Zero- vs. one-dimensional, parametric vs. non-parametric, and confidence interval vs. hypothesis testing procedures in one-dimensional biomechanical trajectory analysis

Todd C. Pataky[1], Jos Vanrenterghem[2], and Mark A. Robinson[2]

[1]Department of Bioengineering, Shinshu University, Japan
[2]Research Institute for Sport and Exercise Sciences, Liverpool John Moores University, UK

March 4, 2015

Abstract

Biomechanical processes are often manifested as one-dimensional (1D) trajectories. It has been shown that 1D confidence intervals (CIs) are biased when based on 0D statistical procedures, and the non-parametric 1D bootstrap CI has emerged in the Biomechanics literature as a viable solution. The primary purpose of this paper was to clarify that, for 1D biomechanics datasets, the distinction between 0D and 1D methods is much more important than the distinction between parametric and non-parametric procedures. A secondary purpose was to demonstrate that a parametric equivalent to the 1D bootstrap exists in the form of a random field theory (RFT) correction for multiple comparisons. To emphasize these points we analyzed six datasets consisting of force and kinematic trajectories in one-sample, paired, two-sample and regression designs. Results showed, first, that the 1D bootstrap and other 1D non-parametric CIs were qualitatively identical to RFT CIs, and all were very different from 0D CIs. Second, 1D parametric and 1D non-parametric hypothesis testing results were qualitatively identical for all six datasets. Last, we highlight the limitations of 1D CIs by demonstrating that they are complex, design-dependent, and thus non-generalizable. These results suggest that (i) analyses of 1D data based on 0D models of randomness are generally biased unless one has explicitly specified an *a priori* 0D variable, and (ii) parametric and non-parametric 1D hypothesis testing provide an unambiguous framework for analysis when one's hypothesis explicitly or implicitly pertains to whole 1D trajectories.

Corresponding Author:
Todd Pataky, tpataky@shinshu-u.ac.jp, T.+81-268-21-5609, F.+81-268–21-5318

Keywords: *bootstrap confidence interval; kinematics; ground reaction force; statistical parametric mapping; random field theory; time series analysis*

1 Introduction

Biomechanical processes are often described using one-dimensional (1D) kinematic and force trajectories. Since these trajectories can be complex, it can be difficult to objectively specify an *a priori* method for analyzing those trajectories. Many studies therefore adopt an *ad hoc* approach: visualize the trajectories and then extract some summary scalar — which was not specified prior to the experiment — to test statistically.

Unfortunately this approach is biased for the following reasons: all statistical analyses require a model of randomness — from that model one computes the probability that random data would produce the observed result (i.e. the p value). If one's *a priori* hypothesis pertains to zero-dimensional (0D) scalars, then a 0D model of randomness is appropriate. However, if one's hypothesis pertains to 1D trajectories, then objectivity obliges one to employ a 1D model of randomness — one which describes how random 1D trajectories behave. Since probabilistic conclusions stemming from 0D and 1D models generally differ (Pataky et al., 2013), it is biased to test a 1D hypothesis using a 0D model.

A randomness model may separately be categorized as either parametric or non-parametric. Parametric models are constructed by first assuming the nature of the random distribution (usually Gaussian), and then computing a small number of parameters (usually the mean and standard deviation — SD) which characterize that distribution and thus its random behavior. In contrast, non-parametric models (Good, 2005) are generally not based on any specific distribution, and are instead constructed using experimental data. When the data do indeed come from a Gaussian distribution, then the parametric and non-parametric models converge (Appendix A), and when the data are non-Gaussian parametric procedures are generally not valid. There are thus four categories of randomness models to consider: 0D parametric, 0D non-parametric, 1D parametric and 1D non-parametric.

In the Biomechanics literature four relevant approaches have emerged: (i) error clouds — often surrounding a mean trajectory (McGinley et al., 2009) (ii) the bootstrap confidence interval (CI) (Olshen et al., 1989; Lenhoff et al., 1999; Peterson et al., 2000; Duhamel et al., 2004), (iii) functional data analysis (Ramsay and Silverman, 2005), and (iv) random field theory (RFT) (Adler and Taylor, 2007; Pataky et al., 2013). Method (ii) is non-parametric and the rest are parametric. Since (iii) and (iv) may be regarded as equivalent from a hypothesis-testing perspective (Appendix B), this paper focusses on only RFT; RFT may be considered simpler because it requires fewer parameters.

The theoretical inadequacy of (i) error clouds, including the SD cloud, is fortunately easy to address: they do not stem from a randomness model. While error clouds objectively quantify trajectory variability, experimental design complexities conspire to dissolve the connection between variability and probability

(Schwartz et al., 2004). Since error clouds cannot support probabilistic claims they must be regarded as descriptive or exploratory in nature.

Unlike error clouds, (ii) CIs do stem from randomness models. Nevertheless it has been shown that 1D CIs are invalid when based on 0D randomness (Lenhoff et al., 1999; Duhamel et al., 2004). The 1D bootstrap CI (Olshen et al., 1989; Lenhoff et al., 1999; Duhamel et al., 2004) is a viable solution because it models 1D randomness in the behavior of the trajectory-wide maximum under random resamplings. The available literature has explored 0D parametric CIs vs. 1D non-parametric CIs (Lenhoff et al., 1999; Duhamel et al., 2004; Gravel et al., 2010; Dixon et al., 2013; Cutti et al., 2014), and has also explored 0D vs. 1D hypothesis testing using RFT (Pataky et al., 2013), but to our knowledge there has previously been no systematic comparison of 0D vs. 1D procedures, parametric vs. non-parametric results, and CIs vs. hypothesis testing.

The primary purpose of this study was to elucidate the theoretical framework of 0D vs. 1D statistical procedures. Specifically, we sought to clarify that choosing 0D vs. 1D procedures is statistically much more important than choosing parametric vs. non-parametric procedures because differences in 0D vs. 1D results are generally much larger than differences in parametric vs. non-parametric results. We also sought to clarify that, in contrast to 1D CIs which are complex and non-generalizable, 1D hypothesis testing results can be presented consistently across all experimental designs.

2 Methods

2.1 Datasets

Three simulated and three experimental datasets consisting of J scalar trajectory responses normalized to Q discrete points were analyzed (Table 1). Since the simulated datasets are artificial readers are encouraged to judge their relevance to real data.

Datasets A and B (Fig.1) mimic a one-sample experiment. These datasets were constructed by adding ten smoothed, amplified Gaussian noise trajectories (Fig.1a) to two true population means (Fig.1b). Dataset B's slightly larger signal at time=80% is evident in both the resulting datasets (Fig.1c,d) and their summary statistics (Fig.1e,f).

Dataset C (Fig.2) mimics a regression design with one independent variable x. To ten true signals (Fig.2a), whose maxima were perfectly correlated with x (Fig.2b), we added smooth Gaussian noise (Fig.2c) to yield the final dataset (Fig.2d). The ten responses were divided into two groups (Fig.2a) to compare categorical and continuous treatments of x.

Dataset D (Fig.3a) (Neptune et al., 1999) consisted of within-subject mean knee flexion trajectories in side-shuffle vs. v-cut tasks during stance. Dataset E (Fig.3b) (Besier et al., 2009) consisted of stance-phase medial gastrocnemius forces during walking in 16 Controls vs. 27 Patello-Femoral Pain (PFP) patients, as estimated by Besier et al. from EMG-driven forward-dynamics simulations. Dataset F (Fig.3c) (Dorn et al., 2012) consisted of left-foot anterior/posterior ground reaction forces (GRF) in one subject running/sprinting at four different speeds.

2.2 General statistical calculations

For simplicity this study focusses on the t statistic. All calculations employed a Type I error rate of α=0.05 and were implemented all in Python 2.7 using Canopy 1.4 (Enthought Inc., Austin, USA) and the open-source software package 'spm1d' (Pataky, 2012).

2.2.1 0D and 1D t statistics

Definitions of 1D t statistics are trivial extensions of their 0D definitions to a 1D domain q, where q represents time in the aforementioned datasets. For example, the 1D one-sample t statistic is:

$$t(q) = \frac{\overline{y}(q)}{s(q)/\sqrt{J}} \tag{1}$$

where $\overline{y}$, s and J are the sample mean, sample standard deviation, and sample size, respectively. This 1D t trajectory can be assembled simply by computing the t statistic value separately at each time point q, thereby approximating the continuous $t(q)$ trajectory just like computing the mean separately at each point approximates the continuous mean trajectory. Definitions of the t statistic for other designs are provided as Supplementary Material (Appendix C).

2.2.2 Parametric 0D and 1D critical thresholds

The critical 0D t statistic $t^*_{0\mathrm{D}}$ is given as the solution to:

$$P(t > t^*_{0\mathrm{D}}) = \int_{t^*_{0\mathrm{D}}}^{\infty} f_{0\mathrm{D}}(x)dx = \alpha \tag{2}$$

where $f_{0\mathrm{D}}(x)$ is the usual 0D t statistic's probability density function (Appendix D) and $P(t > t^*_{0\mathrm{D}})$ is the probability that the t statistic will exceed $t^*_{0\mathrm{D}}$ if the underlying data are 0D Gaussian. In classical hypothesis testing the null hypothesis is rejected if the observed 0D t value exceeds $t^*_{0\mathrm{D}}$.

The critical 1D test statistic t^*_{1D} is given by RFT (Adler and Taylor, 2007) as the solution to:

$$P\Big(t(q)_{\max} > t^*_{1D}\Big) = 1 - \exp\left(-\int_{t^*_{1D}}^{\infty} f_{0D}(x)dx - ED\right) = \alpha \tag{3}$$

where $t(q)_{\max}$ is the maximum value of the 1D t trajectory and where ED is the smoothness-dependent Euler density function (Worsley et al., 2004; Friston et al., 2007). Analogous to the 0D form, Eqn.3 represents the probability that $t(q)_{\max}$ exceeds t^*_{1D} when the underlying data are smooth 1D Gaussian, and in classical hypothesis testing the null hypothesis is rejected if the observed $t(q)_{\max}$ value exceeds t^*_{1D}.

Last, we computed the critical 0D Bonferroni threshold $t^*_{0D\_Bonf}$ as the solution to:

$$P(t > t^*_{0D\_Bonf}) = \int_{t^*_{0D\_Bonf}}^{\infty} f_{0D}(x)dx = 1 - (1-\alpha)^{(1/Q)} \tag{4}$$

Note that the Bonferroni threshold assumes Q independent tests. For smooth 1D data this is clearly a poor assumption because neighboring values in time are correlated. We nonetheless include $t^*_{0D\_Bonf}$ in our initial analyses to demonstrate that it is too extreme; provided the 1D trajectories are smooth, the three critical thresholds are related as follows: $t^*_{0D} < t^*_{1D} < t^*_{0D_{Bonf}}$. Note that although $t^*_{0D_{Bonf}}$ considers the entire 1D domain q, it only uses the 0D probability density function and fails to consider 1D smoothness; therefore only the RFT threshold (Eqn.3) is labeled "1D".

2.2.3 Non-parametric 0D and 1D critical thresholds

Two non-parametric methods — the bootstrap and the permutation method (Good, 2005) — were used to estimate both t^*_{0D} and t^*_{1D}. Descriptions of the 0D bootstrap and permutation methods are provided as Supplementary Material (Appendix E). The 1D bootstrap is described in detail elsewhere (Lenhoff et al., 1999). The 1D permutation method followed Nichols and Holmes (2002) and is summarized in Fig.4a–c. Note that nD parametric and nD non-parametric methods are conceptually identical in that both describe random nD behavior to yield $t^*_{n\mathrm{D}}$. Moreover, the nD non-parametric results are expected to converge to the nD parametric results when the underlying data are nD Gaussian (Appendix A).

