frontiers in
COMPUTATIONAL NEUROSCIENCE

METHODS ARTICLE
published: 20 December 2013
doi: 10.3389/fncom.2013.00185

Model selection for the extraction of movement primitives

Dominik M. Endres**, Enrico Chiovetto and Martin A. Giese

Section Computational Sensomotorics, Department of Cognitive Neurology, CIN, HIH, BCCN, University Clinic Tübingen, Tübingen, Germany

Edited by:
Andrea D'Avella, IRCCS Fondazione Santa Lucia, Italy

Reviewed by:
Stefano Panzeri, Italian Institute of Technology, Italy
Vincent C. K. Cheung, MIT, USA

****Correspondence:***
Dominik M. Endres, Section Computational Sensomotorics, Department of Cognitive Neurology, CIN, HIH, BCCN, University Clinic Tübingen, Otfried-Müller-Str. 25, 72076 Tübingen, Germany
e-mail: dominik.endres@klinikum.uni-tuebingen.de

A wide range of blind source separation methods have been used in motor control research for the extraction of movement primitives from EMG and kinematic data. Popular examples are principal component analysis (PCA), independent component analysis (ICA), anechoic demixing, and the time-varying synergy model (d'Avella and Tresch, 2002). However, choosing the parameters of these models, or indeed choosing the type of model, is often done in a heuristic fashion, driven by result expectations as much as by the data. We propose an objective criterion which allows to select the model type, number of primitives and the temporal smoothness prior. Our approach is based on a Laplace approximation to the posterior distribution of the parameters of a given blind source separation model, re-formulated as a Bayesian generative model. We first validate our criterion on ground truth data, showing that it performs at least as good as traditional model selection criteria [Bayesian information criterion, BIC (Schwarz, 1978) and the Akaike Information Criterion (AIC) (Akaike, 1974)]. Then, we analyze human gait data, finding that an anechoic mixture model with a temporal smoothness constraint on the sources can best account for the data.

Keywords: motor primitives, blind source separation, temporal smoothing, model selection, laplace approximation, bayesian methods, movement primitives

1. INTRODUCTION

In recent years substantial experimental evidence has been provided that supports the hypothesis that complex motor behavior is organized in modules or simple units called movement primitives (Flash and Hochner, 2005; Bizzi et al., 2008). In this framework each module, or motor primitive, consists of a set of movement variables, such as joint trajectories (Santello et al., 1998; Kaminski, 2007) or muscle activations (d'Avella et al., 2006; Chiovetto et al., 2010) acting synergistically over time. By combination of small numbers of these primitives complex motor behaviors can be generated. Several methods have been used so far in the literature for the identification of motor primitives starting from experimental data sets, which include both well-known classical unsupervised learning techniques based on instantaneous mixture models, such as principal component analysis (PCA) and independent component analysis (ICA) (Chiovetto et al., 2010; Dominici et al., 2011), or even more advanced techniques that include the estimation of temporal delays of the relevant mixture components (d'Avella et al., 2006; Omlor and Giese, 2011). On the one hand, all these approaches differ from each other in multiple aspects, such as their underlying generative models or the specific priors imposed on the parameters. On the other hand, however, for all of them the number of primitives to be extracted and subsequently used to approximate the original data has to be set a priori. To our knowledge only very few motor control studies have so far addressed the problem of model selection in a principled way, see e.g., Delis et al. (2013); Hart and Giszter (2013) for notable exceptions. The existing generative models for the extraction of motor primitives have indeed been demonstrated to provide a low-dimensional decomposition of the experimental data, but no clear criterion has been developed to objectively determine which model is best suited for describing the statistical features of the data under investigation. We are concerned with two types of statistical features:

- "hard" constraints, such as the number of primitives. Determining this number is also known as "model order estimation."
- "Soft" constraints, e.g., regularity measures. In other words, a constraint on a parameter is "soft," if it expresses a preference or expectation for the parameter's value, but does allow for deviation from this preference given sufficient evidence. For example, when modeling human walking, we expect a periodic movement with predominantly low frequency components. However, higher frequency components might be critical to capture specific, more complex movement primitives. We therefore would like to allow for the possibility of overriding our initial expectations if the data indicate that this is appropriate. One such regularity measure, *temporal smoothness* quantified by a kernel function, is a novelty of our approach in the context of model selection for blind source separation in motor control.

Concerning the model order selection, several criteria have been developed. Most of them require the computation of the likelihood function (Schwarz, 1978; Akaike, 1987; Basilevsky, 1994; Minka, 2000; Zucchini, 2000) and attempt to determine the right model order as the one that offers the best trade-off between accuracy of data fitting and complexity of the model. Our approach uses this trade-off in a more general setting. Such information criteria were proven to identify with almost no error the model order of noisy data sets when these were corrupted with Gaussian noise, but performances were shown to be noticeably worse when data were corrupted with signal-dependent noise (Tresch et al., 2006), which is actually thought to affect strongly the neural control signals (Harris and Wolpert, 1998). In this article we present a new objective criterion for model-order selection that extends the other classical ones based on information-theoretic and statistical approaches. The criterion is based on a Laplace approximation of the posterior distribution of the parameters of a given blind source separation method, re-formulated as a Bayesian generative model. We derive this criterion for a range of blind source separation approaches, including for the first time the anechoic mixture model (AMM) described in Omlor and Giese (2011).

We provide a validation of our criterion based on an artificial ground truth data set generated in such a way to present well-known statistical properties of real kinematic data. We show in particular that our method performs at least as well as other traditional model order selection criteria [Akaike's Information Criterion, AIC (Akaike, 1974) and the Bayesian Information Criterion, BIC (Schwarz, 1978)], that it works for both instantaneous and delayed mixtures and allows to distinguish between these given moderately sized datasets, and that it can provide information regarding the level of temporal smoothness of the generating sources.

We finally apply the criterion to actual human locomotion data, to find that, differently from other standard synchronous linear models, a linear mixture of time shiftable components characterized by a specific degree of temporal smoothness is a better account of the data-generating process.

1.1. RELATED APPROACHES

The well-known plug-in estimators, BIC and AIC, have the advantage of being easy to use when a likelihood function for a given model is available. Hence, they are often the first choice for model order estimation, but not necessarily the best one. In Tu and Xu (2011) several criteria for probabilistic PCA (or factor analysis) models were evaluated, including AIC, BIC, MIBS (Minka's Bayesian model selection) (Minka, 2000) and Bayesian Ying-Yang (Xu, 2007). The authors found that MIBS and Bayesian Ying-Yang work best. The approach presented in Kazianka and Pilz (2009) corrected the approximations made in MIBS, which yielded improved performance on small sample sizes. This corrected MIBS performed better than all other approaches tested in that paper, including AIC and BIC.

The authors of Li et al. (2007) estimated the number of independent components in fMRI data with AIC and minimum description length [MDL, (Rissanen, 1978)], which boils down to BIC. They showed that temporal correlations adversely affect the accuracy of standard complexity estimators, and proposed a subsampling procedure to remove these correlations. In contrast, we demonstrate below how to deal with temporal dependence as a part of our model. Another MDL-inspired approach, code length relative to a Gaussian prior (CLRG) was introduced in Plant et al. (2010) to compare different ICA approaches and model orders. It was demonstrated to work well on simulated data without the need of choosing additional parameters, such as thresholds, and it was shown that it is able to recover task-related fMRI components better than heuristic approaches.

Such heuristic approaches typically utilize some features of the reconstruction error (or conversely, of the variance-accounted-for (VAF)) as a function of the model order, e.g., finding a "knee" (inflection point) in that function, a procedure which is inspired by the scree test for factor analysis (Cattell, 1966). For example, the authors of Cheung and Xu (1999) experimented with an empirical criterion for ICA component selection. The independent components were ordered according to their contribution to the reduction of reconstruction error. Only those independent components were retained that had a large effect on this error. Similarly, the approach of Sawada et al. (2005) used "unrecovered power," which is basically reconstruction error, to determine which components of a (reverberant) mixture are important. The work in Valle et al. (1999) compared various criteria for PCA component selection on real and simulated chemical reactor data, finding that some of the heuristic reconstruction-error based methods still perform well when PCA model assumptions are violated by the data-generating process.

To distinguish convolutive (but undelayed) mixtures from instantaneous ones, the work in Dyrholm et al. (2007) employed the framework of Bayesian model selection for the analysis of EEG data. Related to our approach, the authors of Penny and Roberts (2001) derived Laplace approximations to the marginal likelihood of several ICA model classes for model selection and model order determination. Their work is conceptually similar to our approach, but we also consider delayed mixtures.

All approaches reviewed so far are deterministic in nature. There are also sampling methods available for model selection purposes, see Bishop (2007) for details. One example is e.g., the work of Ichir and Mohammad-Djafari (2005) which used importance sampling and simulated annealing for model-order selection of L1-sparse mixtures.

2. MATERIALS AND METHODS

We develop our model (order) criterion in the framework of Bayesian generative model comparison Bishop (2007). Let D be observable data, Θ_M a tuple of model parameters for a model indexed by M (the "model index") and Φ a tuple of hyperparameters. Using standard terminology, we denote

$$\text{likelihood} : p\left(D|\Theta_M, \Phi, M\right) \tag{1}$$

$$\text{prior} : p\left(\Theta_M|\Phi, M\right). \tag{2}$$

The likelihood is the probability density of the data given the model parameters, model index and hyperparameters. The parameter prior is the probability density of the model parameters. Then the marginal likelihood of M, or model evidence for M is given by

$$\begin{aligned} p(D|\Phi, M) &= \int d\Theta_M p\left(D, \Theta_M|\Phi, M\right) \\ &= \int d\Theta_M p(D|\Theta_M, \Phi, M) p\left(\Theta_M|\Phi, M\right) \end{aligned} \tag{3}$$

where the second equality follows from the product rule for probability distributions. Strictly speaking, the Φ would have to be integrated out as well after choosing a suitable prior for them. However, to keep the problem tractable we determine their value by maximizing the model evidence with respect to them, finding that this yields sufficiently good approximations for our purposes. Once we have evaluated Equation 3 for all M, we can select that M which maximizes the model evidence, since we have no *a-priori* preference for any M.

