

# Vector field statistical analysis of kinematic and force trajectories

Todd C. Pataky<sup>1</sup>, Mark A. Robinson<sup>2</sup>, and Jos Vanrenterghem<sup>2</sup>

<sup>1</sup>Department of Bioengineering, Shinshu University, Japan

<sup>2</sup>Research Institute for Sport and Exercise Sciences, Liverpool John Moores University, UK

## Abstract

When investigating the dynamics of three-dimensional multi-body biomechanical systems it is often difficult to derive spatiotemporally directed predictions regarding experimentally induced effects. A paradigm of ‘non-directed’ hypothesis testing has emerged in the literature as a result. Non-directed analyses typically consist of *ad hoc* scalar extraction, an approach which substantially simplifies the original, highly multivariate datasets (many time points, many vector components). This paper describes a commensurately multivariate method as an alternative to scalar extraction. The method, called ‘statistical parametric mapping’ (SPM), uses random field theory to objectively identify field regions which co-vary significantly with the experimental design. We compared SPM to scalar extraction by re-analyzing three publicly available datasets: 3D knee kinematics, a ten-muscle force system, and 3D ground reaction forces. Scalar extraction was found to bias the analyses of all three datasets by failing to consider sufficient portions of the dataset, and/or by failing to consider covariance amongst vector components. SPM overcame both problems by conducting hypothesis testing at the (massively multivariate) vector trajectory level, with random field corrections simultaneously accounting for temporal correlation and vector covariance. While SPM has been widely demonstrated to be effective for analyzing 3D scalar fields, the current results are the first to demonstrate its effectiveness for 1D vector field analysis. It was concluded that SPM offers a generalized, statistically comprehensive solution to scalar extraction’s oversimplification of vector trajectories, thereby making it useful for objectively guiding analyses of complex biomechanical systems.

Keywords: *biomechanics, random field theory, Statistical Parametric Mapping, multivariate statistics*

# Glossary

Category	Symbol	Other	Description
Counts	$I$	<b>Index:</b> $i$	Vector components
	$J$	$j$	Responses (i.e. experimental recordings)
	$K$	$k$	Predictor variables
	$N$		Extracted scalars (e.g. maximum force)
	$Q$	$q$	Field measurement nodes (e.g. 100 points in time)
Responses	$y_i$	<b>Mean, variance:</b> $\bar{y}, s^2$	Scalar response (with st.dev.)
	$y_i(q)$	$\bar{y}(q), s^2(q)$	Scalar field response (with st.dev. field)
	$\mathbf{y}(q)$	$\bar{\mathbf{y}}(q), \mathbf{W}(q)$	Vector field response (with covariance field)
Test statistics	$t$	<b>field:</b> SPM $\{t\} \equiv t(q)$	Student's $t$ statistic
	$F$	SPM $\{F\} \equiv F(q)$	Variance ratio (e.g. from ANOVA)
	$T^2$	SPM $\{T^2\} \equiv T^2(q)$	Hotelling's $T^2$ statistic (vector equivalent of $t$ )
	$R$		Canonical correlation coefficient
Probability	$\alpha$		Type I error rate
	$p$		Probability value
Acronyms	CCA		Canonical correlation analysis
	EMG		Electromyography
	GRF		Ground reaction force
	PFP		Patellofemoral pain

# 1 Introduction

Measurements of motion and the forces underlying that motion are fundamental to biomechanical experimentation. These measurements are often manifested as one-dimensional (1D) scalar trajectories  $y_i(q)$ , where  $i$  represents a particular physical body, joint, axis or direction, and where  $q$  represents 1D time or space. Experiments typically involve repeated measurements of  $y_i(q)$  followed by registration (i.e. homologously optimal temporal or spatial normalization) to a domain of 0–100% (Sadeghi et al., 2003). This paper pertains to analysis of registered data  $y_i(q)$ .

Given that many potential sources of bias exist in  $y_i(q)$  analysis (Rayner, 1985; James and Bates, 1997; Mullineaux et al., 2001; Knudson, 2009), a non-trivial challenge is to employ statistical methods that are consistent with one’s null hypothesis. Consider first ‘directed’ null hypotheses: those which claim response equivalence in particular vector components  $i$ , and in particular points  $q$  or windows  $[q_0, q_1]$ :

**Example ‘directed’ null hypothesis:** Controls and Patients exhibit identical maximum knee flexion during walking between 20% and 30% stance.

To test this hypothesis only maximum knee flexion should be assessed, and only in the specified time window. Testing other time windows, joints, or joint axes in a *post hoc* sense would constitute bias. This is because increasing the number of statistical tests increases our risk of incorrectly rejecting the null hypothesis (see Supplementary Material – Appendix A). In other words, it is biased to expand the scope of one’s null hypothesis after seeing the data. We refer to this type of bias as ‘*post hoc* regional focus bias’.

Next consider ‘non-directed’ null hypotheses: hypotheses which broadly claim kinematic or dynamic response equivalence:

**Example ‘non-directed’ null hypothesis:** Controls and Patients exhibit identical hip and knee kinematics during stance phase.

To address this hypothesis both hip and knee joint rotations should be assessed, about all three orthogonal spatial axes, and from 0% to 100% stance (i.e. the entire dataset  $y_i(q)$ ). It would be biased to assess only maximum hip flexion, for example, in a *post hoc* sense but for the opposite reason: it is biased to reduce the scope of one’s null hypothesis after seeing the data.

Non-directed hypotheses expose a second potential source of bias: covariance among the  $I$  vector components. Scalar analyses ignore covariance and are therefore coordinate-system dependent (see Supplementary

Material – Appendix B). This is important because a particular coordinate system — even one defined anatomically and local to a moving segment — may not reflect underlying mechanical function (Kutch and Valero-Cuevas, 2011). Joint rotations, for example, may not be independent because muscle lines of action are generally not parallel to externally-defined axes (Jensen and Davy, 1975). Joint moments may also not be independent because endpoint force control, for example, requires coordinated joint moment covariance (Wang et al., 2000). Under a non-directed hypothesis this covariance must be analyzed because separate analysis of the  $I$  components is equivalent to an assumption of independence, an assumption which may not be justified (see Supplementary Material – Appendix B). We refer to this source of bias as ‘inter-component covariance bias’.

Both *post hoc* regional focus bias and inter-component covariance bias have been acknowledged previously (Rayner, 1985; James and Bates, 1997; Mullineaux et al., 2001; Knudson, 2009). However, to our knowledge no study has proposed a comprehensive solution.

The purpose of this paper is to show that a method called Statistical Parametric Mapping (SPM) (Friston et al., 2007) greatly mitigates both bias sources. The method begins by regarding the data  $y_i(q)$  as a vector field  $\mathbf{y}(q)$ , a multi-component vector  $\mathbf{y}$  whose values change in time or space  $q$  (Fig.1). When regarding the data in this manner, it is possible to use random field theory (RFT) (Adler and Taylor, 2007) to calculate the probability that observed vector field changes resulted from chance vector field fluctuations.

We use SPM and RFT to conduct formalized hypothesis testing on three separate, publicly available biomechanical vector field datasets. We then contrast these results with the traditional scalar extraction approach. Based on statistical disagreement between the two methods we infer that, by definition, at least one of the methods is biased. We finally use mathematical arguments (Supplementary Material) and logical interpretations of the original data to conclude that scalar extraction constitutes a biased approach to non-directed hypothesis testing, and that SPM overcomes these biases.

## 2 Methods

### 2.1 Datasets

We reanalyzed three publicly available datasets (Table 1):

- Dataset A (Neptune et al., 1999) (<http://isbweb.org/data/rn/>): stance-phase lower extremity dynamics in ten subjects performing ballistic side-shuffle and v-cut tasks (Fig.2). Present focus was on

within-subject mean three dimensional knee rotations for the eight subjects whose data were labeled unambiguously in the public dataset.

- Dataset B (Besier et al., 2009) (<https://simtk.org/home/muscleforces>): stance-phase knee-muscle forces during walking and running in 16 Controls and 27 Patello-Femoral Pain (PFP) patients, as estimated from EMG-driven forward-dynamics simulations. Present focus was on walking and absolute forces (newtons) (Fig.3).
- Dataset C (Dorn et al., 2012) (<https://simtk.org/home/runningspeeds>): one subject’s full-body kinematics and ground reaction forces (GRF) during running at four different speeds: 3.56, 5.20, 7.00, and 9.49 ms<sup>-1</sup>. Present focus was on three-dimensional left-foot GRF (Fig.4), for which a total of eight responses were available. We linearly interpolated the GRF data across stance phase to  $Q=100$  time points.

These three datasets were chosen, first, to represent a range of biomechanical data modalities: kinematics, modeled (internal) muscle forces, and external forces. Second, they were chosen to demonstrate how vector field analysis applies to a range of statistical tests: (A) paired  $t$  tests, (B) two-sample  $t$  tests, and (C) linear regression.

## 2.2 Traditional scalar extraction analysis

Two, ten, and three scalars were respectively extracted from the three datasets (Table 1). These particular scalars were chosen either because they appeared to be most affected by the experiment (Datasets A and C), or because they were physiologically relevant (Dataset B: maximum force). As indicated above, Dataset A’s task effects were assessed using paired  $t$  tests, Dataset B’s group effects were assessed using two-sample (independent)  $t$  tests, and Dataset C’s speed effects were assessed using linear regression.

Since we conducted one test for each scalar, we performed  $N=2$ ,  $N=10$  and  $N=3$  tests on Datasets A, B and C, respectively, where  $N$  is the number of extracted scalars. To retain a family-wise Type I error rate of  $\alpha=0.05$  we adopted Šidák thresholds of  $p=0.0253$ ,  $p=0.0051$ , and  $p=0.0170$  respectively, where the Šidák threshold is:

$$p_{\text{critical}} = 1 - (1 - \alpha)^{1/N} \tag{1}$$

These scalar analyses superficially appear to be legitimate analysis options. However, through comparison

with the equivalent vector field analyses (§2.3), we will show how and why scalar extraction is biased for non-directed null hypothesis testing.