2.2.4 0D and 1D confidence intervals

Substituting the critical 0D threshold t^*_{0D} into Eqn.1 yields the height h of the one-sample 0D CI:

$$h_{0D} = t^*_{0D}\frac{s}{\sqrt{J}} \tag{5}$$

CI heights for two-sample and regression-designs similarly follow from the design-dependent definitions of the t statistic (Table 3, Appendix F). Heights of 1D CIs are given simply by substituting t^*_{1D} for t^*_{0D} in CI height calculations.

2.3 Specific dataset analyses

For Datasets A and B we sought to compare 0D vs. 1D, parametric vs. non-parametric, bootstrap vs. permutation and CI vs. hypothesis testing results. We thus computed seven different critical one-sample t values for both datasets: (#1–#3) parametric versions of t^*_{0D}, t^*_{1D} and $t^*_{0D\_Bonf}$, then both bootstrap and permutation versions of both (#4,#5) t^*_{0D} and (#6,#7) t^*_{1D}. We then constructed the associated CIs and qualitatively compared all CIs and hypothesis testing results.

For Dataset C we sought to demonstrate two points: (1) since this is a regression design, neither 0D nor 1D CIs are suitable when the datum is the mean 1D trajectory, (2) unlike 1D CIs, 1D hypothesis testing results can be presented identically across designs. We first constructed the narrowest possible CIs (0D CIs) to emphasize that even these cannot capture probabilistic meaning in regression designs. Next we qualitatively compared two-sample and regression hypothesis testing results for 0D, 1D parametric, 1D non-parametric and 0D Bonferroni thresholds.

For Datasets D–F we sought to emphasize both (i) the generalizability of 1D hypothesis testing procedures, and (ii) the similarities between 1D parametric and 1D non-parametric results in real 1D experimental datasets. Since the bootstrap is unsuitable for arbitrary hypothesis testing (Good, 2005) we conducted only 1D parametric and 1D permutation tests whose results we compared qualitatively.

3 Results

3.1 0D vs. 1D methods (Datasets A and B)

The three 0D CIs were qualitatively identical, and the three 1D CIs were also qualitatively identical, but the 0D CIs were considerably different from both the 1D CIs and the 0D Bonferroni CI (Fig.5a,b). The cause of these differences is the underlying randomness model. The 0D parametric model assumes 0D Gaussian randomness, and the 0D non-parametric procedures discretely approximate the same randomness. Similarly, the 1D (RFT) parametric model assumes 1D Gaussian randomness and describes the behavior of the trajectory maximum, and the 1D non-parametric procedures discretely approximate the same 1D randomness.

The 0D Bonferroni result assumes 0D randomness, but also corrects for Q=101 independent tests across the time domain. Since the data are temporally smooth, adjacent time samples are clearly not independent and thus the Bonferroni correction is overly conservative as has been noted previously (Duhamel et al., 2004).

One-sample hypothesis testing results (Fig.5c,d) mirrored the CI results (Fig.5a,b). In particular, both the 0D CIs and the 0D hypothesis testing results reached significance at approximate times of 15% and 75%. In contrast, the 1D CIs and 1D hypothesis testing results reached significance only for Dataset B and only at 75% time (p=0.037). The Bonferroni-corrected results failed to reach significance in any dataset, emphasizing its overly conservative nature. These results emphasize that CIs are equivalent to one-sample t tests, and also that the threshold-crossing behavior is somewhat clearer for the hypothesis tests (Fig.5d).

3.2 CIs vs. hypothesis tests (Dataset C)

One-sample 0D CIs failed to separate the groups (Fig.6a). Nevertheless both 0D regression (Fig.6b) and a 0D two-sample test (Fig.6c – lower threshold) reached significance. This disagreement is explained by the CI's complex design-dependence (Table 3). In contrast to CIs, both the two-sample test and regression results could be presented in an identical, unambiguous format as a t trajectory with critical thresholds (Fig.6c,d). This result emphasizes that 1D CIs are less generalizable than 1D hypothesis testing.

Note that, even if the 0D CIs in Fig.6a had been constructed more robustly using a 1D two-sample model, the results would be incorrect because the independent variable (x) is continuous. In this case the two-sample results fail to reach significance (Fig.6c) but the regression results do (Fig.6d).

3.3 Parametric vs. non-parametric 1D methods (Datsets D–F)

For each of the experimental datasets, 1D parametric and 1D non-parametric results were qualitatively identical (Fig.7). In particular, (i) the null hypothesis was rejected in all cases, (ii) essentially the same suprathreshold temporal windows were identified, and (iii) similar probabilities were obtained for suprathreshold clusters. Unlike CI results, these results are reportable in an identical manner across arbitrary experimental designs, further emphasizing the generalizability of 1D hypothesis testing.

4 Discussion

4.1 0D vs. 1D methods

This study's results suggest most broadly that choosing between 0D and 1D methods is likely much more important than choosing between parametric and non-parametric methods when analyzing 1D biomechanical data. From the discrepancies amongst the 0D and 1D results (Figs.5–6) it is clear that 0D procedures inaccurately model smooth 1D trajectory variance (Fig.1a, Fig.2c) which characterizes most 1D biomechanical datasets (Duhamel et al., 2004). One may therefore be tempted to ask: "which is the correct method?" That question is important and easy to answer: both 0D and 1D methods are correct, but they cannot both be simultaneously correct. Since a method's validity rests on its assumptions' justifiability, and since 0D and 1D methods make different assumptions (i.e. 0D randomness vs. 1D randomness), they cannot both be valid for the same dataset. A 0D procedure is perfectly justifiable if one formulates a specific 0D hypothesis prior to conducting a 1D experiment, and then analyzes only those specific 0D data (Pataky et al., 2013); in this case 1D probabilistic methods would be unjustified. On the other hand, if one does not have a specific 0D hypothesis, then by definition one's hypothesis implicitly pertains to the whole 1D trajectory, in which case we'd argue that only 1D procedures are justifiable. More simply, one's *a priori* hypothesis must drive one's analysis and not the other way around.

4.2 Parametric vs. non-parametric procedures

The choice between parametric and non-parametric procedures had negligible effects on the current results (Fig.5–7) suggesting that RFT's assumption of 1D Gaussian randomness was a reasonable one. We have separately observed similar agreement between parametric and non-parametric 1D procedures for a much greater variety of 1D Biomechanics data, including EMG time series (Robinson et al., 2015), suggesting that the choice between 0D and 1D models appears to be more important than the choice between parametric and non-parametric models.

The main advantage of (parametric) RFT procedures is that, since they assume an analytical model of 1D randomness, they are very fast. Non-parametric procedures are generally much slower because they build randomness models iteratively based on experimental data. As examples, for the relatively small Dataset B our RFT and 1D permutation implementations required an average of 0.020 s and 0.130 s, respectively. For the larger Dataset E, the durations were 0.023 s and 5.5 s, respectively.

The main disadvantage of RFT procedures is that, like 0D parametric procedures since it assumes a

Gaussian model of randomness, and that assumption may be violated. One should therefore check adherence to the normality assumption when employing parametric procedures, either explicitly through a test for normality, or implicitly by checking for agreement between parametric and non-parametric results. However, such normality checks may be moot: 1D biomechanical trajectories are generally smoothed prior to analysis (Bisseling and Hof, 2006; Kristianslund et al., 2013), and smoothing, by definition, mitigates outliers and drives the data toward normality. Non-parametric 1D procedures are generally valid irrespective of the underlying distribution.

A second disadvantage is that parametric procedures are less flexible than non-parametric procedures. In particular, it has been shown that SD continuum smoothing can enhance the signal:noise ratio because point-by-point SD estimations are generally poor, especially for small sample sizes (Nichols and Holmes, 2002). Such smoothing is valid for non-parametric but not parametric procedures.

4.3 CIs vs. hypothesis tests

The present CI results (Fig.5a,b) agree with previous findings that 1D CIs better model 1D variance than do 0D CIs (Lenhoff et al., 1999; Duhamel et al., 2004; Gravel et al., 2010; Cutti et al., 2014). Although those studies' 1D methods were limited to the 1D bootstrap, our results suggest that the 1D CI can also be constructed in at least two additional ways: parametrically using RFT, and non-parametrically using the permutation procedure of Nichols and Holmes (2002).

Also unlike previous studies, this study's results (Figs.5–6) suggest that 1D CIs are a poorer choice than hypothesis tests, primarily because CIs are suitable only for very simple one- and two-sample designs. Even within those simple designs, CIs have complex design- and datum-dependent interpretations (Table 3), so when reporting 1D CIs graphically one must explicitly specify both the design and the datum one employed to construct the CI. We'd argue that this unnecessarily complicates cross-study comparisons. In contrast, 1D hypothesis testing accommodates arbitrary experimental designs and yet presents 1D results in a much more consistent manner across studies (Fig.7). It has been shown elsewhere that 1D hypothesis testing results can be presented identically for multivariate (vector) trajectories (Pataky et al., 2013) and thus most generally to MANCOVA (Worsley et al., 2004).

The primary advantage of CIs is that they present probabilistic results in the context of the original data, with identical units (Batterham and Hopkins, 2006). This clearly makes the CI valuable for data visualization and exploration. However, since CIs embody no unique probabilistic information relative to hypothesis testing (Table 3), and since CIs are difficult or impossible to interpret in arbitrary experimental

designs (Fig.6), we'd argue that hypothesis testing should preferentially be adopted where possible.

4.4 Limitations of 1D methods

A key assumption of all 1D methods is that trajectories have been appropriately smoothed and registered (i.e. temporally normalized) (Sadeghi et al., 2003). This may be important considering that smoothing algorithm particulars can non-trivially affect biomechanical interpretations (Bisseling and Hof, 2006; Kristianslund et al., 2013), and that nonlinear registration procedures can substantially reduce 1D trajectory variability (Sadeghi et al., 2003). Nevertheless, since these assumptions pertain to data processing and not to statistical inference, they are not unique to 1D analyses, so should be scrutinized for both 0D and 1D analyses. As a rule of thumb, if one is confident that one's mean trajectories are unbiased by smoothing/registration particulars, then by definition 1D inference procedures are valid.

As an anecdotal exploration of (mis-)registration effects, consider that Dataset F appears to contain misregistered early-stance posterior GRF extrema (Fig.3c). In this particular case adopting a nonlinear registration procedure has only moderate quantitative effects on the results and no real qualitative effect (Appendix G). Nevertheless, registration — and more generally the assumption of data homology — requires continued scrutiny for both 0D and 1D analyses.

Partially mitigating both smoothing and registration-related effects is RFT's generalizability to nD continua (Friston et al., 2007; Pataky, 2010). Since both smoothing and registration are generally parameterizable (e.g. smoothing kernel width) the 1D test statistic continuum can be extended to $(K+1)$ dimensions, where K is the number of smoothing/registration parameters (Worsley et al., 1996). Analysis of the resulting $(K+1)$-dimensional test statistic continuum would constitute a systematic sensitivity analysis of smoothing/registration assumptions.

Last, a potentially serious limitation of 1D methods exists for routine biomechanical analyses. Many studies measure a variety of variables including, for example: 3D angles at multiple joints, 3D reaction forces, and electromyographical time series. While 1D methods can handle multivariate trajectories in general (Pataky et al., 2013), the main problem is that statistical power reduces as the number of 0D or 1D variables increases and the sample size remains small. There is no statistical theory of which we are aware that can maintain statistical power in the face of both small sample sizes and an arbitrarily large barrage of 1D measurements. Exploratory analyses (e.g. 1D mean and SD interpretations) may be necessary to formulate specific, feasibly testable hypotheses regarding sub-components of such datasets.

