


Statistical parametric mapping of biomechanical one-dimensional data with Bayesian inference

Ben Serrien ^a, Maggy Goossens^{b,c} and Jean-Pierre Baeyens^{a,b,c}

^aFaculty of Physical Education and Physiotherapy, Vrije Universiteit Brussel, Brussels, Belgium; ^bFaculty of Applied Engineering, Universiteit Antwerpen, Antwerp, Belgium; ^cThim Van Der Laan University College Physiotherapy, Landquart, Switzerland

ABSTRACT

Recent developments in Statistical Parametric Mapping (SPM) for continuum data (e.g. kinematic time series) have been adopted by the biomechanics research community with great interest. The Python/MATLAB package *spm1d* developed by T. Pataky has introduced SPM into the biomechanical literature, adapted originally from neuroimaging. The package already allows many of the statistical analyses common in biomechanics from a frequentist perspective. In this paper, we propose an application of Bayesian analogs of SPM based on Bayes factors and posterior probability with default priors using the *BayesFactor* package in R. Results are provided for two typical designs (two-sample and paired sample *t*-tests) and compared to classical SPM results, but more complex standard designs are possible in both classical and Bayesian frameworks. The advantages of Bayesian analyses in general and specifically for SPM are discussed. Scripts of the analyses are available as supplementary materials.

ARTICLE HISTORY

Received 10 July 2018
Accepted 18 March 2019

KEYWORDS




Bayesian inference; Bayes Factor; posterior probability; Statistical Parametric Mapping; time series; false discovery rate; Q-value

Introduction: Statistical Parametric Mapping

Statistical Parametric Mapping (SPM) was originally developed for statistical inference on neuroimaging data where dependent variables are sampled on a large number of spatially correlated voxels (volume elements) (Friston 2007). The same methodology applies, however, to all spatiotemporally registered and smooth data, and subsequent work from the research group of Todd Pataky has introduced SPM in the biomechanics and human movement science community for analysis of pedobarographic images (Pataky and Goulermas 2008), finite-element simulations (Pataky 2010) and uni/multivariate time series data (kinematics, kinetics, sEMG, etc.) (Pataky 2012; Pataky et al. 2013; Robinson et al. 2015). Recent developments include power analysis and sample size calculations (Pataky 2017) and SPM for cortical bone mapping (Li et al. 2009; Poole et al. 2017; Yu et al. 2017) and pedobarographic videos (Booth et al. 2018). The nomenclature of SPM uses '*nDmD*' to describe the dimensionality of the dataset, where the parameter *n* describes the dimension of the field in which the dependent variable(s) is(are) sampled and the parameter *m* describes the number of dependent variables (Pataky et al. 2016). In the present paper, our focus lies on univariate time series, i.e. *1D1D* data where one dependent variable is sampled continuously over time (a one-dimensional field), but the same principles apply to all *nDmD* data.

SPM offers a couple of strong advantages for biomechanists and movement scientists. The primary advantage is that no abstraction of the originally sampled time series needs to be performed in order to statistically analyze the data. The full *1D* field can be examined in a non-directed hypothesis test without any ad-hoc assumptions regarding the spatiotemporal foci of interest. Since kinematic or kinetic time series can be complex, it can be difficult to objectively specify an a-priori method for analysis and many studies, therefore, adopt an ad-hoc approach: visualize the *1D* time series and extract a summary *0D* scalar (extremum, central tendency, ...) which was not specified a priori to test statistically (Pataky et al. 2015). Accompanying these full-field non-directed hypotheses tests is the ability to visualize the statistical results in the same field as where the data were sampled. For time series data, the statistical result is hence also a time series (e.g. a time series of *t*-values) and allows for better interpretation of data.

SPM uses Random Field Theory (RFT) (Adler and Taylor 2007) to perform topological inference instead of performing separate inferential tests at each time point which would cause an inflation of Type I error. RFT leverages smoothness (local correlation between adjacent time points) to mitigate the multiple testing problem, thereby offering accurate sampling-rate independent control of Type I errors when testing correlated field data. Because biological processes are typically

CONTACT Ben Serrien  ben.serrien@vub.be  Faculty of Physical Education and Physiotherapy, Vrije Universiteit Brussel, Pleinlaan 2, Brussel 1050, Belgium
 Supplemental data for this article can be accessed [here](#).

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

smooth and biomechanical data acquisition samples above the Nyquist criterion, neighboring time samples are not independent and this should be taken into account (Pataky 2010). Rather than computing a p -value at each time sample, a p -value is calculated for clusters of statistics (e.g. t) that cross a critical threshold (t^*). The logic of RFT is that the height and width of supra-threshold clusters produced by smooth random fields are inversely proportional to the probability of their occurrence, making a large supra-threshold cluster the topological equivalent of a large t -value for OD data (Pataky 2010; Appendix A). The definition of the SPM p -values can be stated as: ‘the probability that smooth, random continua would produce a supra-threshold cluster as broad as the observed cluster’ (spm1d.org, © T. Pataky). Critical thresholds are usually calculated with the Type I error $\alpha = 0.05$. Hence, when the observed t -statistic time series crosses the threshold, this cluster has a $p < 0.05$, allowing the researcher to reject the null hypothesis H_0 of no difference between the two time series.

These SPM p -values – as in OD statistics – refer to the probability of the data given that H_0 is true, $P(\text{data} | H_0)$, without recourse to any alternative hypothesis H_1 , which is the classical frequentist approach to inference. In the current implementation of the open-source package *spm1d* (Python and MATLAB versions, spm1d.org, © T. Pataky), there is only the possibility to perform frequentist inference and in this paper, we want to propose a stepping stone towards a Bayesian alternative. In the following sections, we will first briefly introduce the differences between Bayesian and frequentist inference and delineate why the Bayesian alternative can offer additional insights from the data.