## 2.3 Statistical Parametric Mapping (SPM)

SPM analyses (Friston et al., 2007) were conducted using vector field analogs to the aforementioned univariate tests (§2.2). Before detailing SPM procedures, we note that they are conceptually identical to univariate procedures: conducting a one-sample  $t$  test on ten scalar values, for example, is nearly identical to conducting a one-sample  $t$  test on ten vector fields. The only differences are that SPM: (i) considers vector covariance when computing the test statistic, (ii) considers field smoothness and size when computing the critical test statistic threshold, and (iii) considers random field behavior when computing  $p$  values (see Appendix A and B – Supplementary Material).

Ultimately each SPM test results in a test statistic field (e.g. the  $t$  statistic as a function of time), and RFT is used to assess the significance of this statistical field. §2.3.1–§2.3.3 below detail test statistic field computations for the current datasets, §2.3.4 describes RFT computations of critical test statistic values and  $p$  values, and §2.3.5 suggests a *post hoc* procedure for scrutinizing vector field test results.

### 2.3.1 Paired Hotelling’s $T^2$ test (Dataset A)

SPM’s vector field analog to the paired  $t$  test is the paired Hotelling’s  $T^2$  test, which is given by the one-sample  $T^2$  statistic (Cao and Worsley, 1999):

$$\text{SPM}\{T^2\} \equiv T^2(q) = J \bar{\mathbf{y}}(q)^\top \mathbf{W}(q)^{-1} \bar{\mathbf{y}}(q) \quad (2)$$

where  $J$  is the number of vector fields (Table 1) and  $\bar{\mathbf{y}}(q)$  is the mean vector field or — in the case of a paired test — the vector field difference  $\Delta\bar{\mathbf{y}}(q)$  (see Supplementary Material – Appendix C).  $\mathbf{W}$  is the  $(I \times I)$  sample covariance matrix:

$$\mathbf{W}(q) = \frac{1}{J-1} \left( \sum_{j=1}^J (\mathbf{y}_j(q) - \bar{\mathbf{y}}(q)) (\mathbf{y}_j(q) - \bar{\mathbf{y}}(q))^\top \right) \quad (3)$$

representing the variances-within and correlations-between vector components across the  $J$  responses (Supplementary Material - Appendix D).

The notation “SPM $\{T^2\}$ ” (Friston et al., 2007) indicates that the test statistic  $T^2$  varies in continuous time (or space), forming a temporal (or spatial) statistical ‘map’. To clarify: “SPM” refers to the

methodology, and “SPM $\{T^2\}$ ” to a specific variable.

### 2.3.2 Two-sample Hotelling’s $T^2$ test (Dataset B)

SPM’s vector field analog to the two-sample  $t$  test is the Hotelling’s  $T^2$  test (Cao and Worsley, 1999):

$$\text{SPM}\{T^2\} \equiv T^2(q) = \frac{J_1 J_2}{J_1 + J_2} \left( \bar{\mathbf{y}}_1(q) - \bar{\mathbf{y}}_2(q) \right)^\top \mathbf{W}(q)^{-1} \left( \bar{\mathbf{y}}_1(q) - \bar{\mathbf{y}}_2(q) \right) \quad (4)$$

where subscripts “1” and “2” index the two groups. Here  $\mathbf{W}$  is the pooled covariance matrix:

$$\mathbf{W} = \frac{1}{J_1 + J_2 - 2} \left( \sum_{j=1}^{J_1} (\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)^\top + \sum_{j=1}^{J_2} (\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)^\top \right) \quad (5)$$

where the domain “ $(q)$ ” is dropped for compactness.

### 2.3.3 Canonical correlation analysis (Dataset C)

SPM’s vector field analog to linear regression is canonical correlation analysis (CCA) (Hotelling, 1936; Worsley et al., 2004). The goal of CCA is to determine the strength of linear correlation between a set of predictor variables  $\mathbf{x}_j$  ( $K$ -component vectors) and a set of response variables  $\mathbf{y}_j$  ( $I$ -component vectors). We provide a brief technical summary of CCA. An extended discussion is provided as Supplementary Material (Appendix E).

Following Worsley et al. (2004), the test statistic of interest was the maximum canonical correlation ( $R$ ), a single correlation coefficient which varies over  $q$ , and which transforms to the  $F$  statistic via the identity:

$$\text{SPM}\{F\} \equiv F(q) = R(q) \frac{J - 1}{1 - R(q)} \quad (6)$$

To compute  $R$ , one must first assemble three covariance matrices:

- $\mathbf{C}_{XX}$  — the ( $K \times K$ ) predictors covariance matrix
- $\mathbf{C}_{YY}$  — the ( $I \times I$ ) responses covariance matrix
- $\mathbf{C}_{XY}$  — the ( $K \times I$ ) predictor-response covariance matrix

The maximum canonical correlation ( $R$ ) is the maximum eigenvalue of the ( $K \times K$ ) canonical correlation matrix ( $\mathbf{C}$ ) (Worsley et al., 2004):

$$\mathbf{C} = \mathbf{C}_{XX}^{-1} \mathbf{C}_{XY} \mathbf{C}_{YY}^{-1} \mathbf{C}_{XY}^{\top} \quad (7)$$

An equivalent interpretation is that  $R$  is the maximum correlation coefficient obtainable when the predictor and response coordinate systems are permitted to mutually rotate.  $K=2$  predictors (running speed and an intercept) were employed to model the  $I=3$  force vector components of Dataset C.

#### 2.3.4 Statistical inference

To determine the significance of the aforementioned test statistic fields, first field smoothness was estimated from the temporal gradients of the residuals (Friston et al., 2007). Next, given this smoothness, RFT (Adler and Taylor, 2007) was used to determine the critical test statistic threshold that retained a family-wise error rate of  $\alpha=0.05$  (Cao and Worsley, 1999; Worsley et al., 2004). Last, the probability with which suprathreshold clusters could have been produced by chance (i.e. by random fields with the same temporal smoothness) was calculated using analytic expectation (Cao and Worsley, 1999). In other words, rather than controlling the false-positive rate at each point in time, we presently controlled the false-positive rate of the test statistic field’s sample-rate invariant topological features (Friston et al., 2007). For additional details refer to Appendix A (Supplementary Material).

#### 2.3.5 Post hoc scalar field SPM

When testing non-directed hypotheses regarding biomechanical vector fields, we propose that SPM should be implemented in a hierarchical manner, analogous to ANOVA with *post hoc t* testing. One should first use SPM to analyze the entire vector field  $\mathbf{y}(q)$ , and particular vector components (scalar field  $y_i(q)$ ) should only be tested, in a *post hoc* manner, if statistical significance is reached at the vector-field level.

Following vector field analyses, *post hoc* tests were conducted on each vector component separately (i.e. on scalar fields). For scalar fields the aforementioned tests (§2.3.1–§2.3.3) reduce to: the paired  $t$  statistic (Dataset A), the two-sample  $t$  statistic (Dataset B), and the linear regression  $t$  statistic (Dataset C). Each scalar field test produced one  $\text{SPM}\{t\}$ , whose significance was determined as described above §.2.3.1. To maintain a family-wise error of  $\alpha=0.05$ , Sidák thresholds (Eqn.1) of  $p=0.0170$ ,  $p=0.0051$ , and  $p=0.0170$  were used to correct for the  $I=3$ ,  $I=10$ , and  $I=3$  vector components of Datasets A, B, and C, respectively (Table 1). All aforementioned analyses were implemented in Python 2.7 using Enthought Canopy 1.0 (Enthought Inc., Austin, USA).

## 3 Results

### 3.1 Dataset A: knee kinematics

The knee appeared to be comparatively more flexed (Fig.2a) and somewhat more externally rotated (Fig.2c) in the side-shuffle vs. v-cut tasks, with slightly more abduction at 0% stance (Fig.2b). Statistical tests on the extracted scalars found significant differences between the two tasks for both maximal knee flexion ( $t=3.093$ ,  $p=0.018$ ) and abduction at 0% stance ( $t=3.948$ ,  $p=0.006$ ).

SPM vector field analysis (Fig.5) found significant kinematic differences between the two tasks at approximately 1%, 10%, 20%, 30-35% and 95-100% stance. *Post hoc t* tests revealed that the effects over 30-35% and 95-100% stance resulted primarily from increased flexion ( $p=0.015$ ) and increased external rotation ( $p=0.004$ ), respectively, in the side-shuffle vs. v-cut tasks (Fig.6). Apparent discrepancies amongst vector field SPM, scalar field SPM, and scalar extraction (both here and in the remainder of the Results), are addressed in the Discussion.

### 3.2 Dataset B: muscle forces

Most muscles appeared to exhibit higher forces in PFP vs. Controls over most of stance (Fig.3). Nonetheless, none of the statistical tests on the extracted scalars reached significance; the medial gastrocnemius force exhibited the strongest effect ( $t=2.617$ ,  $p=0.013$ ), but like the nine other muscles ( $t<1.91$ ,  $p>0.063$ ) this failed to reach the Šidák significance threshold of  $p=0.0051$ .

In contrast, SPM vector field analyses found significance for the entire stance phase (Fig.7;  $p<0.001$ ). *Post hoc t* tests on individual muscle trajectories found significantly greater forces in PFP only for the medial gastrocnemius, and only over scattered time regions (maximum  $p=0.002$ ) (Fig.8).

### 3.3 Dataset C: ground reaction forces

Forces appeared to increase systematically with running speed, particularly in the vicinity of 30% and 75% stance (Fig.4). Linear regression found that all three extracted scalars surpassed the Šidák threshold for significance ( $p=0.0170$ ); analysis of maximum propulsion, vertical and lateral forces yielded  $r^2=0.951$ , 0.691 and 0.737, and  $p=0.00004$ , 0.001, and 0.006, respectively.

SPM vector field analysis (Fig.9) found that GRF was significantly correlated with running speed in three intervals with approximate windows of: 10–18%, 20–43% and 60–88% stance. *Post hoc* scalar field analysis

revealed that  $GRF_x$  was primarily responsible for the 10–18% and 60–88% effects (Fig.10a), and that  $GRF_y$  was primarily responsible for the 20–43% effect (Fig.10b).

