On the Robustness of Stride Frequency Estimation

Tom Chau*, *Senior Member, IEEE,* and Kim Parker

***Abstract*—The robustness of stride frequency estimation (location and spread) from stride period data is investigated using influence functions. Theoretical analysis reveals that stride frequency estimates by Stokes *et al.* [1] and by direct calculation have unbounded influence functions and zero breakdown points, implying a lack of both local and global robustness. Comparison of estimates obtained from an ensemble of pathological gait stride time series shows that on average, differences among estimators are not statistically significant ($p \geq 0.59$) for long time series (hundreds of strides). Specific circumstances under which nonrobust estimates depart from robust estimates are investigated in terms of outlier influence. We recommend some heuristic rules-of-thumb for prudent selection of nonrobust stride frequency estimators for a given stride time series. The theoretical and empirical estimator comparisons suggest that in general, more research on estimator robustness in quantitative gait analysis is warranted.**

***Index Terms*—Estimation, robustness, stride frequency, stride period.**

Nomenclature

$\mathbf{X}$	Stride period random variable.
$\mathbf{Y}$	Stride frequency random variable.
μ_y	Stride frequency location estimator.
σ_y^2	Stride frequency spread estimator.
F	Probability distribution.
$\hat{f}(\cdot)$	Estimated probability density.
$IF(\cdot)$	Influence function.
$SC(\cdot)$	Sensitivity curve.
$T(\cdot)$	Functional.

I. Introduction

It is known that the stride period varies naturally from stride to stride in able-bodied, neurologically healthy individuals. Recent estimates peg this variability at about 6% of the gait cycle on average in adults and approximately 8.5% in children [2]. Hence, the stride period is a stochastic rather than a deterministic variable. In light of this, Stokes *et al.* [1] remark that simple equations such as $\text{stride frequency} = 1/\text{stride period}$ should be employed with caution. Considering stride period and frequency as random variables, Stokes *et al.* use the fundamental theorem for defining densities of functions of one random variable [3] to estimate the stride frequency location and spread from stride period data. The consequence of not treating the stride variables as random quantities leads to a biased estimate of the stride frequency, as suggested by the Stokes location estimator [1]. This consequence holds regardless of the type of data (e.g., heatlhy, pathological, immature, or elderly gait) under examination. It is important to note that, with stride frequency estimation, like parameter estimation in general [4], an adequate sample of stride period data is required to minimize bias and variance in the estimate. The discussions herein will assume that we have adequately sized samples of stride period data.

In arriving at their estimates, Stokes *et al.* make the fundamental assumption that the stride period fluctuates closely around its mean. We know, however, that this assumption is often violated in the face of gait pathologies, such as cerebral palsy [5], Parkinson's disease [6], and hemiparetic stroke [7]. It is also known that developing gait exhibits higher stride-to-stride variability [8], [2] than does mature gait. This paper scrutinizes the assumption of small fluctuations, both theoretically and empirically. In particular, the purpose of this paper is to determine the robustness of several known stride frequency estimates including those proposed by Stokes *et al.* [1], in the presence of inflated stride period deviations that are often characteristic of pathological and immature gait.

Although the field of robust statistics has seldomly intersected with studies of walking, we argue that in fact, the statistical stability of estimated parameters is very important to quantitative gait analysis. First, many clinically relevant variables are routinely derived from measured quantities. For example, cadence is estimated as the number of steps over time and velocity can be approximated as the distance over time. When measurements are made over extended periods of time, such as in studies of nonlinear dynamics of gait [9], it is imperative that the measured variables be viewed as random variates and any derived quantity, thus, becomes a function of one or more random variates. Failing to take into account the robustness of the derived quantities may lead to inaccurate and clinically meaningless numbers. Second, the use of inductively determined functions of measured and derived variables is an increasingly popular practice in gait data analysis [10]. Recent examples include neural network muscle activation modeling [11], determination of the nonlinear relationship between pain and vertebral motion [12], and gait assessment using evolutionary algorithms [13]. In these examples, the validity of the functional relationships among variables and, hence, the proper interpretation of data are contingent upon the robustness of the derived variables.

In Section II, we briefly review the definitions of some widely applied statistical robustness measures, including the influence function. Subsequently, in Section III, influence functions for

Manuscript received January 24, 2003; revised June 14, 2003. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, in part by the Hospital for Sick Children Foundation, and in part by the Bloorview Childrens Hospital Foundation. *Asterisk indicates corresponding author.*

*T. Chau is with the Department of Rehabilitation Engineering, Bloorview MacMillan Children's Centre, 350 Rumsey Road, Toronto, ON M4G 1R8, Canada, and the Institute of Biomaterials and Biomedical Engineering, University of Toronto, 4 Taddle Creek Road, Toronto, ON M5S 3G9, Canada, (e-mail: tom.chau@utoronto.ca).

K. Parker is with the Department of Rehabilitation Engineering, Bloorview MacMillan Children's Centre, Toronto, ON M4G 1R8, Canada.

Digital Object Identifier 10.1109/TBME.2003.820396

three different stride frequency estimation methods are determined and interpreted in terms of robustness. The practical implications of estimator robustness or lack thereof for pathological gait data is gauged in Section IV. The paper concludes with recommendations on practical stride frequency estimation.

II. Robustness Measures

A. Influence Function

The influence function is a heuristic tool for gauging the impact of a small amount of contamination at any point z on the value of an estimator [14]–[16]. Influence functions have been widely applied to the study of model and parameter robustness. Examples include the determination of influential observations in statistical pattern recognition [17], outlier detection in the growth curve model [18], robust frequency estimation in noise contaminated sinusoidal signals [19], and the assessment of robustness properties of finite-length filters [20]. A robust estimator is characterized by a bounded influence function. Prior to reviewing the definition of the influence function, the contaminated distribution is introduced.

Let F represent an uncontaminated data distribution. Let Δ_z be the probability measure which puts point mass 1 at the point z. Consider a mixture of two distributions $F_{z,\epsilon}$, where an observation is randomly sampled from F with probability $1-\epsilon$ and otherwise from Δ_z, with probability ϵ. In other words, the data is contaminated by a sample z, with probability ϵ. The contaminated distribution is written as [14]

$$F_{z,\epsilon} = (1-\epsilon)F + \epsilon\Delta_z. \tag{1}$$

Let μ and σ represent the mean and standard deviation of the uncontaminated distribution, F. The mean of the mixture distribution, $F_{z,\epsilon}$, is given by [21]

$$\mu_{z,\epsilon} = \int z dF_{z,\epsilon} = (1-\epsilon)\mu + \epsilon z \tag{2}$$

while the variance is determined as

$$\sigma^2_{z,\epsilon} = \int (z-\mu)^2 dF_{z,\epsilon} = (1-\epsilon)\sigma^2 + \epsilon(z-\mu)^2. \tag{3}$$

These equations can be easily obtained by writing $dF_{z,\epsilon} = (1-\epsilon)dF + \epsilon d\Delta_z$ and by noting that $\int g(x)d\Delta_z = g(z)$, where $g(x)$ is any continuous function.

In terms of robustness, we are interested in whether parameters estimated with data from the mixture distribution, $F_{z,\epsilon}$, are very different from those estimated from the uncontaminated distribution, F. It is, thus, useful to think of a parameter or estimator as a function of a distribution. To do so, one makes use of functionals. In the present context, a functional, $T(\cdot)$, is a real-valued function on a vector space of probability distributions [22]. In practice, many estimators can be asymptotically replaced with functionals. Using functionals, one can define a function that describes how an estimator is influenced by an additional observation. This function is aptly named the influence function and is defined as [14]

$$IF(z) = \lim_{\epsilon \to 0} \frac{T(F_{z,\epsilon}) - T(F)}{\epsilon}. \tag{4}$$

The influence function measures the incremental change in the value of the functional T, due to a contamination at z. This contamination usually takes the form of an extreme value. Equivalently, under certain conditions of differentiability [14], (4) can be written as

$$IF(z) = \left.\frac{\partial T(F_{z,\epsilon})}{\partial \epsilon}\right|_{\epsilon=0}. \tag{5}$$

Therefore, the influence function is the first derivative of a functional T at a distribution $F_{z,\epsilon}$. Note that the influence is a measure of local robustness, since it describes the impact of a small amount of contamination at a particular point z, on the value of an estimator.