Bayesian inference

Classical inference answers the inverse question of what researchers usually aim to answer. Above we gave the definition of a frequentist p which is not the same as what we want to know, namely the probability that H_0 or H_1 are good descriptions of the data: $P(\text{data} | H_0) \neq P(H_0 | \text{data})$ (Cohen 1994). In fact, both probabilities are related to each other through Bayes’ theorem:

$$P(H_i | \text{data}) = \frac{P(\text{data} | H_i)P(H_i)}{\sum P(\text{data} | H_j)P(H_j)}$$

where the sum in the denominator (or integral in the limiting case) is taken over the set of all relevant hypotheses j (including i). Additionally, frequentist inference is asymmetric in the sense that: (1) it is only possible to state evidence against H_0 and not evidence in favor of H_0 or in favor of any alternative H_1 and (2) because it does not consider any alternative

hypotheses, the evidence against the null is always overstated (Rouder et al. 2009; Morey and Rouder 2011; Wagenmakers et al. 2018). When a researcher wishes to demonstrate the invariance of some variable’s time series during a movement (invariance with respect to a model prediction, experimental manipulation or group membership), classical inference only allows statements like ‘the data showed no evidence against H_0 during the movement’ which is not the same as the statement which was the aim of the study: ‘the data showed evidence in favor of H_0 and thus an invariance during the movement’. With Bayesian inference, the latter statements are possible. For instance, consider a study where the objective is to show that gait kinematic time series are left-right symmetric, Bayesian inference can quantify the evidence in favor of the H_0 : $\mu_{\text{left}}(t) = \mu_{\text{right}}(t)$. Conversely, a classical approach would be to assume that symmetry exists, calculate a p -value under this assumption and fail to reject the null. But it makes no logical sense to assume something which you want to prove. Interested readers in further contrasts between classical and Bayesian inference are referred to recent tutorial papers on Bayesian statistics (Dienes and Mclatchie 2018; Etz and Vandekerckhove 2018; Kruschke and Liddell 2018a).

Within the Bayesian school of statistics, many related approaches exist, but in this paper, we will focus on an approach to Bayesian SPM based on Bayes Factors and Posterior Probability. A few other Bayesian alternatives are explained in the discussion. Bayes Factors (BF) result from the application of Bayes’ rule and can be linked to the odds of one hypothesis over another:

$$\underbrace{\frac{P(H_1 | \text{data})}{P(H_0 | \text{data})}}_{\text{posterior odds}} = \underbrace{\frac{P(\text{data} | H_1)}{P(\text{data} | H_0)}}_{BF_{10}} \cdot \underbrace{\frac{P(H_1)}{P(H_0)}}_{\text{prior odds}},$$

where BF_{10} is the Bayes Factor with the marginal likelihood of the data under the alternative H_1 in the numerator and the likelihood under H_0 in the denominator. The prior odds reflect the relative belief in both hypotheses before doing the experiment and is often set equal to 1 in order not to favor any hypothesis a priori, in which case BF_{10} reflects the posterior odds of the alternative over the null. However, it is not necessary to set the prior odds to 1, researchers may simply communicate the BF which readers may multiply with any prior odds they hold on the two competing hypotheses to yield a posterior odds. The BF reflects the change in confidence on the two hypotheses after observing the data (Wagenmakers et al. 2018). A BF_{10} of 1 indicates equal evidence for both hypotheses while $0 < BF_{10} < 1$ indicates evidence in favor of H_0 and $BF_{10} >$

1 is evidence in favor of H_1 . For instance, with the prior odds set to 1, a BF_{10} of 5 indicates that $P(H_1 | \text{data}) = 5/6$ and $P(H_0 | \text{data}) = 1/6$, i.e. the probability of the alternative is 5 times higher than that of the null.

To calculate the Bayes Factor, researchers need to specify likelihood functions and associated prior probabilities for both hypotheses. This allows for very flexible analyses and to incorporate any prior knowledge about the specific data from experience, pilot studies, meta-analyses and other sources (Wagenmakers et al. 2018). The resulting BF is naturally sensitive to this choice, and it should be argued why a particular choice is relevant and how robust the results are with respect to reasonable changes in the prior setting. The so-called objective Bayesian school has developed default priors for a variety of typical statistical tests for which the resulting BF has desirable theoretical properties. These properties include: (1) *scale invariance* which means the default BF is unaffected by multiplicative changes of the variables (i.e. independent of the measurement units); (2) *consistency*, which means that in the large sample limit the BF will approach zero or infinity when the effect size is 0 or not zero, respectively; and (3) *consistency in information*, which indicates that the BF will approach the correct limit as the statistic of interest (e.g. t) increases: $\lim_{t \rightarrow \infty} BF_{10} = \infty$, independent of sample size (Rouder et al. 2012). These default priors are general and broadly applicable and are reasonable in most circumstances (Rouder et al. 2012) and we will, therefore, choose them for our proposal of a Bayesian implementation of SPM.

The BF based on default priors is a convenient summary of the evidence but has one disadvantage for SPM, namely the control of the multiple testing problem across the $1D$ field (which typically includes 101 time points; 0–100% of the movement). However, the BF can be converted to posterior probabilities which are better suited to implement a multiple testing control scheme. Posterior probabilities are also easier to interpret for researchers used to classical statistics as they live on the $[0, 1]$ interval whereas a BF exist on the $]0, \infty[$ interval. Given a prior odds of 1, the posterior probabilities (PP) can be calculated as:

$$P_{H_0} = P(H_0 | \text{data}) = \frac{1}{1 + BF_{10}} \text{ and } PP_{H_1} = P(H_1 | \text{data}) \\ = \frac{1}{1 + BF_{01}}$$