## 4 Discussion

The current vector field SPM and scalar extraction results all agreed qualitatively with the data, yet the two approaches yielded different results and even incompatible statistical conclusions. This, by definition, indicates that at least one of the methods is biased. For non-directed hypotheses testing we contend that scalar extraction is susceptible to two non-trivial bias sources:

1. *Post hoc* regional focus bias — Type I or Type II error (i.e. false positives or false negatives) resulting from the failure to consider the entire measurement domain.
2. Inter-component covariation bias — Type I or Type II error resulting from the failure to consider the covariance amongst vector components.

We further contend that vector field testing overcomes both bias sources because it uses the entire measurement domain and all vector components to maintain a constant error rate of  $\alpha$ . The remainder of the Discussion is devoted to justifying these claims.

### 4.1 Bias in scalar extraction analyses

Dataset A exhibited Type I error due to *post hoc* regional focus bias. Scalar extraction analysis of maximum flexion (at ~50% stance) reached significance (§3.1) but neither vector field analysis (Fig.5) nor *post hoc* scalar field analysis (Fig.6a) reached significance in this field region. Similarly, scalar extraction found a significant ab-/adduction effect at 0% stance, but SPM did not. These discrepancies are resolved through multiple comparisons theory (Knudson, 2009); it is highly likely that at least one of Dataset A’s 303 vector field points will exceed an (uncorrected) threshold of  $p=0.05$  simply by chance. By extracting only scalars which appeared to exhibit maximum effects (Fig.2) we effectively conducted 303 tests and then chose to report the results of only two.

The opposite effect (Type II error) was also present in Dataset A. Scalar extraction focussed on only two scalars, and thus failed to identify the other effects present in the dataset (Fig.5), and in particular the large late-stance internal/external rotation effect (Fig.6c). A simple example (Supplementary Material – Appendix

A) clarifies how it is possible for scalar extraction and SPM to yield opposite statistical conclusions, and that the scalar extraction results can't be trusted because they fail to honor the  $\alpha$  error rate.

Dataset B exhibited Type II error due to covariation bias: scalar extraction failed to reach significance (§3.2) even though SPM found substantial evidence for muscular differences between Controls and PFP (Fig.7). This is resolved by correlation amongst muscles like the vasti (Fig.3). A simple example (Supplementary Material – Appendix B) clarifies how it is possible for vector resultant changes to reach significance when vector component changes do not. Scalar analysis of vector data cannot be trusted because it fails to account for vector component covariance.

Scalar extraction analysis of Dataset C exhibited both Type I and Type II error due to regional focus bias. Scalar extraction analyses of lateral forces exhibited Type I error because there is insufficient field-wide evidence to support its conclusion of significance (Fig.10c). Scalar extraction also exhibited Type II error by failing to analyze braking forces and therefore failing to identify positive correlation between running speed and braking forces at 15% stance (Fig.10a).

## 4.2 Bias in scalar field SPM analyses

Scalar field SPM solves regional focus bias (because it tests the entire domain  $q$ ), but it remains susceptible to covariance bias because it separately tests the  $I$  vector components. Scalar field analysis of Dataset A exhibited Type II error by failing to identify all field effects, and particularly the large early-stance effect (Fig.5,6). Appendix B clarifies that this was caused by scalar field analysis' failure to consider inter-component covariance; it regards trajectory variance as a 1D time-varying 'cloud' (Fig.2) when in fact it is an  $I$ -D time-varying hyper-ellipsoid (Fig.1) representing both within- and between-component (co)variance.

In Dataset B, vector field analysis reached significance (Fig.7), so scalar field analysis, had it not been conducted in a *post hoc* manner, would have exhibited Type II error by underestimating the temporal scope of effects (Fig.8). This is also explained by covariance (Appendix B); the effect was manifested more strongly in the resultant 10-component muscle force vector than it was in each muscle independently.

In Dataset C vector field effect timing (Fig.9) agreed with scalar field effect timing (Fig.10), so the latter would not have been biased had they not been conducted in a *post hoc* sense. Nevertheless, by failing to consider covariance the scalar field results fail to capture the full temporal extent of the vector effects.

A separate but notable trend was that Dataset C's covariance ellipses all tended to be narrow and to point toward the origin (Fig.1). This suggests that vector magnitude was far more variable than vector direction. A plausible mechanical explanation is friction: to avoid slipping normal forces must increase when tangential

forces increase. Regardless of the mechanism, this observation reinforces our contention that non-directed hypotheses must consider vector changes.

### 4.3 SPM’s solution to regional focus and covariance bias

SPM solves both regional focus bias and covariance bias by considering the covariance of all vector components ( $i$ ) across the entire measurement domain ( $q$ ), while simultaneously handling the inherent problem of multiple comparisons (Knudson, 2009) in a theoretically robust manner. Specifically, SPM uses a RFT correction (Adler and Taylor, 2007; Worsley et al., 2004) to ensure that no more than  $\alpha\%$  of the points in the ( $I \times Q$ ) vector field reach significance simply by chance; this RFT correction is embodied in the thresholds depicted in Figs.5-10.

Non-RFT corrections like the Šidák correction (Eqn.1) can partially solve the problem of multiple comparisons, but only partially because they fail to consider the (spatiotemporal) smoothness of the measurement domain  $q$ , and therefore overestimate the number of independent tests. This ultimately leads to an overly conservative threshold (i.e. inflated Type II error rate) except for very rough fields (Friston et al., 2007).

Non-RFT corrections also fail to solve covariance bias because they assume that vector components vary independently (Supplementary Material – Appendix B). While covariance bias could partially be solved with a principal axis rotation prior to statistical testing (Cole et al., 1994; Knudson, 2009); we’d argue that: (i) Hotelling’s  $T^2$  and CCA are simpler solutions because their results are identical for all coordinate system definitions, and (ii) principal axis rotations of only the response vectors do not necessarily maximize the mutual correlation between predictors and responses (§2.3.3).

We acknowledge that many additional important sources of bias exist (James and Bates, 1997), (Mullineaux et al., 2001), (Knudson, 2009). However, none of these is unique to SPM. Trajectory mis-registration (Sadeghi et al., 2003) and unit normalization (e.g. absolute vs. relative muscle forces), for example, pose common problems to scalar extraction and vector field analyses. We contend only that SPM addresses two bias sources.

### 4.4 SPM generalizability

Although SPM was originally developed to analyze 3D brain function (Friston et al., 2007), it has been shown that SPM is generalizable to a variety of biomechanical scalar datasets including 1D trajectories (Pataky, 2012), 2D pressure fields and 3D strain fields (Pataky, 2010). The current study is the first, in any scientific field, to have shown that SPM is also applicable to a large class of practical 1D *vector* field

problems. SPM theory suggests that generalizations to biomechanical vector/tensor fields in  $nD$  spaces are also possible (Xie et al., 2010).

SPM encompasses the entire family of parametric hypothesis testing (Worsley et al., 2004; Friston et al., 2007). It also accommodates all non-parametric variants (Nichols and Holmes, 2002; Lenhoff et al., 1999), which may be useful if one’s data do not adhere to the parametric assumption that the residuals are normally distributed (Friston et al., 2007). This hypothesis testing generalization is apparent when one considers the following hierarchy: vector field CCA simplifies to the vector field Hotelling’s  $T^2$  test when the predictors are binary (Worsley et al., 2004), which in turn simplifies to scalar field  $t$  tests when there is only one vector component  $i$ , which in turn simplifies to the univariate Student’s  $t$  test when the scalar field reduces to a single point  $q$ . Thus SPM, through CCA, generalizes to all statistical tests of  $I$  dimensional vectors on arbitrarily sized fields  $Q$  of arbitrary dimensionality (Worsley et al., 2004).

For readers interested in implementing SPM analyses, we note that constructing test statistic trajectories is straightforward; it is trivial to combine mean and standard deviation trajectories to form an SPM $\{t\}$ , for example. The non-trivial step is statistical inference. As a first approximation it is easy to implement a Šidák correction (Eqn.1), which will (very) conservatively reduce the Type I error rate, but which will also unfortunately inflate the Type II error rate. For more precise control of both error rates (via RFT) the reader is directed to the literature (Friston et al., 2007) and open source software packages (Pataky, 2012).

## 4.5 Conclusions

*Ad hoc* reduction of vector trajectories through scalar extraction can non-trivially bias non-directed biomechanical hypothesis testing, most notably via regional focus and coordinate system bias sources. This paper shows that SPM overcomes both sources of bias by treating the vector field as the fundamental, initially indivisible unit of observation. Grounded in random field theory, SPM appears to be a useful, generalized tool for the analysis of often-complex biomechanical datasets.

## Acknowledgments

Financial support for this work was provided in part by JSPS Wakate B Grant#22700465.