To determine influence functions for different stride frequency estimators using (5), we need to first evaluate the functional at the contaminated distribution, i.e., $T(F_{z,\epsilon})$. Generally, we can write

$$T(F_{z,\epsilon}) = T((1-\epsilon)F + \epsilon\Delta_z). \tag{6}$$

The particular form of this expression will differ, depending on the definition of the functional T.

B. Sensitivity Curve

The influence function is based upon large sample asymptotic behavior of estimators. With finite sample sizes, the sensitivity curve [23] captures the sensitivity of an estimator to an additional observation. A definition of the sensitivity curve is [14]

$$SC(z) = N\left\{T(x_1,\ldots,x_{N-1},z) - T(x_1,\ldots,x_{N-1})\right\} \tag{7}$$

where $x_1,\ldots,x_{N-1}$ is the sample, $T(\cdot)$ is the functional for the estimator in question, and z is the additional or contaminating observation. It is known that the sensitivity curve often converges to the influence function as $N \to \infty$[15]. To verify the derivation of an asymptotic influence function, we will thus compare the derived function against the corresponding finite sample sensitivity curve [24], simulated for large N.

C. Breakdown Point

While the influence function measures local robustness, the breakdown point gauges the global robustness of an estimator. The following simple definition will suffice for our purposes here. Consider the functional evaluated at the mixture distribution, $F_{z,\epsilon} = (1-\epsilon)F + \epsilon\Delta_z$. The breakdown point of an estimator is the minimum proportion ϵ of outlier contamination at z for which $T(F_z,\epsilon)$ becomes unbounded at z[16]. In other words, the breakdown point gives us an idea of how much contaminant data an estimator can tolerate before it gives meaningless results. For more rigorous definitions, see for example, [14], [25], and [15].

III. Stride Frequency Estimation

Recall that we are interested in determining the robustness of stride frequency parameters estimated from stride period data. In Section II, we described the influence function as a tool for studying estimator robustness. We now present analytical forms of the influence functions for estimators arising from three different stride frequency estimation methods. For each method, we will first state the stride frequency location and spread estimators. Next, the associated influence functions and their graphical approximation by sensitivity curves will be presented. Based on these plots, we will discuss robustness characteristics of each estimator. The influence function derivations are detailed in the Appendix for the interested reader.

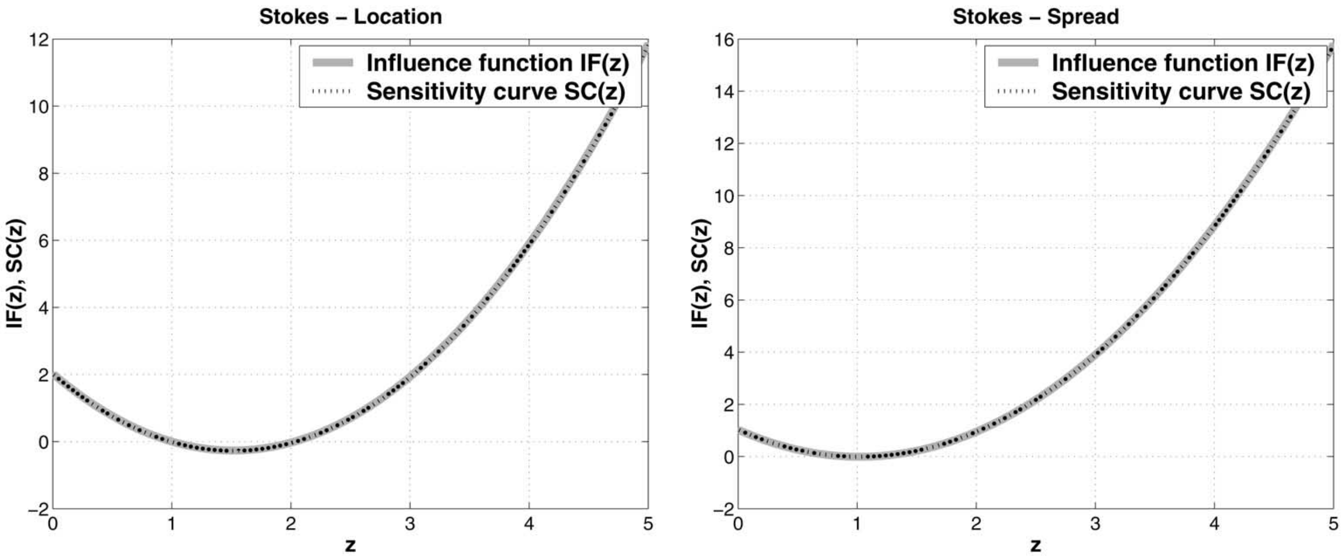

Fig. 1. Stokes estimators. Analytical influence function (solid line) and finite sample sensitivity curve (dotted line) for location (left) and spread estimators (right) proposed by Stokes *et al.*

We will use μ_y and σ_y to denote, respectively, the location and spread estimators for stride frequency.

A. Stokes Method

Let the quantities μ_x and σ_x denote the mean and standard deviation of uncontaminated stride period data, $\mathbf{X} = \{x_1, \ldots, x_N\}$, where x_i represents the i^{th} stride period. The stride frequency location estimator proposed by [1] is given by

$$\mu_y = \frac{1}{\mu_x}\left(1 + \frac{\sigma_x^2}{\mu_x^2}\right). \tag{8}$$

Similarily, the estimator of the spread of the stride frequency is approximated as [1]

$$\sigma_y^2 = \frac{\sigma_x^2}{\mu_x^4}. \tag{9}$$

In brief, these estimators were obtained by evaluating the mean and variance of a function of one random variable as outlined in [3]. The fundamental assumption is that the stride period X is closely centred around its mean. Details of the derivation can be found in [1], [3]. The Stokes estimators are simple formulas and are easy to compute. Further, given the mean μ_x and variance σ_x^2, the actual stride period time series is not required in the estimation.

Equation (8) gives us the form of the functional, $T(\cdot)$ for the location estimator which can be evaluated at the mixture distribution $F_{z,\epsilon}$. Differentiating $T(F_{z,\epsilon})$ with respect to ϵ and subsequently setting $\epsilon = 0$, the influence function for the location estimator is derived as

$$IF_\mu(z) = \frac{1}{\mu_x^3}z^2 - 3\frac{\mu_y}{\mu_x}z + 2\mu_y \tag{10}$$

where $z > 0$ is the contaminant data point, μ_y and σ_y are given by (8) and (9) and as above, μ_x and σ_x pertain to the uncontaminated stride period data. Note that $IF_\mu(z)$ is quadratic in z and unbounded. Using (9) as the functional $T(\cdot)$, and following the same procedure as above, the influence function for the estimator of stride frequency spread is determined as

$$IF_{\sigma^2}(z) = \frac{1}{\mu_x^4}z^2 - 2\left(\frac{\mu_y}{\mu_x^2} + \frac{\sigma_y^2}{\mu_x}\right)z + \frac{1}{\mu_x^2} + 3\sigma_y^2. \tag{11}$$

Again note that the influence function is quadratic in z and unbounded for large positive z. See the Appendix for additional details on the derivation of (10) and (11). To produce plots of the sensitivity curves for verifying the analytical influence functions, we generated "clean" data ($N = 5000$) from a normal distribution with mean $\mu = 1.0$ and standard deviation $\sigma = 0.1$. These values were chosen to coincide with the general order of magnitude of published stride periods for children [2]. Points of contamination, z, are drawn from the range $[1 \times 10^{-2}, 5]$ s in increments of 1×10^{-2} s. This range reflects plausible deviations in stride periods observed in empirical pathological gait data [26]. Note that we do not include 0 s as a contaminant point since mathematically, it cannot be inverted. This omission does not compromise the generality of the analysis, as such data would be filtered away in practice. Both the uncontaminated data and the contaminant points are chosen to ensure that stride periods are strictly positive, i.e., $x, z > 0$.