When PP_{H_1} at a certain time point is, e.g. 0.95, the posterior error probability of classifying this time point as evidence in favor of H_1 is 0.05 ($PEP = 1 - PP$). The false discovery rate (FDR) can be used as a unified multiple-testing framework for Bayesian and classical inference (De Villemeireuil et al. 2014) and is therefore adopted here. A conservative control

of the FDR can be made by thresholding the posterior probability SPM at, e.g. 0.95, keeping the $FDR \leq 0.05$ (Friston and Penny 2003). A less conservative threshold, while still keeping the FDR at the same level is the use of the q -value which is defined as the cumulative mean of the posterior error probabilities (Storey 2003; Käll et al. 2008). A q -value of 0.05 for a certain time point implicates that for all possible thresholds, 5% is the minimal FDR threshold at which this time point will appear in a supra-threshold cluster (Käll et al. 2008). Especially for SPM, the use of the q -value is, we believe, better suited, because it indirectly takes the temporal correlation of adjacent time points into account. Although a q -value is calculated per time point for $1D$ data, it is a property of the entire time series object. When a threshold of $q^* = 0.05$ is used, the first time point to be included in a supra-threshold cluster has at least a posterior probability of 0.95; adjacent time points may fall below that while still keeping below q^* , which is reasonable because adjacent time points are strongly correlated and hence may be categorized in the same cluster.

Classical SPM versus Bayesian SPM

In this section, we will compare the classical SPM with our proposition of a Bayesian version of Statistical Parametric Mapping for $1D1D$ data.

Datasets

We will use example datasets that come included with the open-source `spm1d`-package (`spm1d.org`, © T. Pataky) and one additional dataset from our lab. We will use common statistical tests for demonstration purposes (two-sample t -test and paired-sample t -test), but more complex statistical models are available in the packages for classical and Bayesian analysis (n-way ANOVA, repeated measures, (multi-)linear regression, ...). A description of the three data-sets is given in Table 1.

Classical SPM

The `spm1d` package was used to perform two-tailed $SPM\{t\}$ tests in Spyder (Python 3.6). The classical null hypotheses for the three examples are as follows:

- SimulatedTwoLocalMax: Independent-sample t -test: $H_0: \mu_1(t) = \mu_2(t)$
- PlantarArchAngle: Paired-sample t -tests: $H_0: \mu_1(t) = \mu_2(t)$
- GaitSymmetry: Paired-sample t -tests: $H_0: \mu_{left}(t) = \mu_{right}(t)$

Table 1. Description of example datasets used for the three frequentist and Bayesian SPM tests. The first two datasets are part of the `spm1d`-package (© T. Pataky).

Statistical test	Example dataset
Two-sample t -test	SimulatedTwoLocalMax. Dataset of $2 \times n = 6$ simulated time series of 101 time samples each. The first set are smooth unit Gaussian random trajectories. The second set are also smooth unit Gaussian trajectories, but with bursts at $t = 25$ and $t = 75$.
Paired-sample t -test	PlantarArchAngle (Caravaggi et al., 2010). Dataset of $2 \times n = 10$ experimental time series of the plantar arch angle of the foot at 101 time samples each.
Paired sample t -test	GaitSymmetry. Single subject dataset (healthy male, 28 years, 83 kg, 178 cm) of left and right leg knee flexion angles during 99 gait cycles on a dual-belt treadmill at constant speed of 4.5 km/h, time normalized to 101 time samples. Gait kinematics were recorded with a 6-camera VICON system at 250 Hz. (unpublished data from our lab).

The results of the analyses with $\alpha = 0.05$ are shown in Figure 1, and the details of the supra-threshold clusters are depicted in Table 2.

In all three examples, H_0 may be rejected at a type I error rate of 5%. However, the supra-threshold clusters are quite small and specific to certain phases of the motion which is in contrast to the hypotheses stated above which apparently hold for all time points. In the PlantarArchAngle and especially for the GaitSymmetry example, it is contra-intuitive to completely reject the H_0 because the time series show a strong similarity. The GaitSymmetry dataset is a good example where a marginal effect can result in a statistically significant

difference purely because a large sample size is obtained (but see results below on power). In the Bayesian case, this is not necessarily so and allows a more nuanced statement where the null can also be accepted in other time spans.

Bayesian SPM: posterior probability maps

Similar to the classical approach, we will construct a Statistical Parametric Map of the posterior probability in the time domain where $P(H | \text{data})$ is calculated at each time sample, analogous to an $\text{SPM}\{t\}$. The calculations of the Bayes Factors and posterior probabilities are performed in RStudio with the R-package BayesFactor (Morey et al. 2018), scripts and datasets are available as supplementary materials.

The hypotheses and priors for the t -tests are parameterized in terms of the effect size $\delta = (\mu_1 - \mu_2)/\sigma$, where the indices refer to the two groups or two conditions or left and right legs. Point null hypotheses are very unlikely to be true exactly and trivially small effects may exist that are not of (clinical, theoretical) interest. This does not mean that the null should be abandoned and it may still be preferred for parsimony's sake as a first approximation of the truth (Cohen 1994; Morey and Rouder 2011). Morey and Rouder (2011) provide BF calculations where the null includes trivially small effects around $\delta = 0$. We believe that this is especially appropriate for 1D time series data (and increasingly so

