Propuesta experimental, Luis Luarte

Preámbulo Mi objetivo es definir un modelo del algoritmo de toma de decisiones del cerebro F_{brain}, que mapea un historial de experiencias a una política conductual observable (la estrategia que alguien usa para determinar la siguiente acción basada en su estado actual) y un conjunto de dinámicas neurofisiológicas correspondientes. Propongo que F_{brain} puede ser aproximado por un modelo formal de aprendizaje por refuerzo M_{RL}, como el Q-learning (Sutton & Barto, 2020). Este modelo propone que la conducta es guiada por un conjunto de variables latentes no observables λ, que son actualizadas por computaciones específicas, principalmente por el error de predicción de recompensa (Schultz et al., 1997). Aunque la utilidad del modelo se confirma parcialmente si su política puede ajustarse a la conducta observada, una validación más rigurosa requiere demostrar una correspondencia directa entre sus computaciones latentes y la actividad neural.
Por lo tanto, esta propuesta experimental pone a prueba la hipótesis central de que existe una función de mapeo significativa, G, entre las variables latentes del modelo λ y características cuantificables ϕ extraídas de la forma de onda ERP $\phi(E(t))$. Específicamente, buscamos establecer que $\phi(E(t)) \approx G(\lambda_t)$, donde la amplitud de la Feedback-Related-Negativity en el momento post-decisión $\approx G1(\delta_t)$ y la Entropía-PCA del componente P3 en el momento pre-decisión $\approx G2(\tau_t)$. La existencia de un mapeo G robusto y específico apoyaría la inferencia abductiva de que M_{RL} no es meramente descriptivo de la conducta, sino una representación computacionalmente plausible de los procesos neurofisiológicos instanciados por F_{brain}.
Pregunta de investigación ¿Las características específicas de los ERP's rastrean sistemática y dinámicamente los parámetros latentes clave de un modelo estándar de Q-learning?. ¿La Feedback related negativity se correlaciona de manera confiable con el error de predicción de recompensa (RPE) derivado del modelo?. ¿La entropía derivada de PCA de la forma de onda ERP previa a la decisión se correlaciona con el parámetro de temperatura de incertidumbre en la decisión derivado del modelo?. Entonces, ¿puede un modelo computacional que incorpora estas características de ERP's como entradas trial a trial lograr una predicción superior de la conducta de elección en comparación con un modelo estándar que se basa únicamente en datos conductuales?.
Hipótesis (I) La amplitud de la feedback-related negativity (FRN) se correlacionará significativamente con el error de predicción de recompensa derivado del modelo, confirmando la FRN como un índice neural del RPE. (II) La incertidumbre en la decisión τ_t y la entropía del ERP previa a la decisión mostrarán una correlación positiva, tal que Entropía-PCA $\beta_1\tau_t + \beta_2 R_t$ resultará en $\beta_1 > 0$ después de considerar el tiempo de reacción. (III) Un modelo neuro-informado M_{RL} donde los parámetros α_t, τ_t son modulados por características de ERP's predecirá mejor la conducta de elección que un $M_{standard}$.
Paradigma Experimental El participante realizará una tarea de two-armed bandit (Sutton & Barto, 2020); en esta tarea el objetivo principal es maximizar las recompensas acumuladas obtenidas, eligiendo repetidamente entre dos opciones, cada una entregando recompensas de una magnitud variable extraída de una distribución normal (con un camino aleatorio con pasos de distribución normal a lo largo de los trials). En cada trial, a los participantes se les presenta primero una cruz de fijación durante 3±0.5 segundos, luego aparecen en la pantalla dos símbolos visualmente distintos durante 3±0.5 segundos (fase de pre-decisión), después se imprime en la pantalla una instrucción para hacer una selección mediante una pulsación de tecla, y finalmente se presenta la recompensa obtenida durante 3±0.5 segundos (fase de post-decisión).
Análisis Usando datos de la fase de pre-decisión, computaré la actividad frontal y central con un rango de paso de banda de 0.1 – 30 Hz, desde 300-600 ms después del inicio de la fase, ya que esto corresponde al P3 ERP que ha sido relacionado con la modulación de la evaluación de la elección (Cui et al., 2013). Sin embargo, como el modelo computacional utiliza un parámetro relacionado con la estocasticidad para las etapas de evaluación de la elección, realizaremos un PCA trial a trial, y calcularemos la entropía de Shannon de los valores propios, para determinar cuán bien definida o estocástica es la actividad relacionada con P3 (Entropía-P3-ERP). Para la fase de post-decisión, utilizaré el potencial FRN (Miltner et al., 1997) ya que

