하다. 즉, 임의의 $\varepsilon > 0$ 과 임의의 $f \in L^1(\mathbb{R}^d)$ 에 대해 적당한 $F \in \mathcal{N}$ 가 존재하여 $\int_{\mathbb{R}^d} |F - f| d\mu < \varepsilon$ 이다.

정리 (Kidger & Lyons 2019) 함수 $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ 가 적어도 한 점에서 미분가능하고 그 점에서의 미분이 0이 아닌 nonaffine 연속함수라 하자. 또한, 콤팩트한 $K \subseteq \mathbb{R}^d$ 에 대해 K 에서 연속인 함수 $f : K \rightarrow \mathbb{R}^k$ 의 모임을 $\mathcal{C}(K; \mathbb{R}^k)$ 라 하자. 이제 d 개의 input node와 k 개의 output node를 가지며 각 은닉층이 $d + k + 2$ 개의 node를 가지고, σ 를 활성함수로 하는 FFNN의 모임을 \mathcal{N} 이라 하면 \mathcal{N} 은 $\mathcal{C}(K)$ 에서 uniform norm으로 조밀하다. 즉, 임의의 $\varepsilon > 0$ 과 임의의 $f \in \mathcal{C}(K; \mathbb{R}^k)$ 에 대해 적당한 $F \in \mathcal{N}$ 가 존재하여 $\sup_{\mathbf{x} \in K} \|F(\mathbf{x}) - f(\mathbf{x})\| < \varepsilon$ 이다.

마지막으로 random forest의 경우, 비교적 최근에 이가 KNN과 깊은 관련이 있음이 밝혀졌다. Random forest를 구성하는 개개의 결정트리는 나름의 결정규칙으로 조건부확률 $\mathbf{P}(G = g | \mathbf{X} = \mathbf{x})$ 을 근사하는데, j 번째 결정트리의 근사치는 적당한 가중치 함수 $W_j : \mathcal{X}^2 \rightarrow [0, 1]$ 에 대해 다음과 같이 쓸 수 있다 (Lin & Jeon 2006).

$$\widehat{\mathbf{P}}_j(G = g | \mathbf{X} = \mathbf{x}) = \sum_{i=1}^N W_j(\mathbf{X}_i, \mathbf{x}) G_i \quad (69)$$

이제 전체 random forest는 이를 구성하는 결정트리들의 결과값을 평균하여 최종적인 $\mathbf{P}(G = g | \mathbf{X} = \mathbf{x})$ 에 대한 근사값을 결정하므로 곧 random forest $\widehat{\phi}_{RF}$ 는 다음과 같이 기술할 수 있다.

$$\widehat{\phi}_{RF} : \mathbf{x} \mapsto \mathbf{1} \left\{ \frac{1}{T} \sum_{j=1}^T \sum_{i=1}^N W_j(\mathbf{X}_i, \mathbf{x}) G_i > \frac{1}{2} \right\} \quad (70)$$

이제 위의 식에서의 합을 그 순서를 바꾸어

$$\sum_{i=1}^N \underbrace{\frac{1}{T} \sum_{j=1}^T W_j(\mathbf{X}_i, \mathbf{x})}_{\widetilde{W}} G_i = \sum_{i=1}^N \widetilde{W}(\mathbf{X}_i, \mathbf{x}) G_i \quad (71)$$

로 쓰고, KNN의 정의 (식 66)를

$$W(\mathbf{X}_i, \mathbf{x}) = \frac{1}{k} \mathbf{1}\{\mathbf{X}_i \text{ is } k\text{-nearest neighbor of } \mathbf{x}\} \quad (72)$$

에 대해

$$\hat{\phi}_{\text{KNN}} : \mathbf{x} \mapsto \mathbf{1}\left\{ \sum_{i=1}^k W(\mathbf{X}_i, \mathbf{x}) G_i > \frac{1}{2} \right\} \quad (73)$$

로 쓸 수 있다는 사실로부터 random forest는 가중치를 $1/k$ 로 균등하게 부여하는 KNN에서 가중치의 부여 방식만 바뀌었음을 알 수 있다. 그렇다면 앞서 KNN이 베이즈 분류기의 근사임을 보였으므로 random forest 또한 베이즈 분류기의 근사가 됨은 분명하다.