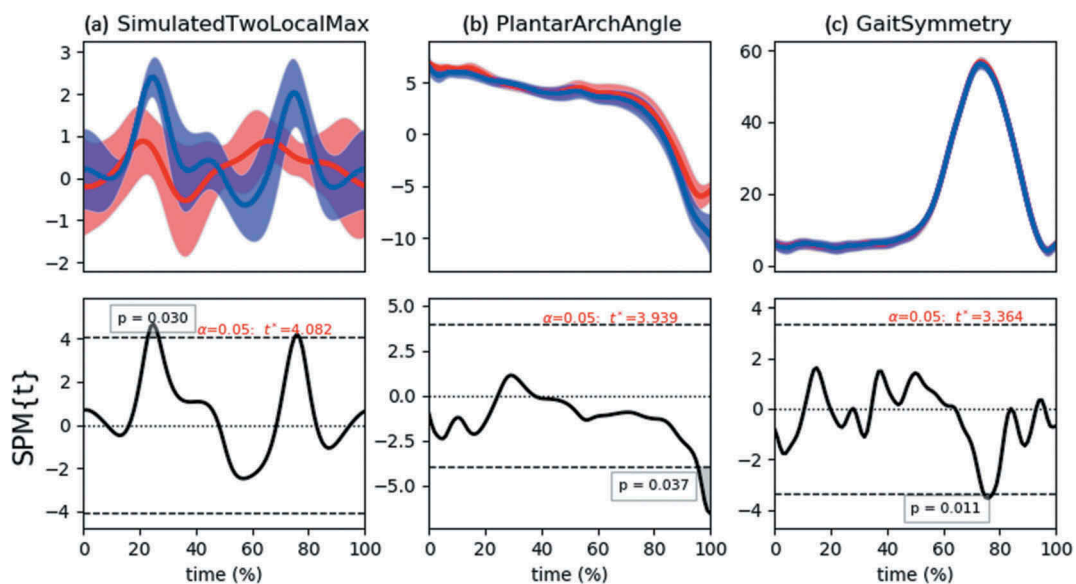


Figure 1. Classical $\text{SPM}\{t\}$ results for the three datasets. Top row shows descriptive statistics for each dataset (Mean \pm 1 SD error cloud). Bottom row shows the frequentist inferences. The horizontal dashed lines depict the critical t^* based on $\alpha = 0.05$ and RFT calculations of residual smoothness. Supra-threshold clusters result in $p < 0.05$. For the GaitSymmetry example, note that these are time series from a single subject, the SD-cloud thus represents within-subject variability instead of between-subject variability. The inference only pertains to this subject.

Table 2. Classical SPM{t} results for the three datasets. Begin and end-points of supra-threshold cluster locations are given as a percentage of the total movement time.

	Evidence against H_0 ($p < 0.05$)	
	Cluster location	p-Value
SimulatedTwoLocalMax	t = 24–27	p = 0.030
Independent-samples t-test	t = 77	p = 0.046
PlantarArchAngle	t = 97–101	p = 0.037
Paired-samples t-test		
GaitSymmetry	t = 74–78	p = 0.011
Paired-samples t-test		

for higher dimensional fields) because point null models would already be unlikely due to technical data registration issues and natural movement variability. In the calculations below, we took $[-0.2, 0.2]$ as an interval of trivially small effects for the difference in the time series data, this choice corresponds to the typical recommendation by (Cohen 1988) for small effects. The null and alternative hypotheses for the Bayes Factor calculations were as follows:

- paired and independent samples t-tests (point H_0):

$$H_0: \delta(t) = 0$$

$$H_1: \delta(t) \sim \text{Cauchy}(r)$$

- paired and independent samples t-tests (interval H_0):

$$H_0: \delta(t) \sim \text{Cauchy}(r) \text{ for } \delta(t) \in [-c, c]$$

$$H_1: \delta(t) \sim \text{Cauchy}(r) \text{ for } \delta(t) \notin [-c, c]$$

Mathematical definitions of the default JZS-priors (Cauchy distributions), their justification and proofs for deriving the BFs can be found in (Rouder et al. 2009; Morey and Rouder 2011). These default priors still allow a flexible scaling of the width of the prior (r) and

a determination of the null interval (c). The scale of the Cauchy prior should be set a priori and should reflect prior knowledge about the effect sizes that are relevant or expected for the variables of interest. The BayesFactor package offers three default options, that are shown in Figure 2. When the effect size is likely to be small to moderate, then the medium scale is a suitable choice, but when very large effects are expected, less probability should be placed in the center and more at the edges. For time series applications, a well-justified prior knowledge may even be reflected in phase-specific priors where the scale is a function of time, $r(t)$. One-sided priors can also be selected for directed alternative hypotheses.

In Figure 3 and Table 3, we show the results of the Bayesian SPM (posterior probability maps). Because we believe the interval H_0 to be the most relevant, we report only these results, but the reader may examine the point H_0 in the R-scripts (supplementary materials).

The results of the Bayesian SPM were partially in line with the classical SPM approach. For the independent-sample t-test, the classical SPM showed evidence against the null at $t = 24-27$ and at $t = 77$. The Bayesian SPM confirmed the existence of both clusters. For the conservative threshold, the results were nearly identical to the classical SPM, while the supra-threshold clusters were a little wider for the q^* based threshold. Note that the posterior probability is below $\frac{1}{2}$ most of the time, which makes sense given the model that created these simulated time series (see Table 1). However, the small sample size was ineffective for claiming strong enough evidence for accepting the null. For both types of threshold, the result was only weakly dependent on the width of the Cauchy scale.

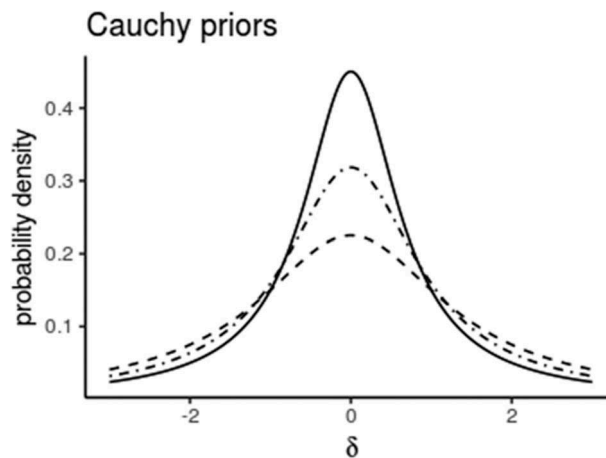


Figure 2. Cauchy priors for the effect size δ with different scales (solid line: $r = \sqrt{2}/2$ (medium), dot-dashed: $r = 1$ (wide) and dashed: $r = \sqrt{2}$ (ultra-wide)). Fifty percent of the probability mass lies between $-r$ and $+r$.

