Selecting the best number of synergies in gait: preliminary results on young and elderly people

Fiorenzo Artoni, Vito Monaco
The Biorobotics Institute
Scuola Superiore Sant'Anna
Pisa, Italy
f.artoni@sssup.it
v.monaco@sssup.it

Silvestro Micera
The BioRobotics Institute
Scuola Superiore Sant'Anna
Pisa, Italy
Translational Neural Engineering Laboratory
Center of Neuroprosthetics
Swiss Federal Institute of Technology
Lausanne, Switzerland
s.micera@sssup.it

***Abstract*—Matrix factorization algorithms are increasingly used to extract meaningful information from multivariate EMG datasets. However a key issue is the selection of the number of synergies (i.e., model order) to retain. In this preliminary work a set of criteria, based on Independent Component Analysis, was developed to determine the number of synergies to extract from a multivariate EMG dataset, and applied on EMG signals acquired from 12 leg muscles during walking at different cadences (40, 60, ..., 140 strides per minute) in young and elderly subjects. The method was tested on ad-hoc created datasets with a predetermined number of embedded sources and amplitude of added noise. Young subjects walking patterns are explained by a number of synergies not significantly different with respect to elderly subjects. The inter-subject variability is greater at high (elderly) and low (young and elderly) cadences suggesting that the walking pattern is more stable at central frequencies. The type of preprocessing influences the number of underlying synergies: an increased number of independent components is needed to explain the variability of unfiltered data. The proposed method could serve as a guideline to scientists in the evaluation of walking performance. Further developments will include a validation of the method and its extension to other factorization algorithms.**

I. Introduction

Several studies found evidence of modifications in gait patterns and muscle activations in healthy people brought about by aging [15], such as reduction in walking speed [11], [21], increased repeatability in EMG signals (suggesting a decreased neural plasticity [18]) and increased fall risks [12]. These findings suggest that neuromuscular adaptation, whether it is due to aging itself, lack of physical activity, or pathologies [6], [13], may be revealed by some features of gait pattern and EMG signals during gait training.

In order to extract meaningful features from EMG signals, several Matrix Factorization (MF) techniques (e.g. Principal Component Analysis - PCA [22], Independent Component Analysis - ICA [3], Factor Analysis - FA and Non-negative Matrix Factorization - NMF [20]) have been developed with the aim of parsing multivariate signals into a set of maximally-informative components, called muscle synergies, which may better represent the main central features of the motor programs. Tresch et al. identified PCA-ICA (ICA performed on the subspace defined by PCA) and pICA (probabilistic ICA with nonnegativity constraints) as the best performing methods for extracting synergies, closely followed by NMF, and FA which demonstrated good robustness across datasets as well [20].

One of the main theoretical issues to overcome when applying ICA (and MF techniques in general) is that of determining the model order, that is the number of components (synergies) that should be retained. In general PCA-based dimensionality reduction needs to be carefully considered, as interesting components may present part of their variance in the low-power region of the eigenspectrum [9]. This choice is often made heuristically as it depends on application-specific and subjective considerations, for instance by adopting the popular $eigenvalue > 1$ criterion or by placing a threshold on the cumulative variance explained by the extracted factors [5], [10], [14], [17]. Other criteria include scree plots or Laplacian information criterion (LIC) for PCA, Bartlett's test for subsets of PCA components, likelihood ratios, projected variance for ICA components, Akaike and Bayesian information (AIC, BIC). D'Avella et al [4] proposed a method to model muscle synergies by explicitly introducing activation delays and to determine their number by identifying the slope change of the total variation explained (R^2) with respect to the number of retained factors. On the whole the basic assumption of these methods is that the random variation attributable to noise is smaller than the structured variation of the synergy combinations. However a great amount of noise could impair the performance of these methods as the MF algorithm may not able to successfully disentangle information from noise anymore (e.g. PCA [20]). In that case the noise would be distributed among the factors thus making it impossible to select the correct number of synergies based on the explained variance.

The aim of this work is to propose a set of criteria to determine the best number of synergies to extract from a EMG dataset with ErpICASSO [1], an improved version of the popular FastICA [8] and ICASSO [7] algorithms. In this preliminary study the ErpICASSO technique was applied on the EMG (12 ipsilateral leg muscles) of young and elderly people while walking over ground in a wide range of cadences with the aim of determining whether aging and pace significantly affect the number of underlying synergies. The possible differences in the results, dependent on the type of

preprocessing of the EMG data, were also determined. The proposed criteria, although not always completely conclusive, are based not only on PCA-explained variance, but also take into account the intrinsic variability of the data by employing trial-to-trial bootstrapping and clustering thus providing a reliability measure and a robust estimate of the number of synergies. These criteria are validated on an ad-hoc constructed dataset (using the recorded data), address the statistical and algorithmic reliability of estimated independent components, and could be used whenever dealing with multivariate EMG walking data.