## Conflict of Interest

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

## References

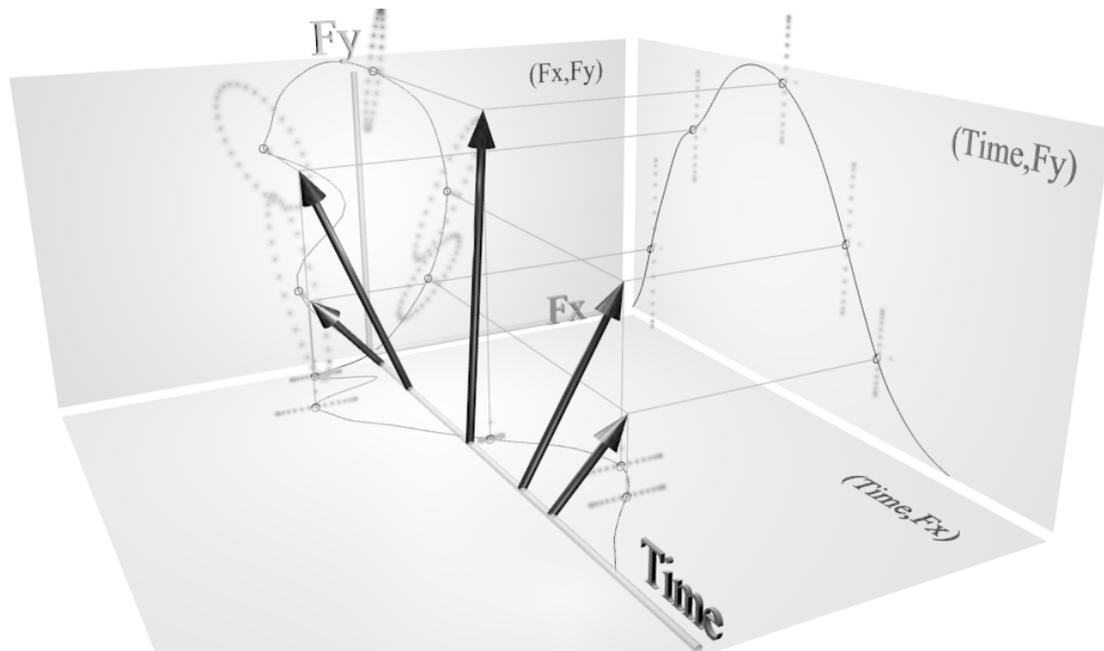
- Adler, R. J. and Taylor, J. E. 2007. *Random Fields and Geometry*, Springer-Verlag, New York.
- Besier, T. F., Fredericson, M., Gold, G. E., Beaupre, G. S., and Delp, S. L. 2009. Knee muscle forces during walking and running in patellofemoral pain patients and pain-free controls, *Journal of Biomechanics* **42(7)**, 898–905, data: <https://simtk.org/home/muscleforces>.
- Cao, J. and Worsley, K. J. 1999. The detection of local shape changes via the geometry of Hotelling's T<sup>2</sup> fields, *Annals of Statistics* **27(3)**, 925–942.
- Cole, D. A., Maxwell, S. E., Arvey, R., and Salas, E. 1994. How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables., *Psychological Bulletin* **115(3)**, 465–474.
- 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>.
- 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.
- Hotelling, H. 1936. Relations between two sets of variates, *Biometrika* **28(3)**, 321–377.
- James, C. R. and Bates, B. T. 1997. Experimental and statistical design issues in human movement research, *Measurement in Physical Education and Exercise Science* **1(1)**, 55–69.
- Jensen, R. H. and Davy, D. T. 1975. An investigation of muscle lines of action about the hip: A centroid line approach vs the straight line approach, *Journal of Biomechanics* **8(2)**, 103–110.
- Knudson, D. 2009. Significant and meaningful effects in sports biomechanics research, *Sports Biomechanics* **8(1)**, 96–104.
- Kutch, J. J. and Valero-Cuevas, F. J. 2011. Muscle redundancy does not imply robustness to muscle dysfunction, *Journal of Biomechanics* **44(7)**, 1264–1270.
- 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.
- Mullineaux, D. R., Bartlett, R. M., and Bennett, S. 2001. Research design and statistics in biomechanics and motor control, *Journal of Sports Sciences* **19(10)**, 739–760.
- 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.

- 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.
- Rayner, J. M. 1985. Linear relations in biomechanics: the statistics of scaling functions, *Journal of Zoology* **206(3)**, 415–439.
- 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.
- Wang, X., Verriest, J. P., Lebreton-Gadegbeku, B., Tessier, Y., and Trasbot, J. 2000. Experimental investigation and biomechanical analysis of lower limb movements for clutch pedal operation, *Ergonomics* **43(9)**, 1405–1429.
- Worsley, K. J., Taylor, J. E., Tomaiuolo, F., and Lerch, J. 2004. Unified univariate and multivariate random field theory, *NeuroImage* **23**, S189–S195.
- Xie, Y., Vemuri, B. C., and Ho, J. 2010. Statistical analysis of tensor fields, *Medical Image Computing and Computer-Assisted Intervention* **13(1)**, 682–698.

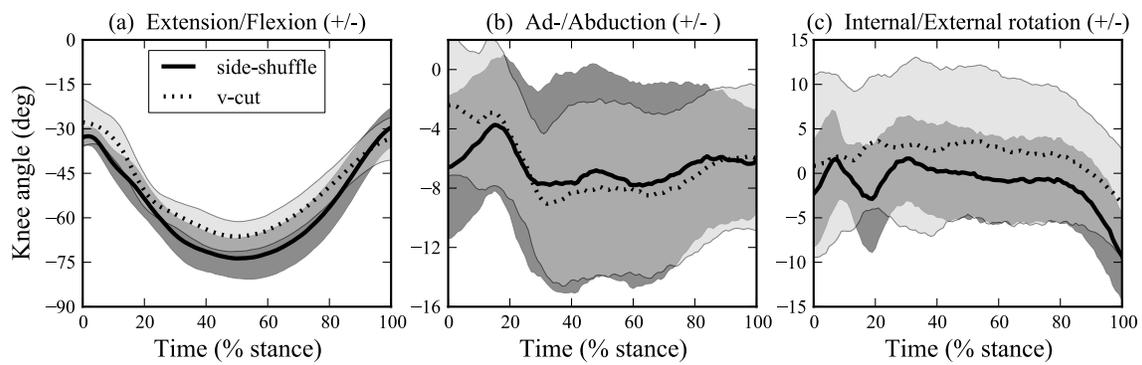
Table 1: Dataset and scalar extraction overview.  $I$ ,  $J$ ,  $Q$  and  $N$  are the numbers of vector components, responses, time points, and extracted scalars, respectively. For vector field analyses, *post hoc* scalar field analyses, and extracted scalar analyses we conducted one,  $I$  and  $N$  tests, respectively. Šidák thresholds of  $p=0.0253$ ,  $p=0.0170$  and  $p=0.0051$  maintained a family-wise error rate of  $\alpha=0.05$  across 2, 3, and 10 tests, respectively (see Eqn.1).

	$I$	$J$	$Q$	$N$	Extracted scalars
Dataset A	3	8	101	2	(1) Max. flexion (at $\sim 50\%$ stance) (2) Ad-abduction at $0\%$ stance
Dataset B	10	43	100	10	Max. force for each muscle ( $J_1=16$ , $J_2=27$ )
Dataset C	3	8	100	3	(1) Max. propulsion force ( $GRF_x$ , $\sim 75\%$ stance) (3) Max. vertical force ( $GRF_y$ , $\sim 30-50\%$ stance) (3) Max. lateral force ( $GRF_z$ , $\sim 15\%$ stance)

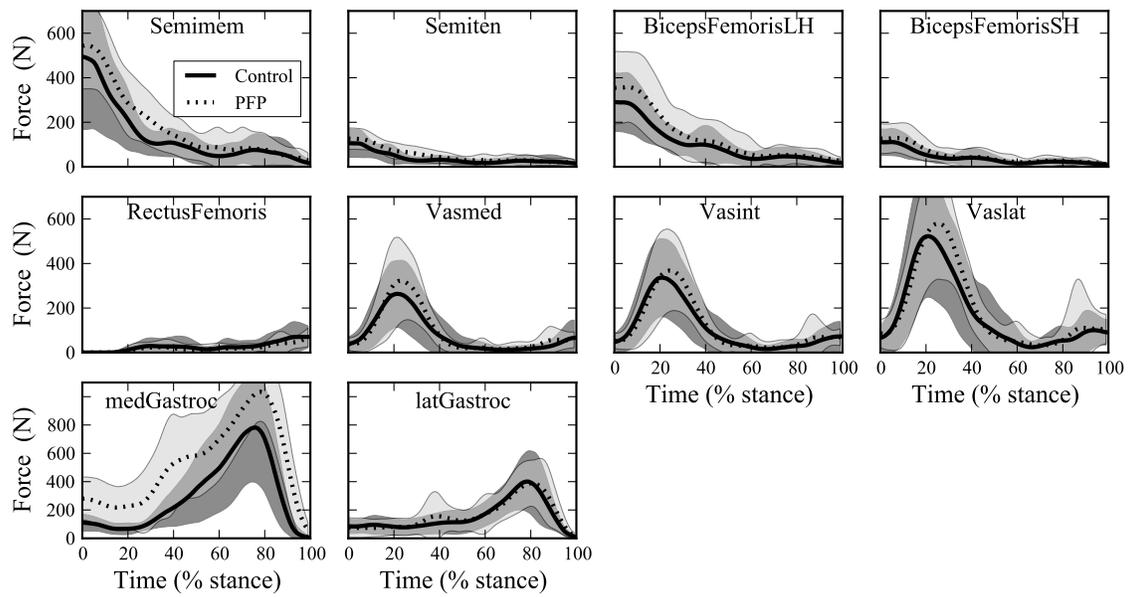
## FIGURES



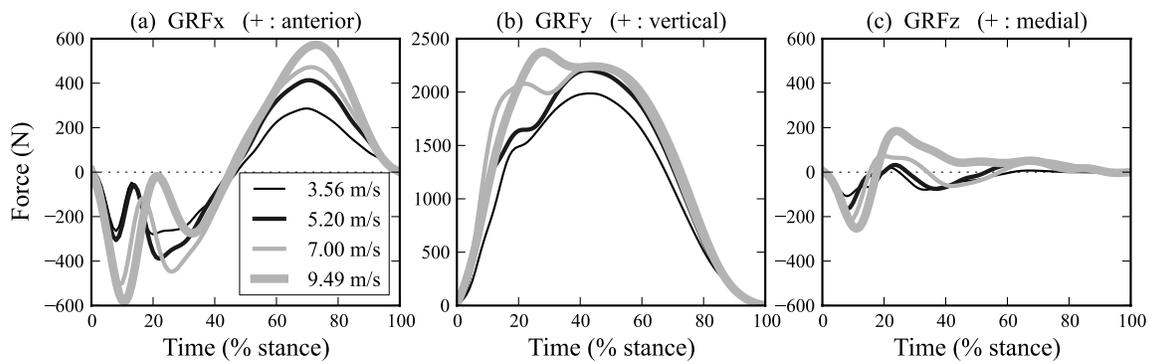
**Figure 1.** Vector field schematic: a two-component vector varying in time. Depicted are mean ground reaction force (GRF) vectors  $\mathbf{F} = [F_x \ F_y]^T$  from one subject during running (Dorn et al., 2012), where  $+x$  and  $+y$  represent the anterior and vertical directions, respectively. These vectors, when projected on the  $(Time, F_x)$  and  $(Time, F_y)$  planes, produce common GRF plots (see Fig. 4a,b); here vertical dotted lines depict standard deviation ‘clouds’. When  $\mathbf{F}$  is projected on the  $(F_x, F_y)$  plane these standard deviations are revealed to arise from covariance ellipses, where ellipse orientation indicates the direction of maximum covariance between  $F_x$  and  $F_y$  (see Supplementary Material - Appendix B).