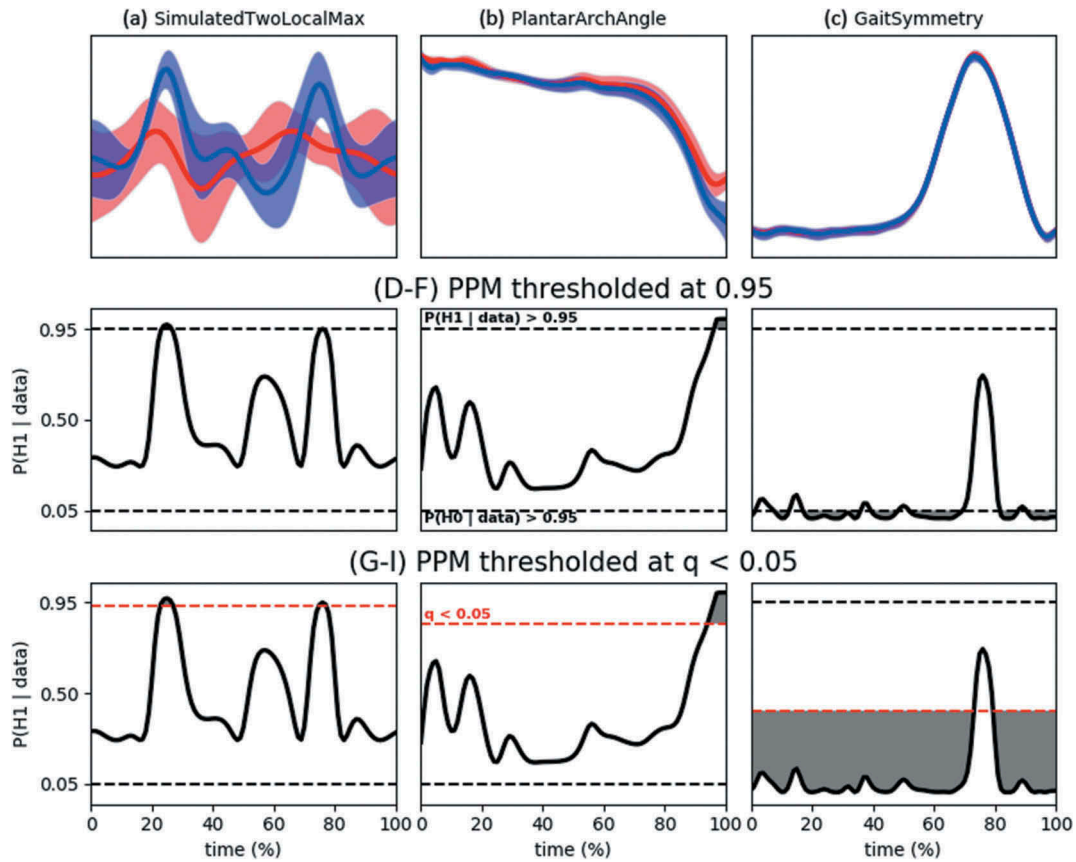


Figure 3. Panels (a), (b) and (c) give descriptive statistics for the three datasets (replicated from Figure 1). Panels (d), (e) and (f) give the posterior probability maps (PPM) for the alternative hypothesis: a time series of $P(H_1 | \text{data})$ (only shown for $r = \sqrt{2}/2$, see Table 3 for comparison to the other scales). The horizontal dashed lines at 0.05 and 0.95 depict the thresholds for which, respectively, $P(H_0 | \text{data}) > 0.95$ and $P(H_1 | \text{data}) > 0.95$ [$P(H_0 | \text{data}) + P(H_1 | \text{data}) = 1$]. Panels (g), (h) and (i) show the same PPM but thresholded using the FDR scheme. The red horizontal dashed line indicates the largest posterior error probability for which $q < 0.05$. It can be seen that no new clusters are created because the minimal posterior probability for either hypothesis must still be 0.95 in order to keep the q below 0.05. Because the cumulative mean is taken, the clusters broaden or in case of the GaitSymmetry dataset, they merge.

For the PlantarArchAngle t -test, the classical SPM showed a significantly different plantar arch angle between $t = 97$ – 101 , while the Bayesian supra-threshold cluster was a little wider ($t = 95/96$ – 101). Similar to the previous example, the sample size was too small to claim strong evidence in favor of the null, although the posterior probability was below $\frac{1}{2}$ most of the time. From a Bayes Factor perspective, the null was more than 4 times more likely than the alternative at large phases of the gait cycle ($BF_{10} < \frac{1}{4}$), but this was not enough to reach the posterior probability thresholds of 0.95. Also, for this example, the sensitivity to the prior scale was very small.

Arguably the biggest difference between both inferential perspectives lies in the GaitSymmetry dataset. From a classical perspective, the (point) H_0 could be rejected between $t = 74$ – 78 ($p = 0.011$), whereas the Bayesian analysis only yields a maximum of 72% probability for the alternative during this time span, which is not convincing evidence for asymmetry. The Bayesian

perspective shows, however, that throughout most of the time $P(H_0 | \text{data}) \geq 0.95$. Using the q^* threshold, we would say that except for a small amount of time (between 1% and 6% of the gait cycle), this subject is left-right symmetric in the knee joint motion. For this dataset, the results are more strongly dependent on the scale factor. Because the wide and ultra-wide settings place a more prior probability on large effect sizes, they are less likely alternatives and thus get penalized in favor of the null hypothesis which results in broader clusters.