II. Materials and Methods

A. Experimental setup and preprocessing

Data collection methods are briefly reported here although they have been previously described [16]. Seven young (4 males, 3 females; 27.3 ± 4.9 yr old) and seven elderly (4 males, 3 females, 69 ± 1.4 yr old) healthy subjects were enrolled with a protocol in accordance with the Local Ethical Committee. The experiments were carried out in a 15-m-long room where subjects walked over ground.

The EMG signals of 12 leg muscles belonging to the right leg [gluteus medius, gluteus maximus, tensor fascia latae, adductor longus, peroneus longus, semitendinosus (ST), gastrocnemius lateralis, soleus, rectus femoris, vastus medialis, tibialis anterior, biceps femoris] were recorded using surface EMG electrodes (NORAXON, Telmyo 2400T, V2), at a sampling rate of $1000Hz$ and 1000 as gain amplifier. The heel strike and toe-off related to the right leg were recorded by means of footswitches. The subjects walked respectively at the cadences of 40, 60, 80, 100, 140 steps/min (in randomized order) at the beat of a metronome to account for the fact that younger people walk with a lower frequency, even if the speed remains unaltered. The data recorded between two consecutive right-leg heelstrikes is a trial.

After rejecting for each subject the first and last three right-side strides [2], [19] the data were high-pass filtered (inverse Chebyshev, $10Hz$, 108^{th} order) and then processed in three different ways to obtain three distinct datasets: $X_1(t)$: the data is left as it is; $X_2(t)$ the data is fully rectified and low-pass filtered (inverse Chebyshev, $5Hz$, 77^{th} order); $X_3(t)$ the data is fully rectified, low-pass filtered (inverse Chebyshev, $5Hz$, 77^{th} order), averaged over trials and time-interpolated over 200 points [10].

B. Selection of the number of Synergies

The n-dimensional dataset $X_i(t)$ $(i = 1, 2, 3)$ was linearly mapped by ErpICASSO [1] onto a set of m maximally informative, independent components (ICs) $S_i(t)$ by means of a mixing matrix A (*n,m*) such that $X_i(t) = AS_i(t)$. ErpICASSO is a ICA-based method that combines (i) algorithm starting point randomization, (ii) trial-to-trial bootstrapping of the input data, (iii) Curvilinear Component Analysis (CCA), to give a measure of the reliability of the ICs extracted.

Within ErpICASSO, FastICA was run 150 times on $X_i(t)$ with (i) symmetrical approach, (ii) "*tanh*" as contrast function, (iii) turned-on stabilization, (iv) stopping criterion $e = 10^{-5}$, (v) 10^4 as maximum number of iterations and (vi) *m* as number of embedded sources. Each time the algorithm starting point was randomized and the data were trial-to-trial bootstrapped with replacement for the datasets $X_1(t)$, $X_2(t)$, and point-to-point bootstrapped with replacement for the dataset $X_3(t)$ (one averaged trial). CCA was used as a multidimensional scaling method to project the points onto a two-dimensional space so as to obtain a similarity map. The columns of the 150 replicates of A were clustered according to their mutual similarities. The ambiguous IC replicates (i.e., IC runs that yielded two or more components belonging to the same cluster) were removed. The IC reliability was defined as the tightness of its cluster (i.e., the quality index Iq) ranging from 0% to 100%. The average reliability was computed as the average Iq ($\bar{Iq}$) of the m clusters. The model order m (number of synergies extracted or, equivalently, the number of retained ICs) varied from 2 to 12 (number of input variables). The final model order m corresponded to that with (i) a rate of ambiguous IC replicates $\xi < 30\%$, (ii) the maximum value of $\bar{Iq} - \xi$ (that is the maximum quality and minimum number of ambiguous IC replicates) and (iii) more than 85% of explained variance. The matrix A is built with the best IC estimates, that is the centroids of the replicates belonging to each cluster.

Fig. 1. The Figure shows the explained variance (top), the number of ambiguous IC replicates (middle) and the average Iq (bottom) while varying the model order m from 2 to 12 for a representative subject. The best performance (dashed red line) is obtained for $m = 7$.

