Supplementary Note on Bayesian analysis

Structured variability of muscle activations supports the minimal intervention principle of motor control

Francisco J. Valero-Cuevas [1,2,3,*] Madhusudhan Venkadesan [3,4,†]
Emanuel Todorov [5]

[1] *Department of Biomedical Engineering, University of Southern California, Los Angeles, CA*
[2] *Division of Biokinesiology & Physical Therapy, University of Southern California, Los Angeles, CA*
[3] *Sibley School of Mechanical & Aerospace Engineering, Cornell University, Ithaca, NY*
[4] *Department of Mathematics, Cornell University, Ithaca, NY*
[5] *Department of Cognitive Science, University of California San Diego, La Jolla, CA*

1 Introduction

In the main text of this manuscript, we estimated the linear mapping from recorded electromyograms (EMGs) to fingertip force outputs using a least squares algorithm (i.e., linear regression). Throughout this supplement and the main text we denote this mapping by A, a 3×7 matrix that transforms muscle EMGs into fingertip forces. Subsequent analysis of variance in muscle EMGs was carried out using this matrix A. Specifically, the row-space and null-space of A were used to parse the variability in EMGs into task-relevant and task-irrelevant portions. However, estimating A using a least squares process could be biased. In particular if the EMGs for some muscles are more noisy than others, the corresponding entries in A could be smaller than the actual moment arms, which in turn could affect our conclusions regarding variability structure. Yet another reason for bias in using a least squares estimate of A arises from the fact that a least-squares procedure specifically tries to find A that minimizes errors in estimated fingertip forces, i.e., noise in EMGs may be selectively dumped into the null-space of A. However, it is possible to work around these forms of bias in estimation by not only estimating the best-fit A, but finding a probability distribution for A. Here, we use Bayesian analysis to estimate the probability distribution for A using a numerical sampling scheme calling the Metropolis algorithm. This is in contrast to the point-estimate we found in the main text. We find that our conclusions about the structure of variability in muscle activations remain the same as when using a point-estimate of A. Therefore, we conclude that the least-squares estimation procedure presented in the main text did not bias the results of this study. We relegated the Bayesian analysis to this supplementary note for ease of reading the main text given greater familiarity with linear regression than with Bayesian techniques for several readers.

2 Methods

We pool together all trials for a given subject, drop the trial index, and also subtract the mean so that the offset $\boldsymbol{b}$ introduced earlier (in the main text) is

*Corresponding author
†Currently at School of Engineering & Applied Sciences, Harvard University, Cambridge, MA, USA

no longer needed. The procedure described below is repeated separately for every subject.

2.1 Bayesian Estimation

Instead of taking the data at face value, we consider the force ($\boldsymbol{f}(t)$) and measured EMG ($\boldsymbol{e}(t)$) as noisy versions of some underlying signals which correspond to the true fingertip force ($\bar{\boldsymbol{f}}(t)$)and true muscle activation ($\bar{\boldsymbol{e}}(t)$):

$$\boldsymbol{f}(t) = \bar{\boldsymbol{f}}(t) + \boldsymbol{\gamma}(t) \tag{1a}$$
$$\boldsymbol{e}(t) = \bar{\boldsymbol{e}}(t) + \boldsymbol{\omega}(t) \tag{1b}$$

where the noise terms $\boldsymbol{\gamma}(t)$ and $\boldsymbol{\omega}(t)$ are independent, zero-mean, multivariate, Gaussian white noise variables with diagonal covariance matrices Q and R. These two matrices along with the matrix A constitute the model parameters which we will estimate from the data. Note that A now relates the underlying signals ($\bar{\boldsymbol{f}}$, $\bar{\boldsymbol{e}}$) and not the measured signals ($\boldsymbol{f}$, $\boldsymbol{e}$):

$$\bar{\boldsymbol{f}}(t) = \mathrm{A}\bar{\boldsymbol{e}}(t) \tag{2}$$

We do not know the underlying signals, but fortunately it is not necessary to know them. Indeed, multiplying the two noise models (Equation 1) by A and subtracting yields:

$$\mathrm{A}\boldsymbol{e}(t) - \boldsymbol{f}(t) = \mathrm{A}\boldsymbol{\omega}(t) - \boldsymbol{\gamma}(t) \tag{3}$$