4.5 Summary

This study's results suggest that 0D methods inaccurately model the behavior of smooth, random 1D trajectories. Since one's primary scientific reporting obligation is to specify the probability with which random data could produce the observed result, these results also suggest that 1D methods should be used to analyze 1D data except when one has a specific 0D hypothesis prior to conducting an experiment. Finally, as compared with 1D CIs, 1D hypothesis tests represent a simpler, more generalizable basis for forming probabilistic conclusions regarding smooth 1D biomechanical trajectories. While parametric 1D (RFT) procedures may be preferable because of their speed, non-parametric 1D procedures may be necessary when deviations from normality are non-negligible.

Acknowledgments

We wish to thank Phil Dixon for helpful discussions pertaining to non-parametric 1D analyses.

Conflict of Interest

The authors report no conflict of interest, financial or otherwise.

References

Adler, R. J. and Taylor, J. E. 2007. *Random Fields and Geometry*, Springer-Verlag, New York.

Batterham, A. M. and Hopkins, W. G. 2006. Making meaningful inference about magnitudes., *International Journal of Sports Physiology and Performance* **1(1)**, 50–57.

Besier, T. F., Fredericson, M., Gold, G. E., Beaupre, G. S., and Delp, S. L. 2009. Knee muscle forces during walking and running in patellofemoral pain patients and pain-free controls, *Journal of Biomechanics* **42(7)**, 898–905, data: `https://simtk.org/home/muscleforces`.

Bisseling, R. W. and Hof, A. L. 2006. Handling of impact forces in inverse dynamics, *Journal of Biomechanics* **39(13)**, 2438–2444.

Cutti, A. G., Parel, I., Raggi, M., Petracci, E., Pellegrini, A., Accardo, A. P., Sachetti, R., and Porcellini, G. 2014. Prediction bands and intervals for the scauplo-humeral coordination based on the bootstrap and two Gaussian methods, *Journal of Biomechanics*, in press.

Dixon, P. C., Stebbins, J., Theologis, T., and Zavatsky, A. B. 2013. Spatio-temporal parameters and lower-limb kinematics of turning gait in typically developing children, *Gait and Posture* **38(4)**, 870–875.

Dorn, T. T., Schache, A. G., and Pandy, M. G. 2012. Muscular strategy shift in human running: dependence of running speed on hip and ankle muscle performance., *Journal of Experimental Biology* **215**, 1944–1956, data: `https://simtk.org/home/runningspeeds`.

Duhamel, A., Bourriez, J., Devos, P., Krystkowiak, P., Destee, A., Derambure, P., and Defebvre, L. 2004. Statistical tools for clinical gait analysis, *Gait and Posture* **20(2)**, 204–212.

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., and Evans, A. C. 1994. Assessing the significance of focal activations using their spatial extent., *Hum Brain Mapp* **1**, 210–220.

Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E., and Penny, W. D. 2007. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, Elsevier/Academic Press, Amsterdam.

Good, P. I. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd ed., Springer, New York.

Gravel, P., Tremblay, M., Leblond, H., Rossignol, S., and de Guise, J. A. 2010. A semi-automated software tool to study treadmill locomotion in the rat: from experiment videos to statistical gait analysis, *Journal of Neuroscience Methods* **190(2)**, 279–2889.

Kiebel, S. J., Poline, J., Friston, K. J., Holmes, A. P., and Worsley, K. J. 1999. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model, *NeuroImage* **10(6)**, 756–766.

Kristianslund, E., Krosshaug, T., and van den Bogert, A. J. 2013. Effect of low pass filtering on joint moments from inverse dynamics: Implications for injury prevention, *Journal of Biomechanics* **45(4)**, 666–671.

Lenhoff, M. W., Santer, T. J., Otis, J. C., Peterson, M. G., Williams, B. J., and Backus, S. I. 1999. Bootstrap prediction and confidence bands: a superior statistical method for analysis of gait data, *Gait and Posture* **9**, 10–17.

McGinley, J. L., Baker, R., Wolfe, R., and Morris, R. E. 2009. The reliability of three-dimensional kinematic gait measurements: a systematic review., *Gait and Posture* **29**, 360–369.

Neptune, R. R., Wright, I. C., and van den Bogert, A. J. 1999. Muscle coordination and function during cutting movements, *Medicine & Science in Sports & Exercise* **31(2)**, 294–302, data: `http://isbweb.org/data/rrn/`.

Nichols, T. E. and Holmes, A. P. 2002. Nonparametric permutation tests for functional neuroimaging a primer with examples, *Human Brain Mapping* **15(1)**, 1–25.

Olshen, R. A., Biden, E. N., Wyatt, M. P., and Sutherland, D. H. 1989. Gait analysis and the bootstrap, *Annals of Statistics* **17(4)**, 1419–1440.

Pataky, T. C. 2010. Generalized n-dimensional biomechanical field analysis using statistical parametric mapping, *Journal of Biomechanics* **43(10)**, 1976–1982.

Pataky, T. C . 2012. One-dimensional statistical parametric mapping in Python, *Computer Methods in Biomechanics and Biomedical Engineering* **15(3)**, 295–301.

Pataky, T. C., Robinson, M. A., and Vanrenterghem, J. 2013. Vector field statistical analysis of kinematic and force trajectories., *Journal of Biomechanics* **46(14)**, 2394–2401.

Peterson, M. G., Murray-Weir, M., Root, L., Lenhoff, M., Daily, L., and Wagner, C. 2000. Bootstrapping gait data from people with cerebral palsy., *Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems*, 57–61.

Ramsay, J. O. and Silverman, B. W. 2005. *Functional Data Analysis*, Springer, New York.

Robinson, M. A., Vanrenterghem, J., and Pataky, T. C. 2015. Statistical parametric mapping for alpha-based statistical analyses of multi-muscle EMG time-series, *Journal of Electromyography and Kinesiology*, in press.

Sadeghi, H., Mathieu, P. A., Sadeghi, S., and Labelle, H. 2003. Continuous curve registration as an intertrial gait variability reduction technique, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **11(1)**, 24–30.

Schwartz, M. H., Trost, J. P., and Wervey, R. A. 2004. Measurement and management of errors in quantitative gait data, *Gait and Posture* **20(2)**, 196–203.

Worsley, K. J., Marrett, S., Neelin, P., and Evans, A. C. 1996. Searching scale space for activation in PET images, *Human Brain Mapping* **4(1)**, 74–90.

Worsley, K. J., Taylor, J. E., Tomaiuolo, F., and Lerch, J. 2004. Unified univariate and multivariate random field theory, *NeuroImage* **23**, S189–S195.

Table 1: Dataset overview. J and Q are the numbers of responses and time nodes, respectively.

	Dataset	J	Q	Model	Link
Simulated	A	10	101	One-sample t test	www.spm1d.org/Downloads.html
	B	10	101	One-sample t test	
	C	10	101	Linear regression	
Experimental	D	8	101	Paired t test	http://isbweb.org/data/rrn/
	E	43	100	Two-sample t test	https://simtk.org/home/muscleforces
	F	8	100	Linear regression	https://simtk.org/home/runningspeeds

Table 2: Statistical procedures and randomness models used in this paper. Here μ and σ are the population mean and standard deviation, respectively. The parameters Q and W are the number of trajectory nodes and the trajectory smoothness, respectively (see text). The Bonferroni procedure assumes Q independent tests, and the RFT procedure assumes Q/W independent trajectory processes.

Number	Class	Procedure	Randomness model	Parameters
1	Parametric	Uncorrected	0D Gaussian	μ, σ
2	Parametric	Bonferroni-corrected	0D Gaussian	μ, σ, Q
3	Parametric	RFT-corrected	1D Gaussian	μ, σ, Q/W
4	Non-parametric	Uncorrected bootstrap	0D empirical	None
5	Non-parametric	Uncorrected permutation	0D empirical	None
6	Non-parametric	Corrected bootstrap	1D empirical	None
7	Non-parametric	Corrected permutation	1D empirical	None

Table 3: Significance threshold definitions for confidence intervals (CIs) and hypothesis tests (see also Appendix F). The Type I error rate α defines the critical threshold t^* which, in turn, defines the design-dependent CI height h that is added to a datum: either one sample's mean ($\overline{y}_A$) or the mean difference ($\Delta\overline{y}$). Paired and two-sample t tests assume $\overline{y}_A \geq \overline{y}_B$. Regression CIs are possible only when the datum is the regression slope or intercept. The key point is that, while the CI height is design dependent and the datum ambiguous, the hypothesis testing threshold is always t^* and its datum is always zero.

	Confidence intervals		Hypothesis tests
Datum:	$\overline{y}_A$	$\Delta\overline{y}$	0
One-sample t test	$\overline{y}_A - h_1 > 0$		$t_1 > t^*$
Paired t test	$\overline{y}_A - h_p > \overline{y}_B$ $\overline{y}_A - \frac{1}{2}h_p > \overline{y}_B + \frac{1}{2}h_p$	$\Delta\overline{y} - h_p > 0$	$t_p > t^*$
Two-sample t test	$\overline{y}_A - h_2 > \overline{y}_B$ $\overline{y}_A - \frac{1}{2}h_2 > \overline{y}_B + \frac{1}{2}h_2$	$\Delta\overline{y} - h_2 > 0$	$t_2 > t^*$
Regression			$t_r > t^*$

FIGURES

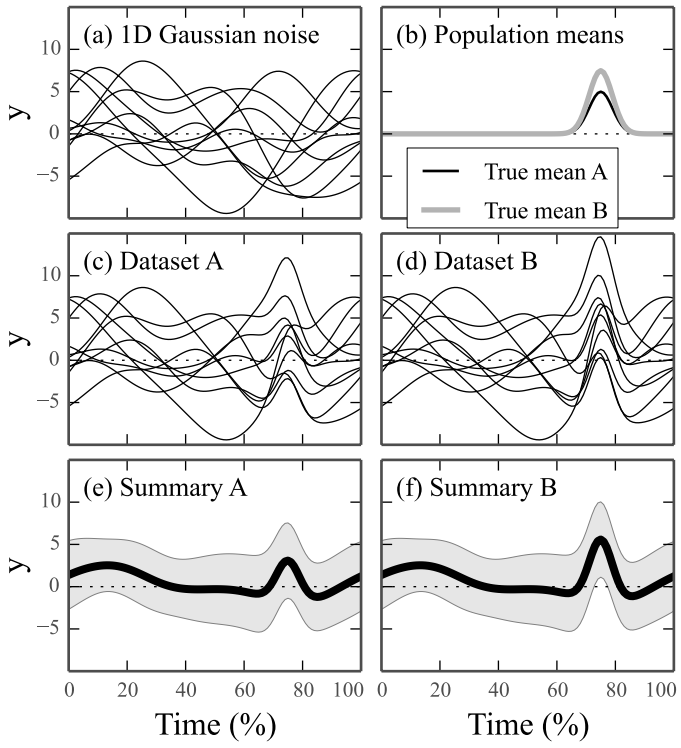

Figure 1. Datasets A and B (both simulated). (a) Smooth 1D Gaussian noise (FWHM=25%). (b) True population mean trajectories. (c,d) Final datasets: sum of true signals and noise. (e,f) Summary statistics: means with SD clouds.