Finally the selection of the number of synergies was performed for each dataset with the three types of aforementioned preprocessing procedures. The effect of cadence (six levels), groups (two levels), type of preprocessing (three levels) and their interaction on the number of extracted synergies was studied with a three-way ANOVA (confidence level 95%).

C. Testing

Although an extensive validation is advisable, testing was performed by selecting a few datasets and computing a number $m < 12$ of synergies for $X_2(t)$ using ErpICASSO. The

IC dataset $S_i(t)$ was then backprojected to obtain a reduced dataset $X_{red}(t) = AS_2(t)$. A dataset of uniformly distributed noise $N(t)$ was then generated with an amplitude equal to 10% of the maximum amplitude of $X_2(t)$, processed in the same way as $X_2(t)$ (i.e., fully rectified and low-pass filtered with an inverse Chebyshev, $5Hz$, 77^{th} order) and added to $X_{red}(t)$. Noticeably the statistical structure of $X_{rec} = X_{red}(t)+N(t)$ is modified artificially by noise with the same frequency content. While frequency-based techniques are not able to reconstruct the informative content of X_{rec}, this constitutes a good dataset to test whether ErpICASSO is able to detect the correct number m of synergies, that is the threshold that separates real process information from noise. Accordingly, to validate the proposed criteria, ErpICASSO was performed both on X_{rec} and X_{red} with the same procedure described in the former section. The data analysis was carried out off-line by means of customized MATLAB (The MathWorks Inc., Cambridge, MA, US) scripts.

III. Results

The EMG signals of 12 leg muscles belonging to the right leg were collected from 14 subjects (7 young and 7 elderly) while walking over ground at different cadences. Each recording (one for each subject and cadence) was processed in three separate ways to obtain the datasets $X_1(t)$ (no preprocessing - raw data), $X_2(t)$ (medium preprocessing - filtering), $X_3(t)$ (heavy preprocessing - filtering and averaging over gaits). ErpICASSO was then applied on each dataset with the criteria explained in section II to determine the number of underlying synergies.

Figure 1 shows the results of ErpICASSO on a representative subject. In particular, from top to bottom (i) the explained variance, (ii) the number of ambiguous IC replicates and (iii) the $\bar{Iq}$ are shown while varying the model order m from 2 to 12 (number of variables). According to the rules described in Section II, in this case the best model order was $m = 7$. As the figure shows, seven components explain more than 85% of the total dataset variance while minimizing the number of ambiguous IC replicates and maximizing the average quality index. Considering all the subjects, m varied from 3 to 12 with an average value of 7.0 ± 1.3, 3.5 ± 1.9, 4.0 ± 1.2, respectively for X_1,X_2,X_3.

In Figure 2 the similarity map shows the 7 clusters corresponding to the ICs selected for the same subject. Each black dot represents a component of a single run estimate. The best estimates (i.e., centroids) for each cluster are circled in blue. The quality index Iq is a measure of the compactness level of the points within each cluster. Ideally a perfect agglomeration would collapse onto a single point. The isolated points far from the main clusters are generally the ones which correspond to FastICA strokes (i.e the algorithm is stuck between two points), included in the number of ambiguous IC replicates.

The results of the testing of the criteria to select the best number of synergies are displayed in Figure 3. X_{rec} was constructed with $m = 4$ synergies and 10% of added noise (see section II). According to expectations, the quality index $\bar{Iq}$ reaches its maximum at the value $m = 4$ and starts decreasing from $m = 5$ onwards. The number of ambiguous IC replicates instead starts to increase significantly from $m = 6$ onwards. The final results are shown in Table 1.