The expression on the left (i.e., the residual) can be computed from the data for an assumed (guessed) A. The expression on the right however, is a Gaussian white random variable with zero mean and covariance matrix $\mathrm{ARA}^{\mathrm{T}} + \mathrm{Q}$. Thus using the formula for the probability density function of a multivariate normal distribution, the likelihood of the measured residual at time t for given (or guessed) model parameters A, R, Q is

$$p_t(\boldsymbol{\theta}) = \frac{e^{\left(-\frac{1}{2}(\mathrm{A}\boldsymbol{e}(t)-\boldsymbol{f}(t))^{\mathrm{T}}(\mathrm{ARA}^{\mathrm{T}}+\mathrm{Q})^{-1}(\mathrm{A}\boldsymbol{e}(t)-\boldsymbol{f}(t))\right)}}{(2\pi)^{\frac{3}{2}}\sqrt{|\mathrm{ARA}^{\mathrm{T}}+\mathrm{Q}|}} \tag{4}$$

where $\boldsymbol{\theta}$, the parameters to be estimated using the data, are simply the elements of A, Q and R strung out as a long vector. Because, A has 21 elements (3×7), Q has 3 elements (a 3×3 diagonal matrix), and R has 7 elements (a 7×7 diagonal matrix), the total number of parameters to be estimated (i.e., size of $\boldsymbol{\theta}$) is 31. Given that our data are in the form of discrete measurements, the total likelihood of the measured data time-series is simply the product over all t of the likelihood given by Equation 4:

$$p(\boldsymbol{\theta}) = \prod_t p_t(\boldsymbol{\theta}) \tag{5}$$

2.1.1 Numerical sampling of the posterior distribution

In the absence of a prior probability distribution[1] for $\boldsymbol{\theta}$ (i.e., a uniform prior), the posterior distribution[2] for $\boldsymbol{\theta}$ (the product of $p(\boldsymbol{\theta})$ and the uniform prior) is then proportional to the likelihood ($p(\boldsymbol{\theta})$). The key idea in Bayesian analyses is to use the posterior distribution instead of a single parameter estimate. When this distribution is too complex to handle analytically, as is the case here, one can use Markov Chain Monte Carlo (MCMC) sampling. The specific algorithm we use is the Metropolis algorithm, which uses the following iterative scheme to generate sample the posterior distribution. Let $\boldsymbol{\theta}^{\mathrm{k}}$ denote the k^{th} parameter sample. Generate a candidate new sample $\boldsymbol{\theta}' = \boldsymbol{\theta}^{\mathrm{k}} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is drawn from a zero-mean multivariate Gaussian distribution. Then, we compute the probability ratio:

$$a = \frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta})} \tag{6}$$

where p is given by Equation 5 and is proportional to the target distribution (posterior) from which we want to sample. If $a \geq 1$, make the deterministic update $\boldsymbol{\theta}^{\mathrm{k+1}} = \boldsymbol{\theta}'$. Otherwise make the stochastic

[1] The prior probability distribution or simply, the *prior*, is the name given to the uncertainty in the variable we want to estimate before any data are taken into account.

[2] The posterior probability distribution or simply, the *posterior*, is the name given to the conditional probability distribution of the variable we want to estimate after taking into account the data and a known prior. According to Bayes' theorem, this is simply the product of the prior and the likelihood of observing the measured data.

update

$$\boldsymbol{\theta}^{\mathrm{k+1}} = \begin{cases} \boldsymbol{\theta}' & \text{with probability } a \\ \boldsymbol{\theta}^{\mathrm{k}} & \text{with probability } 1-a \end{cases} \tag{7}$$

