Reference-free Bayesian model for pointing errors of typein neurosurgical planning

John S H Baxter, Stéphane Croci, Antoine Delmas, Luc Bredoux, Jean-Pascal Lefaucheur, Pierre Jannin

Reference-free Bayesian model for pointing errors-of-type in neurosurgical planning

John S.H. Baxter [0000-0003-3548-4343]1*, Stéphane Croci[2], Antoine Delmas[2], Luc Bredoux[2], Jean-Pascal Lefaucheur [0000-0002-0294-8239]3,4 and Pierre Jannin [0000-0002-7415-071X]1

[1]Laboratoire Traitement du Signal et de l'Image (LTSI - INSERM UMR 1099), Université de Rennes 1, Rennes, France.
[2]SYNEIKA, Rennes, France.
[3]ENT Team, EA4391, Faculty of Medicine,, Paris Est Créteil University, Créteil, France.
[4]Clinical Neurophysiology Unit, Department of Physiology, Henri Mondor Hospital, Hôpitaux de Paris, Créteil, France.

*Corresponding author(s). E-mail(s): jbaxter@univ-rennes1.fr;

Abstract

Purpose: Many neurosurgical planning tasks rely on identifying points of interest in volumetric images. Often, these points require significant expertise to identify correctly as, in some cases, they are not visible but instead inferred by the clinician. This leads to a high degree of variability between annotators selecting these points. In particular, *errors-of-type* are when the experts fundamentally select different points rather than the same point with some inaccuracy. This complicates research as their mean may not reflect any of the experts' intentions nor the ground truth.
Methods: We present a regularised Bayesian model for measuring errors-of-type in pointing tasks. This model is reference-free, in that it doesn't require a priori knowledge of the ground truth point but instead works on the basis of the level of consensus between multiple annotators. We apply this model to simulated data and clinical data from transcranial magnetic stimulation for chronic pain.

Acknowledgements: The authors would like to thank J.-P. N'Guyen and H. Hodaj for their assistance in annotating the chronic pain treatment points along with J.-P. Lefaucheur.

Results: Our model estimates the probabilities of selecting the correct point in the range of 82.6-88.6% with uncertainties in the range of 2.8-4.0%. This agrees with the literature where ground truth points are known. The uncertainty has not previously been explored in the literature, and gives an indication of the dataset's strength.
Conclusions: Our reference-free Bayesian framework easily models errors-of-type in pointing tasks. It allows for clinical studies to be performed with a limited number of annotators where the ground truth is not immediately known, which can be applied widely for better understanding human errors in neurosurgical planning.

Keywords: Error modelling, Surgical planning, Pointing, Localisation, Transcranial magnetic stimulation, Bayesian statistics

1 Introduction

A critical task common in many neurosurgical planning workflows is *pointing* or *localisation* in which some introduced fiducial, anatomical landmark, or target location within a medical image is represented as a single point which is selected by the clinician performing the planning. In the first case, these locations are specifically chosen to be highly distinguishable in the image, reducing the uncertainty for the clinician in selecting them to that introduced by the physical size of the fiducial, the point-spread function of the imaging modality, and any physical error introduced by the pointing system itself, leading to well-behaved and well-understood models for representing pointing errors and their downstream effects in these cases [1, 2]. These models almost uniformly consider the selected point to come from a uni-modal and largely symmetric distribution with the correct location as its mode and expected value, such as a Gaussian distribution [1, 2]. These assumptions are widespread and violations of them are used to detect outliers and erroneous points [3]. They have even been directly experimentally validated in the context of highly visible, unambiguous fiducials [4].

However, anatomical landmarks and target locations are not explicitly engineered, meaning that the clinicians must rely on endogenous contrast and consistent geometry to identify these points which may not be highly visible nor unambiguous. For example, the anterior and posterior commissures (due to their consistent geometry and hyper-intensity compared to the adjacent ventricles) can be used as the basis for describing a consistent co-ordinate system in the deep brain, and selecting these two points has been a frequent task for stereotaxic neurosurgical planning for over 20 years [5]. Despite lacking the aforementioned simplifying assumptions of visibility and unambiguity, simple mean-centred, uni-modal distributions are still used to express the error not only of the human clinician but also automatic pointing algorithms [6].

These distributions largely reflect what is commonly known as an *error-of-degree*. This means that the error associated with a particular annotator for a

particular point is a question of how finely they can visualise and physically target a singular point, leading to a real-valued error, i.e. the distance from their selected point to that point's true location. This is in contrast to an *error-of-type* in which the error is largely a question of selecting between multiple distinct potential points rather than a singular one. In psychophysics, these errors in pointing tasks are often investigated through **multiple alternatives forced choice detection** experiments in which a finite series of items are introduced into a complex or noisy image and a human participant is asked to click on a particular one based on some feature [7–9]. In these experiments, the *distance* to the point selected by the user is secondary to its *choice*.