Note, however, that in the classical SPM, the significant result is most likely caused by an overpowered dataset. This is a relatively common problem in single-subject designs where additional trials are easy to sample. The maximal significant mean difference was only 0.54° which is not clinically relevant to consider asymmetric. The present paper is only demonstrative, but if this were a proper experimental study, an a-priori power analysis may have helped to determine the number of gait cycles. Given

Table 3. Overview of supra-threshold clusters for the Bayesian SPM tests (interval H_0 only). The less conservative $q^* = 0.05$ threshold always yields broader clusters than the $P(H | \text{data})^* = 0.95$ threshold. For the SimulatedTwoLocalMax and PlantarArchAngle datasets, the difference between both thresholds is small. For the GaitSymmetry example, the difference is larger and results in 4–7 separate clusters or 2 broad clusters (for the ultra-wide setting, it is nearly 1 cluster over the entire time span). The GaitSymmetry example also shows sensitivity to the scale of the prior, whereas this sensitivity was negligible in the other two datasets.

	Evidence in favor of H_0		Evidence in favor of H_1	
	$P(H_0 \text{data}) \geq 0.95$	$q(H_0) \leq 0.05$	$P(H_1 \text{data}) \geq 0.95$	$q(H_1) \leq 0.05$
SimulatedTwoLocalMax (independent-samples t -test)				
$r = \text{medium}$	/	/	$t = 25\text{--}27$ $t = 77$	$t = 24\text{--}28$ $t = 76\text{--}78$
$r = \text{wide}$	/	/	$t = 24\text{--}27$ $t = 77$	$t = 24\text{--}28$ $t = 75\text{--}78$
$r = \text{ultra-wide}$	/	/	$t = 24\text{--}27$ $t = 76\text{--}78$	$t = 24\text{--}28$ $t = 75\text{--}79$
PlantarArchAngle (paired-samples t -test)				
$r = \text{medium}$	/	/	$t = 98\text{--}101$	$t = 95\text{--}101$
$r = \text{wide}$	/	/	$t = 98\text{--}101$	$t = 95\text{--}101$
$r = \text{ultra-wide}$	/	/	$t = 98\text{--}101$	$t = 96\text{--}101$
GaitSymmetry (paired-samples t -test)				
$r = \text{medium}$	$t = 1\text{--}2$ $t = 9\text{--}13$ $t = 18\text{--}37$ $t = 41\text{--}49$ $t = 53\text{--}70$ $t = 83\text{--}88$ $t = 92\text{--}101$	$t = 1\text{--}74$ $t = 80\text{--}101$	/	/
$r = \text{wide}$	$t = 1\text{--}2$ $t = 8\text{--}14$ $t = 18\text{--}37$ $t = 40\text{--}50$ $t = 52\text{--}71$ $t = 83\text{--}101$	$t = 1\text{--}75$ $t = 79\text{--}101$	/	/
$r = \text{ultra-wide}$	$t = 1\text{--}3$ $t = 6\text{--}14$ $t = 17\text{--}71$ $t = 83\text{--}101$	$t = 1\text{--}76$ $t = 78\text{--}101$	/	/

a minimal difference of 2° to consider asymmetric, we performed a power analysis with the *power1d* package (Pataky 2017). The python script in the supplementary material explains the construction of the null and alternative models used for the simulations. The results showed that for a power of 0.80, minimal 50 gait cycles should be sampled (Figure 4(a)). Other simulations fluctuated a little around $n = 50$, so we performed the paired-samples t -tests again with the first 55 gait cycles, the results are shown in Figure 4(b) (classical) and 4C (Bayesian). The classical result no longer rejects H_0 while the Bayesian result still provides evidence for H_0 throughout most of the gait cycle. Technically these conclusions are not the same, but from an applied perspective, both conclusions would be in favor of symmetry.

Discussion and future work

In this paper, we have proposed a stepping stone towards a Bayesian alternative to Statistical Parametric Mapping of $1D1D$ data. We have shown results of posterior probability maps in two common statistical tests (two- and paired-sample t -tests) and compared the results to the classical SPM(t). Both similarities and discrepancies are found between both inferential methods. Bayesian methodology in general (not only for $1D$ data and SPM) has a stronger face validity and is not asymmetric like classical inference and takes an explicit alternative hypothesis into account which allows to calculate evidence in favor of a null hypothesis. We used an example of gait analysis to show how a Bayesian approach can statistically demonstrate that knee joint angles are left-right symmetric throughout nearly the entire gait cycle (single subject design). While an appropriately powered classical inference found no evidence of a significant left-right difference, this absence of evidence is not the same as a quantification of the evidence in favor of symmetry like in the Bayesian approach. This ability may also be important in cases like designing neuromuscular models or testing theories where the time series of the prediction is compared to observed time series, $H_0: \mu_{\text{model}}(t) = \mu_{\text{empirical}}(t)$. Also in applied research with biomechanical time series, it may be relevant to examine invariance with respect to certain interventions. When some clinical or sports training intervention is performed in order to change the motion pattern, a frequentist approach could reject (successful intervention) or fail to reject the null (unsuccessful experiment), while a Bayesian approach could also provide evidence that the intervention itself is unsuccessful which is different from an unsuccessful experiment.

Classical sample size calculations for $0D$ -data requires a definition of the minimal effect size that should be detectable with sufficient power and a given alpha level. For $1D$ -data, null and alternative models should be constructed that represent the expected behavior of the time series under the null and alternative models. For well-known signals with simple behavior, this is relatively easy with the *power1d* package (Pataky 2017) but quickly becomes much more difficult for complex, unknown signals. Also, the expectation of the signal variability (within- and between subjects) and how and when the two groups would be different may be difficult to anticipate (effect cluster shape, height, width, location). When the observed effects in the final study are markedly different from the anticipated effects, the study may result in serious over/underpowered conclusions. From the Bayesian