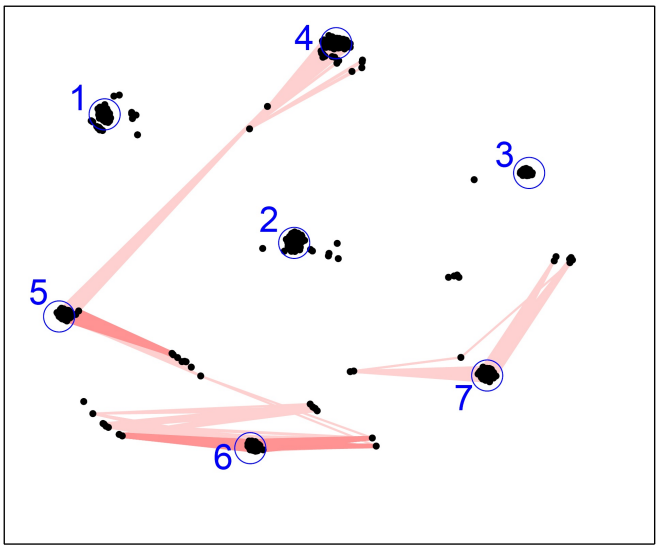

Fig. 2. The Figure shows the similarity map of a representative subject with 7 ICs selected, that is a scatterplot of the similarities between estimates projected in two dimensions by means of the Curvilinear Component Analysis (CCA) used as a multidimensional scaling method. Each black dot represents a single run estimate and the centroid of each cluster is circled in blue. The less sparse is the cluster the more reliable the related ICs are.

For each type of processing (X_1, X_2, X_3), group (elderly, young) and cadence (40, 60, $\ldots$, 140) the number of synergies was studied with a three-way ANOVA (confidence level 95%). Noticeably the type of preprocessing of the data significantly changes the number of ICs that best explain the data ($p < 0.001$). In accordance to expectations, raw data require more synergies (7 on average) than filtered data (3 or 4) due to the greater high-frequency content. The number of ICs is not altered either by cadence or by age. For all the datasets, particularly for the elderly, the variability (standard deviation) is minimum at central frequencies (80-120 strides/min).

IV. Discussion

In this work a set of criteria, based on Independent Component Analysis, was developed to determine the number of synergies to extract from a multivariate EMG dataset and was applied on the data acquired during locomotion (at different cadences) of young and elderly subjects. The proposed technique, although it still requires an extensive validation, aims at providing an answer to the deeply felt issue of selecting the correct model order when dealing with a multivariate dataset [5], [10], [14], [17]. In fact, although in principle MF techniques can be applied without any model-order reduction, in practice, particularly in the case of ICA, the quality of the decomposition may be greatly improved by PCA because the noise level is reduced. This approach however is likely to fail if the threshold that separates information from noise is not correctly identified. On this score ErpICASSO combines bootstrapping and randomization of the algorithm (FastICA) starting point, while changing the dimensionality reduction performed by PCA, thus enabling the user to set up a number of criteria to select the correct model order. Although the procedure described could be wrapped to almost any MF algorithm, ICA was adopted as it guarantees the extraction of

TABLE I. THE TABLE SHOWS THE MEAN AND STANDARD DEVIATIONS OF THE NUMBER OF SYNERGIES OF YOUNG AND ELDERLY SUBJECTS AT DIFFERENT CADENCES (40-140 STRIDES/MIN). THE RESULTS ARE REPORTED FOR THREE DIFFERENT PREPROCESSING LEVELS. RAW-DATA VALUES ARE SIGNIFICANTLY HIGHER ($p < 0.001$) THAN THE FILTERED-DATA ONES (THREE-WAY ANOVA, 95% THRESHOLD).

$X_1(t)$ Raw data							
Strides/min	40	60	80	100	120	140	Average
Young	7.4 ± 0.5	7.1 ± 1.1	7.7 ± 0.8	7.4 ± 0.8	7.4 ± 0.8	6.8 ± 1.1	7.3 ± 0.8
Elderly	6.0 ± 2.1	7.1 ± 1.2	6.6 ± 1.1	7.3 ± 0.8	6.9 ± 1.5	6.6 ± 2.1	6.7 ± 1.5
$X_2(t)$ Filtered data							
Young	3.2 ± 3.5	2.4 ± 3.5	3 ± 3.5	3.8 ± 1.6	5.7 ± 2.4	5.7 ± 0.5	3.4 ± 1.7
Elderly	5.7 ± 3.5	4 ± 2.8	3.4 ± 1.1	2.7 ± 0.4	2.8 ± 0.7	3 ± 3.5	3.6 ± 2.1
$X_3(t)$ Filtered, averaged data							
Young	3.5 ± 1.0	3.4 ± 1.3	4.0 ± 0.6	5.1 ± 2.0	3.9 ± 1.1	4.1 ± 0.9	4.0 ± 1.3
Elderly	4.3 ± 1.5	3.9 ± 0.7	3.9 ± 1.1	4.0 ± 0.6	3.7 ± 1.4	3.5 ± 1.1	3.9 ± 1.1

maximally-independent components by minimizing the mutual information between sources and it is particularly well-suited to EMG datasets [20].