참고문헌

- Agresti, A., J. G. Booth*, J. P. Hobert*, & B. Caffo*, 2000 Random-effects modeling of categorical response data. *Sociological Methodology* 30: 27--80.
- Baker, S. N., 2007 Oscillatory interactions between sensorimotor cortex and the periphery. *Current opinion in neurobiology* 17: 649--655.
- Beyer, K., J. Goldstein, R. Ramakrishnan, & U. Shaft, 1999 When is “nearest neighbor” meaningful? In *International conference on database theory*, pp. 217--235, Springer.
- Bickel, P. J. & K. A. Doksum, 2015 *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117. CRC Press.
- Bohórquez, J. & Ö. Özdamar, 2008 Generation of the 40-hz auditory steady-state response (assr) explained using convolution. *Clinical neurophysiology* 119: 2598--2607.
- Capanu, M., M. Gönen, & C. B. Begg, 2013 An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in medicine* 32: 4550--4566.
- Chicco, D. & G. Jurman, 2020 The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics* 21: 6.
- Cohen, L. T., F. W. Rickards, & G. M. Clark, 1991 A comparison of steady-state evoked potentials to modulated tones in awake and sleeping humans. *The Journal of the Acoustical Society of America* 90: 2467--2479.
- Cone-Wesson, B., R. C. Dowell, D. Tomlin, G. Rance, & W. J. Ming, 2002 The auditory steady-state response: comparisons with the auditory brainstem response. *Journal of the American Academy of Audiology* 13: 173--187.
- Crick, F. & C. Koch, 1990 Towards a neurobiological theory of consciousness. In *Seminars in the Neurosciences*, volume 2, pp. 263--275, Saunders Scientific Publications.
- Cybenko, G., 1989 Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* 2: 303--314.
- Demidenko, E., 2013 *Mixed models: theory and applications with R*. John Wiley & Sons.

- Ehlers, C. & D. Kupfer, 1997 Slow-wave sleep: do young adult men and women age differently? Journal of sleep research 6: 211--215.
- Fisher, R. A., 1936 The use of multiple measurements in taxonomic problems. Annals of eugenics 7: 179--188.
- Friedman, J., T. Hastie, & R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33: 1.
- Galambos, R., S. Makeig, & P. J. Talmachoff, 1981 A 40-hz auditory potential recorded from the human scalp. Proceedings of the national academy of sciences 78: 2643--2647.
- Han, H.-B., K. E. Lee, & J. H. Choi, 2019 Functional dissociation of θ oscillations in the frontal and visual cortices and their long-range network during sustained attention. eNeuro 6.
- Hanin, B. & M. Sellke, 2017 Approximating continuous functions by relu nets of minimal width. arXiv preprint arXiv:1710.11278 .
- John, M. & T. Picton, 2000 Master: a windows program for recording multiple auditory steady-state responses. Computer methods and programs in biomedicine 61: 125--150.
- Kidger, P. & T. Lyons, 2019 Universal approximation with deep narrow networks. arXiv preprint arXiv:1905.08539 .
- Kuwada, S., R. Batra, & V. L. Maher, 1986 Scalp potentials of normal and hearing-impaired subjects in response to sinusoidally amplitude-modulated tones. Hearing research 21: 179--192.
- Lawhern, V., W. D. Hairston, & K. Robbins, 2013 Detect: A matlab toolbox for event detection and identification in time series, with applications to artifact detection in eeg signals. PloS one 8.
- Leshno, M., V. Y. Lin, A. Pinkus, & S. Schocken, 1993 Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural networks 6: 861--867.
- Lin, Y. & Y. Jeon, 2006 Random forests and adaptive nearest neighbors. Journal of the American Statistical Association 101: 578--590.
- McCulloch, C. E. & J. M. Neuhaus, 2011 Prediction of random effects in linear and generalized linear models under model misspecification. Biometrics 67: 270--279.
- Miller, S. & D. Childers, 2012 *Probability and random processes: With applications to signal processing and communications*. Academic Press.
- Mognon, A., J. Jovicich, L. Bruzzone, & M. Buiatti, 2011 Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features. Psychophysiology 48: 229--240.
- Nolan, H., R. Whelan, & R. B. Reilly, 2010 Faster: fully automated statistical thresholding for eeg artifact rejection. Journal of neuroscience methods 192: 152--162.
- Picton, T. W., C. R. Skinner, S. C. Champagne, A. J. Kellett, & A. C. Maiste, 1987 Potentials evoked by the sinusoidal modulation of the amplitude or frequency of a tone. The Journal of the Acoustical Society of America 82: 165--178.

- Pinheiro, J. & D. Bates, 2006 *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.
- Rickards, F. & G. M. Clark, 1982 Steady state evoked potentials to amplitude modulated tones. Scientific publications, vol. 2, 1978-1983, no. 123 .
- Searle, S. R. & C. E. McCulloch, 2001 *Generalized, linear and mixed models*. Wiley.
- Shoker, L., S. Sanei, & J. Chambers, 2005 Artifact removal from electroencephalograms using a hybrid bss-svm algorithm. IEEE Signal Processing Letters 12: 721--724.
- Stoica, P., R. L. Moses, *et al.*, 2005 Spectral analysis of signals .
- Subha, D. P., P. K. Joseph, R. Acharya, & C. M. Lim, 2010 Eeg signal analysis: a survey. Journal of medical systems 34: 195--212.
- Tuerlinckx, F., F. Rijmen, G. Verbeke, & P. De Boeck, 2006 Statistical inference in generalized linear mixed models: A review. British Journal of Mathematical and Statistical Psychology 59: 225--255.
- Van der Vaart, A. W., 2000 *Asymptotic statistics*, volume 3. Cambridge university press.
- Vertes, R. P., 2005 Hippocampal theta rhythm: A tag for short-term memory. Hippocampus 15: 923--935.
- Zhang, D. & X. Lin, 2008 Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In *Random effect and latent variable model selection*, pp. 19--36, Springer.