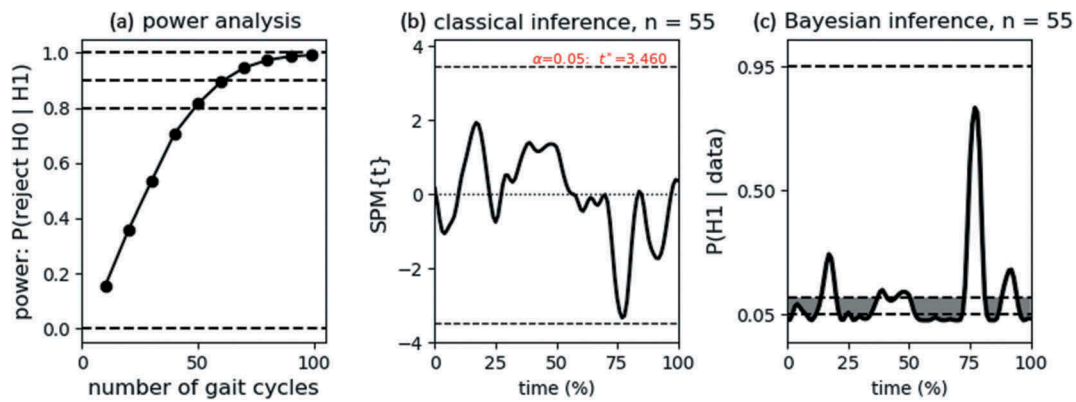


Figure 4. (a) Classical power (omnibus) analysis for calculating the number of trials necessary to reject H_0 given $\alpha = 0.05$ and a minimal 2° difference between the left and right leg (H_1). Horizontal lines show the typical power criterions of 0.80 and 0.90. Panels (b) and (c) give the classical and Bayesian SPMs using the first $n = 55$ gait cycles (for which classical power > 0.80). For the Bayesian SPM, the maximal posterior error probability for which $q < 0.05$ was 0.113.

perspective, this uncertainty may be overcome by using Sequential Bayes Factor Designs (Schönbrodt et al. 2017). Rather than specifying an (unrealistic) alternative hypothesis, researchers may sample sequentially more and more subjects (or trials in a single-subject design) until a pre-defined level of evidence for either the null or alternative or both has been reached. To our knowledge, these Bayesian sampling plans have not been used for 1D-data and will require further investigation. For the presented data, a slight disadvantage of the Bayesian SPM is computation time which is shown in Table 4. The frequentist calculation times are practically zero (analytic solutions exist), while a negligible but non-zero computation time is required for the Bayesian results. The computation time increases for more complex designs because they require the calculation of several Bayes Factor objects corresponding to several potential alternative hypotheses (in the t -test case, only one alternative hypothesis was used). The size of the datasets did not seem to impact the calculations, so for most biomechanical applications (typical sample sizes < 100), the computational burden is expected to be no real problem.

Much further work needs to be performed to explore the validity and theoretical properties of this Bayesian SPM and the FDR control schemes of thresholding on the posterior probability and q -values. Bayesian inference and decision-

making are not based on controlling type I or type II error rates, and the problem of multiple testing is, therefore, less a problem than in frequentist inference (Berry and Hochberg 1999; Kruschke and Liddell 2018b). In OD statistics, p -values are typically less conservative than Bayesian methods. In datasets one and two, we saw that the clusters based on the conservative threshold were indeed smaller than the frequentist clusters, but the q^* -based clusters were a little larger than the classical ones. For the third dataset, the comparison between cluster sizes is difficult because they really signify different conclusions. Friston and Penny (2003) compared Bayesian 95%-thresholded posterior probability maps with SPM (PET and fMRI data) and saw that the Bayesian approach yielded larger supra-threshold clusters than classical SPMs.

In the future work, we should examine the feasibility of hierarchical Bayesian modeling (empirical parametric Bayes) for 1D data (Friston and Penny 2003) which can be used to practically eliminate the problem of multiple testing (Gelman et al. 2012). Future studies should also examine more appropriate priors for spatiotemporally correlated data for SPM applications (Lee et al. 2017; Sidén et al. 2017). In the present proposition, we took a default JZS prior at each point in the time series which does not take the temporal correlation into account. One possibility we see in this respect is the introduction of a ‘dynamic prior’

Table 4. Computation time for frequentist and Bayesian SPM tests.

Statistical test and dataset size	Computing time
	(DELL, i7 processor, Linux Mint 19 operating system)
Independent-sample t -test	Classical SPM (Python): 0.025 s
SimulatedTwoLocalMax [12 x 101]	Bayesian SPM (RStudio): 1.325 s
Paired-sample t -test	Classical SPM (Python): 0.025 s
PlantArchAngle [20 x 101]	Bayesian SPM (RStudio): 1.097 s
Paired-sample t -test	Classical SPM (Python): 0.025 s
GaitSymmetry [198 x 101]	Bayesian SPM (RStudio): 1.580 s

where the posterior density at t_i can serve as the prior for the Bayes Factor calculation at t_{i+1} . Given the temporal correlation of the data, the effect sizes at neighboring time samples are bound to be similar and therefore the posterior at t_i will be a good estimate for that at t_{i+1} .

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Ben Serrien  <http://orcid.org/0000-0001-6538-9051>