To apply this model selection framework, we reformulate three popular blind source separation (BSS) methods, namely probabilistic PCA (pPCA), ICA and anechoic demixing as generative models in section 2.1. This reformulation allows us to evaluate their likelihoods and parameter priors. We then use a Laplace approximation (Laplace, 1774) to compute an approximation to the marginal likelihood of each model. This approximation is derived in section 2.2.

2.1. GENERATIVE MODELS OF BLIND SOURCE SEPARATION METHODS

The BSS methods we consider all assume a linear generative model in discrete time, where observable data $\mathbf{X}$ can be written as a linear superposition of sources $\mathbf{S}$ multiplied by weights $\mathbf{W}$. Let $t = 1, \ldots, t$ be the T (equally spaced) time points, $i = 1, \ldots, I$ the source index, and $j = 1, \ldots, J$ the signal index. Note that J could also be interpreted as a trial index, i.e., one signal repeated J times, or any combination of trials and signals. For the models we consider, there is no formal difference between "trial" and "signal," as opposed to e.g., the time varying synergy model (d'Avella et al., 2006). Then $\mathbf{X}$ is a $(J \times T)$ matrix, $\mathbf{S}$ is $(I \times T)$ and consequently $\mathbf{W}$ must be $(J \times I)$ so that

$$\mathbf{X} = \mathbf{W}\mathbf{S} + \mathbf{\Sigma} \tag{4}$$

$$\Sigma_{jt} \sim \mathcal{N}\left(0, \sigma_n^2\right) \tag{5}$$

where the entries of the noise matrix $\mathbf{\Sigma}$ are drawn independently from a Gaussian distribution with zero mean and variance σ_n^2. In an anechoic (delayed) mixture, the sources additionally depend on the signal index j (see section 2.1.3 for details).

The differences between the BSS approaches can be expressed as priors on $\mathbf{S}$ and $\mathbf{W}$, which we describe in the following.

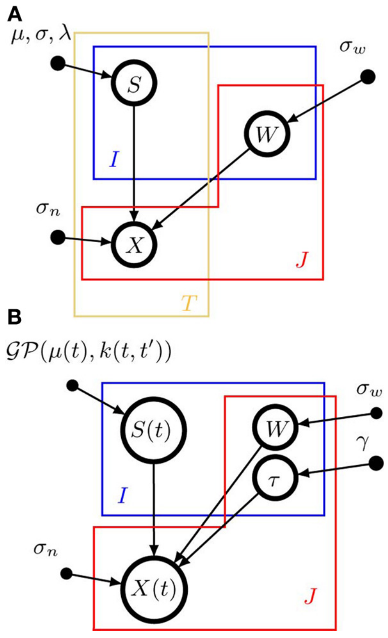

FIGURE 1 | Graphical model representations of the blind source separation algorithms for which we compute a model evidence approximation. We follow standard graphical modeling terminology (see e.g., Bishop, 2007). Open circles represent random variables, which may also be random functions. Filled circles are parameters. Arrows denote conditional dependencies. The *plates* (colored frames) indicate that the enclosed structure is repeated as often as the corresponding letter indicates. Enclosure in multiple plates indicates a product of repetitions. For example, in panel **(A)** there are $I \times T$ random variables S which comprise the source matrix. **(A)** Instantaneous, undelayed mixtures such as pPCA (where $\lambda = 0$) and ICA. $J \times T$ signals X are computed by mixing $I \times T$ sources S with $J \times I$ weights W. σ_W is the standard deviation of the zero-mean Gaussian prior on the weights. σ_n is the noise standard deviation. μ and σ are the parameters of the Gaussian part of the prior on the sources, λ measures the deviation from Gaussianity. **(B)**. Convolutive, delayed mixtures, like the anechoic mixture of Omlor and Giese (2011) with additional temporal smoothness constraints. I source functions $S(t)$ are drawn from a Gaussian process $\mathcal{GP}(\mu(t), k(t, t'))$ with mean function $\mu(t)$ and kernel $k(t, t')$. These sources are shifted in time by $J \times I$ many delays (one per trial and source) drawn from an exponential distribution with parameter γ and mixed with $J \times I$ weights W which are drawn from a zero-mean Gaussian distribution with standard deviation σ_W, to yield J signals $X(t)$. For details, see text.

2.1.1. Probabilistic PCA (pPCA)

PCA is one of the most widely used BSS approaches. In Tipping and Bishop (1999), it was demonstrated how PCA results from a probabilistic generative model: assuming the data have mean zero (i.e., $\forall j : \sum_t \mathbf{X}_{jt} = 0$), and using an independent zero-mean Gaussian prior on the sources, i.e.,

$$\mathbf{S}_{it} \sim \mathcal{N}\left(\mu = 0, \sigma^2\right) \tag{6}$$

the weights $\mathbf{W}$ which maximize the marginal likelihood of $\mathbf{X}$ after integrating out $\mathbf{S}$, are given by the scaled (and possibly rotated) principal I eigenvectors of the $(J \times J)$ data covariance matrix, $\frac{1}{T}\mathbf{X}\mathbf{X}^T$. This model differs from PCA insofar as the sources will only be equal to the PCA factors in the noise-free limit $\sigma_n \to 0$, and is hence referred to as *probabilistic* PCA (Tipping and Bishop, 1999). Similarly, when we put a prior on the weights

$$\mathbf{W}_{ji} \sim \mathcal{N}\left(0, \sigma_w^2\right) \tag{7}$$

and integrate them out, we find that the best I sources $\mathbf{S}$ are the principal eigenvectors of the $(T \times T)$ data covariance matrix $\frac{1}{J}\mathbf{X}^T\mathbf{X}$ (assuming zero mean signals at every time step, i.e., $\forall t : \sum_j \mathbf{X}_{jt} = 0$). We will therefore use both priors for a completely probabilistic pPCA model[1].

A graphical model representation of pPCA is shown in **Figure 1A**. Open circles represent random variables, which may

[1] We will refer to this model interchangeably by pPCA or just PCA in the following.

also be random functions. Filled circles are parameters. Arrows denote conditional dependencies. The *plates* (colored frames) indicate that the enclosed structure is repeated as often as the corresponding letter indicates. Enclosure in multiple plates indicates a product of repetitions. Thus, in a pPCA model, $I \times T$ sources are *a-priori* drawn independently of each other (μ and σ are parameters, not random variables), and source values have no dependencies across time. Likewise, weights have no dependencies across sources or signals. In contrast, data points depend on both weights and sources, as indicated by the arrows converging on **X** from **S** and **W**.

Given the generative model (Equation 4) and the prior specification (Equation 6 and Equation 7), we can now write down the likelihood and prior terms which we need for the evaluation of the model evidence (Equation 3). To this end, we identify the number of sources I with the model index M, and (cf. Equation 3)

$$D = \mathbf{X} \tag{8}$$

$$\Theta_M = (\mathbf{W}, \mathbf{S}) \tag{9}$$

$$\Phi = (\mu, \sigma, \sigma_w, \sigma_n) \tag{10}$$

$$p\,(D|\Theta_M, \Phi, M) = \frac{\exp\left(-\frac{1}{2\sigma_n^2}\,\|\mathbf{X} - \mathbf{WS}\|_F\right)}{\sqrt{2\pi\sigma_n^2}^{JT}} \tag{11}$$

$$p(\Theta_M|\Phi, M) = \frac{\exp\left(-\frac{1}{2\sigma^2}\,\|\mathbf{S}_{it} - \mu \cdot \mathbf{1_{IT}}\|_F\right)}{\sqrt{2\pi\sigma^2}^{IT}} \times \frac{\exp\left(-\frac{1}{2\sigma_w^2}\,\|\mathbf{W}\|_F\right)}{\sqrt{2\pi\sigma_w^2}^{JI}} \tag{12}$$

where $\mathbf{1}_{IT}$ is an $(I \times T)$ matrix with every element being 1, and $\|\mathbf{A}\|_F$ is the Frobenius norm of matrix **A**.

2.1.2. Independent Component Analysis (ICA)

The term ICA refers to a variety of BSS methods which try decompose signals into sources with two main goals:

1. The sources are as statistically independent as possible according to some suitably chosen measure, and
2. the sources allow for a good reconstruction of the signals.

Infomax ICA (Bell and Sejnowski, 1995) tries to achieve these goals by maximizing the mutual information (Cover and Thomas, 1991) between sources and signals, which clearly promotes the second goal. The first one is promoted if the BSS system contains an information bottleneck, e.g., fewer sources than signals. In that case, maximizing mutual information amounts to maximizing the total source entropy, which is achieved if the sources are independent.

The FastICA algorithm (Hyvarinen, 1999) aims directly at minimizing the mutual information between the sources, thereby promoting goal one. Goal 2 is achieved by constraining the (linear) transformation from signals to the sources to be invertible, or at least almost invertible in the noisy or lossy case, such that the signals can be reconstructed using the generative model above (Equation 4). Mutual information is measured via *negentropy*, which is the negative difference between the entropy of a source and the entropy of a variance-matched Gaussian variable, i.e., it is a measure of non-Gaussianity. Maximizing negentropy then minimizes mutual information. To measure negentropy, the authors of Hyvarinen (1999) used the "contrast function" approach developed in Hyvärinen (1998). Contrast functions provide constraints on expectations of probability distributions, in addition to the mean and variance constraints of Gaussians. Consequently, the maximum entropy distributions obeying these constraints have the contrast function(s) as sufficient statistics, with an associated natural parameter, which controls the deviation of the resulting distribution from a Gaussian. For a detailed derivation see Hyvärinen (1998). This motivates the following source prior for probabilistic ICA models: let $G(.)$ be the contrast function, then

$$p\,(\mathbf{S}_{it}) = \frac{1}{Z(\mu, \sigma, \lambda)} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{S}_{it} - \mu)^2 + \lambda G(\mathbf{S}_{it} - \mu)\right) \tag{13}$$

where λ is the natural parameter associated with $G(.)$. The normalization constant $Z(\mu, \sigma, \lambda)$ can be evaluated by numerical integration, since the prior is a density over a one-dimensional random variable. Similar to pPCA, we use a Gaussian prior on the weights. The graphical model representation of ICA is the same as for pPCA (see **Figure 1A**), since there is no a-priori dependency between sources or weights across time.