As shown in Fig. 1, the finite sample sensitivity and asymptotic influence curves agree closely for the simulated sample of Gaussian data, verifying the analytical forms of (10) and (11). As expected, each influence function intersects the line $IF(z) = 0$ at $z = 1$, the value of the true mean of the simulated data. Interestingly, the influence function for the location estimator crosses the line $IF = 0$ twice, suggesting that there is a second point whose influence is completely rejected by the estimator. From Fig. 1, we note that the estimators are more susceptible to the influence of uncharateristically long strides ($z \gg 1$) rather than short ones ($0 < z < 1$). This is due to the lower bound on possible stride periods, $z > 0$. A single outlying value, z, can cause each estimator to go to infinity. Hence, the common finite sample breakdown point is $\epsilon = 1/n$ and in the limit as $n \to \infty$, $\epsilon \to 0$.

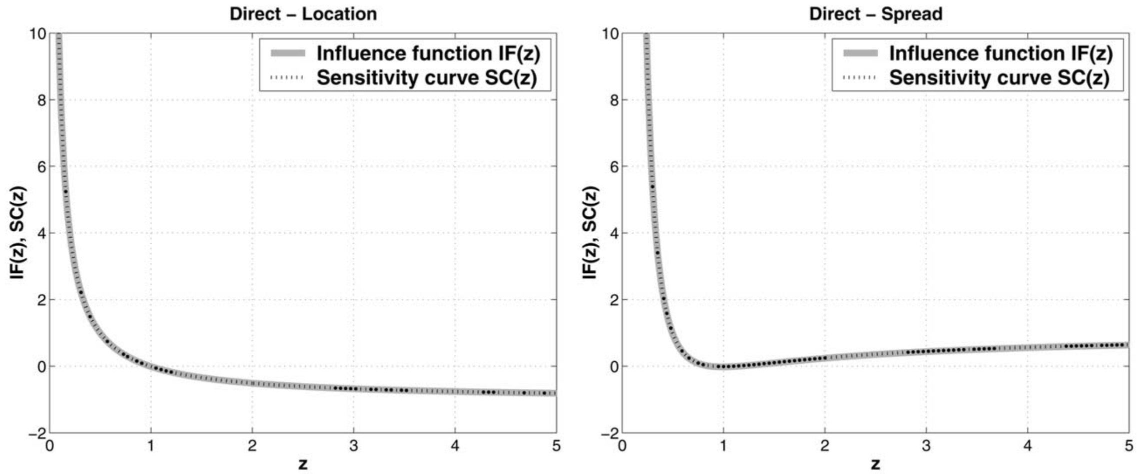

Fig. 2. Direct estimators. Analytical influence function (solid line) and finite sample sensitivity curve (dotted line) for location (left) and spread estimators (right).

In this section, we have shown that theoretically, the estimators for stride frequency location and spread as proposed by [1], are not robust to outlying values, i.e., stride periods which are very different from the mean or central value. In fact, a single outlying point, $z \gg 1$, could make the estimates arbitrarily large.

B. Direct Method

An intuitive way to estimate stride frequency from stride period data is to first invert every nonzero stride period to obtain a series $\mathbf{Y} = \{y_1, \ldots, y_N\}$, where $y_i = 1/x_i$. Next, the arithmetic mean and variance can be calculated with the standard formulas. Specifically, the location estimator would be

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{12}$$

where $y_i = (1/x_i)$. Likewise, the spread estimator would be

$$\sigma_y^2 = \left(\frac{1}{N}\sum_{i=1}^{N} y_i^2\right) - \mu_y^2. \tag{13}$$

We will refer to this straightforward approach as the direct method. The direct method is the most intuitive of the estimation methods studied in this paper. The estimates are based on familiar formulas and are likely the first choice of most readers. The disadvantage to direct estimation is that the entire stride period series is required. In practice, this information may not always be readily available.

As in the Stokes case, the influence functions for the above estimators can be determined by writing $T(F_{z,\epsilon})$ and differentiating with respect to ϵ. Details are outlined in the Appendix . The influence function for the location estimator is given by

$$IF_\mu(z) = \frac{1}{z} - \mu_y \tag{14}$$

where $z > 0$. Likewise, the influence function for the spread estimator is determined as

$$IF_{\sigma^2}(z) = \frac{1}{z^2} - \frac{2\mu_y}{z} + \mu_y^2 - \sigma_y^2 \tag{15}$$

where again $z > 0$. In contrast to the influence functions for the Stokes estimators [1], these influence functions, plotted in Fig. 2, are hyperbolic in z. While the previous estimators were sensitive to increasing positive perturbations, the influence functions for the direct method indicate high sensitivity to near zero perturbations. Indeed, valid stride periods should be nonzero and near vanishing values are likely due to some noise corruption in the measurement process. The heightened sensitivity to near zero perturbations is due to the inversion of stride period values in the direct method.

In the limit of large magnitude deviations, we find that the influence functions approach asymptotic values, $\lim_{z\to\infty} IF_\mu(z) = -\mu_y$ and $\lim_{z\to\infty} IF_{\sigma^2}(z) = \mu_y^2 - \sigma_y^2$. Hence, the direct influence functions can be considered bounded for large z. However, as the influence functions are unbounded as $z \to 0$, these estimators, like those of Stokes *et al.*, are not robust. Further, since only a single contamination $z \ll 1$ is required to cause $IF_\mu, IF_{\sigma^2} \to +\infty$, the direct estimators also have zero breakdown points in the limit as $n \to \infty$.

C. Robust Method

For purposes of comparison, influence functions for known robust location and spread estimators are also determined. For the sake of analytical and computational simplicity, we choose the α-trimmed mean [25] and the squared interquartile range [21] as robust estimators of location and spread, respectively. Many other possibile robust estimators exist (see for example, [16]). As in the direct method discussion, we again consider the frequency data, $Y = \{y_1, \ldots, y_n\}$, where $y_i = 1/x_i$.

The robust stride frequency location estimator is the α-trimmed mean of stride frequencies, given by

$$\mu_y = \frac{1}{1-2\alpha}\int_{y_\alpha}^{y_{1-\alpha}} y dF_Y(y) \tag{16}$$

where $0 < \alpha < 1$ determines the amount of trimming and $y_\alpha = F_Y^{-1}(\alpha)$ is the α-quantile of the distribution $F_Y(y)$. In our simulations, we conveniently chose $\alpha = 0.25$ such that y_α and $y_{1-\alpha}$, corresponded to the first and third quartiles of the distribution, F_Y.