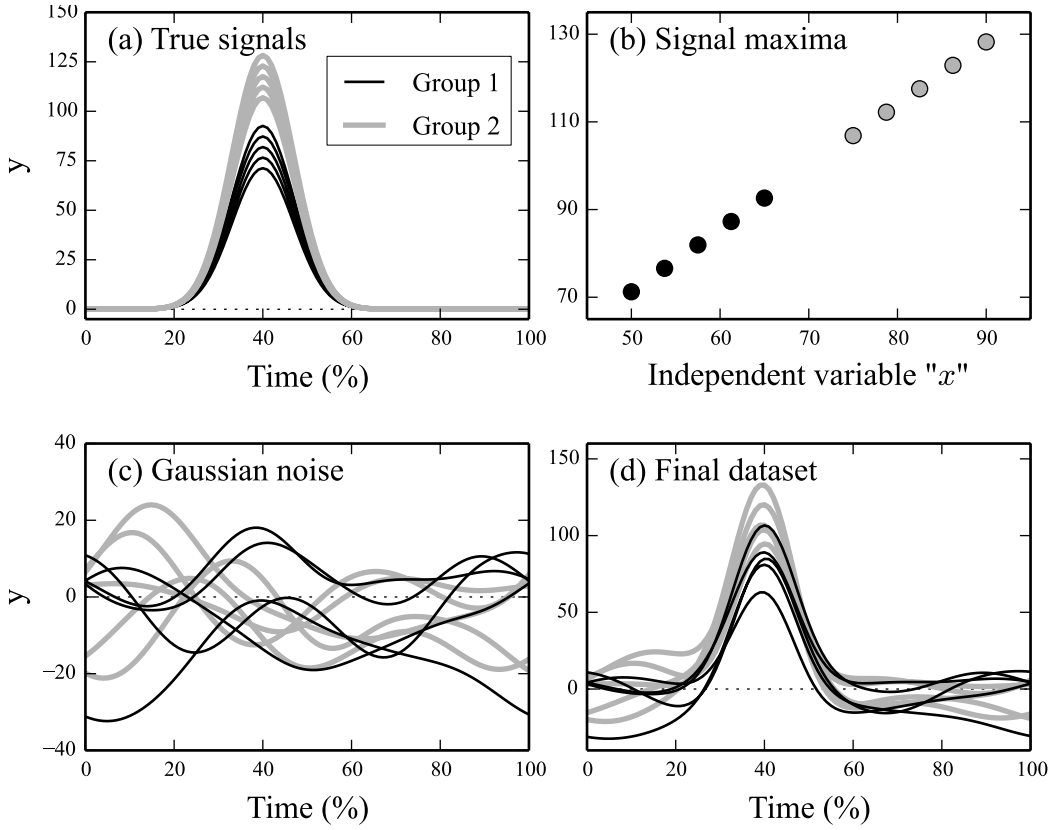

Figure 2. Dataset C (simulated). (a,b) True simulated signals exhibiting a perfect linear correlation between the independent variable and signal maxima; data are divided into two groups for a subsequent comparison between a two-sample *t* test and regression. (c) Smooth 1D Gaussian noise (FWHM=25%). (d) Final dataset: sum of true signals and noise.

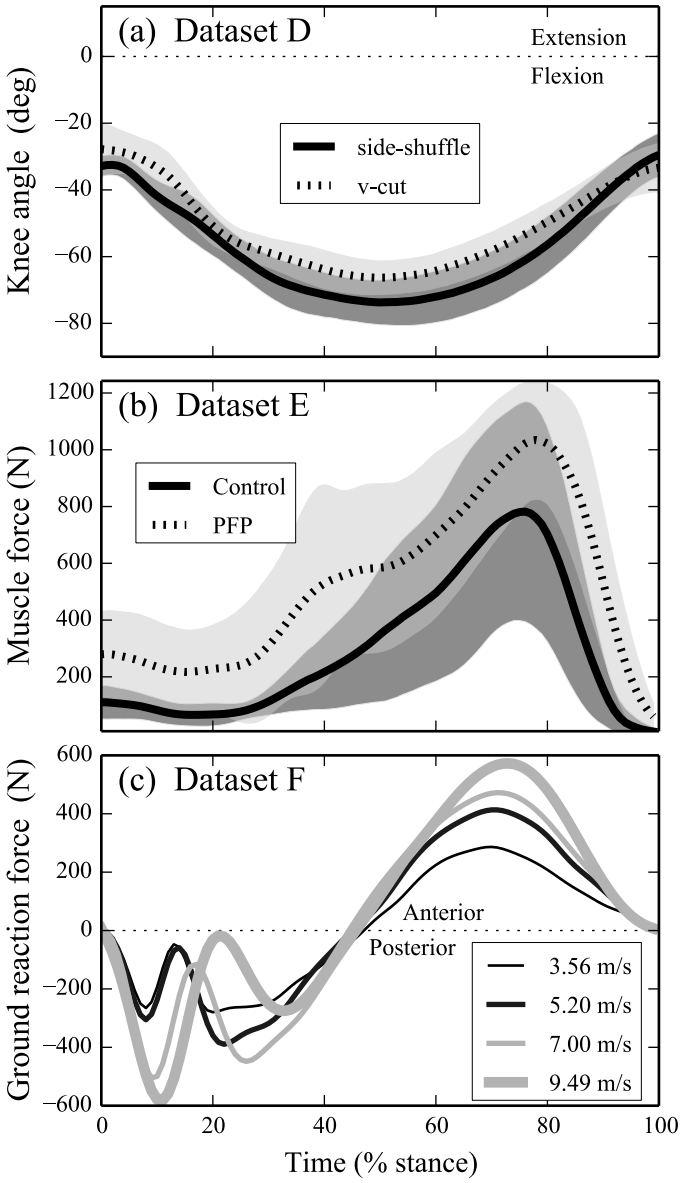

Figure 3. Experimental datasets. (a) Dataset D (Neptune et al. 1999): cross-subject mean knee angle trajectories with SD clouds in side-shuffle vs. v-cut maneuvers. (b) Dataset E (Besier et al. 2009): cross-subject mean medial gastrocnemius force trajectory, as estimated from dynamic simulation, with SD clouds. (c) Dataset F (Dorn et al. 2012): cross-trial horizontal ground reaction force trajectories in one subject during running/sprinting at various speeds.

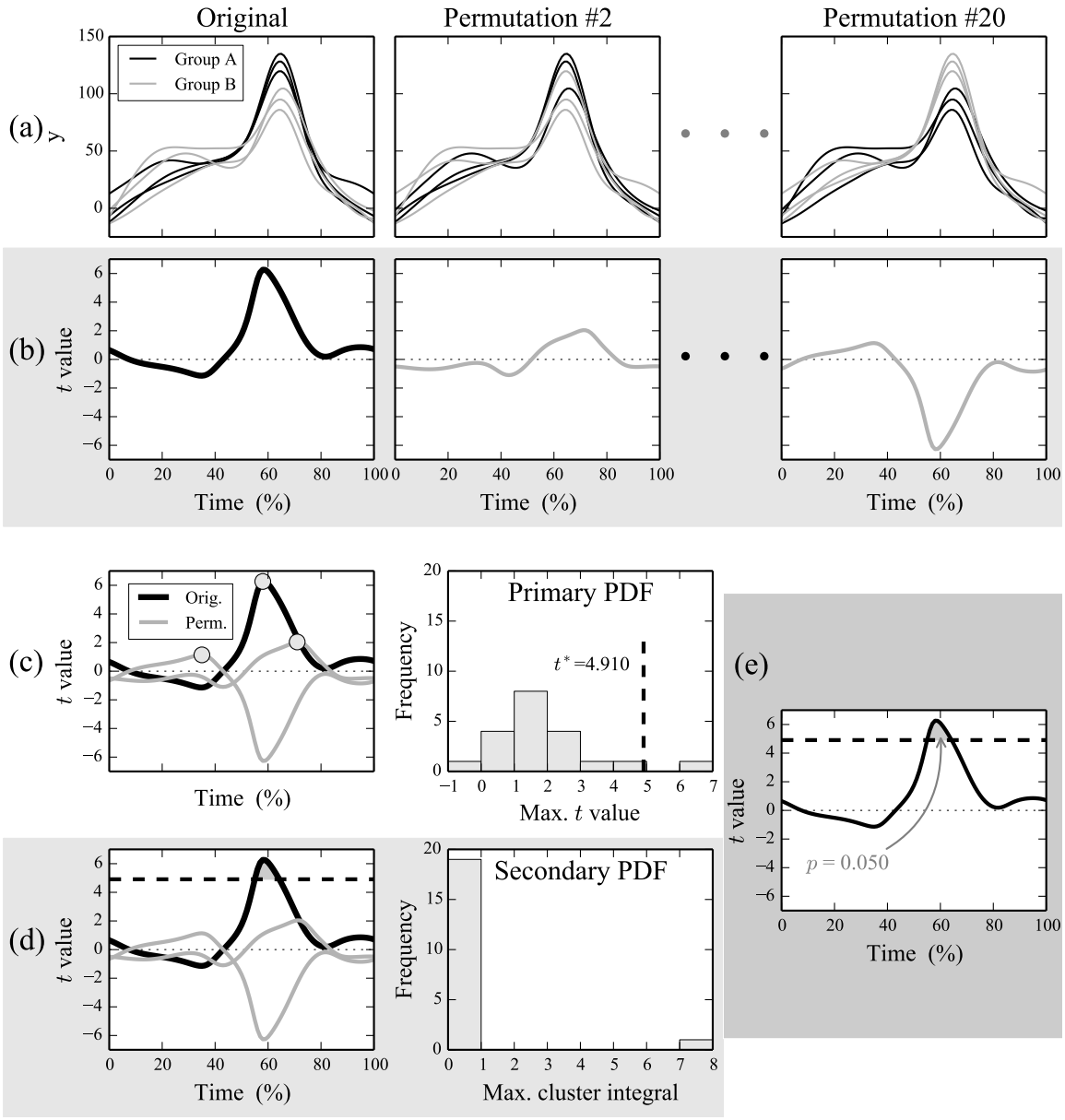

Figure 4. Non-parametric inference overview. (a) Original simulated data and two other of the 20 total permutations; the 20th permutation is the opposite of the original. (b) Test statistic (t) trajectories for each permutation. (c) The maximum t value from each t trajectory forms the primary permutation PDF, from which the critical value t^* was computed as the 95th percentile to ensure that only α=5% of all permutations exceed t^*. (d) The original t trajectory exceeds t^*, which provides sufficient evidence to reject the null hypothesis. To qualify the rejection decision, the maximum suprathreshold cluster integral from each t trajectory was extracted to form a secondary permutation PDF, from which specific cluster-level p values were computed. (e) Final hypothesis testing results. Here the original t trajectory was the only one of all 20 permutations to produce a suprathreshold cluster, so that cluster's p value is 1/20=0.05. Had RFT-based parametric inference been conducted on these data the results would have been: t^*=5.303, p=0.014.

