Hierarchical Bayesian estimation of motor evoked potential recruitment curves yields accurate and robust estimates

Vishweshwar Tyagi[1,2*], Lynda M. Murray[8,9], Ahmet S. Asan[5], Christopher Mandigo[3,6], Michael S. Virk[4], Noam Y. Harel[7,8,9], Jason B. Carmel[1,2,4], James R. McIntosh[1,2,4*]

[1*]Neurology, Columbia University, New York, 10032, NY.
[2]Orthopedic Surgery, Columbia University, New York, 10032, NY.
[3]Neurological Surgery, Columbia University, New York, 10032, NY.
[4]Neurological Surgery, Weill Cornell Medicine, New York, 10065, NY.
[5]Staley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, 02142, MA.
[6]New York Presbyterian, The Och Spine Hospital, New York, 10034, NY.
[7]Neurology, Icahn School of Medicine at Mount Sinai, New York, 10029, NY.
[7]Rehabilitation and Human Performance, Icahn School of Medicine at Mount Sinai, New York, 10029, NY.
[8]James J. Peters Veterans Affairs Medical Center, Bronx, 10468, NY.

*Corresponding author(s). E-mail(s): vt2353@columbia.edu; jrm2263@cumc.columbia.edu;

Abstract

Electrical and electromagnetic stimulation probe and modulate the neural systems that control movement. Key to understanding their effects is the muscle recruitment curve, which maps evoked potential size against stimulation intensity. Current methods to estimate curve parameters require large samples; however, collecting these is often impractical due to experimental constraints. Here, we present a hierarchical Bayesian framework that accounts for small samples, handles outliers, simulates high-fidelity data, and returns posterior distributions over curve parameters that quantify estimation uncertainty. It uses a rectified-logistic function that estimates motor threshold and outranks sigmoidal alternatives in

predictive performance, as demonstrated through cross-validation. In simulations, our method outperforms conventional non-hierarchical models by reducing threshold estimation error on sparse data, and requires fewer participants to detect shifts in threshold compared to frequentist testing. We present two common use cases involving electrical and electromagnetic stimulation data and provide a library for Python, called hbMEP, for diverse applications.

Keywords: hierarchical Bayesian, motor threshold, transcranial magnetic stimulation, spinal cord stimulation, hypothesis testing, mathematical modeling, evoked potential

1 Introduction

Fig. 1 Hierarchical Bayesian estimation of recruitment curves yields parametric estimates for each participant across multiple muscles simultaneously. (a) Example motor evoked potentials (MEPs) recorded at different stimulation intensities from two muscles, abductor pollicis brevis (APB) on left and abductor digiti minimi (ADM) on right. Abscissa represents stimulation intensity which may be specified in different units e.g. μA or mA for Spinal Cord Stimulation, or % maximum stimulator output (MSO) for Transcranial Magnetic Stimulation. **Right panel:** Quantification of MEPs into MEP size using either peak-to-peak amplitude or area under curve. **(b)** Example recruitment curves estimated as 4-parameter sigmoid using mean squared error minimization with the Nelder-Mead method. It provides only point estimates for curve parameters, lacks threshold estimates, and fails to capture sharp deflection from offset. Bottom panels: point estimate of S50. **Top right panel:** point estimate of the saturation parameter. **Bottom right panel:** point estimate of the maximum gradient. **(c)** Example recruitment curves estimated as 5-parameter rectified-logistic function within a hierarchical Bayesian framework. Shading represents the 95% highest density interval (HDI) of posterior predictive distribution. It accurately captures the threshold and S50. Additionally, it returns posterior distributions over curve parameters, quantifying estimation uncertainty using the width of the 95% HDI. Data from multiple participants and muscles is handled simultaneously. Bottom panels: posterior distribution of threshold and S50. Inset: zoom over posterior distribution. **Top right panel:** posterior distribution of the saturation parameter. **Bottom right panel:** posterior distribution of the maximum gradient.