In the context of neurosurgical planning, these tasks can arise when the anatomy to be pointed at is not immediately distinguishable from other proximal similar anatomies. This is especially the case in MRI-guided transcranial magnetic stimulation (TMS) in which a particular gyrus representing a functionally (rather than structurally) defined point is selected as a treatment location [10, 11]. Because of the high degree of variability in cortical gyrification, performing this task is complex and subject to a high degree of error and variability amongst experts even for well-described point targets [12].

The result of this different type of error is that many of the assumptions that previous error-of-degee or error-of-type models have been based upon no longer apply. Unlike in error-of-degree models, using a single annotator will not always result in an admissible estimate of the true location of a particular point as that annotator's point will have compounded uncertainties in terms of their choice of point but also all the continuous visualisation and physical pointing uncertainties, compounded errors-of-type with errors-of-degree. This means that if we had an arbitrarily large number of annotators, the annotations on a single image would likely look more like a series of clusters, one around the ground truth and one for each distractor. This is problematic because it means that the average of the points selected (regardless of the number of annotators) is not a good estimator of the ground truth location, or in fact any of the distractors, but instead would fall somewhere in the space between these hypothetical clusters. This can have large downstream effects on research into pointing tasks in neurosurgical planning, notably machine-learning based automated pointing that relies on the expected value of expert annotations to be correct for purposes of training and evaluating algorithms [13, 14]. This problem is further exacerbated by the small number of expert annotations (often only one or two) for each dataset, thus not even allowing for the clusters to be inferred.

Unlike multiple alternative forced choice models, the number of distractors is unknown as they are not introduced into the image or are so numerous (i.e. considering every bend in every gyrus as a possible distractor) that these models are no longer feasible [8]. Thus, rigorous mathematical models describing these errors-of-type with an unknown number of distractors are necessary for further understanding pointing errors in neurosurgical planning.

Contributions

This paper proposes a framework for modelling errors-of-type in surgical pointing tasks. This framework makes use of our previously published base model [15] as a starting point, interpreting its series of cases as categories for a Dirichlet distribution, thus allowing for the use of Bayes' theorem to derive distributions for model parameters, rather than only best-fit values. In order to solve the problem of these distributions being improper, we introduce and validate an additional regularisation parameter. As the base model is *reference-free*, the framework as a whole does not require the true point location nor even precise knowledge of what points have been selected, which makes it fundamentally different from multiple-alternatives forced-choice models. Instead, it relies on the notion of *agreement* between annotators, using simple information about the number of annotators who appear to be selecting the same point. Our proposed model has also been designed to generalise to any number of annotators, although it is still best suited for cases in which only a few annotators are available. To the best of our knowledge, this is the first reference-free Bayesian model applied to the problem as surgical errors-of-type.

2 Theory

2.1 Reference-free Bayesian model

Our base model comes from our previous work [15]. This model includes a small number of parameters, specifically:

- p, the probability of the annotator choosing the true point; and
- n, the number of *distractors*, or false points which could be chosen by the annotator.

However, as we discovered, this model is *improper* in that it allows for cases such as $n \to \infty$ to take on non-zero values. This would prevent its use in Bayesian modelling as it would prevent the posterior from being *proper*. In order to address this, we have included a third parameter:

- $z > 1$, a regularisation parameter to encourage simpler models with lower values of n, specifically $P(n|z) \propto z^{-n}$.

Given these, the probability of a given annotator selecting point q is:

$$P(q|p,n,z) = \begin{cases} p & \text{if } q \text{ is the true point} \\ \frac{1-p}{n} & \text{else} \end{cases} \tag{1}$$

and for a sequence of k annotators independently picking the sequence of points $[q]$, this formula becomes:

$$P([q]|p,n,z) = p^{\#T([q])} \left(\frac{1-p}{n}\right)^{k-\#T([q])} \tag{2}$$

where $\#T([q])$ is the number of times that the true point appears in the list $[q]$. In theory, we would know where these points are in the image (as well as their number) or, if we had a large number of annotations, this information could be extracted through clustering the annotations. However in practice, we are limited in that we don't know which of the points is correct vs a distractor or even the number of distractors themselves as the number of distractors may even be larger than the number of annotator-selected points. Thus, we need to consider the *agreement* between multiple annotators, which can be determined by grouping those who have chosen roughly the same point, regardless of if that point is the true location or not. This is used to define a series of mutually exclusive *case* (for example, "two annotators chose the same point and the two other annotators chose two different points"), the probabilities of which can be calculated as:

$$P(C|p,n,z) = \sum_{\forall [q] \triangle C} \left(\frac{n!}{(n - \#N([q]))!} \right) P([q]|p,n,z) \tag{3}$$

where $[q]$ is the same as before except it uses placeholders for the distractors (e.g. [true, distractor 1, true, distractor 2] for four annotators), $\#N([q])$ is the number of said placeholders (two in this example), and $[q] \triangle C$ indicates that $[q]$ is compatible with C, that is, C could be used to describe the sequence $[q]$. Note that the number of cases, C, is the number of integer partitions of k, which increases exponentially, making the model only well-suited to cases with a small number of annotators (i.e. fewer than 10). The explicit cases and $P(C|p,n,z)$ formulas for $k = 3$ and $k = 4$ annotators can be found in [15], although the above formula allows for the number k to vary more easily. We have created a Python library that incorporates for the following:

- The creation of a *model* for k annotators which automatically generates a list of all possible values of C along with symbolic formulas for $P(C|p,n,z)$ and $P(p,n|z,[O])$;
- The addition of a sequence of*observations*, $[O]$, of said model, i.e. the number of times each case has been seen in the dataset, which allows for the symbolic formula for $P(p,n|z,[O])$ to be updated as described in the next section; and
- The evaluation of $P(p,n|z,[O])$ as well as its marginals over p and n.

Due to the combinatorial nature of the problem, each distribution can be expressed as a rational mixture of beta distributions multiplied by a simple formula taking into account the regularisation:

$$\left(\sum_i \frac{a_i}{b_i n_i^c} B_{\alpha_i,\beta_i}(p) \right) \left(\frac{z^{-n}}{f(z)} \right) \tag{4}$$

where $B_{\alpha,\beta}(p)$ is the distribution $\frac{(\alpha+\beta+1)!}{\alpha!\beta!} p^\alpha (1-p)^\beta$ over the range $p \in [0,1]$, $a_i, b_i, c_i, \alpha_i, \beta_i$ are all integers (with b_i positive and c_i, α_i, β_i non-negative), and $f(z)$ is the partition function based solely on z. (Note that some of these

elements may be removed if a given variable is specified or marginalised over.) This allows for the formulas to be stored and evaluated symbolically with the exception of $f(z)$ which has no analytic solution for the portions of the mixture where $c_i > 2$. (An approximate solution (a 100^{th}-order polynomial over z^{-1}) is used for those cases.) This library with examples is available open-source at: https://github.com/JSHBaxter/bayes_error_of_type.

2.2 Interpretation via Dirichlet Processes

A simpler method for interpreting this framework is via a Dirichlet distribution which dictates the probability of drawing a number of instances from a set of mutually-exclusive categorical classes. Given that the Bayesian conjugate prior for the Dirichlet distribution is the multinomial distribution, the probability for all of the class rates, $\alpha_1, \alpha_2, \dots \alpha_K$, given the number of instances of each class, $x_1, x_2 \dots x_K$, is:

$$P(\alpha_1, \alpha_2, \dots \alpha_K | [O]) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i \alpha_i^{x_i} \tag{5}$$

In a Dirichlet distribution, the different classes must be distinct, such as the different *cases* in our framework. This however means that the values of $alpha_i$ cannot vary independently of each other, but are linked through p, n, and z via Equation 3, that is: $\alpha_1 = P(C_1|p,n,z), \alpha_2 = P(C_2|p,n,z), \dots \alpha_K = P(C_K|p,n,z)$ where K is the number of cases, i.e. the number of integer partitions of k. We can then translate from the space of α_i to that of p and n:

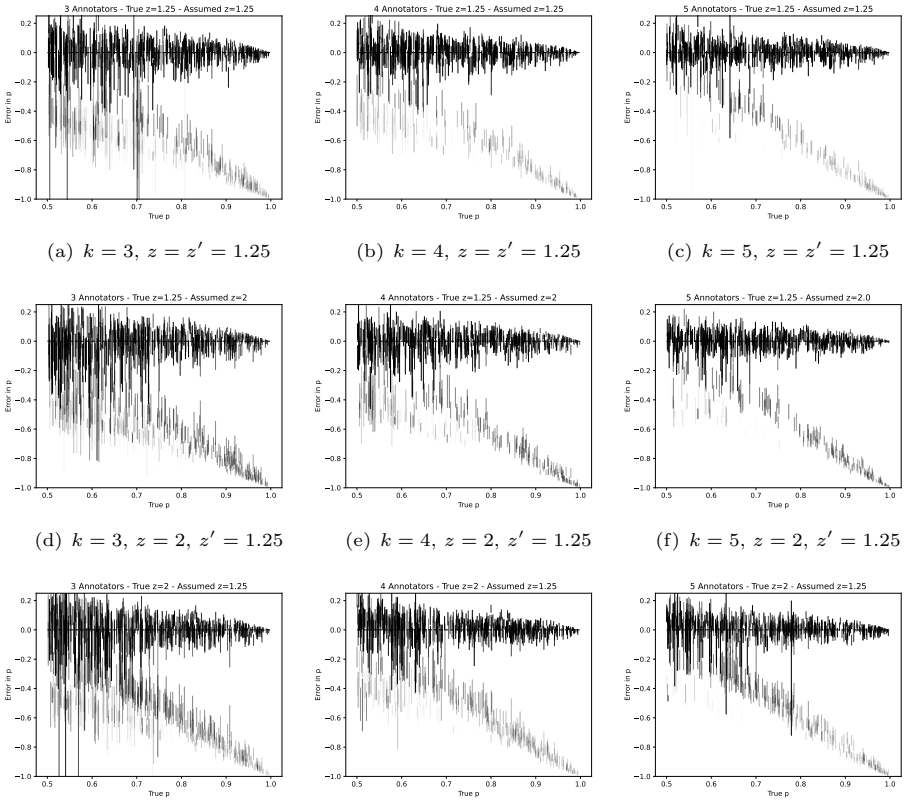
$$\begin{aligned}
&P(p,n|z,[O]) \\
&= \frac{P(\alpha_1 = P(C_1,|p,n,z), \dots \alpha_K = P(C_K,|p,n,z)|z,[O])}{\sum_{n'=1}^{\infty} \int_0^1 P(\alpha_1 = P(C_1,|p',n',z), \dots \alpha_K = P(C_K,|p',n',z)|z,[O]) dp'} \\
&= \frac{\frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i P(C_i,|p,n,z)^{x_i}}{\sum_{n'=1}^{\infty} \int_0^1 \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i P(C_i,|p',n',z)^{x_i} dp'} \\
&= \frac{\prod_i P(C_i,|p,n,z)^{x_i}}{\sum_{n'=1}^{\infty} \int_0^1 \prod_i P(C_i,|p',n',z)^{x_i} dp'}
\end{aligned} \tag{6}$$

Note that each instance of $P(C_i|p,n,z)$ is a rational mixture of beta-distributions (since it is a rationally weighted sum of Equation 2 which is a rational multiple of a beta distribution), which means that the numerator, being simply the product of these distributions is also a rational beta mixture with the denominator acting as the partition function, giving us the Equation 4 again. In practice, our model probabilities are computed via Equation 6 using Equation 4 to simplify the formulas to make it more computationally feasible to solve them exactly.

2.3 Simulation Experiments

In order to verify the model, a series of simulations was performed for the $k = 3$ to $k = 4$ cases. In each simulation, p is selected from a uniform distribution over $[0.5, 1]$, n is selected from a geometric distribution $P(n) \propto z'^{-n}$ for a given $z' > 1$, and the number of tasks is selected from a Poisson distribution with a mean of 19 and then incremented by 1 to ensure positivity. The observations are distributed according to Eq. 1. Due to the importance of p in the literature [12], the focus of the simulation was on evaluating the accuracy of the $P(p|z, [O])$.

Due to the nature of the model as a mixture, it is possible for the final probabilities to be bi- or even multi-modal. Thus, we also created a mode-seeking



(a) $k = 3$, $z = z' = 1.25$ (b) $k = 4$, $z = z' = 1.25$ (c) $k = 5$, $z = z' = 1.25$

(d) $k = 3$, $z = 2$, $z' = 1.25$ (e) $k = 4$, $z = 2$, $z' = 1.25$ (f) $k = 5$, $z = 2$, $z' = 1.25$

(g) $k = 3$, $z = 1.25$, $z' = 2$ (h) $k = 4$, $z = 1.25$, $z' = 2$ (i) $k = 5$, $z = 1.25$, $z' = 2$

Fig. 1 Results of 500 simulations for $k = 3$ (left column), $k = 4$ (middle column), and $k = 5$ (right column) annotators in terms of the distribution $P(p|z, [O])$. The first row (a, b, & c) shows simulations in which the regularisation exactly fits the distribution of n, that is, $P(n|z) \propto z^{-n}$ and $z = z' = 1.25$. The second (d, e, & f) and third (g, h, & i) rows show results when there is a mismatch between the prior and the distribution of n in terms of over- and under-estimating the regularisation, respectively. The simulation code and raw simulation results are available at https://github.com/JSHBaxter/bayes_error_of_type.