We can now identify the number of sources I with the model index M, and furthermore (cf. Equation 3)

$$D = \mathbf{X} \tag{14}$$

$$\Theta_M = (\mathbf{W}, \mathbf{S}) \tag{15}$$

$$\Phi = (\mu, \sigma, \sigma_w, \sigma_n, \lambda) \tag{16}$$

$$p(D|\Theta_M, \Phi, M) = \frac{\exp\left(-\frac{1}{2\sigma_n^2}\,\|\mathbf{X} - \mathbf{WS}\|_F\right)}{\sqrt{2\pi\sigma_n^2}^{JT}} \tag{17}$$

$$p(\Theta_M|\Phi, M) = \prod_{i,t} p(\mathbf{S}_{it}) \times \frac{\exp\left(-\frac{1}{2\sigma_w^2}\,\|\mathbf{W}\|_F\right)}{\sqrt{2\pi\sigma_w^2}^{JI}} \tag{18}$$

2.1.3. Anechoic mixture models (AMM) and smooth instantaneous mixtures (SIM)

AMMs may be seen as an extension of the above BSS approaches to deal with time-shifted sources (Omlor and Giese, 2007a). Such time shifts are obviously useful in motor control, where coordinated movement patterns, such as gaits, might be characterized by opposite joints moving in a similar manner but time-shifted against each other (e.g., the legs during walking); the well-known time-varying synergy model (d'Avella et al., 2006) is a kind of AMM. The generative models of AMMs are linear with additive Gaussian noise (similar to Equation 4), but the sources $S_i(t)$ are shifted by delays τ_{ji}, which are the elements of a $(J \times I)$ matrix τ. We draw these delays from an exponential prior with mean γ, which promotes delays that differ sparsely from zero.

$$\begin{aligned}\mathbf{X}_{jt} &= \sum_i \mathbf{W}_{ji} S_i(t - \tau_{ji}) + \eta_{jt} \\ &= \sum_i \hat{\mathbf{X}} + \eta_{jt} \qquad (19)\end{aligned}$$

$$\eta_{jt} \sim \mathcal{N}(0, \sigma_n) \qquad (20)$$

$$p(\tau_{ji}) = \gamma \exp\left(-\frac{\tau_{ij}}{\gamma}\right) \qquad (21)$$

where we define the matrix of the reconstructed signals $\hat{\mathbf{X}}$ as $\hat{\mathbf{X}}_{ji} = \sum_i \mathbf{W}_{ji} S_i(t - \tau_{ji})$. Moreover, we impose soft temporal regularity constraints on the sources. To this end, we draw the sources from a Gaussian process (GP) (Rasmussen and Williams, 2006) with mean function $\mu(t)$ and covariance (or kernel) function $k(t, t')$. A GP is a prior over functions $S(t)$ where the joint distribution of any finite number of function values at times $t_1, \ldots, t_N$ follows a multivariate Gaussian distribution i.e.,

$$\vec{S} = (S(t_1), \ldots, S(t_N)) \qquad (22)$$

$$\vec{\mu} = (\mu(t_1), \ldots, \mu(t_N)) \qquad (23)$$

$$\mathbf{K}_{mn} = k(t_m, t_n) \qquad (24)$$

$$\vec{S} \sim \mathcal{N}(\vec{\mu}, \mathbf{K}) \qquad (25)$$

Thus, the choice of kernel function determines how much the function values at different points tend to co-vary a priori. Throughout this paper, we will use kernel functions of the form

$$k(t, t') \propto \operatorname{sinc}(2f_0|t - t'|) = \frac{\sin(2\pi f_0|t - t'|)}{2\pi f_0 |t - t'|} \qquad (26)$$

which is also called *wave kernel* (Genton, 2001) in the machine learning literature. This choice is motivated by the observation that the inverse Fourier transform of an ideal low-pass filter with cutoff-frequency f_0 is proportional to this kernel. Thus, functions drawn from a GP with this kernel will vary on timescales comparable to f_0, see **Figure 2** for examples. Note, however, that the regularization provided by the kernel is "soft": when learning sources from small datasets, they will have the smoothness properties given by the kernel. For large datasets, the kernel regularization may be overridden by the data.

With this prior, the matrix of reconstructed signals $\hat{\mathbf{X}}$ and using as model index the tuples $M = (I, f_0)$ we find

$$D = \mathbf{X} \qquad (27)$$

$$\Theta_M = (\mathbf{W}, \tau, S_1(t), \ldots, S_I(t)) \qquad (28)$$

$$\Phi = \left(\mu(t), k_{f_0}(t, t'), \sigma_n, \sigma_w\right) \qquad (29)$$

$$p(D|\Theta_M, \Phi, M) = \frac{\exp\left(-\frac{1}{2\sigma_n^2}\|\mathbf{X} - \hat{\mathbf{X}}\|_F\right)}{\sqrt{2\pi\sigma_n^2}^{JT}} \qquad (30)$$

$$p(\Theta_M|\Phi, M) = \frac{\exp(-\frac{1}{2}\mathbf{S}_i \mathbf{K}^{-1} \mathbf{S}_i^T)}{\sqrt{2\pi}^T \sqrt{|\mathbf{K}|}} \prod_{j,i} \gamma \exp\left(-\frac{\tau_{ji}}{\gamma}\right)$$

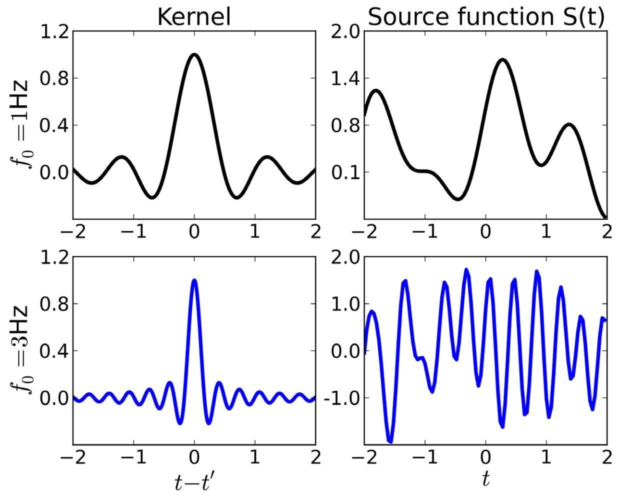

FIGURE 2 | Examples of kernel functions (left) and sources (right) drawn from a Gaussian process prior with the corresponding kernel. Throughout this paper, we use shift-invariant kernels of the form $k(t, t') \propto \operatorname{sinc}(2f_0|t - t'|)$. **Top row:** Kernel function for f_0 = 1Hz (left) and source function drawn from a Gaussian process with that kernel. The source varies rather smoothly on a timescale comparable to f_0. **Bottom row:** Kernel function and source for f_0 = 3Hz.

$$\times \frac{\exp\left(-\frac{1}{2\sigma_w^2}\|\mathbf{W}\|_F\right)}{\sqrt{2\pi\sigma_w^2}^{JI}} \qquad (31)$$

where $\mathbf{S}_i$ is the i-th row of the *undelayed* source matrix, i.e., the matrix of the source functions sampled at times $t = 1, \ldots, T$. A graphical model representation of AMM is shown in **Figure 1B**.

As a special case of the AMM model above, we consider the case $\forall i, j : \tau_{ji} = 0$, i.e., a mixture without delays, but GP-induced temporal regularization. In the following, we refer to this as the *smooth instantaneous mixture*, or SIM.

2.2. LAPLACE APPROXIMATION

We now turn to the evaluation of the model evidence, Equation 3. The difficult part, as usual in Bayesian approaches, is the integral over the model parameters Θ (we drop the index M in the following for notational simplicity, since we evaluate the model evidence for each M separately). Instead of an exact solution, we therefore resort to a *Laplace approximation* (Laplace, 1774; Bishop, 2007). To use this approach, concatenate the Θ into a vector and then construct a saddle-point approximation (Reif, 1995) of intractable integrals of the form

$$\int d\Theta \exp\left(-f(\Theta)\right) \qquad (32)$$

assuming that $f(\Theta)$ has a single, sharply peaked minimum at some $\Theta^* = \operatorname{argmin}_\Theta f(\Theta)$ and is twice continuously differentiable. In this case, only exponents close to the minimal exponent $f(\Theta^*)$ will make noticeable contributions to the integral. Hence, we can approximate $f(\Theta)$ *locally* around Θ^* by a Taylor expansion

$$f(\Theta) \approx f(\Theta^*) + \nabla_{\Theta^*} f(\Theta)^T (\Theta - \Theta^*) + \frac{1}{2}(\Theta - \Theta^*)^T \mathbf{H}\, (\Theta - \Theta^*) \tag{33}$$

where

$$\mathbf{H}_{uv} = \left.\frac{\partial^2 f(\Theta)}{\partial\theta_u \partial\theta_v}\right|_{\Theta^*}$$

is the Hessian matrix of the 2nd derivatives evaluated at Θ^*. Since Θ^* is the location of the minimum of $f(\Theta)$, it follows that $\nabla_{\Theta^*} f(\Theta)^T = 0$ and $\mathbf{H}$ is positive (semi-)definite. Thus

$$f(\Theta) \approx f(\Theta^*) + \frac{1}{2}(\Theta - \Theta^*)^T \mathbf{H}\, (\Theta - \Theta^*) \tag{34}$$

and we can approximate the integral as

$$\begin{aligned}
&\int d\Theta \exp\left(-f(\Theta)\right) \\
&\approx \int d\Theta \exp\left(-f(\Theta^*) - \frac{1}{2}(\Theta - \Theta^*)^T \mathbf{H}\, (\Theta - \Theta^*)\right) \\
&= \exp\left(-f(\Theta^*)\right) \int d\Theta \exp\left(-\frac{1}{2}(\Theta - \Theta^*)^T \mathbf{H}\, (\Theta - \Theta^*)\right) \\
&= \exp\left(-f(\Theta^*)\right) \frac{(2\pi)^{\frac{F}{2}}}{\sqrt{|\mathbf{H}|}}
\end{aligned}$$

where $F = \dim(\Theta)$ is the dimensionality of Θ. For the derivation of our model comparison criterion, we will need the logarithm of this integral:

$$\log\left(\int d\Theta \exp\left(-f(\Theta)\right)\right) \approx -f(\Theta^*) + \frac{F}{2}\log(2\pi) - \frac{1}{2}\log\left(|\mathbf{H}|\right). \tag{35}$$

In summary, the Laplace approximation replaces the intractable integral with differentiation, which is always possible for the models we consider.