After an initial "burn-in" period, the sequence of samples generated in this way is guaranteed to match the target distribution. To ensure that the Markov chain has enough time to converge we allowed a large number of updates (10 million), and furthermore repeated the entire process 3 times with different initial values ($\boldsymbol{\theta}^0$). Visual inspection confirmed that for 7 out of the 8 subjects, the chain always converged in less than 4 million updates. Thus we discarded the first half of each chain. To speed up processing we kept only 1 out of each 100 consecutive samples – which is motivated by the fact that MCMC generates samples that are correlated over time. Combining the results from the 3 chains, we thus analyzed 150,000 samples from the posterior distribution over the parameters for each subject. For one of the 8 subjects we did not observe convergence. Furthermore for this subject the matrix A estimated with linear regression had substantially larger elements compared to the other subjects. Thus we decided to exclude this subject from the present analysis, even though this subject's (unreliable) results were actually consistent with our hypothesis. This left us with a total of 41 trials from 7 subjects.

2.2 Variability ratio

We briefly recall how we quantified the ratio of task-irrelevant to task-relevant variability in EMGs. A detailed description is found in the main text. From the trial EMG data, we calculated three covariance matrices: Λ – the full covariance matrix, D – the diagonal covariance matrix, with diagonal elements identical to Λ and all off-diagonal elements set to zero, and finally S – the diagonal signal dependent noise covariance matrix such that the diagonal terms were proportional to the mean muscle activations squared. For a given A, the three rows of A span the task-relevant subspace of A and the four basis vectors of the nullspace of A span the task-irrelevant subspace of A. For every basis vector $\hat{\boldsymbol{u}}$ and covariance matrix V (where V is one of Λ, D, or S), the projected variance is given by the scalar quantity, $\hat{\boldsymbol{u}}^{\mathrm{T}}\mathrm{V}\hat{\boldsymbol{u}}$. Then, the variability index for each of the task-relevant and -irrelevant subspaces is just the average of the variance projected onto each of their respective basis vectors. The ratio of the variability index of the task-relevant to that of the task-irrelevant subspace is the quantity of interest (we call it the *variability ratio*). In the main text we found that for a least squares estimation of A, this ratio was smallest for Λ and specifically, it was smaller than 1 (both results were statistically significant). Here we repeat the same analysis using the 150,000 samples of A generated by the Metropolis algorithm. In addition to the average value of this ratio, the Bayesian method yields the posterior distribution of the variability ratio for a single trial, thus enabling single-trial hypothesis testing.

3 Results and discussion

We found that samples of the posterior distributions of A, R and Q, and therefore, the posterior distribution of the variability ratio, all converged. In other words, all three chains of the MCMC led to very similar distributions for the estimated mapping from EMGs to forces (A) (see Figure 1a) and the variability index (see Figure 1c). The estimated noise variances (R for EMG and Q for force) however, were typically non-Gaussian and differed to a greater degree between chains (see Figure 1b). This is not surprising because variances are typically harder to estimate. Finally, we found the linear regression estimates to be surprisingly close to the Bayesian estimates.

Using single-trial hypothesis testing we found that for 30 out of 41 trials, the variability ratio was statistically significantly smaller than 1 (filled, blue data points in Figure 2a). Recall that our hypothesis is that the variability index is below 1. For each trial, the Bayesian method gives us a sample from the posterior distribution of the variability ratio. Thus, hypothesis testing is performed by simply counting how many samples agree with the hypothesis and checking that the percentage is above a desired significance level (say 0.95). Of the 41 trials analyzed, in 30 cases