The robust stride frequency spread estimator is the squared interquartile range, written as

$$\sigma_y^2 = (y_{1-\alpha} - y_\alpha)^2 \tag{17}$$

where $y_\alpha = F_Y^{-1}(\alpha)$ is the α-quantile of the distribution F_Y. Although we use $\alpha = 0.25$ throughout this paper, we present the

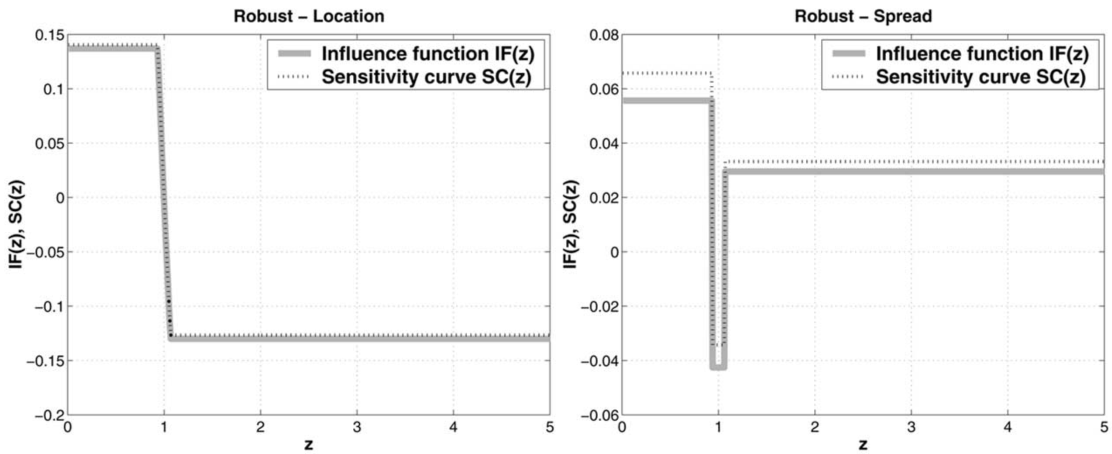

Fig. 3. Robust estimators. Analytical influence function (solid line) and finite sample sensitivity curve (dotted line) for location (left) and spread estimators (right).

equations with α unspecified to emphasize that the results could apply for differently defined interquantile ranges, as is often necessitated by the skewness of the data. These robust estimators are not used as widely as the more intuitive direct alternatives and, like the latter, require access to the entire stride period series. While the robust estimators are constructed to be less sensitive to the influence of outliers, the choice of α may be elusive. In other words, it may be difficult to clinically differentiate outlying values from informative deviations due to pathology or treatment intervention. Before presenting the influence functions, we introduce the following lemma.

Lemma 1 (Stride Period-Frequency Quantile Relationship): Quantiles of the stride period distribution F_x are related to quantiles of the stride frequency distribution F_y through

$$y_\alpha = \frac{1}{x_{1-\alpha}} \tag{18}$$

where $0 < \alpha < 1$ and $F_x^{-1}(\alpha) = x_\alpha$ and $F_y^{-1}(\alpha) = y_\alpha$. Details of the derivation are contained in the Appendix .

The influence function for the robust stride frequency location estimator is adapted from [25] to reflect contaminations in the stride period

$$IF_\mu(z) = \begin{cases} Ay_\alpha - B, & \frac{1}{z} < y_\alpha \\ A\frac{1}{z} - B, & y_\alpha \le \frac{1}{z} \le y_{1-\alpha} \\ Ay_{1-\alpha} - B, & \frac{1}{z} > y_{1-\alpha} \end{cases}. \tag{19}$$

where $A = (1/1-2\alpha)$, $B = \mu_y + \alpha A(y_\alpha + y_{1-\alpha})$, μ_y is given by (16), and y_α is given by (18). This influence function along with the sensitivity curve approximation are graphed on the left panel of Fig. 3. Note that unlike the previous estimators, the influence function is divided into three domains demarcated by the α and $1-\alpha$ quantiles. For a given set of stride frequency data, Y, the influence function is constant below the α quantile and above the $1-\alpha$ quantile. Within these two quantiles, the influence function varies hyperbolically with z. Hence, the influence function is bounded and the estimator is locally robust to regional contamination. The breakdown point for this estimator is α[16]. For $\alpha = 0.25$, this means that μ_y can tolerate a quantity of outliers up to 25% of the total sample size n.

By noting that the functionals for the squared interquartile range and the interquartile range [21] are related upon differentiation, we obtain the following influence function for the robust spread estimator

$$IF_{\sigma^2}(z) = \begin{cases} C\left(\frac{1}{\widehat{f}_Y(y_\alpha)} - D\right), & \frac{1}{z} < y_\alpha \\ -C\cdot D & y_\alpha \le \frac{1}{z} \le y_{1-\alpha} \\ C\left(\frac{1}{\widehat{f}_Y(y_{1-\alpha})} - D\right), & \frac{1}{z} > y_{1-\alpha} \end{cases} \tag{20}$$

where $\alpha = 0.25$, $\widehat{f}_Y(y_\alpha)$ is the estimated stride frequency probability density at the quantile y_α, and $D = \alpha\left(1/(\widehat{f}_Y(y_{1-\alpha})) + 1/(\widehat{f}_Y(y_\alpha))\right)$. Note that the stride frequency density estimate is related to the stride period density estimate through $\widehat{f}_Y(y_\alpha) = 1/y_\alpha^2 \widehat{f}_X(1/y_\alpha)$ [3]. Also, in (20), $C = 2(z_{1-\alpha} - z_\alpha)$ is a factor arising from differentiation of the functional, where $z_\alpha = (F_{z,\epsilon}^{-1}(1-\alpha))^{-1}$ is defined using Lemma 1, as the α quantile for the contaminated frequency distribution. See the Appendix for details of the influence function derivation. As in the robust location estimator, the influence function is also segmented into three domains by the quantiles y_α and $y_{1-\alpha}$. However, the influence function for the spread estimator is independent of z and, hence, constant in all three ranges of z values. This is portrayed in the right panel of Fig. 3. Since the influence function is bounded, this estimator is locally robust. Unlike previous estimators whose influence functions were completely determined by descriptive statistics of the stride period or frequency data, this influence function also requires probability density estimates at the two quantiles. In our simulation, we employed a kernel density estimator based on the shifted histogram [27]. Empirical estimation of the quartiles and the associated probability densities accounted for the differences between the sensitivity curve and the influence function in the right graph of Fig. 3. As with the robust location estimator, the breakdown point of the robust stride frequency spread estimator is α[21].

Note that the influence function for the stride frequency spread estimator is not symmetrical about $z = 1$ as one would expect from the symmetry of the simulated, normally distributed stride periods. This is due to the fact that stride period and frequency densities are related by $f_Y(y) = 1/y^2 f_X(1/y)$[3]. Hence, stride frequency, $\mathbf{Y}$, is not normally distributed, even when stride period $\mathbf{X}$, is normal.

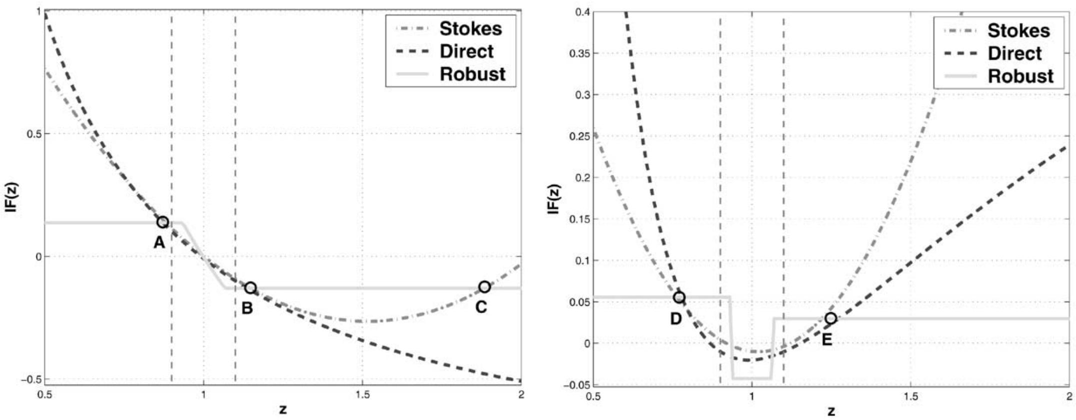

Fig. 4. Comparison of influence functions for location estimates on the left and spread estimates on the right. The labeled points highlight approximate locations where nonrobust estimates (Stokes and direct) begin to deviate monotonically away from robust values. Point B applies exclusively to the direct location estimator while Point C is relevant only to the Stokes location estimator. Points A, D, and E apply to both Stokes and direct estimators. The dashed vertical lines represent the one standard deviation band around the true mean for this simulated data set.