To approximate the model evidence (Equation 3) in this way, let

$$\Theta^* = \mathrm{argmin}_\Theta \left[-\log(p(D|\Theta, \Phi, M)) - \log(p(\Theta|\Phi, M))\right] \tag{36}$$

in other words, Θ^* are the parameters which maximize the likelihood subject to the regularization provided by the parameter prior. Furthermore, denote

$$\begin{aligned}
\mathbf{H}_{uv} = &- \left.\frac{\partial^2 \log(p(D|\Theta, \Phi, M))}{\partial\Theta_u \Theta_v}\right|_{\Theta^*} \\
&- \left.\frac{\partial^2 \log(p(\Theta|\Phi, M))}{\partial\Theta_u \Theta_v}\right|_{\Theta^*}
\end{aligned} \tag{37}$$

and thus

$$\begin{aligned}
p(D|\Phi, M) \approx &\underbrace{\log(p(D|\Theta^*, \Phi, M))}_{\text{log-likelihood}} + \underbrace{\log(p(\Theta^*|\Phi, M))}_{\text{log-prior}} \\
&+ \underbrace{\frac{\dim(\Theta)}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{H}|)}_{\text{log-posterior-volume}}
\end{aligned} \tag{38}$$

which we will refer to as the *LAP* criterion for model comparison: the larger LAP, the better the model. It comprises three parts, which can be interpreted: the log-likelihood measures the goodness of fit, similar to explained variance or VAF. The second term is the logarithm of the prior, which corresponds to a regularization term for dealing with under-constrained solutions for Θ when the datset is small. Finally, the third part measures the volume of the parameter posterior, since $\mathbf{H}$ is the posterior precision matrix (inverse covariance) of the parameters in the vicinity of Θ^*, i.e., it indicates how well the data constrain the parameters (large $|\mathbf{H}|$ means small posterior volume, which means Θ is well-constrained).

We will compare the LAP criterion to two standard model complexity estimators below (see section 3): BIC and AIC. BIC is given by

$$\mathrm{BIC} = -2\left(\log(p(D|\Theta^*, \Phi, M)) - \frac{1}{2}\dim(\Theta)\log(N)\right) \tag{39}$$

where N is the number of data-points. The best model is found by minimizing BIC w.r.t. M. BIC can be obtained from LAP in the limit $N \to \infty$, by dropping all terms from LAP which do not grow with N and multiplying by -2. Assuming that the model has no latent variables (whose number typically grows with N), the terms to be dropped from Equation 38 are the log-prior, the first term of the posterior volume, and the second term of the Hessian (Equation 37). For i.i.d. observations, the determinant of the first term of the Hesssian will typically grow like $Nc^{\dim(\Theta)}$ where c is some constant independent of N. Hence, the BIC follows. While this reasoning is somewhat approximate (a rigorous derivation can be found in Schwarz (1978)), it highlights that we might expect LAP to become more similar to BIC as the dataset increases.

AIC is originally derived from information-theoretic arguments (Akaike, 1987): a good model loses only a small amount of information when approximating (unknown) reality. When information is measured by Kullback-Leibler divergence (Cover and Thomas, 1991), AIC follows. Alternatively, it also obtained by choosing a model complexity prior which depends on N and $\dim(\Theta)$ (Burnham and Anderson, 2004) and is given by

$$\mathrm{AIC} = -2\left(\log(p(D|\Theta^*, \Phi, M)) - \dim(\Theta)\right). \tag{40}$$

Like BIC, a good model has a low AIC score.

2.3. ASSESSMENT OF CRITERION PERFORMANCE

To validate our criterion we assessed its performance on synthesized data sets with well-known statistical properties and on actual kinematic data collected from human participants during a free walking task. We also compared the results with those provided by AIC and BIC. Before applying the model selection criteria we factorized each available data set according to the mixture models Equation 4 and Equation 19. The identification of the parameters Θ was carried out in two phases: first, we applied singular value decomposition to identify the principal components (for PCA), or fastICA (Hyvarinen, 1999) or the anechoic demixing algorithm mentioned above (Omlor and Giese, 2011) to yield weights and sources. Second, we used these solutions to initialize an optimization of the corresponding likelihood function, to determine the optimal parameters Θ^* and hyperparameters Φ needed for the Laplace approximation. The optimization in the second step was carried out using the L-BFGS-B routine in the SciPy package (Jones et al., 2001) for Θ^*, Φ was then reestimated for fixed Θ^*. This second optimization was necessary for two reasons: the statistical reformulations of pPCA and ICA will yield solutions which are very similar, but not identical to the original algorithms, and the AMM method from Omlor and Giese (2011) can not handle temporal smoothness priors. The number of components I identified ranged, for all algorithms, from 1 to 8.

2.3.1. Ground-truth data generation

We simulated kinematic-like data (mimicking, for instance, joint-angle trajectories) based on the generative models Equation 4 and Equation 19 that is linear combinations of I primitives that could be synchronous (SIM) or shifted in time (AMM). For the generation of each primitive $S_i(t)$ we drew 100 random samples from a normal distribution (MATLAB (2010) function "randn") and then we low-pass filtered them with a 6th-order Butterworth filter [MATLAB (2010) functions "butter" and "filtfilt"]. Two cut-off frequencies were used for filtering, respectively, 5 and 10 Hz, to simulate data with two different frequency spectra. Sampling frequency of the data was assumed to be 100 Hz. This procedure allowed to generate band-limited sources mimicking actual kinematic or kinetic trajectories of time duration $T = 1$ s. We generated artificial mixture data by combining a number of sources ranging from 1 to 4. Combination coefficients of the mixing matrix **W** were generated from a uniform continuous distribution in the interval [−10,10]. Temporal delays τ_{ji} were drawn, when needed, from an exponential distribution of mean 20. Sets of noisy data were generated by corrupting noiseless data generated as described above with signal dependent noise. Noise was drawn from a Gaussian distribution of variance $\sigma = \alpha\, |x_i(t)|$, where α is the slope of the relationship between the standard deviation and the noiseless data values $x_i(t)$ (Sutton and Sykes, 1967; Schmidt et al., 1979; van Beers et al., 2004). The slope α was computed though an iterative procedure. Starting from $\alpha = 0$, its value was iteratively increased by a predefined increment until a desired noise-level $1 - R^2$ was reached and stayed constant for at least 10 consecutive computations of $1 - R^2$ given the same value of α. We define R^2 as follows: as the artificial noiseless data sets and their corresponding noisy versions are multivariate time-series, a measure of similarity (typically a ratio of two variances) must be defined using a multivariate measure of data variability. We used the "total variation" (Mardia et al., 1979), defined as the trace of the covariance of the signals, to define a multivariate measure as follows:

$$R^2 = 1 - \frac{\left\| \mathbf{X}_{noiseless} - \mathbf{X}_{noisy} \right\|^2}{\left\| \mathbf{X}_{noiseless} - \overline{\mathbf{X}}_{noiseless} \right\|^2} \tag{41}$$

where $\mathbf{X}_{\mathrm{noiseless}}$ is the matrix of the noiseless data set, $\mathbf{X}_{\mathrm{noisy}}$ the noisy data, and where $\overline{\mathbf{X}}_{\mathrm{noiseless}}$ is a matrix with the mean values of the noiseless data over trials. For each noiseless data set, two datasets were generated with $1 - R^2$ levels equal to 0.15 and 0.3, corresponding to approximate signal-to-noise ratios of 22 dB and 15 dB, respectively. We thus generated 2 models (AMM/SIM) x 2 cut-off frequencies (5 Hz/10 Hz) × 4 number of sources x 3 levels of noise = 48 different data sets. Each of those datasets contained $J \in \{5; 10; 25\}$ (data) trials. A "trial" (one row of the matrix **X** in Equation 4) is a one-dimensional time-series sampled at T points in time. For reliable averages, we drew 20 data sets for each number of trials.

2.3.2. Actual kinematic data

We applied the model selection criteria to select also the model of a second data set consisting of movement trajectories of human actors walking neutrally, or with different emotional styles (happy and sad). This data was originally recorded for the study presented in Roether et al. (2008). The movements were recorded using a Vicon (Oxford, UK) optoelectronic movement recording system with 10 infrared cameras, which recorded the three-dimensional positions of spherical reflective markers (2.5 cm diameter) with spatial error below 1.5 mm. The 41 markers were attached with double-sided adhesive tape to tight clothing, worn by the participants. Marker placement was defined by the Vicon's PlugInGait marker set. Commercial Vicon software was used to reconstruct and label the markers, and to interpolate short missing parts of the trajectories. Sampling rate was set at 120 Hz. We recorded trajectories from six actors, repeating each walking style three times per actor. A hierarchical kinematic body model (skeleton) with 17 joints was fitted to the marker positions, and joint angles were computed. Rotations between adjacent body segments were described as Euler angles, defining flexion, abduction and rotation about the connecting joints. The data for the BSS methods included only the flexion angles of the lower body joints, specifically right and left pelvis, hips, knees and ankles, since the other angles had relatively high noise levels. From each trajectory only one gait cycle was extracted, which was time normalized. This resulted in a data set with 432 samples with a length of 100 time points each. It was already shown previously (Omlor and Giese, 2007a,b) that an anechoic mixture model is more efficient than synchronous models for the representation of such kinematic data. To test the capability of the new LAP criterion to confirm such an observation we applied temporal shift to each trajectory of the data set. Each delay corresponding to a specific trajectory was drawn from a continuous uniform statistical distribution in the interval [−20,20], the sign of the delay determining the shift direction (forwards or backwards), signals were wrapped around at the boundaries of the 100 time-point interval.