**Figure 2.** Dataset A (Neptune et al., 1999) depicting knee kinematics in side-shuffle vs. v-cut tasks. Cross-subject mean trajectories with standard deviation clouds (dark: side-shuffle, light: v-cut) are depicted. Each of the eight subjects has three (scalar) trajectories  $y_i(q)$  for each task, and these were combined into a single ( $I=3$ ,  $Q=101$ ) vector field  $\mathbf{y}(q)$  for each subject and each task.



**Figure 3.** Dataset B (Besier et al., 2009) depicting muscle forces during walking in Control vs. Patello-Femoral Pain (PFP) subjects; 16 and 27 subjects, respectively. Cross-subject mean trajectories with standard deviation clouds (dark: Control, light: PFP). These ten scalar trajectories were combined into a single ( $I=10$ ,  $Q=100$ ) vector field  $y(q)$  for each subject.



**Figure 4.** Dataset C (Dorn et al., 2012) depicting ground reaction forces (GRF) during running/sprinting at various speeds. Single-subject cross-trial means; standard deviation clouds are not depicted in interest of visual clarity. These data form one ( $I=3$ ,  $Q=100$ ) vector field  $\mathbf{y}(q)$  for each trial.

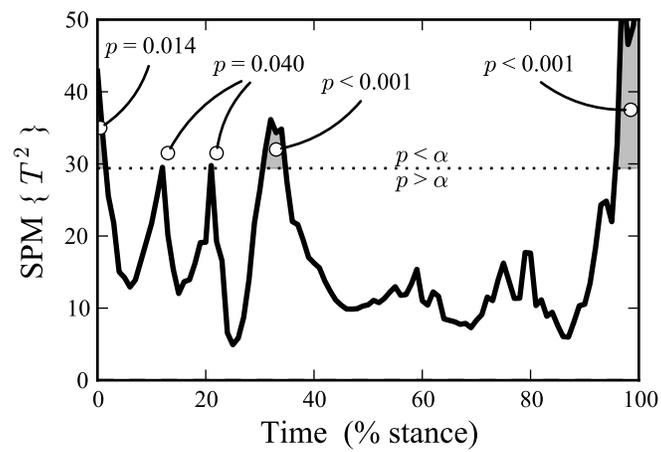
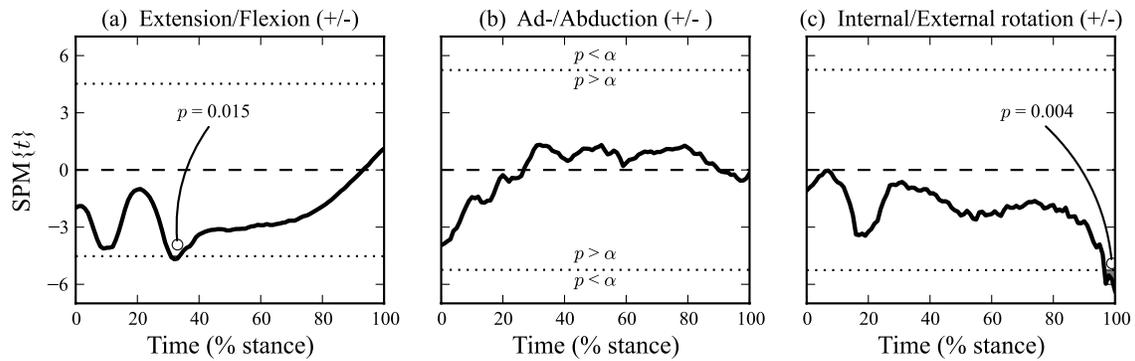


Figure 5. Dataset A, Hotelling's  $T^2$  trajectory ( $SPM\{T^2\}$ ). The horizontal dotted line indicates the critical random field theory threshold of  $T^2 = 29.39$ .



**Figure 6.** Dataset A, *post hoc* scalar field  $t$  tests ( $SPM\{t\}$ ), depicting where side-shuffle angles were greater (+) and less (-) than v-cut angles. At a Sidak threshold of  $p=0.017$  (Eqn.1), the thin dotted lines indicate the critical RFT thresholds for significance:  $|t| > 4.52, 5.24, 5.26$  for (a), (b), and (c) respectively. The thresholds are different because each vector component has different temporal smoothness (Fig.2); less smooth trajectories have higher thresholds because there are more ‘processes’ present between 0 and 100% time. Probability ( $p$ ) values indicate the likelihood with which each suprathreshold cluster is expected to have been produced by a random field process with the same temporal smoothness.

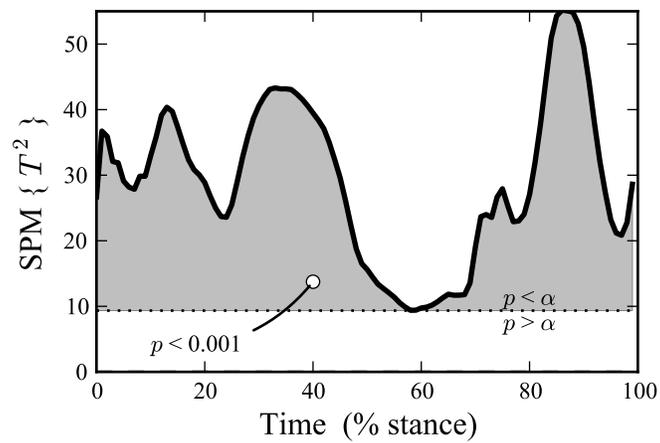


Figure 7. Dataset B, Hotelling's  $T^2$  trajectory ( $SPM\{T^2\}$ ), depicting where muscle forces differed between Controls and PFP. The horizontal dotted line indicates the critical RFT threshold of  $T^2 = 9.35$ . The entire trajectory has exceeded the threshold, so the single suprathreshold cluster has a very low  $p$  value.

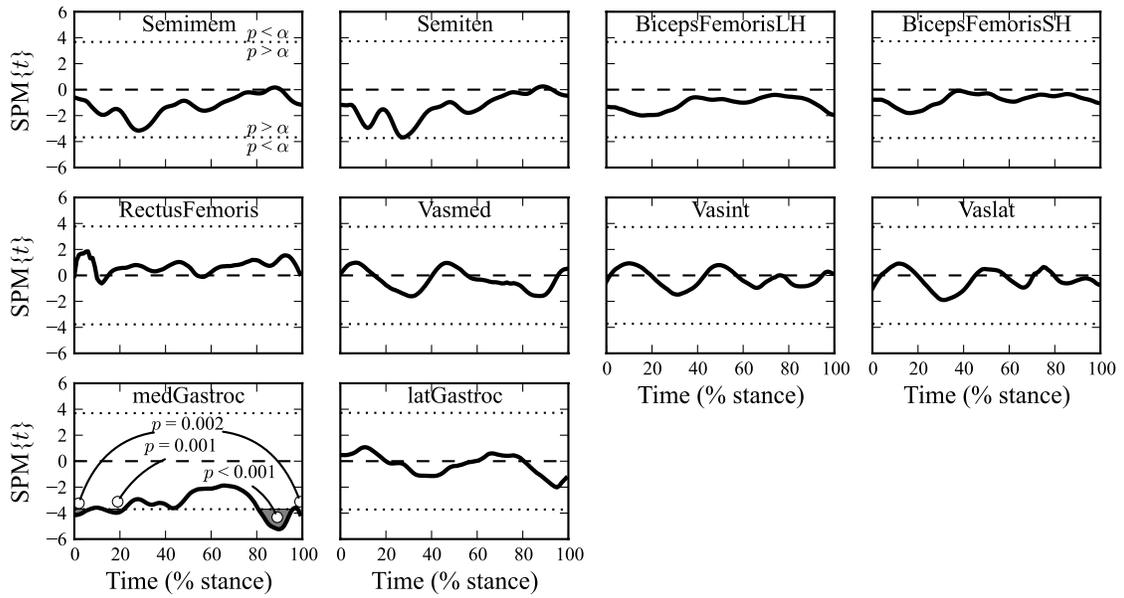


Figure 8. Dataset B, *post hoc* scalar trajectory  $t$  tests ( $SPM\{t\}$ ), depicting where Control forces were greater than (+) and less than (-) PFP forces. Thin dotted lines indicate the critical RFT thresholds for significance.

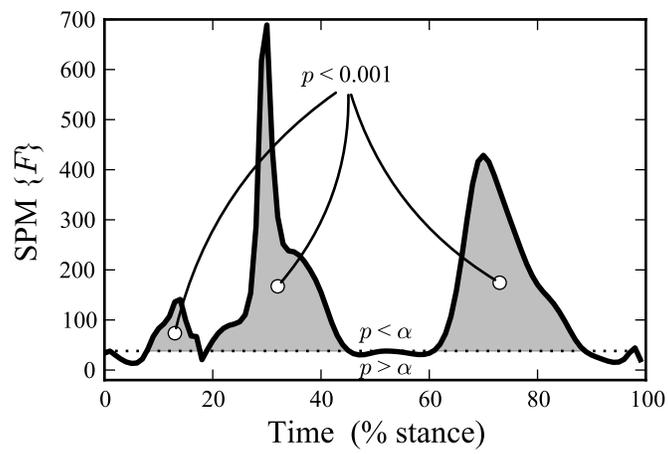
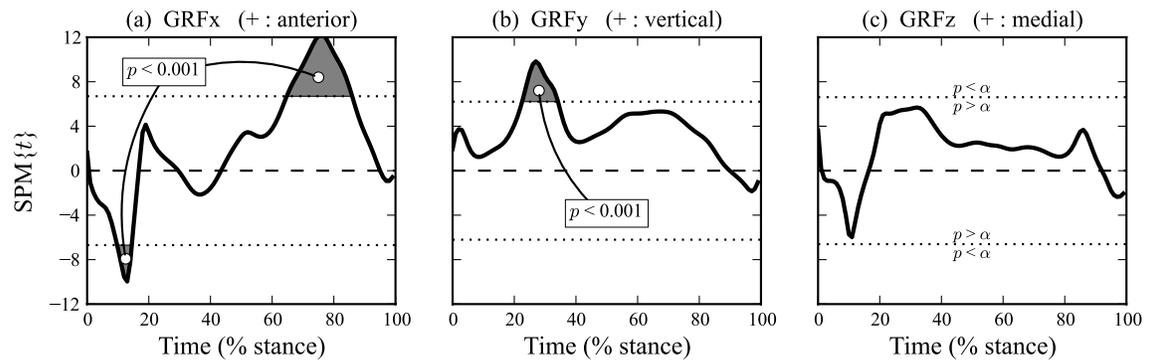


Figure 9. Dataset C, canonical correlation analysis results, with  $SPM\{F\}$  depicting where ground reaction forces were correlated with running speed. Critical RFT threshold:  $F = 38.1$ .



