Hi, REG folks,

the following is a brief rationale of my strategy of global assessment of 'multichannel data', i. e. data recorded simultaneously from multiple spatially distributed sites but measured on the same physical scale. The background for the strategy presented here was the need to condense information from several seconds of multichannel EEG data to few physically meaningful state variables. A copy of my original article where the strategy was presented is enclosed; note, however, that there's a lot of garbage not worth of reading (I had to explain to general audience the meaning of simple mathematical operations, so the text is a compromise between a tutorial and a research paper). The following information should be enough to implement computation of the global descriptors.

Assume we have a series of N vectors, each consisting of simultaneous measurements at K sites (e. g. electrodes on the scalp or REGs on the Earth surface): $\{\vec{u}^{(0)}, \vec{u}^{(1)}, ..., \vec{u}^{(N-1)}\}$, sampled in equidistant time steps $\Delta t = 1/f_{samp}$. In other words, $u_i^{(n)}$ is the value recorded at the i-th measuring site at time $t = n \cdot \Delta t$.

Now, compute the following two mean values

$$m_0 = \frac{1}{N} \sum_{n=0}^{N-1} \|\vec{u}^{(n)}\|^2 \quad (1)$$

$$m_1 = \frac{1}{N-1} \sum_{n=1}^{N-1} \left\| \frac{\Delta \vec{u}^{(n)}}{\Delta t} \right\|^2 \quad (2)$$

where $\|\vec{u}\|^2 \equiv \sum u_i^2$ is the squared euclidean norm of vector $\vec{u}$, and $\Delta \vec{u}^{(n)} \equiv \vec{u}^{(n)} - \vec{u}^{(n-1)}$ is the difference between the n-th and the previous data vector.

The two descriptors, Σ und Φ, are then defined as follows

$$\Sigma = \sqrt{\frac{m_0}{K}} \quad (3)$$

$$\Phi = \frac{1}{2\pi} \sqrt{\frac{m_1}{m_0}} \quad (4)$$

Obviously, Σ^2 is a measure of the global (averaged over space and time) variance — or 'power' if the input data are electrical measurements (EEG) — while Φ is a measure of dominant frequency of the changes (to get it more intuitively, imagine that the time series of vectors of measurements draws a circular trajectory in the K-dimensional state space). If the physical dimension of the measurements is X, then the dimension of Σ is X as well, and the dimension of Φ is time^{-1} (the $1/2\pi$ factor in the second equation above serves just to transform the circular frequency [radians/time] to a wavenumber frequency 1/time.)

The third descriptor, Ω, is a measure of spatial synchronization among the K measuring sites. First, compute the covariance matrix

$$\mathbf{C} = \frac{1}{N} \sum_{n=0}^{N-1} \vec{u}^{(n)} \vec{u}^{(n)\prime} \quad (5)$$

that is

$$c_{ij} = \frac{1}{N} \sum_{n=0}^{N-1} u_i^{(n)} u_j^{(n)} \quad (6)$$

Then compute the eigenvalues $\lambda_1, \ldots, \lambda_K$ of the matrix $\mathbf{C}$, and normalize the eigenvalues to a unit sum

$$\lambda'_i = \frac{\lambda_i}{T}, \text{ where } T = \sum_{i=1}^{K} \lambda_i \tag{7}$$

Note that the sum of eigenvalues T is equal to the trace of the covariance matrix tr $\mathbf{C}= \sum c_{ii}$ and thus also to the quantity m_0 defined above in (1). Finally, compute

$$\log \Omega = -\sum_{i=1}^{K} \lambda'_i \log \lambda'_i \tag{8}$$

In my article of 1995, Ω is defined as exp() of the expression above; the (silly) idea behind was to scale Ω from 1 thru K, to make it more intuitive to physiologists who don't like logarithms too much. Well, forget it — $\log \Omega$ is just fine, attains values from 0 (maximal synchronization = minimal complexity) through $\log K$ (minimal synchronization = maximal complexity).

Introducing the Ω complexity into EEG was no tremendous intellectual achievement, of course; it is just one of many members of a family of so-called *covariance complexities*, used in pattern recognition business etc. What was new was the use of those three global descriptors together to characterize dynamical properties of a spatially distributed systems; thus, the triples (Σ, Φ, Ω) represent points in (macro)state space. Physiologically similar states project to subregions of this space, and changes in the working mode of the brain manifest themselves as paths, making fine variations in the global brain functional states to appear as filamentar or layered manifolds. For example, if a (putative) generator in the brain becomes dominant over the other ongoing activity, the total variance Σ increases *and* the Ω decreases as the activities recorded from different locations become more and more correlated — thus such a change will display a different transition path than if the same goes on in the brain, just with higher voltage. The idea that emerged in our talks with Roger was to try to apply this methodology to REGG data.