algorithm for finding each of these modes and representing them as a Gaussian in order to facilitate their interpretation. Figure 1 shows a summary of 500 simulated trials in terms of $P(p|z,[O])$. Each bar shows the mean plus or minus one standard deviation for the Gaussian representing that mode and the darkness of the line refers to the amount of the distribution it captures. That is, darker bars indicate that more of the distribution belongs to that singular mode. Lighter ones represent modes that capture little of the distribution. For high values of p (above 70%), there is always a clear mode correctly estimating p (i.e. where the standard deviation overlaps with the 0% error line) that captures well over 50% of the distribution with potentially another mode centred at approximately $1-p$. The existence of this second string of modes corresponds to the cases in which there are (at most) two clusters (i.e. the annotators are split into two groups selecting two different points). In this scenario, the model thinks there is a possibility that $n=1$ and, for that value, it cannot distinguish the group that chose the one correct point from the group that chose the one incorrect point, hence generating an estimate at p and another at $1-p$. The proportion of cases in which this can happen increases linearly with k, but the total number of cases increases exponentially, meaning that these secondary modes become less likely as the number of annotators increases. This is qualitatively supported by Figure 1 in which the salience of the light band at $1-p$ decreases as the number of annotators increases. At lower values of p, additional bands can be seen which represent other types of modes which become markedly less likely to occur as p or k increases.

The simulations show that the model is most utile when p is greater than 75% as, in that case, the potential erroneous second peak is easily separated from the primary peak. In practice, the value of p for expert users tends to exceed 80% [12, 15] This observation also appears to be robust in terms of over- and under-estimation of the regularisation parameter, z. The other observation one can make from the simulation is that small mismatches in the z parameter do not seem to have a noticeable effect on the accuracy of the primary mode (in terms of increased error or uncertainty) but do have an effect on the frequency and weight of the secondary erroneous modes.

3 Experiment

3.1 Patient Images and Annotations

44 patient T_1-weighted MR images (1mm isotropic resolution) were collected. As the patient base comes from multiple hospital centres, there is heterogeneity in terms of the MRI manufacturer (database includes Phillips Acheiva, Siemens Verio, and GE Signa HDxt) and protocol (T1 3D N NAV, MPRAGE, and CRANE STANDARD/20).

The images were annotated by a set of three expert neurologists / neurosurgeons in order to determine six points in the primary motor cortex that are often used for treating chronic pain. The agreement results determined by a fourth expert neurologist can be found in Table 1. This last neurologist was

only asked to identify which points referred to the same area as opposed to different areas. As there are three annotators, the three possible cases are:

- C_1: All annotators agree on a single point,
- C_2: Two of the three annotators agree on a single point,
- C_3: Each annotator has picked a different point.

Note that the presence of cases C_2 and C_3 demonstrate the existence of errors-of-type in this particular neurosurgical pointing task.

3.2 Model

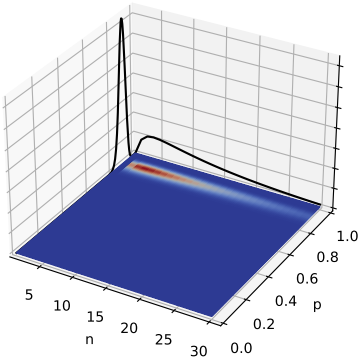
For the model, the regularisation parameter $z = 1.1$ was chosen in order to cohere with our previous simulation experiments. A separate model was created for each of the six cortical points. Due to the possibility of a secondary mode, we calculated our descriptive statistics for p and n using only the parts of the distribution where $p > 0.5$, thus avoiding averaging together two disparate modes. However, for visualisation purposes, the entire probability distribution was calculated and sampled at 0.5% intervals for p and for all n from 1 to 30.

4 Results

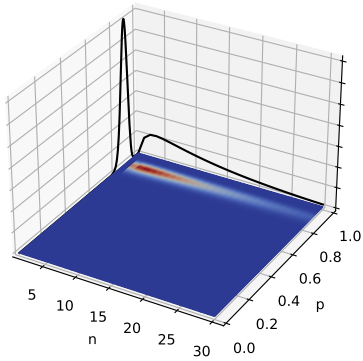
The distributions and resulting descriptive statistics for each of the six primary motor cortex points are shown in Figure 2. (Projections of the 3D distribution are given to visualise the marginal distributions.) The data appears to fall into two groups based on the number of points in C_3 (i.e. all three annotators picking distinct points). For the right upper and lower limb areas (RULMC and RLLMC), this particular case did not occur (likely due to randomness rather than a distinct left-right difference) leading to the bimodal distribution described in Section 2.3 whereas the remaining cortical targets avoided this. However, it is worth noting that the lack of cases with full disagreement also allowed the model to greatly reduce the potential spread in terms of the number of distractors even for the correct mode which is confirmed by the reduced standard deviation in the descriptive statistics for n. The model's results are very coherent with that of the TMS literature where the identity of the ground truth point is known and thus p can be more directly measured [12] and the uncertainty measurements are also within a reasonable range if one considers each of the six problems to be equally difficult, largely being within a single

Acronym	Region	C_1#	C_2#	C_3#
LFMC	Facial region of the left PMC	32	10	2
RFMC	Facial region of the right PMC	31	11	2
LLLMC	Lower limb region of the left PMC	30	12	2
RLLMC	Lower limb region of the right PMC	28	16	0
LULMC	Upper limb region of the left PMC	30	11	3
RULMC	Upper limb region of the right PMC	25	19	0

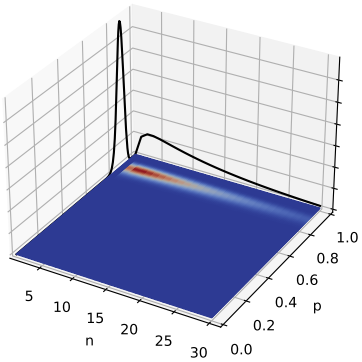
Table 1 Chronic pain treatment stimulation points in the primary motor cortex (PMC) used in the TMS dataset, reproduced from [15].