D. Comparison of Influence Functions

In Fig. 4, the influence functions for the different stride frequency estimators are superimposed. The ranges of the axes have been reduced to focus on the behavior near the known true mean period of the simulated data, namely, 1 s. The dashed vertical lines delineate one standard deviation (± 0.1 s) around the mean (1 s). Note that within the one standard deviation band, all three estimators have comparable influence function values, suggesting that with small deviations around the mean, the choice of estimator will not be critical in terms of robustness.

For small magnitude contaminations $z \ll 1$, both the direct location and spread influence functions become unbounded at higher rates than their Stokes counterparts. On the other hand, for large values of z, influence functions for direct estimators approach asymptotic values while those for Stokes extend to infinity. Practically, this suggests that the direct estimates are better suited to large scale deviations while Stokes estimators more aptly handle small magnitude contaminations.

It is perhaps not surprising that both estimators due to Stokes *et al.* [1] and the direct method are theoretically not robust since they are both functions of known nonrobust quantitites, the arithmetic mean and standard deviation. Due to different manipulations in their computation (see Appendix for details), the estimators are characterized by qualitatively different influence functions.

IV. Practical Implications: Pathological Gait

While idealized, Gaussian data facilitated the study of the theoretical influence of the different estimators, real stride time series can be non-Gaussian [26]. We, therefore, study the practical differences between robust and nonrobust stride frequency estimates using real gait time series. In particular, we examine estimation differences across a group of subjects and within individual subjects.

A. Pathological Gait Data

The data consist of stride period time series from a 10-min walk at a self-selected pace, collected from 13 mobile children with spastic diplegic cerebral palsy. Details of data collection are reported in [26]. This previous study by Chau and Rizvi [26] suggested that a sample size of 10 children with cerebral palsy is adequate for detecting differences in a variety of temporal gait measures. While we are concerned with different measures here, we use the suggested sample size as a rough guideline. The time series range in length from 169 to 567 strides. We choose these data as representatives of potential "worse case" data because it is known that gait variability is greater in children than in adults and is often further inflated in the face of pathology [2], [5], [8]. It is, therefore, anticipated that differences among estimators would be most clearly elucidated by stride period data from pathological, pediatric gait. By the same argument, we would expect that estimator choice would not be critical with low variability healthy gait data. This latter expectation is supported by the fact that the theoretically derived influence functions are comparable in value for small deviations around the mean (Fig. 4). We, therefore, omit healthy gait from consideration in the present study. An example of an empirical stride period time series and the accompanying stride frequency series obtained by inverting every sample, are shown in Fig. 5.

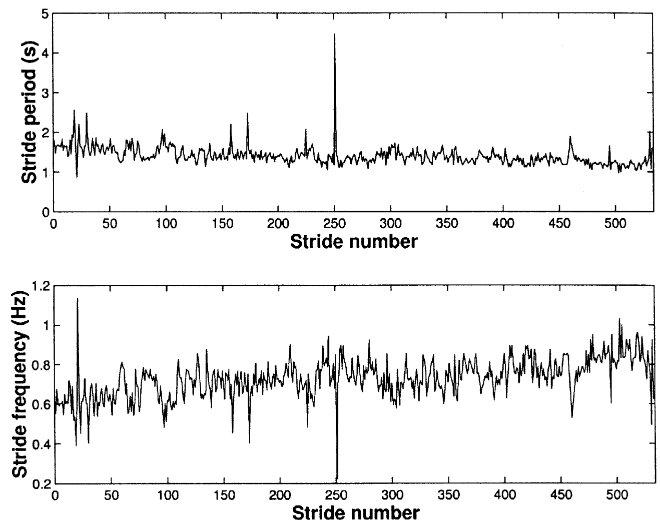

Fig. 5. A typical stride period time series (top) and the corresponding stride frequency time series (bottom).

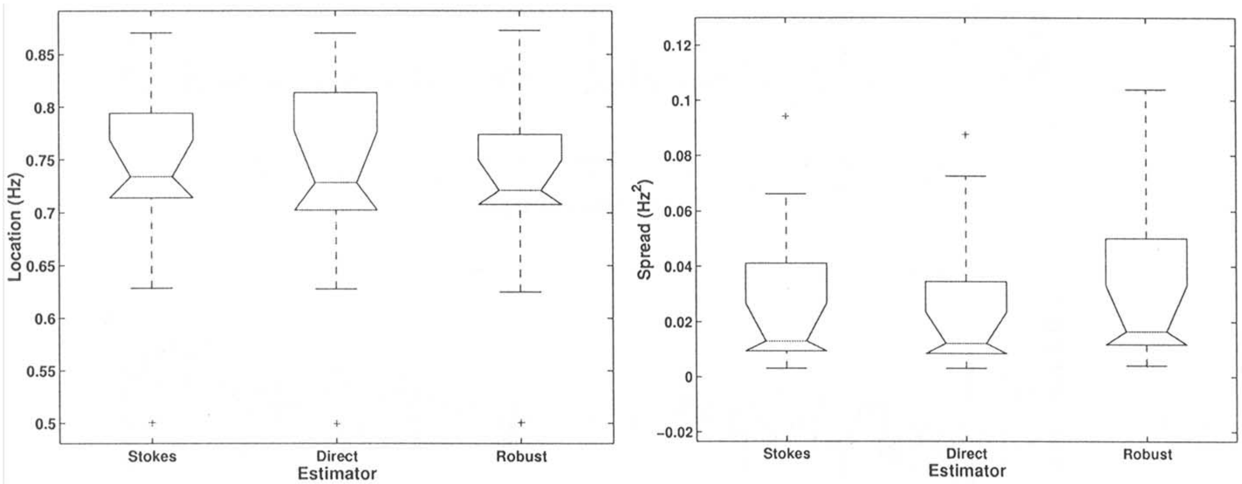

Fig. 6. Boxplots comparing distributions of location (left) and spread (right) estimates for Stokes, direct and robust estimators.

Note the presence of multiple extreme values in both time series, which may unduly influence the nonrobust estimates.

B. Group (Across Subject) Estimates

We first look at differences in estimating a group stride frequency value for a sample of subjects. Stride frequency location and spread estimators are computed by the Stokes, direct and robust methods, for each of the 13 stride period time series. We then use a Kruskal–Wallis test to compare the distribution of location and spread estimates obtained by each of the three methods. The results are captured in the box plots of Fig. 6. Surprisingly, we find no statistical differences between the nonrobust (Stokes and direct) and robust estimates, for both location ($p = 0.86$) and spread ($p = 0.59$). The overall agreement between nonrobust and robust methods is somewhat unexpected given that the stride period variability in this data ($22 \pm 11\%$ of the gait cycle) is well above that of healthy paediatric gait (cf. $\approx 8.5\%$[2]). This finding suggests that the nonrobust methods should not be dismissed when computing a group mean or variance across a number of subjects, even for pathological gait.

C. Within-Subject Estimates

Given the lack of differences in the estimates when taken as a group, we now investigate the differences among robust and nonrobust estimates for each individual time series.