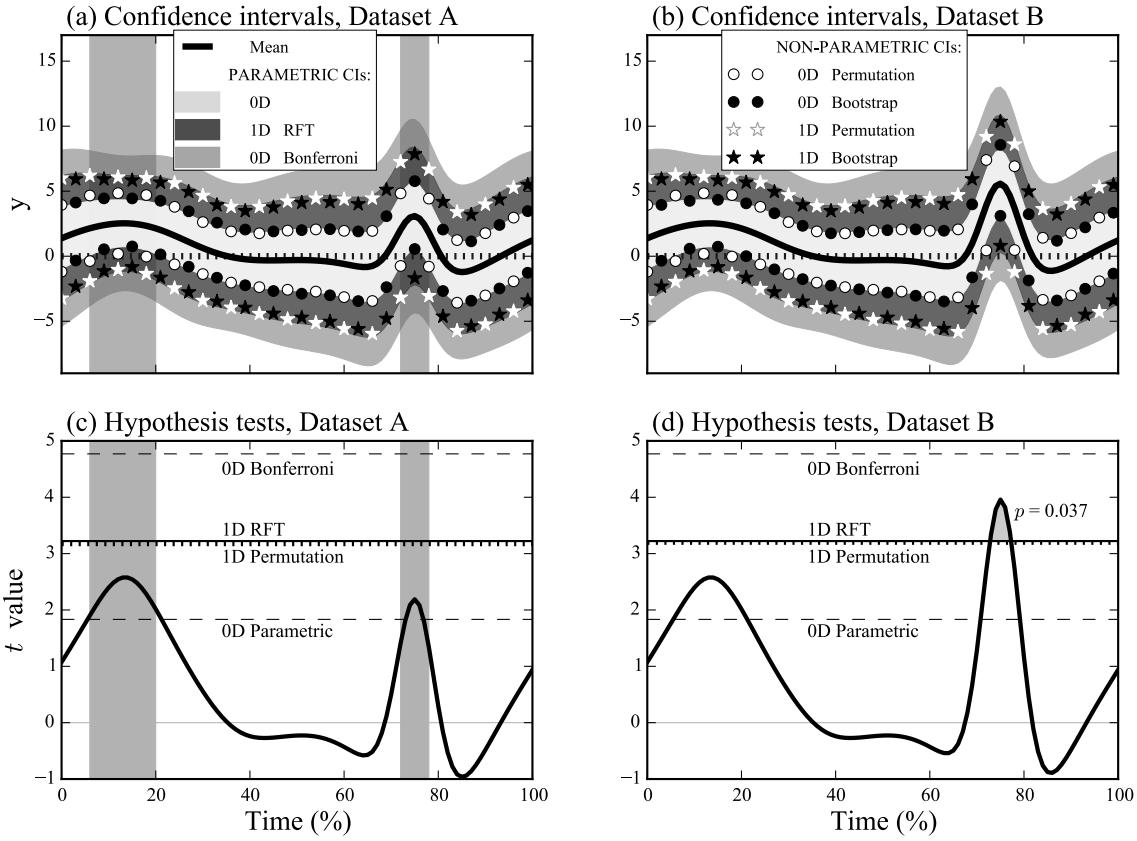

Figure 5. Results for Datasets A and B. (a,b) Seven different confidence intervals (CIs) are depicted (see Table 1) as labeled in the two legends. Dark vertical bars highlight key temporal windows discussed in the text. (c,d) Hypothesis testing results for four different tests; the null hypothesis is rejected at α=0.05 if the test statistic trajectory (thick black line) traverses the depicted threshold. In panel (d), the *p* value is the RFT result, describing the frequency with which smooth Gaussian trajectories are expected to produce a supra-threshold cluster of that temporal extent.

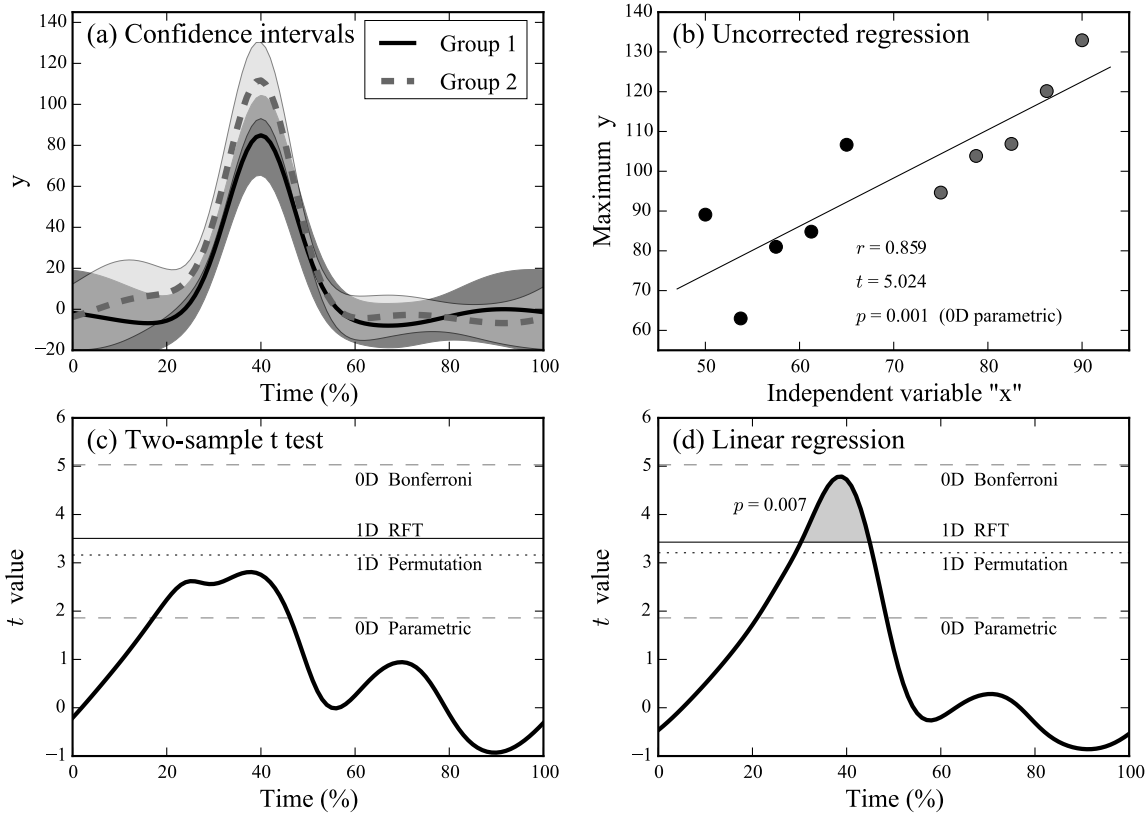

Figure 6. Results for Dataset C. (a) Separate one-sample CIs for Groups 1 and 2, using an uncorrected threshold. (b) Regression results on only y maxima; these results are uncorrected and therefore invalid if the null hypothesis pertains to the whole trajectory. (c) A two-sample t test comparing Group 1 and 2 means. (d) Linear regression between x and $y(q)$.

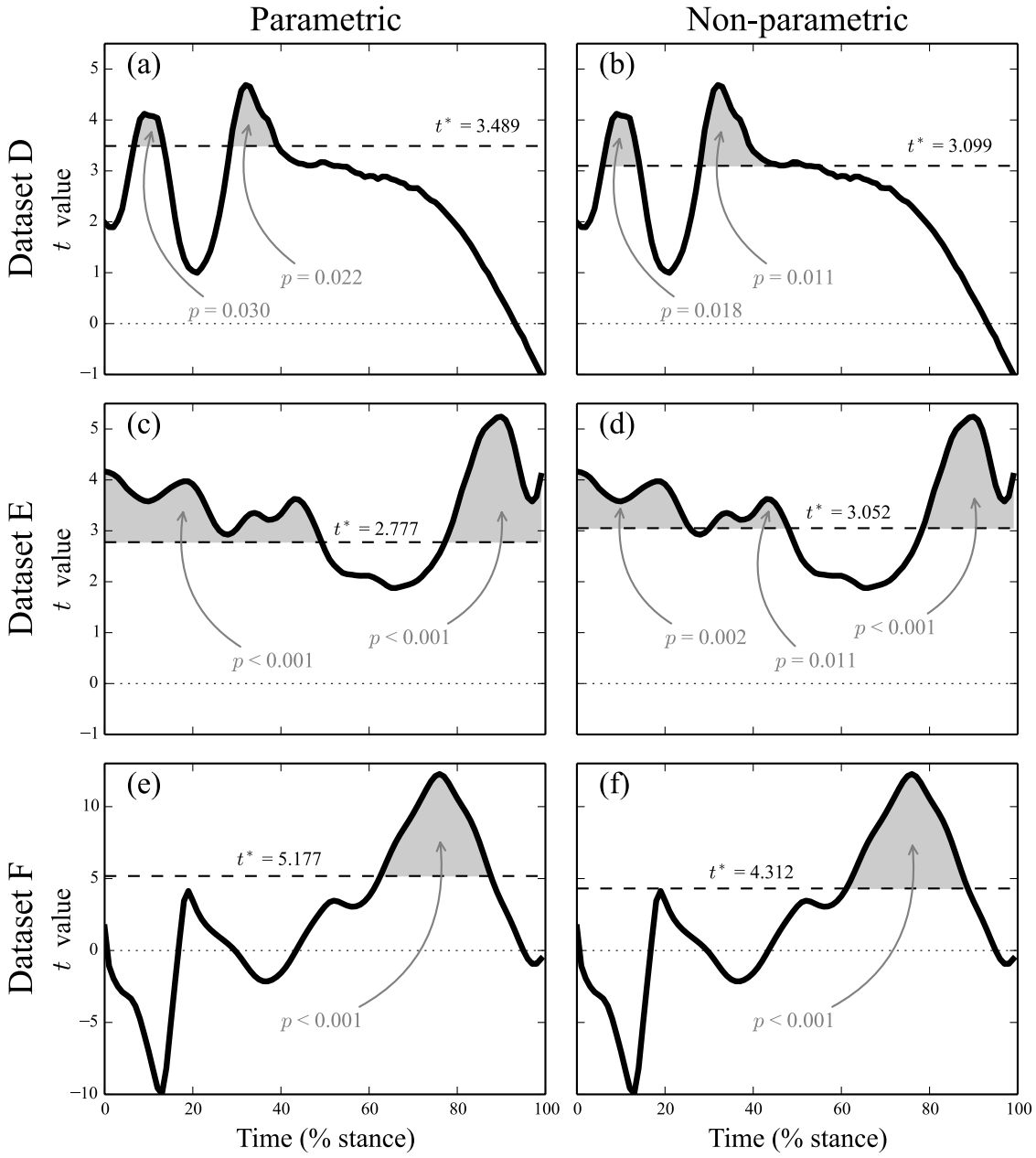

Figure 7. Hypothesis testing results for the experimental data. The top, middle and bottom panel rows depict results for Datasets D (paired), E (two-sample) and F (regression), respectively (see Fig.4). The left and right panel columns depict parametric and non-parametric results, respectively. Critical thresholds (t^*) and cluster-level probability values (p) are shown. The main points are that: (i) parametric and non-parametric results are qualitatively identical, and (ii) unlike CIs, hypothesis testing results can be presented identically for all experimental designs.

SUPPLEMENTARY MATERIAL

Note to readers:

This Supplementary Material was peer-reviewed along with the main manuscript, but has not been edited by the journal. Sections appear in the order in which they are cited in the main manuscript.

Appendix A Parametric vs. non-parametric hypothesis testing

The main difference between parametric and non-parametric hypothesis testing is that the former parameterizes probability density functions (PDFs) (Appendix D) and the latter does not. This distinction exists at two levels:

- Experimental data: parametric hypothesis testing assumes that the data are drawn from a population with a known, parameterizable PDF (usually the Gaussian distribution), but non-parametric procedures generally makes no such assumption.
- Test statistic: parametric procedures base inferences on parameterized test statistic PDFs which are analytically derived from the population PDF, but non-parametric procedures generally base inferences on empirically derived test statistic PDFs.

Below we consider these points in detail.

Parametric PDFs

The fundamental PDF upon which most parametric inference is based is the normal (Gaussian) distribution, which is parameterized by the true population mean μ and true population standard deviation σ (Fig.A1):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{A.1}$$

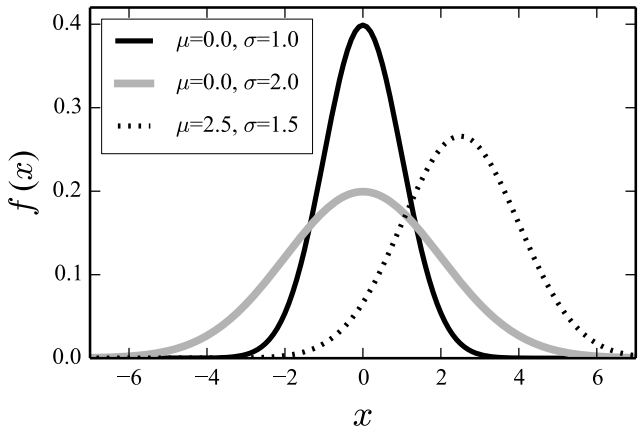

Figure A1: Gaussian probability density functions.