A preliminary testing of the proposed criteria was performed by artificially creating a dataset of four synergies and by adding noise with overlapping frequency characteristics. Figure 3 gives evidence that PCA alone is not able to completely disentangle information from noise: according to PCA in fact, in the testing dataset 4 synergies explain almost 100% of the variance, indicating that the extracted factors also captured the artificially-added noise. Notwithstanding this, it was possible to identify the correct number of underlying factors. With respect to the variance explained alone, the Iq and the number of ambiguous IC replicates add a significant amount of information on the data structure.

Regarding the results on the experimental data, ErpICASSO was used to determine the effects that age and cadence have on the number of synergies that best account for the muscular activations during walking. The results may also help in assessing walking performance, e.g., with prosthetics.

Considering $X_2(t)$, it can be observed that the variability (standard deviation) across elderly subjects is greater at high (140 steps/min) and low (40, 60 steps/min) cadences. This is also true for young subjects but only at a low frequency. It could be surmised that young subjects are more likely to be able to cope with high walking speeds, while the elderly subjects' performance depends strongly on physical condition. The high variability at lower and higher cadences suggests that diagnostic experiments on locomotion should be performed at cadences between 80 and 120 steps/min.

The results presented are in accordance to those obtained in the previous work which analyzes these data [16]. Monaco et al. extracted a fixed number of synergies (five) from all the datasets, as suggested by Ivanenko et al. [10] and adopted three metrics to assess the effect of cadence and age on synergies, namely (i) the scalar product of the weight coefficient vectors after factorization and normalization with respect to their own norms, (ii) the Pearson correlation coefficient and (iii) the phase lag between factors (i.e. temporal offset). They showed that the synergies extracted at slower cadences (40 and 60 steps/min) correlated poorly to those extracted at the reference speed of 100 steps/min, had more variable phase-lags (reduced to zero with increasing speed) and had lower similarity. They

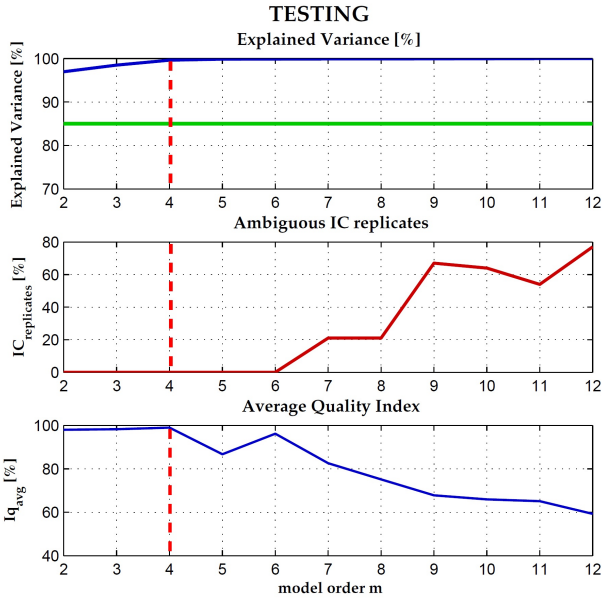

Fig. 3. The Figure shows the explained variance (top), the number of ambiguous IC replicates (middle) and the average Iq (bottom) while varying the model order m from 2 to 12 on the Testing dataset.

also showed that the correlation of some synergies between the groups was significantly higher at faster cadences, whereas when cadence decreased, primitive signals related to the young and elderly were less correlated with each other.

The criteria to determine the number of synergies presented in this work, though not always conclusive, may also be valuable in conjunction with other methods and approaches (e.g. d'Avella et al. [4]) to determine the number of synergies. Further work in this direction is desirable, in fact determining the optimal number of factors can bring further information on the underlying walking processes as it gives evidence on the complexity of the walking pattern of a subject which may be linked to neural plasticity [18]. It is reasonable to suppose that the number of synergies reflects the walking patterns physiological variability, which may be different across subjects and conditions. Fixing an *a-priori* number of factors [10] could impair results as low-power components may be contaminated by noise.

On a different note, the results of this work may be also relevant to the scientific community as they suggest that different preprocessing methods can considerably alter final results: the number of factors is not significantly altered by age or by cadence but is considerably reduced by processing. The reduction of number of synergies brought about by filtering suggests caution in rectifying, filtering, averaging over one gait cycle and time-interpolating data [10].