References

- Adler RJ, Taylor JE. 2007. Random fields and geometry. New York (NY): Springer-Verlag.
- Berry DA, Hochberg Y. 1999. Bayesian perspectives on multiple comparisons. *J Stat Plan Inference*. 82:215–227.
- Booth BG, Keijsers NLW, Sijbers J, Huysmans T. 2018. STAPP: spatiotemporal analysis of plantar pressure measurements using statistical parametric mapping. *Gait Posture* [Internet]. 63:268–275. doi:10.1016/j.gaitpost.2018.04.029
- Cohen J. 1988. Statistical power analysis for the behavioral sciences. 2nd ed. New York (NY): Lawrence Erlbaum Associates.
- Cohen J. 1994. The earth is round ($p < .05$). *Am Psychol*. 49:997–1003.
- De Villemereuil P, Frichot E, Bazin E, François O, Gaggiotti OE. 2014. Genome scan methods against more complex models: when and how much should we trust them? *Mol Ecol*. 23:2006–2019.
- Dienes Z, Mclatchie N. 2018. Four reasons to prefer Bayesian analyses over significance testing. *Psychon Bull Rev*. 25:207–218.
- Etz A, Vandekerckhove J. 2018. Introduction to Bayesian inference for psychology. *Psych Bull Rev*. 25(1):5–34. doi:10.3758/s13423-017-1262-3.
- Friston K. 2007. A short history of SPM. In: Friston K, Ashburner J, Kiebel S, Nichols T, Penny W, editors. *Statistical Parametric Mapping – the analysis of functional brain images*. Amsterdam: Elsevier.
- Friston KJ, Penny W. 2003. Posterior probability maps and SPMs. *Neuroimage*. 19:1240–1249.
- Gelman A, Hill J, Yajima M. 2012. Why we (Usually) don't have to worry about multiple comparisons. *J Res Educ Eff*. 5:189–211.
- Käll L, Storey JD, Maccoss MJ, Noble WS. 2008. Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res*. 7:40–44.
- Kruschke JK, Liddell TM. 2018a. Bayesian data analysis for newcomers. *Psych Bull Rev*. 25(1):155–177. doi:10.3758/s13423-017-1272-1.
- Kruschke JK, Liddell TM. 2018b. The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psych Bull Rev*. 25(1):178–206. doi:10.3758/s13423-016-1221-4.
- Lee K-J, Hsieh S, Wen T. 2017. Spatial Bayesian hierarchical model with variable selection to fMRI data. *Spat Stat* [Internet]. 21:96–113. doi:10.1016/j.spasta.2017.06.002
- Li W, Kornak J, Harris T, Keyak J, Li C, Lu Y, Cheng X, Lang T. 2009. Identify fracture-critical regions inside the proximal femur using statistical parametric mapping. *Bone*. 44(4):596–602.
- Morey RD, Rouder JN. 2011. Bayes factor approaches for testing interval null hypotheses. *Psychol Methods*. 16:406–419.
- Morey RD, Rouder JN, Jamil T, Urbanek S, Forner K, Ly A. 2018. BayesFactor 0.9.12-4.2. *Compr R Arch Netw*.
- Pataky TC. 2010. Generalized n-dimensional biomechanical field analysis using statistical parametric mapping. *J Biomech*. 43:1976–1982.
- Pataky TC. 2012. One-dimensional statistical parametric mapping in Python. *Comput Methods Biomech Biomed Engin*. 15:295–301.
- Pataky TC. 2017. Power1D: a Python toolbox for numerical power estimates in experiments involving one-dimensional continua. *PeerJ Comput Sci*.
- Pataky TC, Goulermas JY. 2008. Pedobarographic statistical parametric mapping (pSPM): A pixel-level approach to foot pressure image analysis. *J Biomech*. 41:2136–2143.
- Pataky TC, Robinson MA, Vanrenterghem J. 2013. Vector field statistical analysis of kinematic and force trajectories. *J Biomech*. 46:2394–2401.
- Pataky TC, Vanrenterghem J, Robinson MA. 2015. Zero- vs. one-dimensional, parametric vs. non-parametric, and confidence interval vs. hypothesis testing procedures in one-dimensional biomechanical trajectory analysis. *J Biomech* [Internet]. 48:1277–1285. doi:10.1016/j.jbiomech.2015.02.051
- Pataky TC, Vanrenterghem J, Robinson MA. 2016. The probability of false positives in zero-dimensional analyses of one-dimensional kinematic, force and EMG trajectories. *J Biomech* [Internet]. 49:1468–1476. doi:10.1016/j.jbiomech.2016.03.032
- Poole KES, Skingle L, Gee AH, Turmezei TD, Johannesdottir F, Blesic K, Rose C, Vindlacheruvu M, Donell S, Vaculik J, et al. 2017. Focal osteoporosis defects play a key role in hip fracture. *Bone* [Internet]. 94:124–134. doi:10.1016/j.bone.2016.10.020
- Robinson MA, Vanrenterghem J, Pataky TC. 2015. Statistical Parametric Mapping (SPM) for alpha-based statistical analyses of multi-muscle EMG time-series. *J Electromyogr Kinesiol* [Internet]. 25:14–19. <http://www.ncbi.nlm.nih.gov/pubmed/25465983>.
- Rouder JN, Morey RD, Speckman PL, Province JM. 2012. Default Bayes factors for ANOVA designs. *J Math Psychol* [Internet]. 56:356–374. doi:10.1016/j.jmp.2012.08.001
- Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev*. 16:225–237.
- Schönbrodt FD, Wagenmakers EJ, Zehetleitner M, Perugini M. 2017. Sequential hypothesis testing with Bayes factors: efficiently testing mean differences. *Psychol Methods*. 22:322–339.
- Sidén P, Eklund A, Bolin D, Villani M. 2017. Fast Bayesian whole-brain fMRI analysis with spatial 3D priors.

- Neuroimage [Internet]. 146:211–225. doi:[10.1016/j.neuroimage.2016.11.040](https://doi.org/10.1016/j.neuroimage.2016.11.040)
- Storey JD. 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat.* 31:2013–2035.
- Wagenmakers EJ, Marsman M, Jamil T, Verhagen J, Love J, Selker R, Gronau QF, Epskamp S, Matzke D, Rouder JN, et al. 2018. Bayesian inference for psychology. Part I : theoretical advantages and practical ramifications. *Psych Bull Rev.* 25(1):35–57. doi:[10.3758/s13423-017-1343-3](https://doi.org/10.3758/s13423-017-1343-3).
- Yu A, Carballido-Gamio J, Wang L, Lang TF, Su Y, Wu X, Yang M, Wei J, Yi C, Cheng X. 2017. Spatial differences in the distribution of bone between femoral neck and trochanteric fractures. *J Bone Miner Res.* 32:1672–1680.