3. RESULTS

We first present the evaluation of the three model selection criteria, LAP, BIC and AIC on the ground truth data described above in section 2.3.1. The evaluation is done with respect to three questions: how well can the generator type be detected (AMM or SIM), how accurate is the number of sources I estimation, and whether the amount of temporal smoothness [i.e., f_0 in Equation 26] can be determined. Second, we analyze the human gait data.

3.1. GROUND TRUTH EVALUATION

3.1.1. Model type detection

We measure the accuracy with which the generating model can be detected by the classification rate, averaged across generating and estimated number of sources, the estimated f_0 and the 20 data sets per condition. It is given by

$$\text{classification rate} = \frac{\text{number of correct detections}}{\text{total number of trials}} \tag{42}$$

The results are summarized in**Table 1** for each number of trials J. LAP clearly outperforms BIC and AIC, particularly for small J. To understand where this difference comes from, **Figure 3** shows a detailed analysis of the results for $J = 10$ trials. The anechoic generator is correctly detected by both LAP and BIC in most cases, whereas AIC often mistakes it for a pPCA model. The SIM generator, on the other hand, is only detected by LAP, both BIC and AIC mistake it for a pPCA model. Hence, LAP achieves very high classification rates, BIC is wrong about half the time, and AIC is even worse. This is due to the terms in BIC and AIC which punish complex models (second terms in Equation 39 and Equation 40, respectively): they only depend on the *number* of degrees of freedom and the number of data-points, but do not measure the effects of any "soft" constraints. Since such soft constraints will have a reducing effect on the likelihood, BIC and AIC will prefer models without such soft constraints over those with constraints. Consequently, BIC and AIC select pPCA over SIM. Note that in the limit of $f_0 \to \infty$, the kernel of the SIM model will give rise to a diagonal covariance matrix $\mathbf{K}$, and thus uncorrelated sources, whereas the $\mathbf{K}$ for finite f_0 will impose a correlational constraint. Thus, the SIM model will turn into a pPCA model in this limit.

In contrast, the LAP criterion measures the effect of the source correlations via the log-prior and log-posterior-volume terms. If the posterior is concentrated in a region of parameter space where the prior is high, the effects of the reduced likelihood can be counterbalanced. Since we evaluated the LAP criterion for $f_0 \in \{5\,\text{Hz}, 10\,\text{Hz}\}$, one of the tested SIM models will match the generator and have a correspondingly high LAP score.

3.1.2. Estimating the number of sources

Next, we looked at how well the criteria are suited for estimating the number of sources. **Table 2** shows the average difference between estimated and generating number of sources, averaged across noise levels, number of generating sources and f_0s. An empty cell indicates that this model would have been picked by the above model type detection only very infrequently.

Particularly for a small number of trials J, LAP is closer to the correct number of sources than BIC or AIC. For larger number of trials, the results between BIC and LAP become more similar, which is to be expected, even though BIC does not detect the correct model type. Moreover, the average number of sources estimated by LAP is always within one standard deviation of 0, and these standard deviations are mostly smaller than those of BIC and AIC.

Table 1 | Model type classification rates of the three tested criteria, for number of trials between 5 (top) and 25 (bottom).

$1 - R^2$	LAP	BIC	AIC
NUMBER OF TRIALS J: 5			
0.00	▶0.899 ± 0.017	0.003 ± 0.003	0.003 ± 0.003
0.15	▶0.944 ± 0.013	0.003 ± 0.003	0.003 ± 0.003
0.30	▶0.908 ± 0.016	0.003 ± 0.003	0.003 ± 0.003
NUMBER OF TRIALS J: 10			
0.00	▶0.947 ± 0.012	0.463 ± 0.028	0.214 ± 0.023
0.15	▶0.981 ± 0.008	0.494 ± 0.028	0.189 ± 0.022
0.30	▶0.972 ± 0.009	0.484 ± 0.028	0.342 ± 0.026
NUMBER OF TRIALS J: 25			
0.00	▶0.994 ± 0.004	0.500 ± 0.028	0.481 ± 0.028
0.15	▶0.978 ± 0.008	0.484 ± 0.028	0.388 ± 0.027
0.30	▶0.947 ± 0.012	0.469 ± 0.028	0.314 ± 0.026

$1 - R^2$ is the noise level from Equation 41.

*▶ indicates the best criterion for each row. LAP consistently outperforms BIC and AIC, mostly because the latter two are unable to distinguish between a smooth instantaneous mixture and a pPCA model (see also **Figure 3**). Furthermore, for 5 trials BIC and AIC tend mistake an anechoic mixture for an ICA model, leading to model type classification rates which are virtually zero.*

The dependency of the estimated number of sources on the noise level is depicted in **Figure 4** for $J = 10$. Also unsurprisingly, the estimated number of sources decreases with increasing noise level, since noisy data contain less information about the generating process.

3.1.3. Temporal smoothness constraints

In section 3.1.1, we showed that LAP is the only criterion which can detect the presence of temporal smoothness constraints. Now we investigate whether it can also identify the amount of smoothness, i.e., f_0 in Equation 26. To this end, we computed the LAP score for 16 smoothness settings: {1Hz, 2Hz, ..., 15Hz}, and also without smoothness constraint (i.e., effectively a pPCA model). We select the optimal smoothness setting for each dataset, and compute the average deviation to the generator smoothness (either 5 or 10 Hz) across all numbers of generating and estimated sources. The results are summarized in **Table 3** for all noise levels, **Figure 5** shows the detailed distributions for $J = 10$ trials. Except for the noiseless anechoic case, the correct temporal smoothness is found with average deviations near zero and standard deviations < 1.5Hz. We have as of yet no explanation for the overestimation in the noiseless anechoic case, but speculate that it is due to some jitter in the estimated delays of the anechoic model, which can be "explained away" by allowing for high-frequency components in the sources. As soon as noise is present in the data, this effect disappears.

3.2. HUMAN GAIT ANALYSIS

Having confirmed the validity of the LAP criterion on the synthetic ground truth, we now turn to real data. Since we are interested in smoothness properties as well as model types, we carried out a comparison between PCA and ICA; and SIM and AMM models with different $f_0 \in \{1\text{Hz}; \ldots; 12\text{Hz}\}$. The results

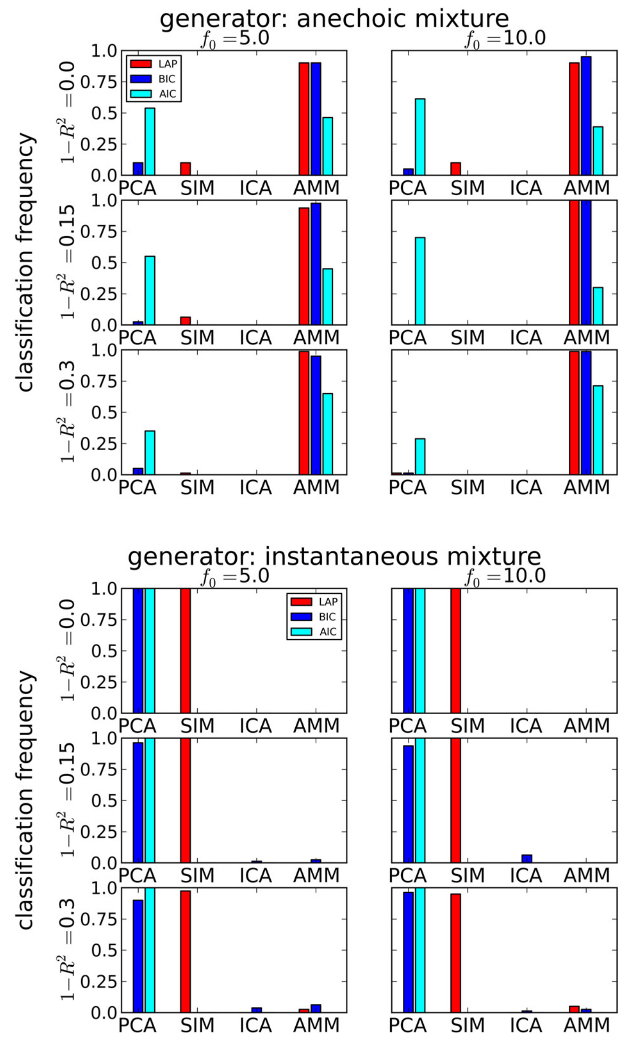

FIGURE 3 | Determining if the model type is a smooth instantaneous mixture (SIM) or an anechoic delayed mixture (AMM) from 10 trials, after marginalizing the estimated number of sources. Top: Anechoic ground truth. BIC and LAP perform comparably well, AIC often mistakes an AMM for a pPCA model. **Bottom**: Ground truth from smooth instantaneous mixture. Only the Laplace approximation criterion (LAP) correctly detects the SIM model in the majority of cases, BIC and AIC confuse SIM with pPCA due to their inability of handling soft constraints. For details, see text.

are summarized in **Figure 6**, top, where the simple “Anechoic” model (dark blue) is an AMM without smoothness constraints. As might be expected, AMMs are the best models (within our tested models) for this kind of data. Furthermore, a correctly chosen f_0 increases the LAP score significantly, i.e., the soft constraint provided by the smoothing kernel is an important feature of these kinematic data, see **Figure 6**, bottom. The best AMM has 3 sources, whereas the best SIM model needs 5, and has a lower score (see **Figure 6**, bottom).