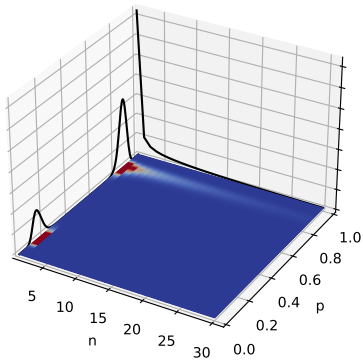
(a) LFMC ($p = 86.2 \pm 3.0\%$, $n = 11.0 \pm 7.1$)



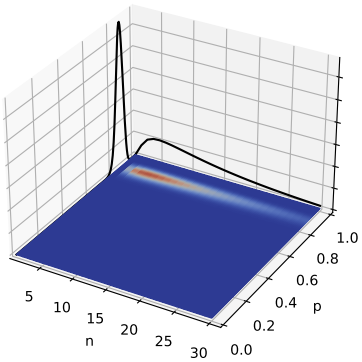
(b) RFMC ($p = 87.8 \pm 2.9\%$, $n = 10.8 \pm 7.0$)



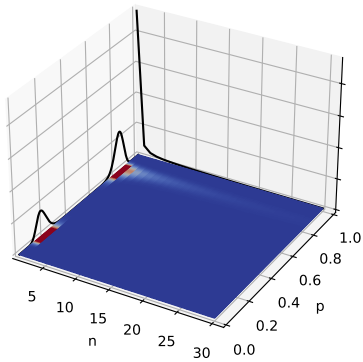
(c) LLLMC ($p = 87.0 \pm 3.0\%$, $n = 10.6 \pm 7.0$))



(d) RLLMC ($p = 86.1 \pm 3.3\%$, $n = 3.7 \pm 5.4$)



(e) LULMC ($p = 86.2 \pm 3.0\%$, $n = 11.5 \pm 7.1$)



(f) RULMC ($p = 82.6 \pm 4.0\%$, $n = 2.3 \pm 3.8$)

Fig. 2 Full parameter distribution $P(p, n|z, O)$ for six primary motor cortex points including the left facial region (a), right facial region (b), left lower limb region (c), right lower limb region (d), left upper limb region (e), and right upper limb region (f). For each figure, the marginal distributions $P(p|z, [O])$ and $P(n|z, [O])$ are given on the left and right walls of the graph respectively. The full distribution is colour-mapped to the floor of the graph.

standard deviation of each other. Lastly, the use of regularisation and uncertainty has removed the bias of n towards extreme values previous seen with our previous non-Bayesian approach [15].

5 Discussion

Despite the introduction of regularisation, the model remains sufficiently simple that it can even be symbolically stored and evaluated. In addition, the two primary model parameters, p and n, are both immediately interpretable in terms of human error (p) and a standard element of psychophysics and perceptual evaluation (i.e. number of distractors, n). This allows it to be readily extended to include another p-like parameter for a particular identified annotator and thus evaluate their performance for surgical training or evaluation purposes on real data without the need for ground truth annotations. (This would be conceptually similar to the STAPLE algorithm in segmentation [16] in that it could individually assess how close each annotator is to the consensus determined from all annotators as a whole.) Such an approach could allow for a large increase in the number and diversity of images used in such surgical training environments.

For problems with an *a priori* unknown number of options (e.g. distractors), Dirichlet processes are often used as models as they extend Dirichlet distributions, allowing for an infinite number of classes. These processes can be used to model many different problems in clustering [17] and few-shot learning [18] in which the number of classes are not known in advance. One of the powerful aspects of Dirichlet processes is that they can provide a prior for these types of problems similar to our regression parameter z. However, due to the nature of our model not being able to distinguish the ground truth point from the distractors and the possibility for more distractors than annotators, significant theoretical work will need to be performed to fit this particular problem into the framework of Dirichlet processes.

5.1 Future Work

There are a few immediate avenues for future development. The biggest and likely most difficult would be to replace the remaining expert-dependant task of determining which annotators selected the same cluster, especially when given a small number of annotators, which would prevent the use of clustering algorithms. This is a difficult task as it not only has to take into account the geometric and imaging elements of the anatomy, but also the distribution of a small and highly variable point set. Having a machine intelligence capable of determining this clustering would be ideal for the extension of these models towards there use in training neurosurgeons as mentioned above, but also for training other machine learning models.