Fig. 2 Rectified-logistic recruitment curve outranks traditionally used alternatives based on Bayesian cross-validation while having the unique advantage of estimating threshold and S_{50}. (a) Example recruitment curve (RC) fitted to rat epidural Spinal Cord Stimulation (SCS) biceps data (gray dots) using 3-parameter rectified-linear function. Black line shows the RC. Gray shading represents the 95% highest density interval (HDI) of the posterior predictive distribution. It undershoots data at low intensities due to curvature in data, and subsequently overshoots, failing to capture saturation which results in wide 95% HDI. It estimates threshold but not S_{50}, since it doesn't saturate. **(b)** 4-parameter sigmoid (logistic-4) captures the saturation but fails to capture the sharp deflection from offset due to its symmetry about the inflection point. It estimates S_{50} and lacks threshold estimate. **(c)** 5-parameter sigmoid (logistic-5) function improves upon this by better capturing the deflection. It estimates S_{50} and lacks threshold estimate. **(d)** 5-parameter rectified-logistic function is flexible enough to accurately capture the deflection, curvature, and saturation, resulting in narrower 95% HDI. It estimates both threshold and S_{50}. **(e)** Predictive performance measured with expected log-pointwise predictive density (ELPD) leave-one-out cross-validation on rat epidural SCS dataset, consisting of 3 muscles, 450 RCs, and 23,028 motor evoked potentials (MEPs). Black circles represent mean ELPD score, black bars represent standard error of ELPD mean, gray triangles represent mean of pairwise ELPD differences from the best-ranked rectified-logistic model, and gray bars represent standard error of pairwise differences. **(f)** As for (e), but on human Transcranial Magnetic Stimulation dataset, consisting of 6 muscles, 114 RCs, and 8,490 MEPs. **(g)** As for (e), but on human epidural SCS dataset, consisting of 3 muscles, 78 RCs, and 3,453 MEPs. The predictive performance of rectified-logistic function is significantly better than logistic-5 on the largest tested rat SCS dataset, and it is comparable to logistic-5 on human TMS and SCS datasets.

2 Results

2.1 Choice of recruitment curve function

2.2 Robustness & efficiency

2.2.1 Simulated data and posterior predictive checks

2.2.2 Accuracy on sparse data

2.2.3 Statistical power

2.3 Common use cases

2.3.1 Within-participants comparison

2.3.2 Between-groups comparison

2.3.3 Optimizing experimental design

3 Discussion

4 Methods

4.1 Modeling MEP size

The various choices for modeling recruitment curves include 3-parameter rectified-linear **????** (Eq. 4.1.1, Fig. 2a), 4-parameter logistic-4 **??** (Eq. 4.1.2, Fig. 2b). Additionally, 5-parameter logistic-5 **?** (Eq. 4.1.3, Fig. 2c) is a more generalized version of logistic-4 and contains an extra parameter v to control near which asymptote (lower L or upper $L + H$) the maximum growth or point of inflection occurs. In contrast to