LAP is an approximation of the marginal log-probability of the data [cf. Equation 3]. The best SIM model and the best AMM differ by a LAP score of ≈ 46, which translates into a probability ratio of $\frac{P(\text{AMM})}{P(\text{SIM})} > 10^{19}$. The best PCA model (5 sources) has a LAP score which is lower than the 7 Hz AMM score by ≈ 600.

Table 2 | Estimating the number of sources, marginalized across all noise levels and cutoff frequencies f_0.

Gen.	Anal.	LAP	BIC	AIC
NUMBER OF TRIALS *J*:5				
AMM	AMM	-0.37 ± 0.91		
AMM	ICA		1.17 ± 1.58	1.17 ± 1.58
AMM	PCA		2.15 ± 1.34	2.15 ± 1.34
AMM	SIM	0.94 ± 1.30		
SIM	ICA		0.93 ± 1.55	0.93 ± 1.55
SIM	PCA		1.88 ± 1.26	1.88 ± 1.26
SIM	SIM	-0.44 ± 0.76		
NUMBER OF TRIALS *J*:10				
AMM	AMM	▶0.19 ± 1.11	-0.60 ± 1.13	1.06 ± 1.44
AMM	PCA			3.44 ± 1.60
SIM	PCA		-0.48 ± 0.87	▶0.21 ± 0.86
SIM	SIM	-0.04 ± 0.52		0.21 ± 0.65
NUMBER OF TRIALS *J*:25				
AMM	AMM	▶0.32 ± 1.34	-0.99 ± 1.37	2.11 ± 1.72
AMM	PCA			5.31 ± 1.36
SIM	PCA		▶-0.18 ± 0.91	5.26 ± 1.25
SIM	SIM	0.69 ± 1.25		

Shown is the difference between the best number of sources determined with a given criterion (LAP, BIC, or AIC) and the number of sources in the generator. Gen. is the generating model, either anechoic (AMM) or smooth instantaneous (SIM) mixture. Anal. is the analysis algorithm. An empty cell indicates that a model comparison criterion (LAP, BIC, AIC) would have picked the corresponding analysis algorithm with a chance of less than 10% (cf. ***Figure 3*** *and* ***Table 1****). For rows with more than one entry,* ▶ *indicates best criterion.*

Note that individual datasets consisted of $J = 8$ trials (one per joint angle), therefore models with more than 8 sources are a priori too complex. This fact is also detected correctly by LAP, which assigns a roughly linearly decreasing score (exponentially decreasing in marginal probability) to models with ≥ 8 sources.

4. DISCUSSION

In this study, we attempted to develop a more objective probabilistic criterion for motor primitive model selection. Our criterion turned out to be more reliable than other already existing classical criteria (cf. sections 1.1 and 3) in selecting the generative model underlying a given data set, as well as in determining the corresponding dimensionality. The criterion can moreover provide accurate information about soft constraints, here the smoothness of the temporal evolution of the signals.

We tested LAP performance on synthesized, kinematic-like data and on actual motion capture trajectories. However, motor primitives have also been identified at the muscle level (Bizzi et al., 2008), where usually the signals are rectified after collection. As we tested LAP only on data with unconstrained signs, its applicability to positive-only data, such as EMG recordings, is a subject for further investigations.

The application of the criterion to emotional gait trajectories suggested the anechoic model as the most suitable description of the data. This result is in agreement with previous findings from our lab (Omlor and Giese, 2007b), where it was demonstrated that the anechoic model can represent emotional gait data more

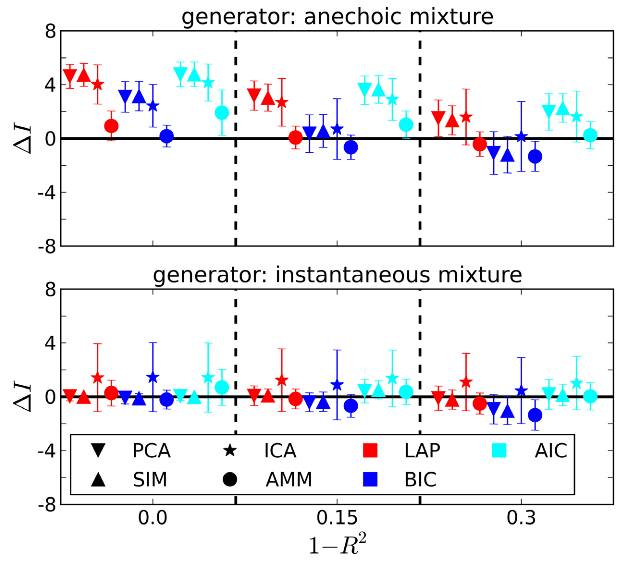

FIGURE 4 | Estimating the number of sources in the ground truth from $J = 10$ trials. ΔI =estimated-generating number of sources. $1 - R^2$: noise level from Equation 41. Error bars are ± one standard deviation. Symbol shapes stand for analysis algorithms, colors indicate the selection criterion. **Top panel**: Anechoic generator. If an AMM is used for analysis, BIC and LAP perform comparably well within the error bars, AIC tends to overestimate. For incorrect analysis models (SIM/ICA/PCA), all criteria overestimate the number of sources, since the extra variability provided by the time shifts needs to be explained via additional sources in instantaneous mixture models. Note, however, that a model type determination step based on BIC or LAP would have ruled out an instantaneous mixture with high probability (cf. **Figure 3** and **Table 2**). **Bottom panel:** For the instantaneous mixture (bottom panel) all three criteria give good results when using PCA or SIM models.

efficiently (in terms of data compression) than other classical synchronous models. Also the best number of primitives determined by LAP is in line with Omlor and Giese (2007b), where three components were found capable to explain about 97% of the total data variation. Interestingly, the criterion suggested a temporal smoothness regularization with $f_0 = 7$ Hz. Such a value may at first seem to be in contradiction with the step frequency of normal walking behavior that tends to be around 2 Hz (Pachi and Ji, 2005). The higher frequency value found by LAP can however be justified by multiple reasons. First, our data comprised also happy walks, which are known to be characterized by higher movement energy (Omlor and Giese, 2007b; Roether et al., 2009) and higher average movement velocity when compared to neutral or sad walks (Omlor and Giese, 2007a; Roether et al., 2009). Therefore, the average frequency power spectrum of the walking trajectories shows indeed considerable power within the band ranging from 0 to 10 Hz, with a peak at 5 Hz. In addition, in **Figure 6** the maximum LAP score occurs at $f_0 = 7$ Hz. However, taking the error bars into account, the LAP score associated with the optimal frequency is not statistically different from that associated with any score in the range [3 Hz, 10 Hz], in agreement with the power spectrum. Another factor contributing to $f_0 = 7$ Hz might be the tendency of LAP to overestimate the cutoff frequency slightly for nearly noise-free datasets, see **Figure 5**.

Table 3 | Mean temporal smoothness estimation accuracies and standard deviations for LAP criterion, marginalized across number of source of both generator and analysis model.

	$1 - R^2$		
Gen.	**0.00**	**0.15**	**0.30**
NUMBER OF TRIALS J:5			
SIM	0.526 ± 0.584	−0.171 ± 0.605	−0.526 ± 0.803
AMM	2.224 ± 1.793	0.137 ± 1.469	−0.553 ± 0.956
NUMBER OF TRIALS J:10			
SIM	0.694 ± 0.487	0.019 ± 0.518	−0.325 ± 0.638
AMM	1.762 ± 1.656	0.275 ± 1.475	−0.346 ± 1.441
NUMBER OF TRIALS J:25			
In	0.806 ± 0.494	0.275 ± 0.536	0.006 ± 0.553
An	2.150 ± 1.848	0.562 ± 1.288	0.106 ± 0.795

Estimation accuracy is given by best estimated f_0 (see Equation 26) as determined by LAP minus actual cutoff frequency (either 5 or 10 Hz). Gen. is the generative model: anechoic (AMM) or smooth instantaneous mixture (SIM). $1 - R^2$ is the noise level (Equation 41). Except for the zero-noise anechoic generator, f_0 can be determined to within 1 Hz of its true value. For details, see text.

Additional and more advanced models of motor primitives, corresponding to a multivariate version of the anechoic mixture model considered in this study, have been developed (d'Avella et al., 2003, 2006) to describe the modular organization associated with EMG data sets. As we have not computed the LAP for these models, they are not among the possible model selection options yet. Future work will therefore aim to formulate the priors and generative models which would allow for the application of LAP to EMG data.

An interesting feature of LAP is its capability to discriminate between instantaneous vs. anechoic mixtures. The importance of introducing temporal delays in the model of a motor behavior has revealed to be crucial in some cases such as, for instance, in the modeling of emotional movements or facial expressions (Roether et al., 2008; Giese et al., 2012). LAP is to our knowledge the first model selection criterion explicitly designed for this.

Another remarkable feature of LAP is its capability, thanks to the addition a smoothness prior, to identify the amount of smoothness in the data, in other words to select the frequency f_0 in Equation 26 based on the available data. While a Fourier analysis would also reveal where the power spectrum drops off, LAP has the advantage of providing a principled, quantitative trade-off between smoothness and goodness-of-fit, which allows for a more objective selection of f_0. However, computing the power spectrum could be a first step to determine the range of f_0s across which to search for the optimum.