If we assign numerical values to μ and σ, then we can compute arbitrary probabilities using Eqns.D.1 and D.2 (Appendix D). For example, if μ=0 and σ=1, then the survival function (Eqn.D.2) predicts $P(x>0.0)$=0.500 and $P(x>2.0)$=0.023. These probabilities respectively imply that 50% of random values drawn from this distribution are expected to be greater than zero, and only 2.3% are expected to be greater than 2.0. To re-emphasize the meaning of 'parametric', we note that two simple parameters (μ and σ) completely specify the probabilistic behavior of Gaussian data.

The Gaussian PDF (Eqn.A.1) is nevertheless seldom used directly when conducting statistical inference. One reason is that the Gaussian PDF describes a random variable x, which is analogous to the raw data we measure experimentally. Most experiments are less interested in x itself than in averages (one-sample tests), average differences (two-sample tests), and correlations between x and an independent variable (regression tests). To address these empirical pursuits, the parametric approach funnels the Gaussian PDF into a particular experimental design, and generates predictions regarding what Gaussian data would do in that particular setting, over an infinite number of identical experiments.

Another reason the Gaussian PDF is not used directly for statistical inference is that μ and σ are true population parameters, but we rarely know these true values because we rarely have access to the entire population. We instead have to estimate μ and σ using a sample drawn from that population, but those estimates are imperfect, especially if the data are not sampled randomly. Even when the data are sampled randomly, estimates of μ and σ worsen as sample size decreases (Fig.A2), and parametric inference must account for this sample size-dependent behavior. Student solved this problem in 1908 through use of PDF which depends only on sample size:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{A.2}$$

Here ν is the degrees of freedom and Γ is the gamma function. The ν parameter specifies the number of values which can vary freely in a particular statistic's computation. For example, in the one-sample t test (Table F2) there are J responses, but not all response values can vary freely. In particular, after one estimates the mean, there are only $(J-1)$ responses which can vary freely to produce the same mean, so the SD estimate is normalized using $(J-1)$ rather than J (Table F2).

Equation A.2 is the analytical result obtained when Gaussian data (Eqn.A.1) are funneled into t statistic equations (Table F2). In other words, Gaussian data behave in a sample-size dependent manner (Fig.A3) when the sample is smaller than the population size.

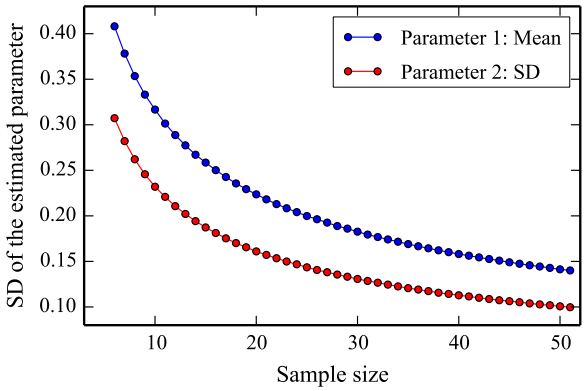

Figure A2: Variability of population parameter estimates as a function of sample size. The true mean and SD were 0 and 1, respectively. These results were constructed by simulating 10^6 samples of each sample size, computing each sample's mean and SD, then computing the SD of each parameter across all 10^6 samples.

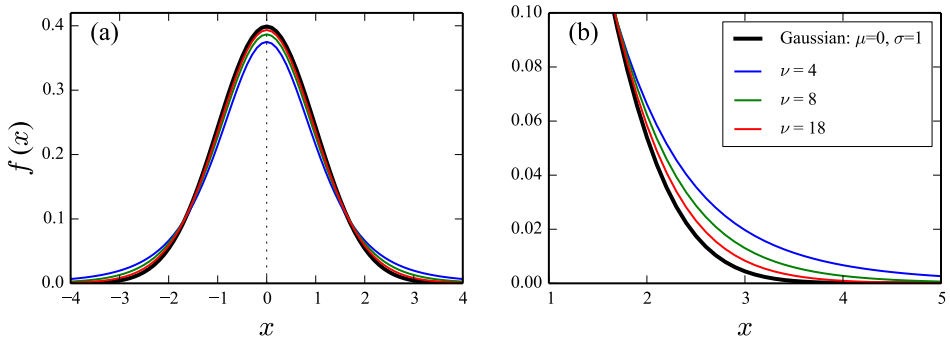

Figure A3: Comparison of various t PDFs with the standard normal PDF (μ=0, σ=1). The PDFs in panels (a) and (b) are identical, but panel (b) zooms in on one part of the PDF for clarity.

The t PDF approaches the standard normal PDF (μ=0, σ=1) as ν increases (Fig.A3b). Equivalently and conversely, large t values become increasingly likely as sample size decreases. Although the effect of ν may appear small in Fig.A3, consider the following numerical results: $P(x>3.0)$=0.020 when ν=4, but $P(x>3.0)$=0.00384 when ν=18. This implies that Gaussian data are approximately five times more likely to produce t values larger than 3.0 for ν=4 vs. ν=18.

Last, let us consider a full numerical example, which we shall repeat with non-parametric analyses below. Imagine that an experiment yields Group A and Group B responses of {1.14,

1.21, 1.25, 1.43, 1.57} and {1.37, 1.52, 1.61, 1.74, 1.54}, respectively. A two-sample independent t test (ν=8) yields t=2.378. From Eqns.D.2 and A.2 we may conclude that Gaussian data are expected to produce a t value this large with a probability of p=0.022 over many random samplings.

To summarize, the t statistic's PDF (Eqn.A.2) is completely specified by one parameter: ν, and that PDF is derived from the Gaussian PDF (Eqn.A.1), which is also parametric. More generally, parametric procedures use a small number of parameters to specify both the PDF from which experimental data are assumed to have been randomly drawn, and the test statistic PDF upon which statistical inference is based.

Non-parametric PDFs

Non-parametric PDFs are identical to parametric PDFs in the sense that they describe the behavior of randomly sampled data. The main difference is that non-parametric PDFs generally make no assumptions regarding the distribution from which data are drawn, and instead build PDFs empirically, directly from experimental data. If the underlying data are in fact Gaussian distributed, then non-parametric PDFs converge to parametric PDFs (Fig.A4) and non-parametric results converge to parametric results (Appendix E). If experimental data deviate from Gaussian behavior then parametric approaches based on the Gaussian PDF (like the t PDF) are generally not valid.

To emphasize these points it is sufficient to describe one non-parametric approach to PDF construction. Below we describe a simple two-sample permutation procedure similar to the one used in the main manuscript, but somewhat different from the one-sample procedure described in Appendix E . Returning to the numerical example above, the two-sample permutation approach starts by labeling the original data as follows:

Label	A	A	A	A	A	B	B	B	B	B
Value	1.14	1.21	1.25	1.43	1.57	1.37	1.52	1.61	1.74	1.54

As we saw before, this particular labeling (AAAAA–BBBBB) yields t=2.378. To build the permutation PDF, we simply permute these ten labels and recompute the t statistic for each permutation. For example, labels of BAAAA–ABBBB and BBAAA–AABBB yield t=1.208 and t=0.154, respectively. Repeating for many or all label permutations builds a permutation PDF (Fig.A4). In this example there are ten labels, but once we choose positions for the five A labels, the positions of the five B labels are decided. There are thus $\binom{10}{5} = 10!/(5!5!) = 252$ unique permutations. Assembling all or a large number of permutation t values forms a permutation PDF (or empirical PDF), from which probability values can be computed as follows:

$$\mathrm{P}(t \geq u) = \frac{\text{Number of permutation values greater than or equal to } u}{\text{Number of permutations}} \tag{A.3}$$

Since this example has 252 permutations, the minimum possible p value is $1/252 = 0.004$. Of those 252 permutations, this example yields a total of eight which satisfy $t \geq u$, including a maximum t value of 4.804 for a labeling of: AAAAB–ABBBB. Thus the p value is 8/252=0.0318, which is similar to the parametric p value of 0.022. This indirectly suggests that the parametric approach's assumption of normality is a reasonable one.

Which p value is correct, the parametric or non-parametric one? Both are correct, but their meanings are different. The interpretation of the parametric p value is as follows: if there were truly no difference between Groups A and B and if the population data are Gaussian distributed then a t value as large as the observed value would be expected in 2.2% of an infinite number of identical experiments. The interpretation of the non-parametric p value is: if there were truly no difference between Groups A and B and the group labels were assigned randomly to the data then only 3.18% of relabelings would yield a t value as large as the observed value. The primary difference between the two approaches is thus that the parametric p value assumes that the population distribution is Gaussian but the non-parametric p value does not.

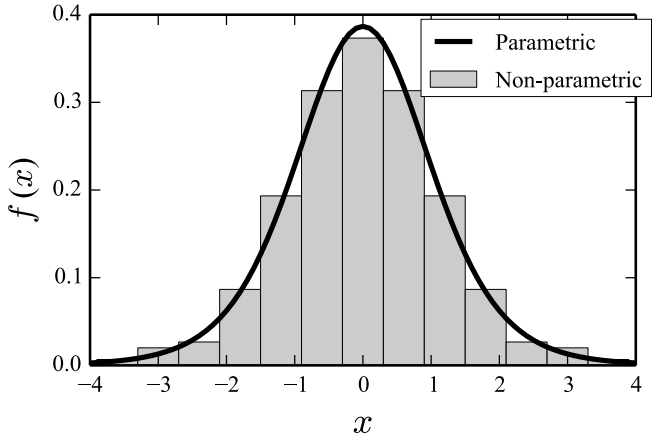

Figure A4: Comparison of parametric and non-parametric PDFs for the two-sample t test example described in the text. Here ν=8 completely parameterizes the parametric PDF. The non-parametric PDF is a histogram of the t values computed from all 252 permutations.

Appendix B Functional data analysis and random field theory

Functional data analysis (FDA) (Ramsay and Silverman 2005) emerged in the 1990s as a tool for statistically analyzing one-dimensional continua or "functions". By regarding experimentally sampled continua as continuous functions, FDA shows that experimental data can be well-approximated by a set of mathematically precise basis functions, including for example: splines and Fourier series. Representing the data in this manner opens up a wide range of analysis possibilities for describing continua, covariance between continua, etc. Although FDA was initially developed primarily as an exploratory tool of 1D continuum variance, over the years it has expanded to a wide array of statistical uses including classical hypothesis testing in arbitrary experimental designs through a variety of inference techniques.

Random field theory (RFT) (Adler and Taylor 2007) was initially developed in the 1970s to extend the (0D) Gaussian distribution to n-dimensional continua with arbitrary geometrical bounds. RFT shows, for example, how smooth 1D Gaussian continua exhibit particular geometric features (like maximum continuum height) with known probability. Statistical Parametric Mapping (SPM) emerged in the 1990s to apply RFT to experimentally measured continua (Friston et al. 2007). In the case of unbroken 1D continua, SPM estimates just one parameter more than is estimated for common 0D analyses — the ratio of continuum length to smoothness — then uses RFT to make probabilistic conclusions, like the probability that smooth 1D Gaussian data will yield a t continuum which reaches a height of 3.0 in a two-sample experiment. Directly related to classical hypothesis testing, SPM can use RFT to compute the critical height t^* above which only $\alpha\%$ of t continua would reach if those t continua were produced by smooth 1D Gaussian continua in an infinite number of identical experiments.