Tailoring this model towards use in evaluating machine learning algorithms in the presence of annotation errors could have a large effect on machine-learning based neurosurgical pointing models for particularly challenging tasks

[13]. This however places an interesting requirement on the cluster determination task as it would require the result to be meaningfully differentiable in order to be used in current gradient-descent–based machine learning frameworks. An alternative and easier approach would be to determine mathematically the probability of each annotation as being reflective of the ground truth or not, using that as weights in the loss function. The naive probability estimate (that is, one that doesn't use any information about the other points) would be p, although one can easily imagine this probability increasing for points selected by multiple annotators (for example, if all k annotators agree, then the probability of the consensus point being the true location rises to $\frac{n^k p^k}{n^k p^k + n(1-p)^k}$ which is greater than p under the very mild assumption that $p > \frac{1}{n+1}$, i.e. the correct point is more likely than any given distractor) and decreasing if the point is not part of an agreed-upon cluster. This may still experience issues with the expected minima of weighted loss functions still not corresponding with the ground truth point, although more research would be needed to verify or falsify this.

From the simulation experiments, we can surmise that the model does have a significant dependence on the regularisation parameter in terms of the weight and existence of secondary modes. However, more investigation is needed to characterise precisely how it would effect the accuracy of the primary mode as it is possible that it could introduce some slight bias into the model results with respect to p. In addition, there could be other meaningful regularisation schemes other than exponential decay that could assist in mitigating secondary erroneous modes.

Lastly, it should be noted that the model is not completely scaleable for large number of observations or a large number of annotators. In the case of the latter, easy simplifying assumptions can be made for estimating p, specifically by considering the true point to be the one with the majority vote, which is the most probable explanation in the model for a high number of annotators under mild assumptions. In the case of large numbers of observations, some additional work can be done in merging together large numbers of beta distributions given the identity:

$$\begin{aligned}(\alpha+\beta+2)B_{\alpha,\beta}(p) =&(\beta+1)B_{\alpha,\beta+1}(p)\\ &+(\alpha+1)B_{\alpha+1,\beta}(p)\end{aligned} \tag{7}$$

or its extension:

$$\begin{aligned}\frac{(\alpha+\beta+n+1)!}{(\alpha+\beta+1)!}B_{\alpha,\beta}(p) =&\\ \sum_{c=0}^{n}&\frac{(\alpha+n-c)!(\beta+c)!n!}{\alpha!\beta!c!(n-c)!}\\ &\times B_{\alpha+n-c,\beta+c}(p)\end{aligned} \tag{8}$$

This identity may allow for a more compact representation of the distribution, possibly with negative terms rendering it no longer a proper mixture. At the moment, the model's memory consumption appears to grow linearly with the number of images, which is considerably higher than the constant memory required by most other approaches that estimate a final number of parameters and thus a potential issue for machine learning on large datasets. Some degree of trade-off will need to be found to balance the exactness of the model with its computational demands.

6 Conclusions

This paper presents a simple reference-free model for understanding errors-of-type in tasks where the number of distractors nor the identity of the correct point is known in advance. This framework uses Bayesian statistics to find the model parameters which include the probability of selecting the correct point, which is of high interest in the literature. Our simulation experiments show that the model is usable in cases where p is known to be relatively high (i.e. greater than 75% for 3 or 4 annotators) which is largely the case for neurosurgical pointing tasks. Due to the relative novelty of quantification methods for errors-of-type, it is unknown to what degree these models are applicable more generally in surgical planning. The model produces results that are highly coherent with similar literature in human errors in TMS cortical point targeting, although this is the first model that provides uncertainty measurements for said results.

Statements and Declarations

Funding: No funding was received to assist in the preparation of this manuscript.

Competing interests: S. Croci, A. Delmas and L. Bredoux are employees of SYNEIKA. J.S.H. Baxter, J.-P. Lefaucheur, and P. Jannin have no financial or non-financial conflicts of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Availability of data and material: Data is available for this study at https://github.com/JSHBaxter/bayes_error_of_type.

Code availability: Code is available for this study at https://github.com/JSHBaxter/bayes_error_of_type.