Moreover, incorporating smoothness priors in time and/or space might be a viable extension of LAP to make it suitable to distinguish between a low-dimensional generative model based on time-invariant primitives vs. a model based on space-invariant primitives. Muscle synergies, for instance, have indeed been presented in the literature in those terms. Among

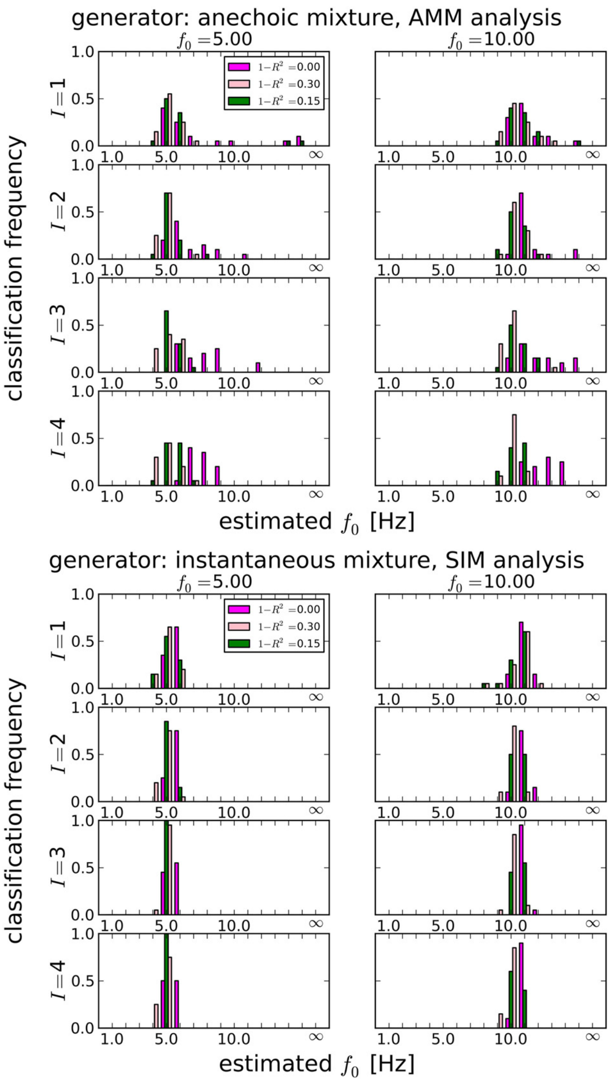

FIGURE 5 | Estimating the best temporal smoothness regularization cutoff frequency f_0 (Equation 26) for $J = 10$ trials. Results obtained with LAP criterion. This estimation can not be done with AIC or BIC: temporal smoothness, while it reduces the effective degrees of freedom (DF), cannot be expressed in BIC or AIC, because these criteria need an integer number for the DF, which would not change with a continuous regularization like smoothness. We tested 16 smoothness settings: {1Hz, 2Hz, ..., 15Hz} and no smoothness constraint, indicated by "∞" in the plots (this is equivalent to a pPCA model). Left column: 5 Hz ground truth, right column: 10 Hz ground truth. Estimating the smoothness works well for both anechoic **(Top)** and instantaneous **(Bottom)** mixtures, except for the zero-noise anechoic case, where f_0 is overestimated by ≈ 2 Hz on average.

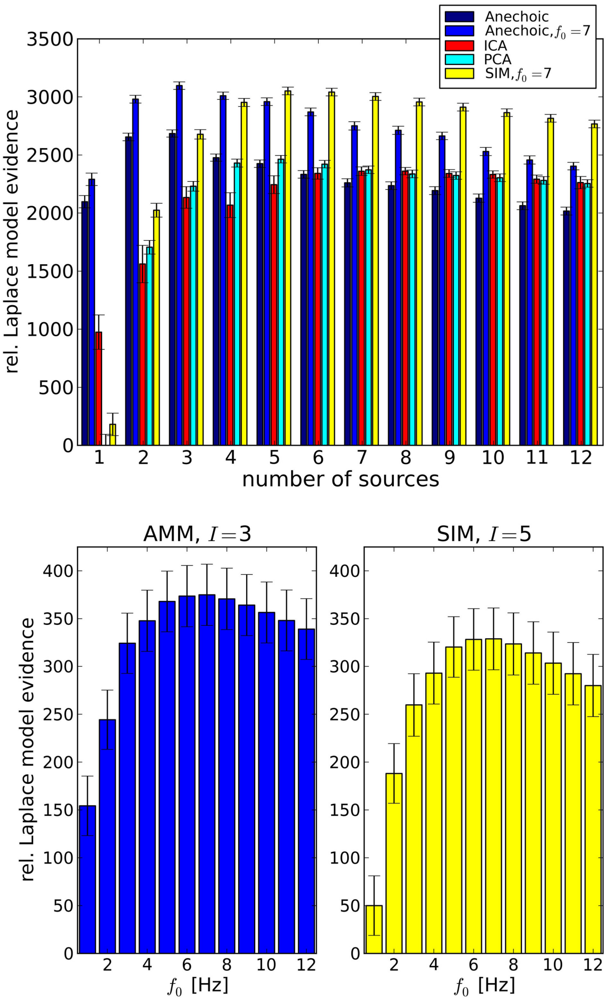

FIGURE 6 | Top: Analysis of emotional gait data from Roether et al. (2008) with PCA, ICA and Anechoic demixing for different numbers of sources. The bars represent model evidences computed with Laplace approximation, relative to the lowest observed model evidence (PCA, 1 source). The anechoic analyses were carried out either without smoothing (black, $f_0 \rightarrow \infty$) or with the optimal $f_0 = 7$ Hz (blue) for the wave kernel (see Equation 26). Error bars are standard errors, computed across trials. The best model (highest evidence) is the anechoic mixture with three sources and $f_0 = 7$ Hz, followed by the SIM model with $f_0 = 7$ Hz. PCA and ICA are significantly worse for any number of sources. **Bottom**: Detailed cutoff frequency analysis of the AMM model (left) and the SIM model (right), at their respective best number of sources I. The LAP score (relative to the SIM model with $I = 1$ set to 50) for both models peaks at $f_0 = 7$Hz. However, the best AMM model's approximate posterior probability is larger than the best SIM posterior by a factor of $\approx 10^{19}$. For details, see text.

them, "synchronous" synergies (Cheung et al., 2005; Ting and Macpherson, 2005; Torres-Oviedo et al., 2006) have been described as stereotyped co-varying groups of muscles activations, with the EMG output specified by a temporal profile determining the timing of each synergy during task accomplishment. This definition of synergies reflects the idea of invariance across space (namely the space spanned by the muscles) mentioned above. "Temporal" synergies (Ivanenko et al., 2004, 2005; Chiovetto et al., 2010, 2012), are instead defined as temporal activation profiles that can be simply linearly combined together to reconstruct the actual activity of each muscle. Such a definition of synergies is therefore incorporating a notion of invariance across time. Also more "hybrid" definition of primitives have been given. "Time-varying" synergies (d'Avella et al., 2003, 2006), for instance, are defined as spatio-temporal pattern of muscle activations, with corresponding EMG output determined by the scaling coefficients and time delays associated with each synergy. Chiovetto et al. (2013) already showed heuristically what movement features these definitions of synergies are describing. Although this knowledge can surely help to decide, dependent on the kind of analysis that one needs to carry out, which kind of synergies to extract from a given EMG data set, it however, does not provide a systematic criterion for such a decision. An extension of LAP might help here, too.

To apply LAP to a given source extraction method, it is necessary to (re)formulate this method in the language of generative probabilistic models. Only when the joint probability of the data and all latent variables (such as **W** or **S**) is available can Equation 38 be evaluated. Furthermore, since LAP results from a second-order approximation to the exponent of that joint probability, LAP will only yield (approximately) correct answers if such an approximation is valid. While the possibility of reformulating a given method can usually be decided *a-priori*, the validity of the second-order approximation typically needs testing on ground-truth data.

In conclusion, we presented an innovative and objective criterion that can be used to reliably select an adequate factorization model to explain the variance associated with kinematic/dynamic data and its corresponding dimensionality. We showed LAP to perform better than two plug-in estimators, BIC and AIC. It needs, however, to be extended to be used in the future for additional types of data, such as EMG data.

ACKNOWLEDGMENTS

This work was supported by the EU Commission, 7th Framework Programme: EC FP7-ICT-249858 TANGO, EC FP7-ICT-248311 AMARSi, FP7-ICT-2013-10/611909 Koroibot, Deutsche Forschungsgemeinschaft: DFG GI 305/4-1, DFG GZ: KA 1258/15-1, German Federal Ministry of Education and Research: BMBF, FKZ: 01GQ1002A, European Commission, Fp 7-PEOPLE-2011-ITN(Marie Curie): ABC PITN-GA-011-290011, The FP7 Human Brain Flagship Project. All graphics were prepared with matplotlib (Hunter, 2007). We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of Tübingen University.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Cont.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* 52, 317–332. doi: 10.1007/BF02294359

Basilevsky, A. T. (1994). *Statistical Factor Analysis and Related Methods: Theory and Applications*. New York, NY: Wiley. doi: 10.1002/9780470316894

Bell, A. J., and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 7, 1129–1159. doi: 10.1162/neco.1995.7.6.1129

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Bizzi, E., Cheung, V. C. K., D'Avella, A., Saltiel, P., and Tresch, M. (2008). Combining modules for movement. *Brain Res. Rev.* 57, 125–133. doi: 10.1016/j.brainresrev.2007.08.004

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* 33, 261–304. doi: 10.1177/0049124104268644

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behav. Res.* 1, 245–276. doi: 10.1207/s15327906mbr0102_10

Cheung, V. C. K., D'Avella, A., Tresch, M. C., and Bizzi, E. (2005). Central and sensory contributions to the activation and organization of muscle synergies during natural motor behaviors. *J. Neurosci.* 25, 6419–6434. doi: 10.1523/JNEUROSCI.4904-04.2005

Cheung, Y. M., and Xu, L. (1999). "An empirical method to select dominant independent components in ICA for time series analysis," in *Proceedings of 1999 International Joint Conference on Neural Networks (IJCNN'99)*, (Washington DC), 3883–3887.