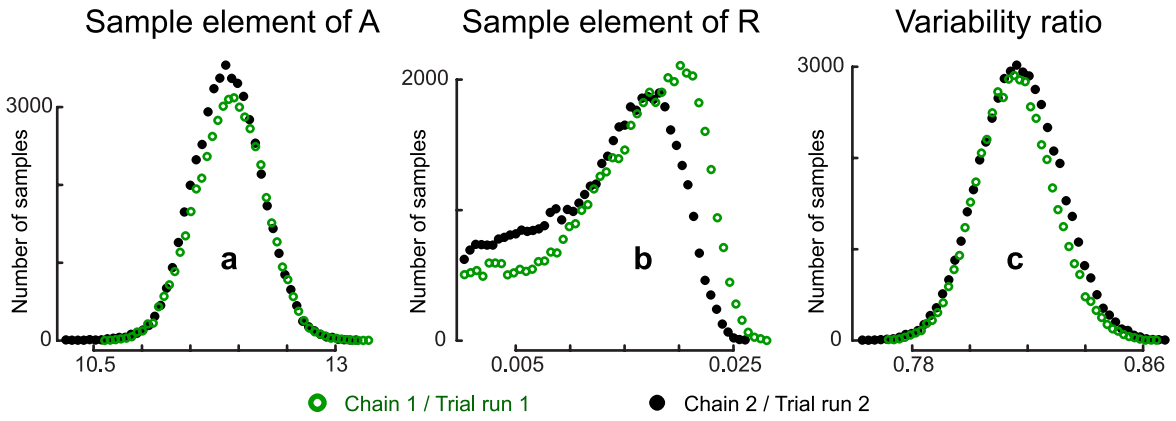

Figure 1: Examples of posterior distributions of A, R (EMG variance), and the calculated variability ratio. (**a**) Histograms generated using 50,000 samples (number of samples per chain) of one element of A for one subject and two different chains (different initial values) of the Metropolis algorithm. Note that the two distributions are almost identical and near-Gaussian. This was typical for all 21 elements of A. For this element, the linear regression estimate was 11.4, which is close to the mean of the histograms (11.9). (**b**) Histograms from two different chains for one of the EMG noise magnitudes (expressed as standard deviation rather than variance) for one subject, one trial. Here the distributions are no longer Gaussian and tend to differ more between chains. (**c**) Histograms of the variability ratio for one trial in one subject for two different chains of the sampling algorithm. Again we have similar and near-Gaussian shapes. For this trial/subject, the entire posterior distribution lies below 1, supporting our hypothesis about task-relevant:task-irrelevant variability.

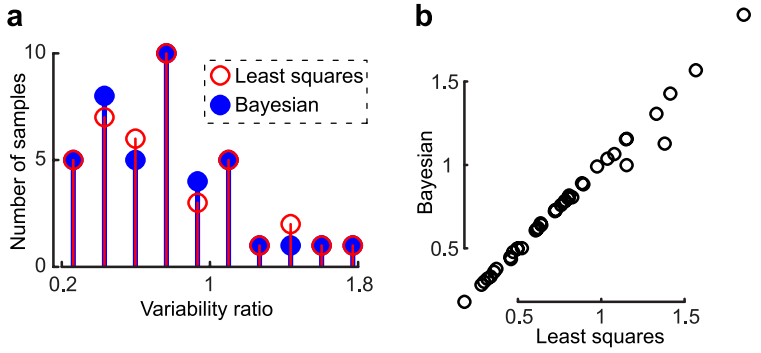

Figure 2: Variability ratios for Bayesian vs. least squares estimates of A. (**a**) Histogram of the variability index from 41 trials for the Bayesian (filled-in, blue circles) and the least squares (open, red circles) estimates. In case of the Bayesian estimates, the histogram represents the mean of the posterior distribution of the variability ratio. (**b**) Scatter plot comparing the variability ratio found using least squares vs. that found using the Bayesian method.

the index was significantly smaller than 1, in 9 cases it was significantly larger than 1, and in the remaining 2 cases the result was not significant (meaning that the distribution was centered near 1 and the spread was too large to reach significance in either direction). Note that compared to standard statistical tests, Bayesian tests such as this are much more intuitive as well as accurate. See Mackay (2003) for a discussion of sampling and Bayesian hypothesis testing.

Figure 2b shows a scatter plot comparing the variability ratio computed using least squares and the Bayesian method. Note the almost perfect agreement. The mean was 0.775 for the least squares method and 0.765 for the Bayesian method. This rules out the potential confounds of biased least-squares estimation. To be sure the estimated noise magnitudes differed, but these differences did not interact with our hypothesis. The average (RMS) noise magnitude was 3% MVC for EMGs and 0.18 N for fingertip forces.

References

D. Mackay. *Information theory, inference and learning algorithms*. Cambridge University Press, Cambridge, U.K., 2003. URL http://www.inference.phy.cam.ac.uk/mackay/itprnn/book.pdf.