We define the relative error in estimating location by the Stokes method as, $(\mu_S - \mu_R)/\mu_R \times 100\%$, where the subscripts S and R denote the Stokes and robust estimation methods, respectively. Likewise, the relative error in estimating spread by the Stokes method is, $(\sigma_S^2 - \sigma_R^2)/\sigma_R^2 \times 100\%$. These relative errors reflect the amount by which the Stokes estimates for each signal deviate from the corresponding robust estimates. Relative errors for the direct estimates are defined similarly. Fig. 7 depicts the errors in Stokes and direct estimates with respect to robust estimates.

Note that for location estimation, the errors are generally less than or equal to 5% for all signals except for subject #3, implying that the choice of estimator is not critical in most instances. To understand the massive discrepancy in estimate #3, we use influence functions to study the impact of outliers in

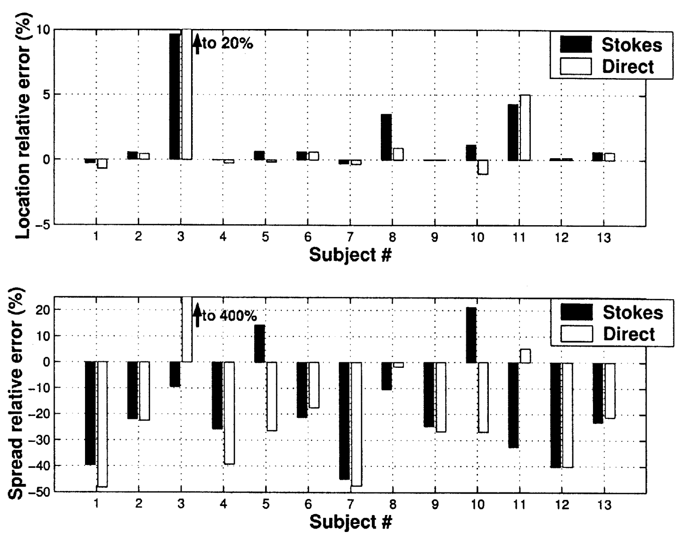

Fig. 7. Errors in nonrobust (Stokes and direct) location (top) and spread (bottom) estimates relative to robust location estimates.

the signals. For simplicity, outliers are defined as points situated more than 1.5 times the interquartile range away from the median. More formal discordancy tests for outliers can be found in [28]. We use (10) and (14) to compute the influence of every outlier in each signal on Stokes and direct location estimates. Subsequently, we sum the individual influence values to obtain a value representing the collective influence of all outliers in the signal. Plotting these collective influences in the top graph of Fig. 8, we immediately observe that the outliers found in signal #3 bear the greatest influence on location estimation, for both Stokes and direct methods. In fact, signal #3 contains the greatest fraction of extreme outliers (more than 3 interquartile ranges away from the median) among all the signals. Similarily, other large errors can also be explained in terms of collective outlier influence. Remarkably, the collective influence values are closely correlated to the relative location errors, for both Stokes ($r = 0.84$) and direct ($r = 0.76$) location estimators. Here, r denotes the Pearson correlation coefficient [29]. This strong correlation confirms that the influence of outliers in the signal directly impacts the agreement between robust and nonrobust estimates.

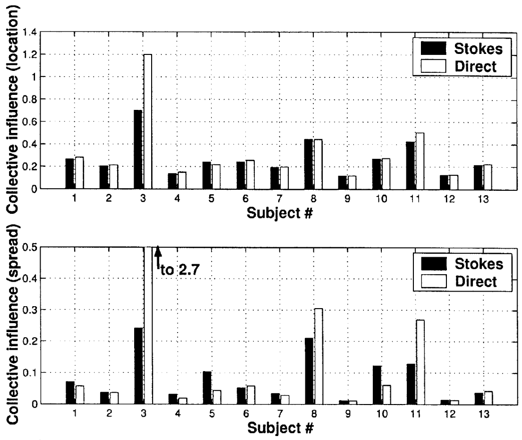

Fig. 8. Collective influence of the outliers in each signal for Stokes and direct location (top) and spread (bottom) estimates.

In contrast to errors in location estimation, the bottom graph of Fig. 7 indicates that relative errors in spread estimation are generally much larger and typically negative. As above, relative errors (bottom plot of Fig. 7) are strongly correlated to the collective influence of outliers in each signal on spread estimates (bottom plot of Fig. 8), for both Stokes ($r = 0.74$) and direct ($r = 0.997$) spread estimates. To interpret the unusually large error for subject #3 by direct estimation, recall that the direct spread estimator is more susceptible to outliers below the central value. Indeed, the time series for subject #3 exhibited the highest fraction (5% of all samples) of outliers below the median. Given that the direct spread estimate has a zero breakdown point, it is not surprising that an apparently small fraction of outliers drives up the relative error by an order of magnitude.

The negativity of the errors bears an important practical implication. Since a negative error arises when the nonrobust estimate is less than the corresponding robust estimate, direct and Stokes methods tend to underestimate the variability in the stride frequency data. This increases the chance of a Type I error when statistically comparing distributions of stride frequency estimates obtained by these methods. In other words, hypothesis testing would tend to incorrectly suggest the presence of statistically significant differences. This underestimation tendency is also apparent in the box plot on the right hand side of Fig. 6. We note however, that despite the large relative errors in individual spread estimates, due to the wide variability of spread estimates across subjects, the differences among estimation methods are not statistically significant, as reported in Section IV-B.

D. Guidelines for Estimator Choice

Based on the empirical and theoretical comparisons, we outline some rough guidelines for estimator selection. These guidelines stem directly from the concepts of breakdown point and influence function. When the time series contains even a small fraction ($< 5\%$ of total sample) of extreme outliers (more than 3 interquartile ranges away from the median), then robust alternatives are recommended.

Below this level of outlier extremity, nonrobust methods may be admissible. In particular, the direct method is preferred over the Stokes method when mild outliers (between 1.5 and 3 interquartile ranges away from the median) occur principally above the central value of stride period. This can be visualized in the left plot of Fig. 4, where point C, marking the start of the monotonic ascent of the Stokes influence function is further to the right of the monotonic descent of the direct influence function at point B. On the other hand, Stokes' estimation is recommended when there are mild outliers below the central value of stride period. Again, this heuristic rule is supported by the plots of Fig. 4 where at points A and D, the direct influence function climbs to infinity at a higher rate than the Stokes curves.

In the absence of extreme outliers, within-subject Stokes, direct and robust estimates are comparable. Examining Fig. 4, we see that this means that the majority of the signal falls within the one standard deviation band, within which all influence functions are comparable in value. Thus, the empirically determined rules-of-thumb generally corroborate the behavior theoretically indicated by the influence functions.

Note that while familiar statistics such as skewness and kurtosis are typically used to indicate distributional deviation from normality, we do not employ them in guiding stride frequency estimator selection. The reason is that skewness and kurtosis have influence functions which differ from those of the estimators under study and, hence, behave differently in the presence of outliers. In fact, with our pathological data set, skewness and kurtosis are only weakly correlated to the relative errors of the nonrobust estimators.

V. Concluding Remarks

A. Limitations and Extensions

For computational convenience, sensitivity and influence plots were based on normally distributed data. However, the pathological stride period time series used in this paper actually exhibited a statistically significant positive skewness, something also observed in other fractal physiological time series (see, for example, [30] and [31]). Hence, use of positively skewed data such as that generated from a χ^2 density, may have produced influence and sensitivity plots more reflective of real data. For skewed data, one-sided trimming [16], [32] for both robust location and spread estimates may be more appropriate.