Some questions arise, though; the following is a brief informal discussion, no dogmatic opinion:

What is the input data? With multichannel EEG, the primary data consists of a series of voltage measurements in μvolts. With REGG data, we can use bitsums, relative frequencies of 1's, Z values, chi-squares, whatever...
IMHO the best candidates are relative frequencies of 1's, i. e. r = bitsum/bitcount, as these are invariant (provided the REG behaves constantly) against change of the bit count per time slice. If time series are present in a form of Z scores, these should be transformed back to relative frequencies.

Data centering. In the formulas above, data is tacitly assumed to be centered to zero baseline; the question is, whether we should center the data by subtracting the empirical mean value over given epoch, or subtracting the theoretical expected value; I strongly prefer the latter option, i. e. to subtract 0.5 from the observed relative frequencies; it is deviations from the theoretical baseline what we want to assess.

Sampling frequency. This is closely related to the abovementioned issue of bit counts per time slice and, on the other hand, to the expected *rate* of changes we want to assess. Here I have no clue; we'd touched the latter issue in a table talk in Halifax, but as far as I remember without any definite conclusion. I guess the best way would be to sum up trials to combined

time slices over a wide time scale, say, ranging from seconds to hours.
(There was no problem with EEG: the sampling frequency is more or less enforced by the frequency content of brain's 'humming' – I have no idea at which frequency (if any) Gaia may be humming.)

Epoch length. While the previous comment was dealing with combining simultaneous REG data to a single data vector and the 'base frequency' of the observed fluctuations, this one relates to the issue of *stationarity* of the fluctuations and also of *stability* of our estimates.
With EEG, we were using epochs of length from a second to few seconds (brain's electrical activity is notoriously known to be non-stationary over longer time periods what regards its frequency contents as well as the spatial distribution of activities). Using typical sampling frequencies of order of 10^2 simultaneous measurements in a second, we had $N \approx 10^2 \div 10^3$.
With REG data, I have no *a priori* idea and think we have to experiment a bit again. There must be a lower bound for N, anyway; the number of data vectors have to be large enough to allow for a good estimate of the covariance matrix $\mathbf{C}$, and thus for a reliable estimate of its eigenvalues. I'm no expert in multivariate statistics, so I can't tell the rule. If anyone of you will want to elaborate this a bit, I'd be highly grateful!

Statistics. If there is no structure in the single REG data, the deviations of relative 1's frequencies r from 0.5 will be normally distributed and uncorrelated in time (gaussian white noise). Confidence intervals should be specified for Σ and Φ, given the count of data vectors per epoch N and the bit count per data point. This should be an easy excercise in χ^2 statistics (admittedly, I never did it as we *know* that EEG is no gaussian noise).
If there is no structure in the multiple REG data, the theoretical covariance matrix $\mathbf{C}$ is a diagonal matrix with equal values along the diagonal, thus the normalized spectrum of eigenvalues is uniform ($\lambda'_1 = \cdots \lambda'_K = \frac{1}{K}$) and the complexity attains its maximum $\log K$; however, we are dealing with sample estimates of $\mathbf{C}$ based on a finite number N of data vectors and, consequently, the distribution of the eigenvalues of the sample covariance matrix $\hat{\mathbf{C}}$ will not be exactly uniform. In other words, we have to expect *lower* empirical values of $\log \Omega$ even if these don't indicate a *significant* departure from 'sphericity' of the data (no correlations between distant sites). Would someone be so kind and derive the proper formula for a confidence interval for $\log \Omega$ with given N and K? I'd be much grateful again!

Presentation. The easy way are the three traces, Σ, Φ and Ω along time axis, whatever our time resolution may be. However, as I alluded to above, the strategy occured to be most useful with 3D displays of a 'macro-state space' with coordinates Σ, Φ and Ω, showing the state representations grouping and clustering. This might be a useful approach even here; as for the axes, I incline to $\log \Sigma$, $\log \Phi$, $\log \Omega$; as argued above, $\log \Omega$ is more 'natural' than Ω by its original definition; for the former two, I prefer logarithms just because transforming changes by multiplicative factors to equidistant shifts.

That's it...I think it's worth trying (and not abandoning too early, perhaps we need a certain minimum number of REGs distributed around to observe any reasonable patterns). I'll be looking (well, with one eye, the lab business takes a lot of time) forward to the results.

Jiri Wackermann `<jw@igpp.de>`