**Figure 10.** Dataset C, *post hoc* scalar trajectory linear regression tests ( $SPM\{t\}$ ), depicting the strength of positive (+) and negative (-) correlation between ground reaction forces (GRF) and running speed.

## Appendix A. Scalar extraction vs. scalar field statistics

The purpose of this Appendix is to demonstrate how scalar extraction can bias non-directed hypothesis testing. To this end we developed and analyzed an arbitrary dataset (Fig.S1). We caution readers that we have constructed these data specifically to demonstrate particular concepts. The reader is therefore left to judge the relevance of this discussion to real (experimental) datasets.

The specific goal of this Appendix is to scrutinize the similarities and differences between: (a) a typical univariate two-sample  $t$  test, and (b) a scalar field two-sample  $t$  test.

Consider the simulated scalar field dataset in Fig.S1. In Fig.S1a, arbitrary true mean fields are defined for two experimental conditions: “Cond A” and “Cond B”. The Cond B mean was produced using a half sine cycle. The Cond A mean was produced by adding a small Gaussian pulse (at time= 85%) to the Cond B mean. This Gaussian pulse is evident in the true mean field difference (Fig.S1b).

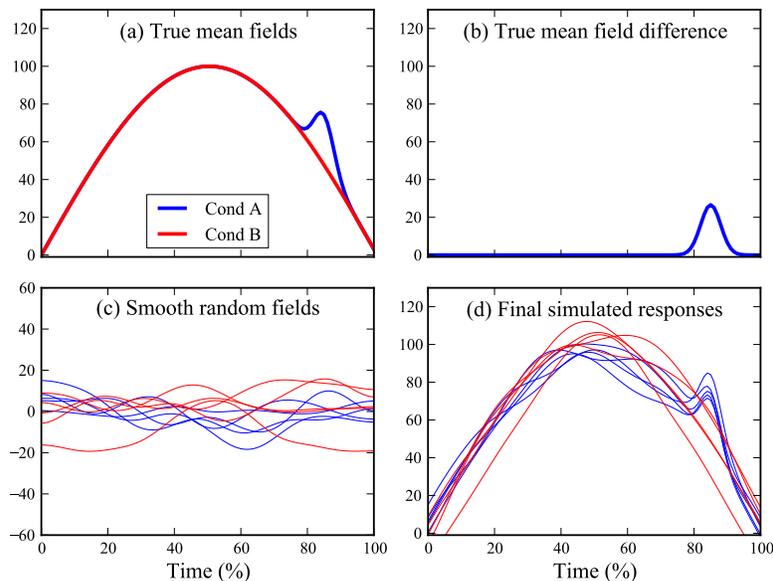


Figure S1: Simulated scalar field dataset depicting two experimental conditions: “Cond A” and “Cond B” (arbitrary units).

We next simulate smooth random fields: five for each condition (Fig.S1c). These random fields were constructed by generating ten fields, each containing 100 random, uncorrelated and normally distributed numbers, then smoothing them using a Gaussian kernel. Adding the random fields to the true field means (Fig.S1a) produced the final simulated responses (Fig.S1d). For interpretive convenience, let us assume that these data represent joint flexion.

Imagine next that we wish to test the following (non-directed) null hypothesis: “Cond A and Cond B yield identical kinematics”. Consider first scalar extraction: after observing the data (Fig.S1d) one might decide to extract and analyze the maximum flexion, which occurs near time = 50%:

$$y_A = [ 100.0 \quad 91.2 \quad 92.2 \quad 95.5 \quad 97.1 ]$$

$$y_B = [ 97.2 \quad 101.9 \quad 104.8 \quad 106.3 \quad 111.7 ]$$

A two-sample  $t$  test on these data yields:  $t=3.16$ ,  $p=0.013$ . We would reject the null hypothesis at  $\alpha=0.05$ , and we would conclude that Cond B produces significantly greater maximal flexion than Cond A.

An alternative is to use Statistical Parametric Mapping (SPM) (Fig.S2). The SPM procedures are conceptually identical to univariate procedures (Table S1). The only apparent difference is that SPM uses a different probability distribution (Steps 4 and 5). This probability distribution is actually not different because it reduces to the univariate distribution when  $Q=1$  (i.e. if there is only one time point).

SPM results find significant differences between the two conditions near time = 85% (Fig.S2d). We would therefore reject our null hypothesis, with the caveat that significant differences were only found near time = 85%.

Although univariate  $t$  testing and SPM  $t$  testing are conceptually identical, they have yielded (effectively) opposite results. The univariate test found significantly greater maximal flexion in Cond B, but SPM found significantly greater flexion in Cond A (near time=85%).

Table S1: Comparison of computational steps for univariate and SPM two-sample  $t$  tests (“st.dev.” = standard deviation).

Step	(a) Univariate two-sample $t$ test	(b) SPM two-sample $t$ test	Figure
1	Compute mean values $\bar{y}_A$ and $\bar{y}_B$ .	Compute mean fields $\bar{y}_A(q)$ and $\bar{y}_B(q)$	S2(b)
2	Compute st.dev. values $s_A$ and $s_B$ .	Compute st.dev. fields $s_A(q)$ and $s_B(q)$	S2(b)
3	Compute the $t$ test statistic: $t = \frac{\bar{y}_B - \bar{y}_A}{\sqrt{\frac{1}{J}(s_A^2 + s_B^2)}}$	Compute the $t$ test statistic field: $\text{SPM}\{t\} \equiv t(q) = \frac{\bar{y}_B(q) - \bar{y}_A(q)}{\sqrt{\frac{1}{J}(s_A^2(q) + s_B^2(q))}}$	S2(c)
4	Conduct statistical inference. First use $\alpha$ and the univariate $t$ distribution to compute $t_{\text{critical}}$ . If $t > t_{\text{critical}}$ , then reject null hypothesis.	Conduct statistical inference. First use $\alpha$ and the random field theory $t$ distribution to compute $t_{\text{critical}}$ . If $\text{SPM}\{t\}$ exceeds $t_{\text{critical}}$ , then reject null hypothesis for the suprathreshold region(s).	S2(d)
5	Compute exact $p$ value using $t$ and the univariate $t$ distribution.	Compute exact $p$ value(s) for each suprathreshold cluster using cluster size and random field theory distribution(s) for $\text{SPM}\{t\}$ topology.	S2(d)

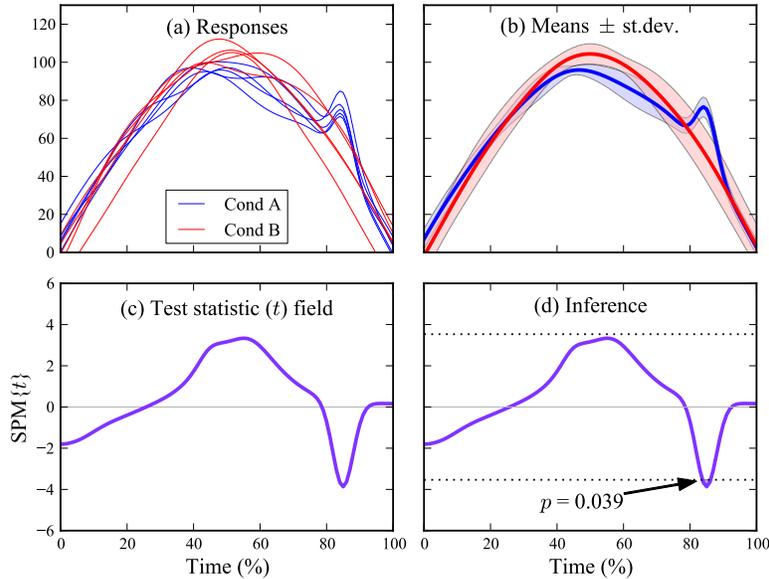


Figure S2: Scalar field analysis using Statistical Parametric Mapping (SPM). In panel (d) the thin dotted lines depict the critical random field theory threshold of  $|t_{\text{critical}}|=3.533$ . The (incorrect) Šidák threshold is  $|t_{\text{critical}}|=5.595$ .

This discrepancy can be resolved through standard probability theory regarding multiple comparisons, through a consideration of ‘corrected’ and ‘uncorrected’ thresholds. First consider conducting one statistical test at  $\alpha=0.05$ . The choice: “ $\alpha=0.05$ ” means that we are accepting a 5% chance of incorrectly rejecting the null hypothesis, or, equivalently, a 5% chance of a ‘false positive’. If we conduct more than one test, there is a greater-than 5% chance of a false positive. Specifically, if we conduct  $N$  statistical tests, the probability of at least one false positive is given by the family-wise error rate  $\bar{\alpha}$ :

$$\bar{\alpha} = 1 - (1 - \alpha)^N$$

For  $N=2$  tests, there is a  $\bar{\alpha}=9.75\%$  chance that at least one test will produce a false positive. For  $N=100$  tests,  $\bar{\alpha}=99.4\%$ .

To protect against false positives, and to maintain a constant family-wise error rate of  $\bar{\alpha}=0.05$ , we must adopt a corrected threshold. One option is the Šidák threshold:

$$p_{\text{critical}} = 1 - (1 - \bar{\alpha})^{1/N}$$

For  $N=2$  and  $N=100$  tests, the Šidák thresholds are  $p_{\text{critical}}=0.0253$  and  $p_{\text{critical}}=0.000513$ , respectively.