Fig. 3 (a–d) Generative capability of the hierarchical Bayesian framework enables simulation of high fidelity synthetic data. (a) Example participant from Transcranial Magnetic Stimulation (TMS) data that is used by hierarchical Bayesian model to estimate the participant-level and population-level parameters. **(b)** Data simulated from the model conditioned on estimated participant-level parameters. The model can replicate the observed participants. **(c)** Data simulated from the model conditioned on estimated population-level parameters. The model can simulate new participants for subsequent model comparison. **(d)** A principal component analysis plot shows a large overlap between the new simulated parameters (blue dots) and those estimated from observed TMS data (orange dots). A large overlap signifies that the generated synthetic data closely matches real TMS data. Green dots represent parameters simulated from the prior predictive distribution, which is weakly informative as indicated by its large spread. **(e–f) Hierarchical Bayesian structure improves threshold estimation accuracy over non-Bayesian and non-hierarchical models on simulated data. (e)** For a single participant, both non-hierarchical Bayesian (nHB) and hierarchical Bayesian (HB) models produce the same mean absolute error (MAE), which is lower than traditionally used maximum likelihood (ML) and Nelder-Mead (NM) methods. With an increase in the number of participants, the HB model uses partial pooling across participants to further reduce error. This simulation consisted of 48 equispaced stimulation intensities between 0–100% maximum stimulator output (MSO), and was repeated 2000 times. Error bars represent standard error of the mean. **(f)** For eight simulated participants, the HB model outperforms other non-hierarchical methods at all tested number of stimulation intensities. Advantage of the HB model is largest for low number of intensities, i.e., on sparse data. This simulation consisted of 8 participants and was repeated 2000 times. **(g–h) Hierarchical Bayesian estimation is more powerful and produces fewer false positives when detecting shifts in threshold on simulated data. (g)** Comparison of statistical power of Bayesian estimation versus non-hierarchical models to detect a negative shift in threshold parameter from pre- to post-intervention. Bayesian estimation is more powerful and requires approximately 35% fewer participants to achieve 80% statistical power compared to nHB. The threshold differences were simulated from a normal distribution Normal $(-3, 1.5)$, where the null hypothesis (no threshold shift) is false. Bayesian estimation rejects the null if the 95% highest density interval is entirely left of zero. This was compared with a one-sided Wilcoxon signed-rank test on point estimates of pairwise threshold differences estimated using non-hierarchical models (a t-test was not applicable due to non-normality of estimated differences indicated by Shapiro-Wilk test). The significance level was set at 5% for the signed rank test. This simulation was repeated 2000 times. **(h)** Comparison of false positive rates of Bayesian estimation against the set significance level of 5%. Bayesian estimation is more conservative in falsely rejecting the null hypothesis and maintains a false positive rate between 0 and 2.5%. The differences were simulated from Normal $(0, 1.5)$, where the null hypothesis holds true. This simulation was repeated 2000 times.

Fig. 4 Within-participant comparison, midline versus lateral stimulation example on human epidural Spinal Cord Stimulation data. (a) Example participant with lateral (light) and midline (dark) stimulation. Inset: zoom to show presence of threshold, despite small MEP size. **Bottom panels:** Posterior distribution of threshold. Inset: zoom over posterior distribution. **(b)** Posterior distribution of difference between midline and lateral thresholds summarized across 13 participants who underwent intraoperative surgery. A priori, the model assumes no difference between midline and lateral thresholds, indicated by a flat prior (gray distribution) centered at zero. The 95% highest density interval (HDI) of the posterior suggests strong evidence in favor of lateral stimulation being more effective.

logistic-4, the logistic-5 is not necessarily symmetrical about its inflection point.

$$\text{rectified-linear} \qquad \forall a, b, L > 0 \quad x \mapsto L + \max\{0, b\,(x - a)\} \tag{4.1.1}$$

$$\text{logistic-4} \qquad \forall a, b, L, H > 0 \quad x \mapsto L + \frac{H}{1 + e^{-b(x-a)}} \tag{4.1.2}$$

Fig. 5 Between-groups comparison, SCI versus uninjured participants example on Human Transcranial Magnetic Stimulation (TMS) data. (a) Example uninjured participant. Inset: zoom to show presence of threshold, despite small MEP size. Bottom panels: Posterior distribution of threshold. **(b)** participant with spinal cord injury (SCI). **(c)** Posterior distribution of threshold difference summarized between 12 uninjured and 7 SCI participants. A priori, the model assumes no difference, indicated by a flat prior (gray distribution) centered at zero. Although not significant, the mass of posterior distribution indicates with high probability that spinal cord injury is associated with higher thresholds.

$$\text{logistic-5} \qquad \forall a, b, v, L, H > 0 \quad x \mapsto L + \frac{H}{\left\{1 + (2^v - 1)\, e^{-b(x-a)}\right\}^{1/v}} \tag{4.1.3}$$

Note that in Eq. 4.1.1, parameter a models the threshold, and in Eq. 4.1.2, 4.1.3 it models the S_{50}. L represents the offset MEP size, $(L + H)$ defines the saturation, and b is the growth rate. The logistic functions do not have a direct representation of threshold since they are smooth, and the rectified-linear function does not have a direct representation of S_{50} since it doesn't saturate.

4.1.1 Observation model

We introduce a rectified-logistic function (Eq. 4.1.4) that can estimate both the threshold and S_{50}. Fig. 6a–f shows the effect of varying its parameters. Parameters L, H, b have similar interpretation as in the logistic-4 function, and a models the threshold. Eq. 4.1.5 gives the S_{50} of rectified-logistic function. Similar to logistic-5, there is an additional parameter ℓ (Fig. 6e) that controls the location of inflection point, or the point of maximum gradient, whether near the offset L or saturation $(L + H)$.