From a classical hypothesis testing perspective for 1D data, there is thus only one difference between FDA and RFT. Whereas FDA inference procedures are widely flexible, with a varying number of parameters, RFT inference is based on a single parameter: the continuum length-to-smoothness ratio. Since they can both describe random 1D continua, they may be regarded as equivalent for the purposes of the present paper. The main manuscript focusses on separate issues: 0D vs. 1D, parametric vs. non-parametric, and confidence interval vs. hypothesis testing procedures. While we could have used FDA address these issues, we opted for RFT simply because we find RFT easier to describe.

Appendix C Extending the t statistic to the time domain

The 1D t statistic is assembled simply by computing the 0D t statistic separately at each time point q. Since all 1D t statistic definitions are therefore trivial extensions of their 0D definitions to the 1D domain q, they are listed here only for completeness. The t statistic continua for the one-sample, paired and two-sample designs are respectively:

$$t(q) = \frac{\overline{y}(q)}{s_1(q)/\sqrt{J}}$$

$$t(q) = \frac{\overline{\Delta y}(q)}{s_p(q)/\sqrt{J}} \tag{C.1}$$

$$t(q) = \frac{\Delta \overline{y}(q)}{s_2(q)\sqrt{\frac{1}{J_A} + \frac{1}{J_B}}}$$

For regression against a continuous independent variable x, the model is:

$$y(q) = \beta_1(q)x + \beta_0(q) + \varepsilon(q)$$

where β_1, β_0 and ε are the slope, intercept and prediction error, respectively. Least-squares estimates of the slope and intercept (denoted $\hat{\beta}_1$ and $\hat{\beta}_0$, respectively) produce the following prediction for the jth response:

$$\hat{y}_j(q) = \hat{\beta}_1(q)x_j + \hat{\beta}_0(q)$$

and the standard error is:

$$s_\beta(q) = \frac{\sqrt{\frac{1}{J-2}\sum(y_j(q) - \hat{y}_j(q))^2}}{\sum(x_j - \overline{x})^2}$$

Finally, the regression t statistic is:

$$t(q) = \frac{\hat{\beta}_1(q)}{s_\beta(q)} \tag{C.2}$$

Appendix D Probability density functions (PDFs)

A PDF is a continuous function $f(x)$ which, when integrated over an interval $[x_0,\ x_1]$, specifies the probability that a random variable x adopts a value in that interval:

$$\mathrm{P}(x_0 < x < x_1) = \int_{x_0}^{x_1} f(x)dx \tag{D.1}$$

The probability that x adopts a specific value $\hat{x}$ is zero because there are an infinite number of other values it could adopt. The probability that x lies in the interval $[x_0,\ x_1]$ is at least zero and at most one. All PDFs additionally share the trivial constraint that x lies in the interval $[-\infty,\ \infty]$. These three constraints can be expressed as follows:

$$\begin{aligned} P(x = \hat{x}) &= 0 \\ 0 \leq P(x_0 < x < x_1) &\leq 1 \\ P(-\infty < x < \infty) &= 1 \end{aligned}$$

The key probability for classical hypothesis testing is the survival function — the probability that x exceeds (or 'survives') an arbitrary threshold u:

$$\mathrm{P}(x > u) = \int_{u}^{\infty} f(x)dx \tag{D.2}$$

When Eqn.D.2 is set to α, then u becomes a "critical threshold"; an experimentally observed value $\hat{x}$ which exceeds this threshold leads to null hypothesis rejection.

Random Field Theory (RFT) (Adler and Taylor, 2007) provides the foundation for generalizing Eqn.D.2 to the case of Gaussian nD continua. An important RFT probability is:

$$\mathrm{P}(x_{\max} > u) = \int_{u}^{\infty} f(x)dx \tag{D.3}$$

where $x_{\max}$ is the maximum continuum value. For classical hypothesis testing on 1D continua, setting Eqn.D.3 to α and solving for u yields the critical threshold for the null hypothesis rejection decision.

Appendix E Bootstrap and permutation techniques

The purpose of this appendix is to clarify (a) the similarities and differences between the bootstrap and permutation confidence intervals (CIs), and (b) the role of both techniques in the broader context of parametric and non-parametric hypothesis testing. Note that the bootstrap has been advocated in the Biomechanics literature for trajectory-level analysis. The permutation technique is used in the main manuscript because it is more generalizable than the bootstrap. Interested readers may wish to consult Good (2005) for a more thorough treatment of these topics for 0D datasets, and to Nichols and Holmes (2002) for a discussion of how these techniques extend to 1D and higher-dimensional data.

Sections E.1 and E.2 below analyze the following eight-response dataset:

117 104 110 122 119 90 110 97

Section E.1 computes CIs for this dataset using three different techniques, and Section E.2 conducts one-sample hypothesis testing using four different techniques. Table E1 below summarizes the results of those analyses. Considering these results briefly, it is clear that all techniques produce similar, albeit non-identical CIs and p values. To emphasize why these results are similar but not identical, Section E.3 repeats the three CI techniques for thousands of random (Gaussian) datasets to demonstrate why all techniques may be regarded as theoretically equivalent when the data are normally distributed. This Appendix thus shows that it is sufficient in the main manuscript to compare a single parametric technique (which assumes normality) to a single non-parametric technique (which does not assume normality).

Table E1: Confidence intervals (CI) and one-sample hypothesis tests computed using four different techniques, based on the dataset above.

Class	Technique	95% CI	One-sample test
Non-parametric	Bootstrap	[98.4, 117.5]	$p = 0.07559$
Non-parametric	Permutation	[98.9, 118.3]	$p = 0.06250$
Non-parametric	Wlicoxon		$p = 0.06735$
Parametric	Student's t	[99.3, 117.9]	$p = 0.06411$

E.1 Confidence intervals (CIs)

E.1.1 Bootstrap CI

A simple bootstrap CI can be computed as follows:

(a) Compute the sample mean (in this case: 108.625).

(b) Label the responses as follows:

A	B	C	D	E	F	G	H
117	104	110	122	119	90	110	97

(c) Resample with replacement: select a random set of labels, allowing labels to repeat, then compute the mean for the resampled data. For example, a labeling of "AABBBCDE" has responses: [117, 117, 104, 104, 104, 110, 122, 119], and a sample mean of: 112.125.

(d) Repeat (c) many times and store all sample means. Stop either when (i) all possible resamplings have been made (i.e. AAAAAAAA through HHHHHHHH), or when (ii) a specified number of iterations (e.g. 1000) has been completed.

(e) After all sample means have been accumulated, find the value C_{upper} above which only 2.5% of all estimates traverse, and the value C_{lower} below which only 2.5% of all estimates traverse. The CI is $[C_{\text{lower}}, C_{\text{upper}}]$.

As specified in Table E1 above this procedure yields a CI of [98.4, 117.5].

E.1.2 Permutation CI

A simple permutation CI can be computed as follows:

(a) Compute the sample mean (in this case: 108.625) .

(b) Subtract the sample mean from all responses, then label each observation as "+1":

+1	+1	+1	+1	+1	+1	+1	+1
8.375	–4.625	1.375	13.375	10.375	–18.625	1.375	–11.625

(c) Resample without replacement: permute using either a "+1" or a "–1" label for each response, then multiply each response by each label. For example, a labeling of "+1 +1 +1 –1 –1 –1 +1 –1" produces the new sample "8.375 –4.625 1.375 –13.375 –10.375 18.625 1.375 11.625". For each new sample compute the one-sample t statistic. If there are n responses, there are 2^n possible labelings (256 in this case).

(d) Repeat (c) many times and store all t statistic values for all resamplings. Stop either when (i) all possible resamplings have been made (i.e. "+1 +1 +1 +1 +1 +1 +1 +1" through "–1 –1 –1 –1 –1 –1 –1 –1"), or when (ii) a specified number of iterations (e.g. 1000) has been completed.

(e) After all t statistic values have been accumulated, find the critical height above which only 2.5% of t statistic values traverse, then compute the CI according to Appendix F.

This results in a CI of [98.9, 118.3] (Table E1).

E.1.3 Parametric CI

The parametric CI can be computed using the critical height h^*, which is defined via the one-sample t statistic distribution (see Appendix F, and in particular the "One-sample" row of Table F3). This procedure yields a CI of [99.3, 117.9], which is very similar to both the bootstrap and permutation results.

E.1.4 Comparison of CI results

All three techniques yield similar, but non-identical results. Since the parametric technique assumes that the data come from a normal (Gaussian) distribution, all CI techniques should, by definition, converge to the identical value when (a) the data are normal and (b) the sample size is large. The different techniques will only produce precisely the same result when the sample size is infinitely large, as we will see in Section E.3 below. Investigators must therefore judge whether the discrepancies amongst the techniques is negligible or non-negligible. Evidence of departure from normality, for example, would be a good reason to choose one of the non-parametric techniques. For the results above (Table E1), the discrepancies amongst the different CI techniques are likely negligible for most applications. The main point is that the three CI techniques are theoretically equivalent when the data are normally distributed.

E.2 Hypothesis tests

Thorough descriptions of one-sample hypothesis tests using the bootstrap, permutation, Wilcoxon and parametric (one-sample t test) techniques can be found in many statistics text-books so in interest of brevity are not repeated here. Additionally, as will be shown in Appendix F, CIs are equivalent to one-sample hypothesis tests, so re-describing the techniques here would be redundant. This section therefore just focusses on the results in Table E1 above.

Like the CI results, all four hypothesis test procedures produce similar, but non-identical p values. For classical hypothesis testing, the null hypothesis would not be rejected for any of the four tests at α=0.05. The next section explores why these four approaches are theoretically equivalent (when the data are normal) even when the results are not precisely equivalent.

E.3 Convergence of CIs

Repeating the bootstrap, permutation and parametric CI procedures on thousands of random datasets (drawn from the Gaussian distribution) of increasingly larger sample sizes yields the results in Fig.E1. The two non-parametric CIs clearly converge to the parametric CI as sample size increases, implying theoretical equivalence amongst the three procedures (when the data are normally distributed).

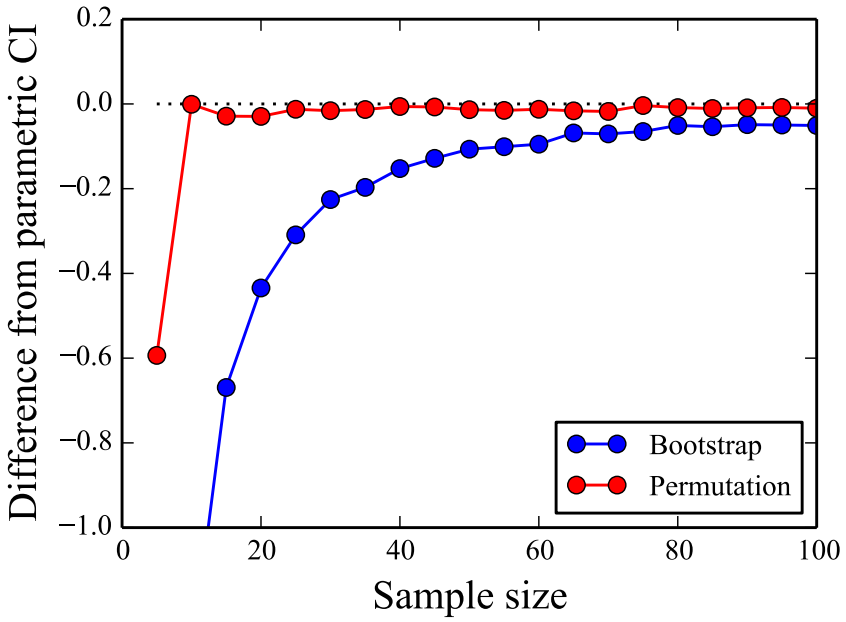

Figure E1: Convergence of the CI for three estimation procedures. These results were obtained by: (i) producing a random sample from the Gaussian distribution of the given sample size, with a mean of 100 and a variance of 10, (ii) estimating the CI using the three procedures indicated (Bootstrap, Permutation, Parametric), and (iii) repeating 500 times for each sample size. Single results depict the mean values across the 500 repetitions.