Herein lies one problem: our scalar extraction analysis has used an uncorrected threshold of  $p_{\text{critical}}=0.05$ . Even though we have formally conducted only one statistical test, the data were extracted from a dataset that is 100 times as large. Since we observed the data before choosing which scalar to extract, we effectively conducted  $N=100$  tests, albeit visually, then chose to focus on only one test. By failing to adopt a corrected threshold, we have biased our analyses.

Although the Šidák correction helps to avoid false positives, it is not generally a good choice because it assumes that there are 100 independent tests (i.e. one for each time point in our dataset). The points in this dataset are clearly not independent because the curves are smooth, changing only gradually over time. Thus the Šidák correction is too severe, lowering  $\bar{\alpha}$  well

below 0.05. An overly severe threshold produces the opposite bias: an increased chance of false negatives.

SPM employs a random field theory (RFT) correction to more accurately maintain  $\bar{\alpha}=0.05$ . The precise threshold is based not only on field size ( $Q=100$ ), but also on field smoothness — which is estimated from temporal derivatives. Computational details for RFT corrections are provided in the SPM literature.

Unfortunately, even if our scalar analysis had employed a corrected threshold, it still would have been biased, but for a separate reason. By focussing only on maximal flexion (which did not appear in our null hypothesis), we have neglected to consider the signal at time = 85%, and have therefore not detected the true field difference (Fig.S1a). In contrast, SPM was able to uncover the true signal because it both adopted a corrected threshold and considered the entire field simultaneously (Fig.S1d).

The aforementioned sources of bias — (1) failing to adopt a corrected threshold, and (2) failing to consider the entire field — are referred to collectively in the main manuscript as ‘regional focus bias’.

Last, we reiterate that this Appendix is relevant only to non-directed hypotheses. If we had formulated a (directed) hypothesis regarding only maximal flexion — prior to observing the data — then our scalar extraction analyses would not have been biased because our null hypothesis would not have pertained to the entire time domain 0–100%.

In summary, regional focus bias can be avoided by:

1. Specifying a directed null hypothesis — before observing the data — and then extracting only those scalars which are specified in the null hypothesis.
2. Analyzing the data using SPM or another field technique which both considers the entire temporal domain and which adopts a corrected threshold.

## Appendix B. Univariate vs. vector analysis

The purpose of this Appendix is to demonstrate how univariate testing of vector data can bias non-directed hypothesis testing. To this end we developed and analyzed an arbitrary dataset (Table S2). As in Appendix A, we caution readers that we have constructed these data specifically to demonstrate particular concepts. The reader is therefore left to judge the relevance of this discussion to real (experimental) datasets.

The specific goal of this Appendix is to compare and contrast the (univariate)  $t$  test and its (multivariate) vector equivalent: the Hotelling's  $T^2$  test.

Table S2: A simulated dataset exhibiting biased univariate testing. (a) Two-component force vector responses  $\mathbf{F} = [F_x, F_y]^\top$ . (b)-(d) Scalar (univariate) testing. (e)-(g) Vector (multivariate) testing. Sources of bias and further details are discussed in the text. Technical overviews of covariance matrices ( $\mathbf{W}$ ) and the Hotelling's  $T^2$  statistic are provided in Appendix D and §2.3 (main manuscript), respectively.

		Group A	Group B	Inter-Group
	(a) Responses	$\mathbf{F}_{A1} = [159, 719]^\top$ $\mathbf{F}_{A2} = [115, 762]^\top$ $\mathbf{F}_{A3} = [177, 681]^\top$ $\mathbf{F}_{A4} = [138, 694]^\top$ $\mathbf{F}_{A5} = [98, 697]^\top$	$\mathbf{F}_{B1} = [143, 759]^\top$ $\mathbf{F}_{B2} = [172, 734]^\top$ $\mathbf{F}_{B3} = [161, 735]^\top$ $\mathbf{F}_{B4} = [195, 733]^\top$ $\mathbf{F}_{B5} = [168, 706]^\top$	
Univariate	(b) Means	$(\overline{F_x})_A = 137.4$ $(\overline{F_y})_A = 710.6$	$(\overline{F_x})_B = 167.8$ $(\overline{F_y})_B = 733.4$	$\Delta\overline{F_x} = 30.4$ $\Delta\overline{F_y} = 22.8$
	(c) St.dev.	$(s_x)_A = 28.6$ $(s_y)_A = 28.5$	$(s_x)_B = 16.8$ $(s_y)_B = 16.8$	$s_x = 23.5$ $s_y = 23.4$
	(d) $t$ tests			$t_x=1.832; p_x=0.104$ $t_y=1.380; p_y=0.205$
Vector	(e) Means	$\overline{\mathbf{F}}_A = [137.4, 710.6]^\top$	$\overline{\mathbf{F}}_B = [167.8, 733.4]^\top$	$\Delta\overline{\mathbf{F}} = [30.4, 22.8]^\top$
	(f) Covariance	$\mathbf{W}_A = \begin{bmatrix} 817.8 & -323.2 \\ -323.2 & 809.8 \end{bmatrix}$	$\mathbf{W}_B = \begin{bmatrix} 283.8 & -131.9 \\ -131.9 & 281.8 \end{bmatrix}$	$\mathbf{W} = \begin{bmatrix} 550.8 & -227.6 \\ -227.6 & 545.8 \end{bmatrix}$
	(g) $T^2$ test			$T^2=7.113; p=0.028$

In Table S2(a) above there are five force vector responses ( $\mathbf{F} = [F_x, F_y]^\top$ ) for each of two groups: “A” and “B”. Their means and standard deviations are shown in Table S2(b)-(c). In Table S2(d) we see that  $t$  tests pertaining to both  $F_x$  and  $F_y$  fail to reach significance;  $p$  values are greater than (even an uncorrected) threshold of  $p = 0.05$ . An adequate interpretation is that the mean force component changes ( $\Delta\bar{F}_x$  and  $\Delta\bar{F}_y$ ) are not unexpectedly large given their respective variances (i.e. standard deviations:  $s_x$  and  $s_y$ ).

We next jump ahead to the final results of the vector procedure in Table S2(g): here we see that the Hotelling’s  $T^2$  test reached significance ( $p = 0.032$ ). An adequate interpretation is that the mean force vector change ( $\Delta\bar{\mathbf{F}}$ ) was unexpectedly large given its (co)variance ( $\mathbf{W}$ ). Let us now backtrack and consider why the univariate and vector procedures yield different results.

The first step of the vector procedure is to compute mean vectors; in Table S2(e) we can see that the vector means have the same component values as the univariate means from Table S2(b). However, there is already one critical discrepancy to note: the vector procedure assesses  $\Delta\bar{\mathbf{F}}$ , which is the **resultant** vector connecting the Group A and Group B means (Fig.S3). From Pythagoras’ theorem:

$$|\Delta\bar{\mathbf{F}}|^2 = \Delta\bar{F}_x^2 + \Delta\bar{F}_y^2 \quad (\text{B.1})$$

it is clear that the magnitude of the resultant will always be greater than the magnitude of its components — except in the experimentally unlikely cases of  $\Delta\bar{F}_x=0$  and/or  $\Delta\bar{F}_y=0$ . This is non-trivial for two reasons. First, since the vector procedure assesses the maximum difference between the two groups, it is more robust to Type II error than univariate procedures (note: the univariate tests in Table S2 exhibit Type II error by failing to reach significance). Second, the vector technique’s assessment of differences is independent of the  $xy$  coordinate system definition; whereas the component effects ( $\Delta\bar{F}_x$  and  $\Delta\bar{F}_y$ ) can change when the  $xy$  coordinate system definition changes, both the resultant and the variance along the resultant direction will always have the same magnitude. This may have non-trivial implications for biomechanical datasets that employ difficult-to-define coordinate systems (e.g. joint rotation axes).

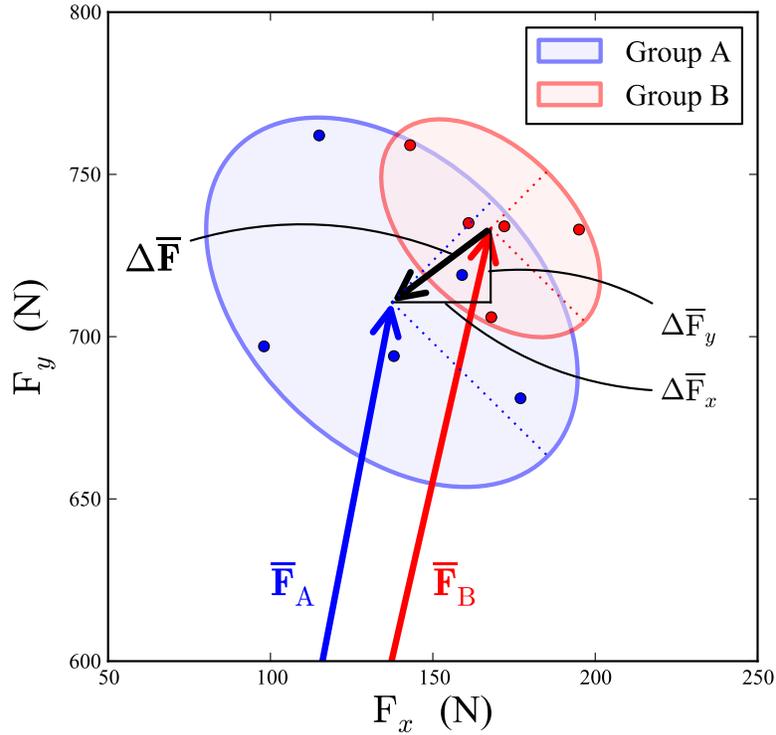


Figure S3: Graphical depiction of the data from Table S2. Small circles depict individual responses. Thick colored arrows depict the mean force vectors for the two groups. The thick black arrow depicts the (vector) difference between the two groups, and thin black lines indicate its  $x$  and  $y$  components. The ellipses depict within-group (co)variance; their principal axes (thin dotted lines) are the eigenvectors of the covariance matrices in Table S2(f). Here covariance ellipse radii are scaled to two principal axis standard deviations (to encompass all responses).