For $a, b, L, \ell, H > 0$, define $\mathcal{F} : \mathbb{R} \to \mathbb{R}^+$ as

$$\mathcal{F}(x) = L + \max\left\{0, -\ell + \frac{H + \ell}{1 + \left(\frac{H}{\ell}\right) e^{-b(x-a)}}\right\} \tag{4.1.4}$$

$$S_{50}(\mathcal{F}) = a - \frac{1}{b} \ln\left(\frac{H\ell}{H(H + 2\ell)}\right) \tag{4.1.5}$$

We use a gamma (in shape-rate parametrization) observation model (Eq. 4.1.6–4.1.8) to model the relationship between MEP size (y) and stimulation intensity (x). Note that in Eq. 4.1.7, we model the expected MEP size as a rectified-logistic function of intensity, since $\mathbb{E}(y \mid x, \Omega, c_1, c_2) = \mu = \mathcal{F}(x \mid \Omega)$.

For $c_1, c_2 > 0$

$$y \mid x, \Omega, c_1, c_2 \sim \text{Gamma}(\mu \cdot \beta, \beta) \tag{4.1.6}$$

$$\mu = \mathcal{F}(x \mid \Omega) \qquad \Omega = \{a, b, L, \ell, H\} \tag{4.1.7}$$

Fig. 6 (a–f) Effect of varying parameters of rectified-logistic function. (a) Varying a shifts the threshold. **(b)** b changes the growth rate. L changes the offset MEP size. **(c)** H changes the distance between offset and saturation $(L + H)$ **(d)** Varying H changes the saturation. **(e)** ℓ affects where inflection point or maximum gradient occurs, near offset L, or saturation $(L + H)$. **(f)** Varying b, ℓ simultaneously. **(g–h) Hierarchical Bayesian model structure. (g)** Graphical model of hierarchical Bayesian model used to estimate population-level parameters of Transcranial Magnetic Stimulation (TMS) data. The model yields parameter estimates for each participant across multiple muscles simultaneously. Circular nodes represent random variables. Filled circular nodes represent observed data. Diamonds represent deterministic variables. Arrows represent that the child node is informed by the distribution of the parent node. Plates denote the re-instantiation of nodes. **(h)** Bayesian model specification with participant- and population-level parameters, and weakly-informative hyperpriors. Weakly informative hyperpriors help combine data from multiple participants in a principled way making the model less vulnerable to overfitting.

$$\beta = \frac{1}{c_1} + \frac{1}{c_2 \cdot \mu} \tag{4.1.8}$$

We chose a gamma distribution to capture the long tailed heteroskedasticity noise spread (Goetz and before **??**) around the recruitment curve. In Eq. 4.1.8 we specify the rate parameter (β) as a linear function of the reciprocal of expected MEP size $\left(\frac{1}{\mu}\right)$ with positive weights $\left(\frac{1}{c_1}, \frac{1}{c_2}\right)$, which allows capturing the increase in noise spread as MEP size increases. The Bayesian framework enables inferring posterior distributions over curve parameters. Estimation uncertainty can be quantified using the width of 95% highest density interval (HDI) of posterior distributions.

4.1.2 Recruitment Curves

More generally, in Eq. 4.1.7, $\mathcal{F}$ is called the activation function, or, the recruitment curve in the context of modeling MEP size, which transforms the input stimulation intensity (x), and links it to the expected MEP size $\mathbb{E}(y \mid x)$. $\mathcal{F}$ can be replaced by other available choices, including rectified-linear, logistic-4, or logistic-5.