The estimator comparisons of this paper were based on long time series. With shorter time series, the analytical curves and heuristic ranges may not readily apply, and one would need to study finite sample robustness [14] more closely. Conclusions about a lack of statistical difference among estimates were based on a sample size of 13 time series. Given the heightened variability of pathological gait, larger sample sizes would help to further delineate the practical robustness characteristics of the different estimators. In fact, a power analysis of the empirical data collected here suggests that a sample size of roughly 55 children with spastic diplegia is required to achieve 80% power in detecting a medium effect difference between robust and nonrobust estimators. The fact that differences were revealed among estimators for pathological pediatric gait, suggests that in a future study, it might be interesting to isolate the differences due to gait immaturity from those due to pathology.

B. Conclusion

We have shown that influence functions can be a useful qualitative tool for studying the robustness of quantities derived from gait and possibly other physiological time series. We have examined both the theoretical (via influence functions) and practical robustness (via real gait data) of three different stride frequency estimation methods. Influence functions derived for the nonrobust estimators accurately predicted estimator performance with empirical gait data. Generally, the choice of estimator did not lead to statistically different ensemble estimates, even for pathological gait data. However, prudence in estimator selection is warranted in the presence of extreme outliers. Indeed, the most intuitive estimator may not always be the most appropriate choice and nonrobust estimators need not be entirely discarded when analyzing pathological gait data. To practically choose a stride frequency estimator, the extremity and plurality of outliers in the data can serve as useful guidelines. Estimator comparisons suggest that closer attention to and more research on estimator robustness in gait analysis are warranted.

APPENDIX

Here, we prove Lemma 1 and highlight the key steps involved in the derivation of the influence functions for the Stokes, direct and robust stride frequency estimators.

A. Lemma 1

The stride period density is related to the stride frequency density by $f_Y(y) = 1/y^2 f_X(1/y)$[3]. Integrating both sides with respect to y, we obtain, $F_Y(y) = -F_X(1/y) + C$, where C is a constant of integration. To evaluate C, we note that $F_Y(\infty) = 1$ by definition of the distribution and $F_X(0) = 0$ for the stride period since $x > 0$. Applying these conditions, we find $C = 1$ or

$$F_Y(y) = 1 - F_X\left(\frac{1}{y}\right). \tag{21}$$

By definition of the quantile, $F_Y(y_\alpha) = \alpha$. Using (21) to substitute for $F_Y(y_\alpha)$ in this definition and rearranging, yields

$$F_X\left(\frac{1}{y_\alpha}\right) = 1 - \alpha. \tag{22}$$

Isolating the argument on the left-hand side gives the desired relationship

$$\frac{1}{y_\alpha} = F_X^{-1}(1-\alpha) \tag{23}$$

where we note that $F_X^{-1}(1-\alpha) = x_{1-\alpha}$.

B. Influence Functions

1) Stokes Estimates: Equation (8) gives us the form of the functional for the stride frequency location estimator proposed by [1]. Hence, for the contaminated distribution we can write

$$T(F_{z,\epsilon}) = \frac{1}{\mu_{z,\epsilon}}\left(1 + \frac{\sigma_{z,\epsilon}^2}{\mu_{z,\epsilon}^2}\right). \tag{24}$$

Next, substitute for the mean and variance of the mixture distribution using (2) and (3). Simplifying the resulting algebraic equation, we obtain

$$T(F_{z,\epsilon}) = \frac{\epsilon^2(z-\mu_x)^2 + \epsilon(z^2 - \mu_x^2 - \sigma_x^2) + \sigma_x^2 + \mu_x^2}{(\epsilon(z-\mu_x)+\mu_x)^3}. \tag{25}$$

Recall that $IF(z) = \partial T(F_{z,\epsilon})/\partial\epsilon$. Differentiating $T(F_{z,\epsilon})$ with respect to ϵ, subsequently setting ϵ to 0, and simplifying algebraically, we obtain

$$\frac{\partial T(F_{z,\epsilon})}{\partial\epsilon} = \frac{z^2}{\mu_x^3} - 3z\left(\frac{\mu_x^2+\sigma_x^2}{\mu_x^4}\right) + 2\left(\frac{\mu_x^2+\sigma_x^2}{\mu_x^3}\right). \tag{26}$$

Finally, we substitute for μ_S and σ_S^2 using (8) and (9), where appropriate, to arrive at the influence function of (10).

Following the same procedure as above, the functional at the mixture distribution for the spread estimator is written as

$$T(F_{z,\epsilon}) = \frac{\epsilon\left[(z-\mu_x)^2 - \sigma_x^2\right] + \sigma_x^2}{\left[\epsilon(z-\mu_x)+\mu_x\right]^4}. \tag{27}$$

Differentiating with respect to ϵ, taking the limit as $\epsilon \to 0$, and simplifying, we get,

$$\frac{\partial T(F_{z,\epsilon})}{\partial\epsilon} = \frac{z^2}{\mu_x^4} - 2z\left(\frac{1}{\mu_x^3} - \frac{2\sigma_x^2}{\mu_x^5}\right) + \left(\frac{1}{\mu_x^2} + 3\frac{\sigma_x^2}{\mu_x^4}\right). \tag{28}$$

Upon substitution for μ_S and σ_S^2, the influence function of (11) is obtained.

2) Direct Estimates: The functional for the estimator in (12) can be written in terms of the distribution F_x

$$T(F_x) = \int \frac{1}{x} dF_x. \tag{29}$$

To derive the influence function for the direct location estimator, we start by evaluating the functional at the mixture distribution

$$\begin{aligned} T(F_{z,\epsilon}) &= T((1-\epsilon)F_x + \epsilon\Delta_z) && (30)\\ &= (1-\epsilon)T(F_x) + \epsilon T(\Delta_z) && (31)\\ &= (1-\epsilon)\mu_Y + \epsilon\int \frac{1}{z} d\Delta_z && (32)\\ &= (1-\epsilon)\mu_Y + \frac{\epsilon}{z} && (33)\end{aligned}$$

where μ_Y is given by (12). The above follows from linearity of the functional. Differentiating with respect to ϵ, we obtain

$$\frac{\partial T(F_{z,\epsilon})}{\partial\epsilon} = \frac{1}{z} - \mu_Y \tag{34}$$

which is the functional for the direct location estimator. Similarily, for the spread estimator, we write the functional for (13) as

$$T(F_x) = \int \left(\frac{1}{x} - \mu_Y\right)^2 dF_x. \tag{35}$$

Evaluating the functional at the mixture distribution yields

$$\begin{aligned} T(F_{z,\epsilon}) &= (1-\epsilon)T(F_x) + \epsilon T(\Delta_z) && (36)\\ &= (1-\epsilon)\sigma_Y^2 + \epsilon(\frac{1}{z} - \mu_Y)^2 && (37)\end{aligned}$$

where σ_Y is given by (13). Differentiating this last expression with respect to ϵ yields the influence function in (15).

C. Robust Estimates

The influence function for the α-trimmed mean is given in various standard texts including [14], [16] and [21]. The functional for (16) can be rewritten in terms of the stride period distribution

$$T(F_x) = \frac{1}{1-2\alpha}\int_{(F_x^{-1}(1-\alpha))^{-1}}^{(F_x^{-1}(\alpha))^{-1}} \frac{1}{x} dF_x \tag{38}$$

where we have used Lemma 1 in the integral limits and the integrand has been substituted using $y = 1/x$. Comparing this to the standard form of the functional for the trimmed mean [21] and its corresponding influence function [25] and noting that $y_\alpha = 1/x_{1-\alpha}$ for stride period and frequency, we can write the influence function (19) by inspection.