The next step of the vector procedure is to compute covariance matrices  $\mathbf{W}$  (Appendix D). The diagonal elements of  $\mathbf{W}_A$  and  $\mathbf{W}_B$  in Table S2(f) are simply the variances (i.e. squared standard deviations)  $s_x^2$  and  $s_y^2$  from Table S2(c). The off-diagonal terms are equal and represent the covariance (i.e. correlation) between  $F_x$  and  $F_y$ . If  $F_x$  tends to increase when  $F_y$  increases then the off-diagonal terms would be positive, but in this case they are negative, indicating that  $F_x$  tends to decrease when  $F_y$  increases. This tendency can be seen in the raw data (small circles) in Fig.S3.

The presence of non-zero off-diagonal terms thus has a critical implication: changes in  $F_x$

and  $F_y$  are not independent. This is critical because univariate tests implicitly assume that  $F_x$  and  $F_y$  are independent.

To appreciate this point it is useful to recognize that covariance matrices may be interpreted geometrically as ellipses: the eigenvectors of  $\mathbf{W}$  represent the ellipse's principal axes, and its eigenvalues represent the variance along each principal direction. This is perfectly analogous to inertia matrices: the eigenvectors of an inertia matrix define a body's principal axes of inertia, and eigenvalues specify the principal moments of inertia.

The importance of this geometric interpretation becomes clear when visualizing covariance ellipses. In Fig.S3 we can see that the principal axes of the covariance matrices are not aligned with the  $xy$  coordinate system, implying that changes in  $F_x$  and  $F_y$  are not independent. Critically, we can also see that the direction of minimum variance is very similar to the direction of  $\Delta\bar{\mathbf{F}}$ . Thus the standard deviations  $s_x$  and  $s_y$  (used in the univariate analyses) are larger than the standard deviation in the direction of  $\Delta\bar{\mathbf{F}}$ .

In summary, vector statistical testing more objectively detects vector changes because : (a) it is coordinate system-independent, (b) it considers both the maximum difference between groups (i.e. the resultant difference) and the variation along this direction. This Appendix has demonstrated how univariate testing of vector data can lead to Type II error. With a different dataset it would also be possible to demonstrate Type I error, but in interest of space we end here. The most important point, the main paper contends, is that non-directed hypothesis testing mustn't assume vector component independence.

## Appendix C. Mean vector field calculation

An  $I$ -component vector  $\mathbf{y}$  which varies over  $Q$  points in space or time may be regarded as an  $(I \times Q)$  vector field response  $\mathbf{y}(q)$ . Given  $J$  responses, the mean vector field is:

$$\bar{\mathbf{y}}(q) = \frac{1}{J} \sum_{j=1}^J \mathbf{y}_j(q) \quad (\text{C.1})$$

For the paired Hotelling's  $T^2$  test (Dataset A: §2.3.1, main manuscript), one must first compute pairwise differences:

$$\Delta \bar{\mathbf{y}}_j(q) = \mathbf{y}_{Bj}(q) - \mathbf{y}_{Aj}(q) \quad (\text{C.2})$$

where “A” and “B” represent the two tasks (v-cut and side-shuffle) and  $j$  indexes the subjects. A paired Hotelling's  $T^2$  test is thus equivalent to a one-sample Hotelling's  $T^2$  test conducted on the pairwise differences  $\Delta \bar{\mathbf{y}}(q)$ . The same is true in the univariate case: a paired  $t$  test is equivalent to a one-sample  $t$  test on pairwise differences.

## Appendix D. Covariance matrices

Although the concepts presented below apply identically to vector fields, for brevity present discussion is limited to simple vectors.

Consider a two-component force vector response  $\mathbf{F}$ :

$$\mathbf{F}_j = [F_{xj} \quad F_{yj}]^\top \quad (\text{D.1})$$

where  $j$  indexes the responses, and there are a total of  $J$  responses. After computing the mean force vector  $\bar{\mathbf{F}}$  as:

$$\bar{\mathbf{F}} = \begin{bmatrix} \bar{F}_x \\ \bar{F}_y \end{bmatrix} = \frac{1}{J} \sum_{j=1}^J \mathbf{F}_j \quad (\text{D.2})$$

the covariance matrix  $\mathbf{W}$  can be assembled as follows:

$$\mathbf{W} = \begin{bmatrix} W_{xx} & W_{xy} \\ W_{yx} & W_{yy} \end{bmatrix} \quad (\text{D.3})$$

where the elements of  $\mathbf{W}$  are:

$$W_{xx} = \frac{1}{J-1} \sum_{j=1}^J (F_{xj} - \bar{F}_x)^2 \quad (\text{D.4})$$

$$W_{yy} = \frac{1}{J-1} \sum_{j=1}^J (F_{yj} - \bar{F}_y)^2 \quad (\text{D.5})$$

$$W_{xy} = W_{yx} = \frac{1}{J-1} \sum_{j=1}^J (F_{xj} - \bar{F}_x)(F_{yj} - \bar{F}_y) \quad (\text{D.6})$$

Thus the diagonal elements  $W_{xx}$  and  $W_{yy}$  are the intra-component variances (i.e. squared

standard deviations), and the off-diagonal elements  $W_{xy}$  and  $W_{yx}$  are the inter-component covariances between  $F_x$  and  $F_y$  over multiple responses. Importantly, changes in  $F_x$  and  $F_y$  are completely uncorrelated if and only if  $W_{xy}=0$ .

One contention of this paper is that separate (univariate) analysis of  $F_x$  and  $F_y$  is biased when testing non-directed hypotheses. The main reason is that  $F_x$  analysis considers only  $W_{xx}$  and  $F_y$  analysis considers only  $W_{yy}$ . This is equivalent to assuming  $W_{xy}=0$ , an assumption which may not be valid (Appendix B).

A geometric interpretation of  $\mathbf{W}$  is useful both for visualizing vector variance (Fig.S3) and for appreciating canonical correlation analysis (Appendix E). Consider that  $\mathbf{W}$  represents an ellipse whose geometry is defined by the solutions to the eigenvalue problem:

$$\mathbf{W}\mathbf{v} = \lambda\mathbf{v} \tag{D.7}$$

Here  $\mathbf{v}$  and  $\lambda$  are the eigenvectors and eigenvalues, respectively, and there are two unique eigensolutions unless both  $(W_{xx} = W_{yy})$  and  $(W_{xy} = 0)$ , in which case there is only one eigensolution and  $\mathbf{W}$  represents a circle. When there are two solutions the eigenvectors represent the ellipse axes (or equivalently: principal axes), and the eigenvalues represent the axes' lengths (or variance in the direction of the principal axes). An equivalent interpretation is that one eigenvector of  $\mathbf{W}$  represents the direction of maximum variance within the dataset. This means that we can rotate our original coordinate system  $xy$  to a new coordinate system  $x'y'$  so that variance along the new  $x'$  axis is the maximum possible variance obtainable for all possible  $x'$ .

## Appendix E. Canonical correlation analysis (CCA)

CCA aims to quantify the amount of variance that a multivariate predictor (i.e. vector)  $\mathbf{X}$  can explain in a multivariate response  $\mathbf{Y}$ . One type of CCA useful for hypothesis testing is to find the maximum possible correlation coefficient that can be obtained when the coordinate systems defining  $\mathbf{X}$  and  $\mathbf{Y}$  are permitted to mutually rotate.

Consider a response variable  $\mathbf{Y}$  that describes three orthogonal force components  $F$ :

$$\mathbf{Y}_j = [F_{1j} \ F_{2j} \ F_{3j}]^\top \quad (\text{E.1})$$

where “1”, “2” and “3” represent orthogonal axes and where  $j$  indexes a total of  $J$  responses. Next consider a predictor variable  $\mathbf{X}$  that describes the rotations  $\theta$  about two orthogonal axes at a given joint:

$$\mathbf{X}_j = [\theta_{1j} \ \theta_{2j}]^\top \quad (\text{E.2})$$

where “1” and “2” indicate the two joint axes. The relevant null hypothesis is:  $\mathbf{X}$  and  $\mathbf{Y}$  are not linearly related.

To test this hypothesis one needs to assemble three covariance matrices. The first is a  $(3 \times 3)$  response covariance matrix  $\mathbf{W}_{YY}$  which describes variance within and the co-variation between the three force components (see Appendix D). The second is a  $(2 \times 2)$  predictor covariance matrix  $\mathbf{W}_{XX}$  which describes the variance and covariance of the two joint angles. The third is a  $(2 \times 3)$  predictor-response covariance matrix  $\mathbf{W}_{XY}$  which describes how each of the predictor variables co-varies with each of the response variables.

The predictor-response covariance matrix  $\mathbf{W}_{XY}$  is relevant to the null hypothesis because it embodies the strength of linear correlation between  $\mathbf{X}$  and  $\mathbf{Y}$ . For completion, in the example above  $\mathbf{W}_{XY}$  has six elements, corresponding to:

1. The linear correlation between  $\theta_1$  and  $F_1$
2. The linear correlation between  $\theta_1$  and  $F_2$
3. The linear correlation between  $\theta_1$  and  $F_3$
4. The linear correlation between  $\theta_2$  and  $F_1$
5. The linear correlation between  $\theta_2$  and  $F_2$
6. The linear correlation between  $\theta_2$  and  $F_3$

Initially these correlations refer only to  $\mathbf{X}$ 's and  $\mathbf{Y}$ 's original coordinate systems. Since arbitrary coordinate systems can bias non-directed hypothesis testing (Appendix B), we must allow the coordinate systems to rotate in order to most objectively test our null hypothesis.

One CCA solution is to choose the  $\mathbf{X}$  and  $\mathbf{Y}$  coordinate systems that mutually maximize a single correlation coefficient. The logic is that all other coordinate systems underestimate correlation strength. In other words, as the coordinate systems rotate the elements of  $\mathbf{W}_{XY}$  change, and one (not necessarily unique) coordinate system combination maximizes an element of  $\mathbf{W}_{XY}$ . CCA solves this problem efficiently using the maximum eigenvalue of the canonical correlation matrix (Eqn.7, main manuscript).

As an aside, we note that the  $K=2$  model in the main manuscript is equivalent to a  $K=1$  model (i.e. only a running speed regressor) because only one (diagonal) element of  $\mathbf{W}_{XX}$  is non-zero. For generalizability the main manuscript treats CCA in its  $K > 1$  form.