4.2 Hierarchical Bayesian Model

4.2.1 Default model

The simplest form of a standard 3-stage hierarchical Bayesian model (Eq. 4.2.1–4.2.3) in the context of modeling MEP size can be described as follows. Let there be $N_M \times N_P$ exchangeable sequences $\left\{(x_i{}^p, y_i{}^{p,m})_{i=1}^{n(p)} \mid p = 1 \dots N_P, m = 1 \dots N_M\right\}$ of MEP sizes $y_i{}^{p,m} \in \mathbb{R}^+$ recorded at stimulation intensity $x_i{}^p \in \mathbb{R}^+ \cup \{0\}$ from muscle m of participant p, for a total of N_M muscles and N_P participants. Here $n(p)$ denotes the number of intensities tested for participant p and $y_i{}^{p,m}$ is the MEP size recorded simultaneously from muscles $m = 1, \dots N_M$ at a given intensity $x_i{}^p$ for participant p.

The first stage of hierarchy is the participant-level (Eq. 4.2.1). It specifies the parametric models $P(y_i{}^{p,m} \mid x_i{}^p, \theta^{p,m})$ for each of the N_M muscles of N_P participants, and models the MEP size $y_i{}^{p,m}$ as a function of intensity $x_i{}^p$ and participant-level parameters $\theta^{p,m}$. In the second stage (Eq. 4.2.2), the participant-level parameters $\theta^{p,m}$

are assumed to be generated from a common distribution $P\left(\theta^{p,m} \mid \gamma\right)$ with population-level hyper-parameters γ. In the third stage (Eq. 4.2.3), the population-level hyper-parameters γ are assumed to be unknown and assigned a weakly-informative prior $P\left(\gamma\right)$, also called hyperprior.

$$\text{Stage I} \qquad {y_i}^{p,m} \sim P\left({y_i}^{p,m} \mid {x_i}^{p}, \theta^{p,m}\right) \tag{4.2.1}$$

$$\text{Stage II} \qquad \theta^{p,m} \sim P\left(\theta^{p,m} \mid \gamma\right) \tag{4.2.2}$$

$$\text{Stage III} \qquad \gamma \sim P\left(\gamma\right) \tag{4.2.3}$$

Figure 6g,h specifies the default model used on human TMS data in Results 2.1, 2.2.1.

In Results 2.1 (Fig. 2e–g), we used the default model structure (Fig. 6g) for cross-validation **?**. The participant-level parameters of the rectified-logistic function were replaced with those of logistic-5, logistic-4, and rectified-linear. Supplementary Fig. **??**–**??** specify the models for each function on rat epidural SCS, human TMS and human epidural SCS datasets. For the rat epidural SCS data, 150 recruitment curves (RCs) were simultaneously fit on three muscles: biceps, extensor carpi radialis longus (ECR), and triceps–for a total of 450 RCs. For the human TMS data, [****update****] 19 RCs were simultaneously fit on six muscles: abductor digiti minimi (ADM), abductor pollicis brevis (APB), biceps, extensor carpi radialis (ECR), flexor carpi radialis (FCR), and triceps—for a total of 114 RCs. For the human SCS data, 26 RCs were simultaneously fit on three muscles: ADM, APB, and triceps—for a total of 78 RCs. An arviz **?** implementation of cross-validation **?** was used to compute expected log-pointwise predictive density (ELPD) scores and pairwise differences between models.

In Results 2.2.1 (Fig. 3a–d), we used the default model (Fig. 6g,h) to estimate participant-level and population-level parameters from TMS data in Results 2.2.1. We used data from the target APB muscle, which consisted of 19 participants. The model was conditioned on estimated participant-level parameters to replicate the observed participants (Fig. 3b). It was conditioned on estimated population-level parameters to simulate new participants (Fig. 3c). A principal component analysis (PCA) plot was used to visualize the participant-level parameters on a 2D plane. The PCA map was fit on parameters estimated from TMS participants (Fig. 3d, orange dots). The map was used to transform parameters of new simulated participants (Fig. 3d, blue dots), and participant-level parameters simulated from prior predictive (Fig. 3d, green dots).

4.2.2 Within-participants comparison

This section presents a hierarchical model that is useful for modeling shifts in curve parameters. This is applicable in settings where the same set of participants are tested for multiple experimental conditions, such as pre- and post-intervention phases, stimulation location (e.g. midline or lateral), or stimulation parameters (e.g. electrode position, stimulation frequency).