Chiovetto, E., Berret, B., Delis, I., Panzeri, S., and Pozzo, T. (2013). Investigating reduction of dimensionality during single-joint elbow movements: a case study on muscle synergies. *Front. Comput. Neurosci.* 7:11. doi: 10.3389/fncom.2013.00011

Chiovetto, E., Berret, B., and Pozzo, T. (2010). Tri-dimensional and triphasic muscle organization of whole-body pointing movements. *Neuroscience* 170, 1223–1238. doi: 10.1016/j.neuroscience.2010.07.006

Chiovetto, E., Patanè, L., and Pozzo, T. (2012). Variant and invariant features characterizing natural and reverse whole-body pointing movements. *Exp. Brain Res.* 218, 419–431. doi: 10.1007/s00221-012-3030-y

Cover, T., and Thomas, J. (1991). *Elements of Information Theory*. New York, NY: Wiley. doi: 10.1002/0471200611

d'Avella, A., Portone, A., Fernandez, L., and Lacquaniti, F. (2006). Control of fast-reaching movements by muscle synergy combinations. *J. Neurosci.* 26, 7791–7810. doi: 10.1523/JNEUROSCI.0830-06.2006

d'Avella, A., Saltiel, P., and Bizzi, E. (2003). Combinations of muscle synergies in the construction of a natural motor behavior. *Nat. Neurosci.* 6, 300–308. doi: 10.1038/nn1010

d'Avella, A., and Tresch, M. C. (2002). "Modularity in the motor system: decomposition of muscle patterns as combinations of time-varying synergies," in *Advances in Neural Information Processing Systems 14*, eds M. I. Jordan, M. J. Kearns and S. A. Solla (Cambridge, MA: MIT Press), 141–148.

Delis, I., Berret, B., Pozzo, T., and Panzeri, S. (2013). Quantitative evaluation of muscle synergy models: a single-trial task decoding approach. *Front. Comput. Neurosci.* 7:8. doi: 10.3389/fncom.2013.00008

Dominici, N., Ivanenko, Y. P., Cappellini, G., d'Avella, A., Mondì, V., Cicchese, M., et al. (2011). Locomotor primitives in newborn babies and their development. *Science* 334, 997–999. doi: 10.1126/science.1210617

Dyrholm, M., Makeig, S., and Hansen, L. K. (2007). Model selection for convolutive ICA with an application to spatio-temporal analysis of eeg. *Neural Comput.* 19, 934–955. doi: 10.1162/neco.2007.19.4.934

Flash, T., and Hochner, B. (2005). Motor primitives in vertebrates and invertebrates. *Curr. Opin. Neurobiol.* 15, 660–666. doi: 10.1016/j.conb.2005.10.011

Genton, M. G. (2001). Classes of kernels for machine learning: a statistics perspective. *J. Mach. Learn. Res.* 2, 299–312. Available online at: http://www.crossref.org/jmlr_DOI.html

Giese, M., Chiovetto, E., and Curio, C. (2012). Perceptual relevance of kinematic components of facial movements extracted by unsupervised learning. *Perception* 41:150. doi: 10.1068/v120635

Harris, C. M., and Wolpert, D. M. (1998). Signal-dependent noise determines motor planning. *Nature* 394, 780–784. doi: 10.1038/29528

Hart, C. B., and Giszter, S. (2013). Distinguishing synchronous and time varying synergies using point process interval statistics: motor primitives in frog and rat. *Front. Comput. Neurosci.* 7:52. doi: 10.3389/fncom.2013.00052

Hunter, J. D. (2007). Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* 9. 90–95. doi: 10.1109/MCSE.2007.55

Hyvärinen, A. (1998). "New approximations of differential entropy for independent component analysis and projection pursuit," in *Advances in Neural Information Processing Systems 10*, eds M. I. Jordan, M. J. Kearns, and S. A. Solla (Cambridge, MA: MIT Press), 273–279.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* 10, 626–634. doi: 10.1109/72.761722

Ichir, M. M., and Mohammad-Djafari, A. (2005). "Determination of the number of sources in blind source separation," in *25th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conf. Proc. 803*, (San Jose, CA: American Institute of Physics), 266–273.

Ivanenko, Y. P., Cappellini, G., Dominici, N., Poppele, R. E., and Lacquaniti, F. (2005). Coordination of locomotion with voluntary movements in humans. *J. Neurosci.* 25, 7238–7253. doi: 10.1523/JNEUROSCI.1327-05.2005

Ivanenko, Y. P., Poppele, R. E., and Lacquaniti, F. (2004). Five basic muscle activation patterns account for muscle activity during human locomotion. *J. Physiol.* 556(Pt 1), 267–282. doi: 10.1113/jphysiol.2003.057174

Jones, E., Oliphant, T., Peterson, P., et al. (2001). *SciPy: Open source scientific tools for Python.* Available online at: http://www.scipy.org/scipylib/citing.html

Kaminski, T. R. (2007). The coupling between upper and lower extremity synergies during whole body reaching. *Gait Posture* 26, 256–262. doi: 10.1016/j.gaitpost.2006.09.006

Kazianka, H., and Pilz, J. (2009). A corrected criterion for selecting the optimum number of principal components. *Aust. J. Stat.* 38, 135–150.

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événemens. *Savants étranges* 6, 621–656.

Li, Y., Adali, T., and Calhoun, V. (2007). Estimating the number of independent components for functional magnetic resonance imaging data. *Hum. Brain Mapp.* 28, 1251–1266. doi: 10.1002/hbm.20359

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* London, UK: Academic Press.

MATLAB. (2010). *MATLAB version 7.10.0 (R2010a)*. Natick, MA: The MathWorks Inc.

Minka, T. (2000). *Automatic Choice of Dimensionality for pca.* Cambdrige, MA: Technical report, M.I.T. Media Laboratory Perceptual Computing Section.

Omlor, L., and Giese, M. (2007a). "Blind source separation for over-determined delayed mixtures," in *Advances in Neural Information Processing Systems 19*, eds B. Schölkopf, J. Platt, and T. Hoffman (Cambridge, MA: MIT Press), 1049–1056.

Omlor, L., and Giese, M. A. (2007b). Extraction of spatio-temporal primitives of emotional body expressions. *Neurocomputing* 70, 1938–1942. doi: 10.1016/j.neucom.2006.10.100

Omlor, L., and Giese, M. A. (2011). Anechoic blind source separation using wigner marginals. *J. Mach. Learn. Res.* 12, 1111–1148. Available online at: http://www.crossref.org/jmlr_DOI.html

Pachi, A., and Ji, T. (2005). Frequency and velocity of people walking. *Struct. Eng.* 83, 36–40.

Penny, W., and Roberts, S. (2001). "ICA: Model order selection and dynamic source models," in *Independent Component Analysis*, eds S. Roberts and R. Everson (Cambridge, UK: Cambridge University Press), 299–314.

Plant, C., Theis, F. J., Meyer-Baese, A., and Böhm, C. (2010). "Information-theoretic model selection for independent components," in *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation, LVA/ICA'10,* (Berlin, Heidelberg: Springer-Verlag), 254–262.

Rasmussen, C. E., and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Reif, F. (1995). *Fundamentals of Statistical and Thermal Physics.* New York, NY: McGraw-Hill.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica* 14, 465–471. doi: 10.1016/0005-1098(78)90005-5

Roether, C., Omlor, L., and Giese, M. A. (2008). Lateral asymmetry of bodily emotion expression. *Curr. Biol.* 18, R329–R330. doi: 10.1016/j.cub.2008.02.044

Roether, C. L., Omlor, L., Christensen, A., and Giese, M. A. (2009). Critical features for the perception of emotion from gait. *J. Vis.* 9, 15.1–1532. doi: 10.1167/9.6.15

Santello, M., Flanders, M., and Soechting, J. F. (1998). Postural hand synergies for tool use. *J. Neurosci.* 18, 10105–10115.

Sawada, H., Mukai, R., Araki, S., and Makino, S. (2005). Estimating the number of sources using independent component analysis. *Acoust. Sci. Technol.* 26, 450–452. doi: 10.1250/ast.26.450

Schmidt, R. A., Zelaznik, H., Hawkins, B., Frank, J. S., and Quinn, J. Jr. (1979). Motor-output variability: a theory for the accuracy of rapid motor acts. *Psychol. Rev.* 47, 415–451. doi: 10.1037/0033-295X.86.5.415

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sutton, G. G., and Sykes, K. (1967). The variation of hand tremor with force in healthy subjects. *J. Physiol.* 191, 699–711.

Ting, L. H., and Macpherson, J. M. (2005). A limited set of muscle synergies for force control during a postural task. *J. Neurophysiol.* 93, 609–613. doi: 10.1152/jn.00681.2004

Tipping, M. E., and Bishop, C. M. (1999). Probabilistic principal component analysis. *J. R. Stat. Soc. B* 61, 611–622. doi: 10.1111/1467-9868.00196

Torres-Oviedo, G., Macpherson, J. M., and Ting, L. H. (2006). Muscle synergy organization is robust across a variety of postural perturbations. *J. Neurophysiol.* 96, 1530–1546. doi: 10.1152/jn.00810.2005

Tresch, M. C., Cheung, V. C. K., and d'Avella, A. (2006). Matrix factorization algorithms for the identification of muscle synergies: evaluation on simulated and experimental data sets. *J. Neurophysiol.* 95, 2199–2212. doi: 10.1152/jn.00222.2005

Tu, S., and Xu, L. (2011). An investigation of several typical model selection criteria for detecting the number of signals. *Front. Electr. Electron. Eng. China* 6, 245–255. doi: 10.1007/s11460-011-0146-y

Valle, S., Li, W., and Qin, S. (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Ind. Eng. Chem. Res.* 38, 4389–4401. doi: 10.1021/ie990110i

van Beers, R. J., Haggard, P., and Wolpert, D. M. (2004). The role of execution noise in movement variability. *J. Neurophysiol.* 91, 1050–1063. doi: 10.1152/jn.00652.2003

Xu, L. (2007). Bayesian ying yang learning. *Scholarpedia* 2, 1809. doi: 10.4249/scholarpedia.1809

Zucchini, W. (2000). An introduction to model selection. *J. Math. Psychol.* 44, 41–61. doi: 10.1006/jmps.1999.1276

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 September 2013; paper pending published: 18 November 2013; accepted: 05 December 2013; published online: 20 December 2013.

Citation: Endres DM, Chiovetto E and Giese MA (2013) Model selection for the extraction of movement primitives. Front. Comput. Neurosci. **7***:185. doi: 10.3389/fncom.2013.00185*

This article was submitted to the journal Frontiers in Computational Neuroscience.