Summary

This Appendix has shown that there is fundamentally little difference between the bootstrap and permutation approaches, but that they might produce non-negligibly different numerical results in certain situations, like when sample sizes are very small. The larger point is that the bootstrap procedure is not particularly unique, as has been implied in the literature. Instead the bootstrap procedure yields a solution which can also be obtained using other techniques,

and its scope is also relatively limited in the broader context of generalized hypothesis testing (Fig.E2).

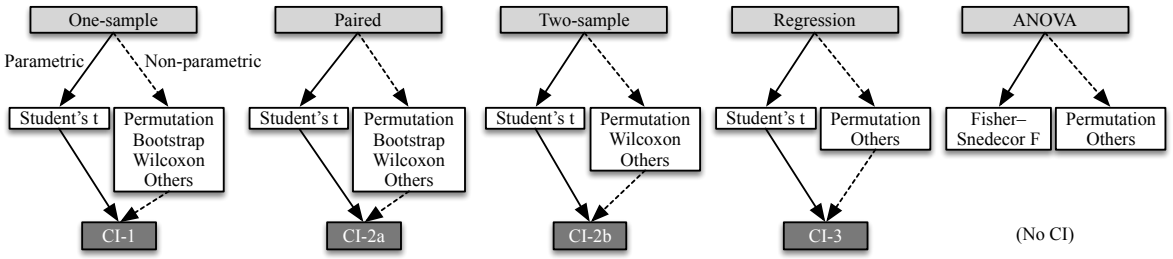

Figure E2: Context of the bootstrap (for both 0D and 1D tests). Light grey, white and dark grey boxes respectively depict: experimental designs, statistical inference procedures, and confidence intervals (CI).

Appendix F Confidence interval design dependence

Confidence intervals (CIs) are defined as:

$$CI = y_0 \pm h^* \tag{F.1}$$

where y_0 is a datum and h^* is the design-dependent critical height. More specifically, h^* is given by a critical t value and simple algebraic manipulation of the design-dependent t statistic definition.

To clarify, first consider that design-dependent mean and SD definitions (Table F1) yield design-dependent t statistic definitions (Table F2). Next, given t^* one may compute the design-dependent h^* (Table F2). Last, after choosing a datum y_0, there are various acceptable null hypothesis rejection criteria (Table F3).

The main point is that hypothesis testing employs a single unambiguous criterion: $(t > t^*)$, irrespective of the particular design, making it easy to compare results across experiments. In contrast, CIs are both design- and datum-dependent.

Clearly h^* is valuable for data visualization because it represents the null hypothesis rejection criterion in the same units as the original data. However, it is also clear that h^* must be computed with careful attention to both the datum and the design, and can only be interpreted by readers if the precise datum and design are made explicit. The main manuscript therefore argues that hypothesis testing is simpler.

Table F1: Mean and standard deviation (SD) definitions for one-sample, paired and two-sample designs. For simplicity equal variance is assumed in the two-sample case.

Design	Mean	SD
One-sample	$\overline{y} = \frac{1}{J}\sum y_j$	$s_1 = \sqrt{\frac{1}{J-1}\sum(y_j - \overline{y})^2}$
Paired	$\overline{\Delta y} = \frac{1}{J}\sum(y_{Aj} - y_{Bj})$	$s_p = \sqrt{\frac{1}{J-1}\sum\left((y_{Aj} - y_{Bj}) - \overline{\Delta y}\right)^2}$
Two-sample	$\Delta\overline{y} = \overline{y}_A - \overline{y}_B$	$s_2 = \sqrt{\frac{(J_A-1)s_A^2 + (J_B-1)s_B^2}{J_A + J_B - 2}}$

Table F2: Design-dependence of the CI's critical height h^* .

Design	t	Mean	h^*
One-sample	$t_1 = \dfrac{\overline{y}}{s_1/\sqrt{J}}$	$\overline{y} = t\dfrac{s_1}{\sqrt{J}}$	$h_1^* = t^*\dfrac{s_1}{\sqrt{J}}$
Paired	$t_p = \dfrac{\overline{\Delta y}}{s_p/\sqrt{J}}$	$\overline{\Delta y} = t\dfrac{s_p}{\sqrt{J}}$	$h_p^* = t^*\dfrac{s_p}{\sqrt{J}}$
Two-sample	$t_2 = \dfrac{\Delta\overline{y}}{s_2\sqrt{\frac{1}{J_A}+\frac{1}{J_B}}}$	$\Delta\overline{y} = t_2 s_2\sqrt{\frac{1}{J_A}+\frac{1}{J_B}}$	$h_2^* = t^* s_2\sqrt{\frac{1}{J_A}+\frac{1}{J_B}}$

Table F3: Design- and datum-dependence of h^*-based null hypothesis rejection criteria. All criteria assume $\overline{y}_A \geq \overline{y}_B$.

Design	Datum (y_0)	Criterion: zero	Criterion: mean	Criterion: tail
One-sample	$\overline{y}$	$\overline{y} - h_1^* > 0$	—	—
Paired	$\overline{\Delta y}$	$\overline{\Delta y} - h_p^* > 0$	—	—
	$\overline{y}_A$	—	$\overline{y}_A - h_p^* > \overline{y}_B$	$\overline{y}_A - \dfrac{h_p^*}{2} > \overline{y}_B + \dfrac{h_p^*}{2}$
Two-sample	$\Delta\overline{y}$	$\Delta\overline{y} - \dfrac{h_2^*}{2} > 0$	—	—
	$\overline{y}_A$	—	$\overline{y}_A - h_2^* > \overline{y}_B$	$\overline{y}_A - \dfrac{h_2^*}{2} > \overline{y}_B + \dfrac{h_2^*}{2}$

Appendix G Dataset F reanalyses

Here we reanalyze the Dataset F dataset using (i) two-tailed inference, and (ii) nonlinear registration followed by two-tailed inference. Regarding two-tailed inference: the results in the main manuscript (Fig.7De,f) are based on one-tailed inference at $\alpha = 0.05$. One-tailed inference has only a single positive critical threshold (i.e. $t^* > 0$). Two-tailed inference has two thresholds: $+t^*$ and $-t^*$, and excursion beyond either threshold warrants null hypothesis rejection. Moreover, $+t^*$ is higher than in a one-sample test; $+t^*$ for two-tailed inference at α is equivalent to t^* for one-tailed inference at $\alpha/2$. Results suggest that, although two-tailed found highlighted a temporal of significant negative correlation between GRF and walking speed (Fig.G2a), two-tailed inference did not affect the main manuscript's null-hypothesis rejection decision.

Regarding nonlinear registration: we used a simple piecewise linear registration approach (Kneip et al. 2000) to align the first two local extrema in Datasets F, which occurred between approximately 10% and 25% stance (Fig.G1a), resulting in reduced temporal variability of those extrema (Fig.G1b). Non-linear registration also produced slightly amplified supra-threshold t signals with respect to the original data. However, this affected neither the null hypothesis rejection decision nor the general biomechanical interpretation. In particular, both original and registered results suggest negative correlation between posterior GRF and running speed in early stance, and positive correlation in late stance. Misregistration effects may be non-negligible in other datasets (Sadeghi et al.2003).

References:

Kneip A, Li X, MacGibbon KB (2000). Curve registration by local regression, Canadian Journal of Statistics 28(1): 19–29.

Sadeghi H, Mathieu PA, Sadeghi S, Labelle H (2003). Continuous curve registration as an intertrial gait variability reduction technique, IEEE Transactions on Neural Systems and Rehabilitation Engineering 11(1): 24–30.

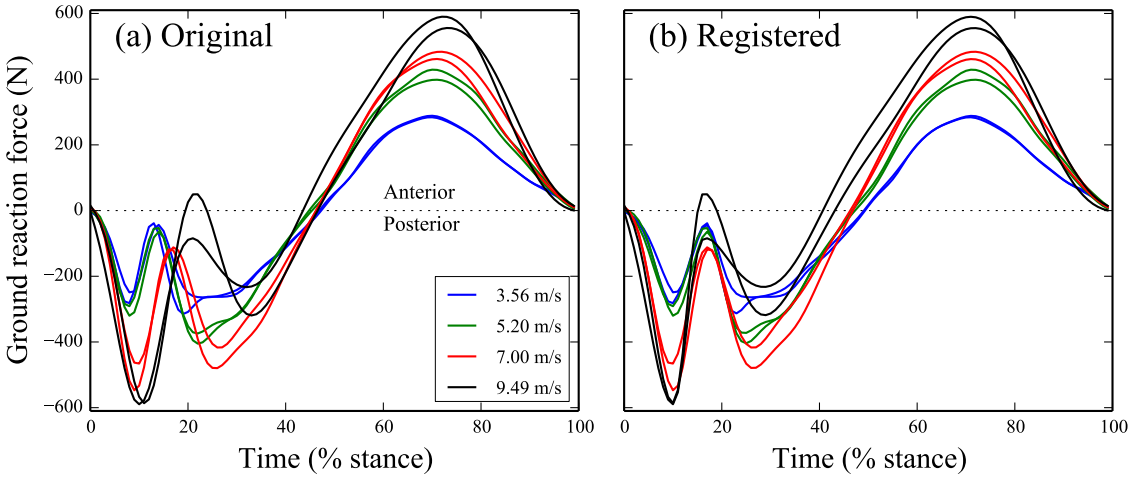

Figure G1: Dataset F, (a) original and (b) registered trajectories.

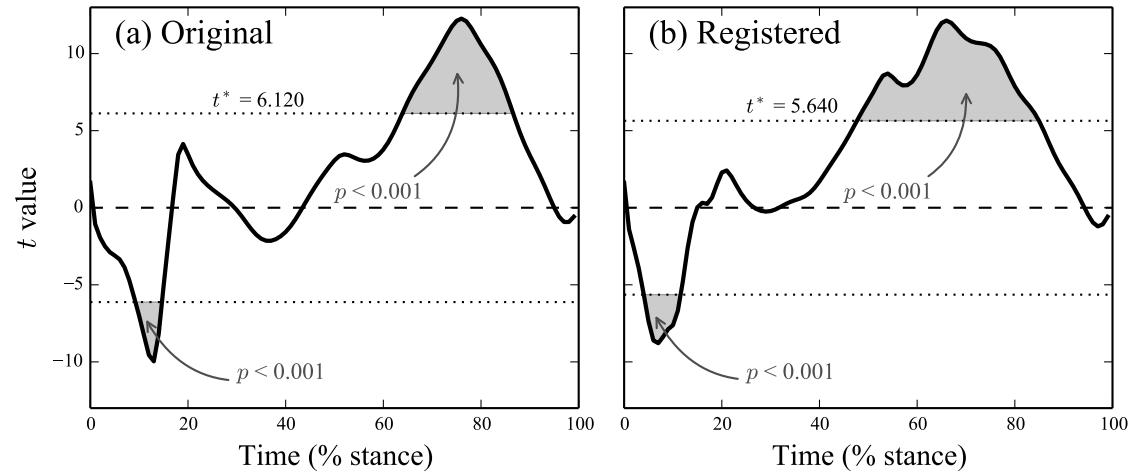

Figure G2: Dataset F, two-tailed results for (a) original and (b) registered data.