Supplementary Fig. **??** gives the graphical representation of such a model used to summarize differences in the threshold parameter. Here, we have the threshold $a^{p,k,m}$

of participant p, muscle m, at tested condition k given by,

$$a^{p,k,m} = \begin{cases} {a_{\text{fixed}}}^{p,m} & k = 1 \\ {a_{\text{fixed}}}^{p,m} + {a_\Delta}^{p,k,m} & k > 1 \end{cases} \tag{4.2.4}$$

The threshold is broken (Eq. 4.2.4) into a fixed component ($k = 1$) and a shift component ($\forall k > 1$) that measures the difference from the fixed component. This shift component is parametrized by condition-muscle-level location $\left({\mu_{a_\Delta}}^{k,m}\right)$ and population-level scale (σ_{a_Δ}) hyperparameters. The location parameter summarizes the shift within participants of each additional tested condition ($\forall k > 1$) from the fixed component ($k = 1$) for each muscle. The scale parameter measures the overall variability in the estimated shifts.

Additionally, this location hyperparameter $\left({\mu_{a_\Delta}}^{k,m}\right)$ is partially pooled across conditions and muscles and given location $\left(\mu_{\mu_{a_\Delta}}\right)$ & scale $\left(\sigma_{\mu_{a_\Delta}}\right)$ hyperparameters. This partial pooling is done to account for multiple comparisons **??** across tested conditions and muscles.

A priori we assume there is no shift from the fixed component and $\mu_{\mu_{a_\Delta}}$ is given a flat prior which is symmetric about zero. Once the model is fit, the 95% highest density interval (HDI) of ${\mu_{a_\Delta}}^{k,m}$ posterior is used to assess the strength of shift from the fixed component at every muscle. Additionally, the 95% HDI of $\mu_{\mu_{a_\Delta}}$ posterior is used to assess the overall shift across all muscles and conditions.

4.2.3 Between-groups comparison

Supplementary Fig. **??** gives a hierarchical model used to compare threshold between different groups of participants. The parameter of interest, threshold in this case, is parameterized by group-muscle-level location $({\mu_a}^{g,m})$ and population-level scale (σ_a) hyperparameters. The location parameter summarizes the thresholds of all participants belonging to a group. The scale parameter summarizes the overall variability.

Additionally, this location hyperparameter $({\mu_a}^{g,m})$ is partially pooled across groups and muscles and given location (μ_{μ_a}) & scale (σ_{μ_a}) hyperparameters. Once the model is fit, the 95% HDI of ${\mu_a}^{g,m}$ posterior is used to compare the parameter between groups.

4.2.4 Extension to mixture model

The models discussed so far can be extended to handle outliers by replacing the gamma distribution (Eq. 4.1.6) with a 2-component mixture of the gamma and a half-normal distribution. The resultant observation model is given as,

$$y \mid x \sim (1 - q_y) \cdot \text{Gamma}\left(\mu \cdot \beta, \beta\right) + q_y \cdot \text{HalfNormal}\left(\sigma_{\text{outlier}}\right) \tag{4.2.5}$$

$$q_y \sim \text{Bernoulli}\left(p_{\text{outlier}}\right) \tag{4.2.6}$$

$$p_{\text{outlier}} \sim \text{Uniform}\left(0, C_{p_{\text{outlier}}}\right) \tag{4.2.7}$$

$$\sigma_{\text{outlier}} \sim \text{HalfNormal}\left(C_{\sigma_{\text{outlier}}}\right) \tag{4.2.8}$$

where $C_{p_{\text{outlier}}}, C_{\sigma_{\text{outlier}}} > 0$ are constants and $C_{p_{\text{outlier}}} < 1$ is chosen to be small, usually in the range $0.01 - 0.05$. Intuitively, this means that we expect roughly $1\% - 5\%$ of outliers to be captured by the half-normal distribution. Supplementary Fig. **??** shows an extension of the default model (Methods 4.2.1) to the mixture model.

In Supplementary Fig. **??**, the default model (Methods 4.2.1) with rectified-logistic function and Gamma observation was cross-validated against its mixture extension. Supplementary Fig. **??**–**??** specify the mixture model for the three datasets that were used in Results 2.1.

5 Additional information

6 Supplementary information