The influence function for the interquartile range, $IF_{IQR}(z)$, can be found in [21]. Here, we simply relate the influence function of the squared interquartile range to $IF_{IQR}(z)$. The functional at the mixture distribution can be written as

$$T(F_{z,\epsilon}) = \left[F_{z,\epsilon}^{-1}(1-\alpha) - F_{z,\epsilon}^{-1}(\alpha)\right]^2. \tag{39}$$

Define $z_\alpha = (F_{z,\epsilon}^{-1}(1-\alpha))^{-1}$ as the α quantile for the contaminated data distribution $F_{z,\epsilon}$. Differentiating (39) with respect to ϵ yields

$$\frac{\partial T(F_{z,\epsilon})}{\partial \epsilon} = 2(z_{1-\alpha} - z_\alpha)\frac{\partial}{\partial \epsilon}\left[F_{z,\epsilon}^{-1}(1-\alpha) - F_{z,\epsilon}^{-1}(\alpha)\right]. \tag{40}$$

The derivative term on the right is simply the definition of the influence function for the interquartile range. Hence, we have

$$\frac{\partial T(F_{z,\epsilon})}{\partial \epsilon} = 2(z_{1-\alpha} - z_\alpha) IF_{IQR}(z) \tag{41}$$

which upon substitution for $IF_{IQR}(z)$ yields (20).

References

[1] V. P. Stokes, A. Thorstensson, and H. Lanshammar, "From stride period to stride frequency," *Gait Posture*, vol. 7, no. 1, pp. 35–38, 1998.

[2] H. Stolze, J. P. Kuhtz-Buschbeck, C. Mondwurf, K. Johnk, and L. Friege, "Retest reliability of spatiotemporal gait parameters in children and adults," *Gait Posture*, vol. 7, no. 2, pp. 125–130, 1998.

[3] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw Hill, 1991.

[4] C. R. Rao, *Linear Statistical Inference and its Application*, 2nd ed. New York: Wiley, 1973.

[5] G. Steinwender, V. Saraph, S. Scheiber, E. B. Zwick, C. Uitz, and K. Hackl, "Intrasubject repeatability of gait analysis data in normal and spastic children," *Clin. Biomech.*, vol. 15, no. 2, pp. 134–139, 2000.

[6] J. M. Hausdorff, J. Balash, and N. Giladi, "Cognitive challenge increases gait variability and fall risk in patients with Parkinson's disease," *J. Amer. Geriatrics Soc.*, vol. 50, no. 4, p. 104, 2002.

[7] M. L. Harris-Love, L. W. Forrester, R. F. Macko, K. H. C. Silver, and G. V. Smith, "Hemiparetic gait parameters in overground versus treadmill walking," *Neurorehab. Neural Repair*, vol. 15, no. 2, pp. 105–112, 2001.

[8] J. M. Hausdorff, L. Zemariy, C. K. Peng, and A. L. Goldberger, "Maturation of gait dynamics: stride-to-stride variability and its temporal organization in children," *J. Appl. Physiol.*, vol. 86, no. 3, pp. 1040–1047, 1999.

[9] J. M. Hausdorff, A. Lertratanakul, M. E. Cudkowicz, A. L. Peterson, D. Kaliton, and A. L. Goldberger, "Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis," *J. Appl. Physiol.*, vol. 88, no. 6, pp. 2045–2053, 2000.

[10] T. Chau, "A review of analytical techniques for gait data, Part 2: Neural network and wavelet methods," *Gait Posture*, vol. 13, no. 2, pp. 102–120, 2001.

[11] S. D. Prentice, A. E. Patla, and D. A. Stacey, "Artificial neural network model for the generation of muscle activation patterns for human locomotion," *J. Electromyogr. Kinesiol.*, vol. 11, no. 1, pp. 19–30, 2001.

[12] J. P. Dickey, M. R. Pierrynowski, D. A. Bedriar, and S. X. Yang, "Relationship between pain and vertebral motion in chronic low-back pain subjects," *Clin. Biomech.*, vol. 17, no. 5, pp. 345–352, 2002.

[13] W. L. Wu, F. C. Su, Y. M. Cheng, and Y. L. Chou, "Potential of the genetic algorithm neural network in the assessment of gait patterns in ankle arthrodesis," *Ann. Biomed. Eng.*, vol. 29, no. 1, pp. 83–91, 2001.

[14] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley, 1986.

[15] G. L. Shevlyakov and N. O. Vilchevski, *Robustness in Data Analysis: Criteria and Methods*: VSP, 2002.

[16] R. G. Staudte and S. J. Sheather, *Robust Estimation and Testing*. New York: Wiley, 1990.

[17] W. K. Fung, "The influence of observations on misclassification probability in multiple discriminant analysis," *Commun. Statist.- Theory Meth.*, vol. 25, no. 8, pp. 1917–1930, 1996.

[18] J. X. Pan and K. T. Fang, "Influential observation in the growth curve model with unstructured covariance matrix," *Computational Statist. Data Anal.*, vol. 22, no. 1, pp. 71–87, 1996.

[19] S. G. Oh and R. L. Kashyap, "A robust approach for high resolution frequency estimation," *IEEE Trans. Signal Processing*, vol. 39, pp. 627–643, Mar. 1991.

[20] S. Peltonen, P. Kuosmanen, and J. Astola, "Output distributional influence function," *IEEE Trans. Signal Processing*, vol. 49, pp. 1953–1960, Sept. 2001.

[21] R. R. Wilcox, *Introduction to Robust Statistics and Hypothesis Testing*. New York: Academic, 1997.

[22] E. Kreyszig, *Introductory Functional Analysis*: Wiley, 1989.

[23] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

[24] C. Goodall, "M-estimators of location: An outline of the theory," in *Understanding Robust and Exploratory Data Analysis*, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, Eds. New York: Wiley, 1983, pp. 339–403.

[25] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[26] T. Chau and S. Rizvi, "Automatic stride interval extraction from long, highly variable and noisy gait timing signals," *Human Movement Sci.*, vol. 21, no. 4, pp. 495–514, 2002.

[27] D. W. Scott, *Multivariate Density Estimation*. New York: Wiley, 1992.

[28] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. New York: Wiley, 1994.

[29] B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*, 2nd ed. New York: Harper Collins, 1989.

[30] C. D. Lewis, G. L. Gebber, P. D. Larsen, and S. M. Barman, "Long-term correlations in the spike trains of medullary sympathetic neurons," *J. Neurophysiol.*, vol. 85, no. 4, pp. 1614–1622, 2001.

[31] J. H. Carlson and S. L. Foote, "Oscillation of interspike interval length in substantia-nigra dopamine neurons – Effects of nicotine and the dopamine-d2 agonistly-163 502 on electrophysiological activity," *Synapse*, vol. 11, no. 3, pp. 229–248, 1992.

[32] K. Kafadar, J. Rice, and C. Spiegelman, "One-sided trimming in small samples with asymmetric contamination," *Commun. Statist.- Theory Meth.*, vol. 12, no. 4, pp. 477–496, 1983.

Tom Chau (S'92–M'97–SM'03) received the B.A.Sc. degree in engineering science and the M.A.Sc. degree in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 1992 and 1994, respectively. He received the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1997.

He is with Bloorview MacMillan Children's Centre, Toronto, where he investigates intelligent data analysis methods, augmented environments, and automatic pattern recognition, for the rehabilitation of children with physical and cognitive disabilities. He is also an Adjunct Assistant Professor of biomedical engineering and rehabilitation science at the University of Toronto.

Kim Parker received the B.A.Sc. degree in mechanical engineering from Queen's University, Kingston, ON, Canada, in 1993 and the M.A.Sc. degree in mechanical engineering from the Institute of Biomedical Engineering at the University of Toronto, Toronto, ON, Canada, in 1995.

She is currently a researcher with the Human Movement Laboratory and the Rehabilitation Engineering Department, Bloorview MacMillan Children's Centre, Toronto. Her research interests include the measurement of movement and prosthetic and orthotic product development.

Ms. Parker is a registered Professional Engineer in the province of Ontario.