evalúa directamente los errores de estimación de recompensa. Como el parámetro relacionado con el modelo es direccional, usaré la amplitud del ERP como la variable principal, y para obtener estimaciones trial a trial, usaré filtrado espaciotemporal y el criterio de máxima correntropía usando información de múltiples canales para generar estimaciones resistentes al ruido (Li et al., 2007). Posteriormente, se utilizarán modelos lineales mixtos para dar cuenta de la estructura jerárquica de los ensayos dentro de los sujetos, con los parámetros de M_{RL} como variables dependientes y las características derivadas de ERP's como variables independientes, τ $Entropía - P3ERP$ y RPE $amplituddeFRN$, respectivamente. Para la comparación de modelos, se ajustarán dos modelos de Q-learning, el modelo nulo utilizará solo datos conductuales (elecciones y recompensas), mientras que el modelo neuro-informado utilizará la Entropía-P3-ERP y la amplitud de FRN dentro del procedimiento de ajuste del modelo.

Suplementario

Procedimiento de ajuste del modelo Q-learning estándar ($M_{standard}$)

El modelo Q-learning estándar se ajustará a los datos conductuales de cada participante para encontrar el conjunto de parámetros que maximice la verosimilitud de sus elecciones observadas. El modelo asume una tasa de aprendizaje α estática y una temperatura de decisión τ estática. En cada ensayo t, después de elegir la acción c_t y recibir la recompensa R_t, la función de valor de acción $Q_t(c_t)$ se actualizara según el RPE, δ_t:

$$\delta_t = R_t - Q_t(c_t)$$
$$Q_{t+1}(c_t) = Q_t(c_t) + \alpha \cdot \delta_t$$

La probabilidad de seleccionar una acción se determinara mediante la función softmax:

$$P(c_t) = \frac{e^{Q_t(c_t)/\tau}}{\sum_{j=1}^{2} e^{Q_t(a_j)/\tau}}$$

Los parámetros libres (,) se estimarán para cada sujeto minimizando la log-verosimilitud negativa de los datos de elección usando optimización no lineal acotada.

Procedimiento de ajuste del modelo Q-learning neuro-informado (M_{RL})

Para este modelo, la tasa de aprendizaje se modelará como una función logístico-sigmoidal de la amplitud estandarizada de FRN del ensayo anterior (FRN_{t-1}):

$$\alpha_t = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \mathrm{FRN}_{t-1})}}$$

La temperatura de decisión se modelará como una función exponencial de la Entropía-P3-ERP estandarizada previa a la decisión:

$$\tau_t = e^{(\gamma_0 + \gamma_1 \cdot \mathrm{Entropy}_t)}$$

Los parámetros libres para este modelo serán los pesos de las funciones de mapeo, $(\beta_0, \beta_1, \gamma_0, \gamma_1)$. Estos se estimarán utilizando el mismo procedimiento de máxima verosimilitud que Mstandard. La comparación de modelos entre $M_{standard}$ y M_{RL} se realizará utilizando el Criterio de Información de Akaike (AIC) para tener en cuenta la diferencia en el número de parámetros libres.

Referencias

[1] Cui, J., Chen, Y., Wang, Y., Shum, D. H. K., & Chan, R. C. K. (2013). Neural correlates of uncertain decision making: ERP evidence from the Iowa Gambling Task. *Frontiers in Human Neuroscience, 7*. https://doi.org/10.3389/fnhum.2013.00776

[2] Li, R., Principe, J. C., Bradley, M., & Ferrari, V. (2007). Robust single-trial ERP estimation based on spatiotemporal filtering. *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 5206-5209*. https://doi.org/10.1109/IEMBS.2007.4353515

[3] Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-Related Brain Potentials Following Incorrect Feedback in a Time-Estimation Task: Evidence for a "Generic" Neural System for Error Detection. *Journal of Cognitive Neuroscience, 9*(6), 788-798. https://doi.org/10.1162/jocn.1997.9.6.788

[4] Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.), 275*(5306), 1593-1599. https://doi.org/10.1126/science.275.5306.1593

[5] Sutton, R. S., & Barto, A. (2020). *Reinforcement learning: An introduction* (Second edition). The MIT Press.