References

[1] Fitzpatrick, J.M., West, J.B.: The distribution of target registration error

in rigid-body point-based registration. IEEE transactions on medical imaging **20**(9), 917–927 (2001)

[2] Wiles, A.D., Likholyot, A., Frantz, D.D., Peters, T.M.: A statistical model for point-based target registration error with anisotropic fiducial localizer error. IEEE transactions on medical imaging **27**(3), 378–390 (2008)

[3] Luo, J., Frisken, S., Machado, I., Zhang, M., Pieper, S., Golland, P., Toews, M., Unadkat, P., Sedghi, A., Zhou, H., Mehrtash, A., Preiswerk, F., Cheng, C.-C., Golby, A., Sugiyama, M., Wells, W.M.: Using the variogram for vector outlier screening: application to feature-based image registration. International journal of computer assisted radiology and surgery **13**, 1871–1880 (2018)

[4] Bardosi, Z., Freysinger, W.: Estimating FLE_{image} distributions of manual fiducial localization in CT images. International Journal of Computer Assisted Radiology and Surgery **11**(6), 1043–1049 (2016)

[5] Francel, P.C., Jackson, T.R., Kamiryo, T., Laws, E.R.: Optimizing accuracy in magnetic resonance imaging—guided stereotaxis: a technique with validation based on the anterior commissure—posterior commissure line. Journal of neurosurgery **90**(1), 94–100 (1999)

[6] Prakash, K.B., Hu, Q., Aziz, A., Nowinski, W.L.: Rapid and automatic localization of the anterior and posterior commissure point landmarks in mr volumetric neuroimages1. Academic radiology **13**(1), 36–54 (2006)

[7] Eckstein, M.P., Abbey, C.K., Bochud, F.O.: Visual signal detection in structured backgrounds. iv. figures of merit for model performance in multiple-alternative forced-choice detection tasks with correlated responses. JOSA A **17**(2), 206–217 (2000)

[8] Elangovan, P., Mackenzie, A., Dance, D.R., Young, K.C., Cooke, V., Wilkinson, L., Given-Wilson, R.M., Wallis, M.G., Wells, K.: Design and validation of realistic breast models for use in multiple alternative forced choice virtual clinical trials. Physics in Medicine & Biology **62**(7), 2778 (2017)

[9] Roxin, A.: Drift–diffusion models for multiple-alternative forced-choice decision making. The Journal of Mathematical Neuroscience **9**(1), 1–23 (2019)

[10] Lefaucheur, J.-P., André-Obadia, N., Antal, A., Ayache, S.S., Baeken, C., Benninger, D.H., Cantello, R.M., Cincotta, M., de Carvalho, M., De Ridder, D., Devanne, H., Di Lazzaro, V., Filipović, S.R., Hummel, F.C., Jääskeläinen, S.K., Kimiskidis, V.K., Koch, G., Langguth, B., Nyffeler, T.,

Oliviero, A., Garcia-Larrea, L.: Evidence-based guidelines on the therapeutic use of repetitive transcranial magnetic stimulation (rtms). Clinical Neurophysiology **125**(11), 2150–2206 (2014)

[11] Kim, W.J., Min, Y.S., Yang, E.J., Paik, N.-J.: Neuronavigated vs. conventional repetitive transcranial magnetic stimulation method for virtual lesioning on the broca's area. Neuromodulation: Technology at the Neural Interface **17**(1), 16–21 (2014)

[12] Mylius, V., Ayache, S., Ahdab, R., Farhat, W., Zouari, H., Belke, M., Brugières, P., Wehrmann, E., Krakow, K., Timmesfeld, N., Schmidt, S., Oertel, W.H., Knake, S., Lefaucheur, J.P.: Definition of dlpfc and m1 according to anatomical landmarks for navigated brain stimulation: inter-rater reliability, accuracy, and influence of gender and age. Neuroimage **78**, 224–232 (2013)

[13] Baxter, J.S.H., Bui, Q.A., Maguet, E., Croci, S., Delmas, A., Lefaucheur, J.-P., Bredoux, L., Jannin, P.: Automatic cortical target point localisation in mri for transcranial magnetic stimulation via a multi-resolution convolutional neural network. International Journal of Computer Assisted Radiology and Surgery **16**(7), 1077–1087 (2021)

[14] Vakharia, V.N., Sparks, R., Pérez-García, F., Granados, A., Miserocchi, A., McEvoy, A., Ourselin, S., Duncan, J.S.: Machine learning for stereotactic neurosurgery: A prospective implementation and validation. Hugh Cairns Prize Essay (2019)

[15] Baxter, J.S.H., Croci, S., Delmas, A., Bredoux, L., Lefaucheur, J.-P., Jannin, P.: Errors of type or errors of degree? cortical point targeting in transcranial magnetic stimulation. In: Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 12034, p. 1203403 (2022). SPIE

[16] Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)

[17] Kulis, B., Jordan, M.I.: Revisiting k-means: new algorithms via bayesian nonparametrics. In: Proceedings of the 29th International Coference on International Conference on Machine Learning, pp. 1131–1138 (2012)

[18] Allen, K., Shelhamer, E., Shin, H., Tenenbaum, J.: Infinite mixture prototypes for few-shot learning. In: International Conference on Machine Learning, pp. 232–241 (2019). PMLR