Despite the promising results, the proposed criteria require an extensive validation and finer tuning on different simulated datasets with various types of noise and datasets. Future works will also aim at extending ErpICASSO to other widely used MF techniques such as NMF and FA.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Alessio Ghionzoli for his help providing the experimental data used in this work. This work was partly supported by CYBERLEGs "The CYBERnetic LowEr-Limb CoGnitive Ortho-prosthesis" (ICT 287894); I-DONT-FALL "Integrated prevention and Detection sOlutioNs Tailored to the population and Risk Factors associated with FALLs" (GA 297225).

REFERENCES

[1] F. Artoni, A. Gemignani, L. Sebastiani, R. Bedini, A. Landi, and D. Menicucci. Erpicasso: A tool for reliability estimates of independent components in eeg event-related analysis. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 368–371, 2012.

[2] S.A. Bus and A.D. Lange. A comparison of the 1-step, 2-step, and 3-step protocols for obtaining barefoot plantar pressure data in the diabetic neuropathic foot. *Clinical Biomechanics*, 20(9):892–899, 2005.

[3] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[4] A. D'Avella, A. Portone, L. Fernandez, and F. Lacquaniti. Control of fast-reaching movements by muscle synergy combinations. *Journal of Neuroscience*, 26(30):7791–7810, 2006.

[5] B.L. Davis and C.L. Vaughan. Phasic behavior of emg signals during gait: Use of multivariate statistics. *Journal of Electromyography and Kinesiology*, 3(1):51–60, 1993.

[6] P. DeVita and T. Hortobagyi. Age causes a redistribution of joint torques and powers during gait. *Journal of Applied Physiology*, 88(5):1804–1811, 2000.

[7] J. Himberg, A. Hyvarinen, and F. Esposito. Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage*, 22(3):1214–1222, 2004.

[8] A. Hyvarinen. Fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters*, 10(1):1–5, 1999.

[9] A. Hyvarinen, P.O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[10] Y.P. Ivanenko, R.E. Poppele, and F. Lacquaniti. Five basic muscle activation patterns account for muscle activity during human locomotion. *Journal of Physiology*, 556(1):267–282, 2004.

[11] D.C. Kerrigan, M.K. Todd, U. Della Croce, L.A. Lipsitz, and J.J. Collins. Biomechanical gait alterations independent of speed in the healthy elderly: Evidence for specific limiting impairments. *Archives of Physical Medicine and Rehabilitation*, 79(3):317–322, 1998.

[12] T. Masud and R.O. Morris. Epidemiology of falls. *Age and Ageing*, 30(SUPPL. 4):3–7, 2001.

[13] C.A. McGibbon. Toward a better understanding of gait changes with age and disablement: Neuromuscular adaptation. *Exercise and Sport Sciences Reviews*, 31(2):102–108, 2003.

[14] L.A. Merkle, C.S. Layne, J.J. Bloomberg, and J.J. Zhang. Using factor analysis to identify neuromuscular synergies during treadmill walking. *Journal of Neuroscience Methods*, 82(2):207–214, 1998.

[15] V. Monaco, A. Ghionzoli, P. Dario, and S. Micera. Muscle synergies during walking: Comparison between young and elderly people. preliminary results. *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS'08 - "Personalized Healthcare through Technology"*, pages 5370–5373, 2008.

[16] V. Monaco, A. Ghionzoli, and S. Micera. Age-related modifications of muscle synergies and spinal cord activity during locomotion. *Journal of Neurophysiology*, 104(4):2092–2102, 2010.

[17] K.S. Olree and C.L. Vaughan. Fundamental patterns of bilateral muscle activity in human locomotion. *Biological Cybernetics*, 73(5):409–414, 1995.

[18] A. Schmitz, A. Silder, B. Heiderscheit, J. Mahoney, and D.G. Thelen. Differences in lower-extremity muscular activation during walking between healthy older and young adults. *Journal of Electromyography and Kinesiology*, 19(6):1085–1091, 2009.

[19] A. Silder, B. Heiderscheit, and D.G. Thelen. Active and passive contributions to joint kinetics during walking in older adults. *Journal of Biomechanics*, 41(7):1520–1527, 2008.

[20] M.C. Tresch, V.C.K. Cheung, and A. D'Avella. Matrix factorization algorithms for the identification of muscle synergies: Evaluation on simulated and experimental data sets. *Journal of Neurophysiology*, 95(4):2199–2212, 2006.

[21] D.A. Winter, A.E. Patla, J.S. Frank, and S.E. Walt. Biomechanical walking pattern changes in the fit and healthy elderly. *Physical Therapy*, 70(6):340–347, 1990.

[22] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.