

# applying MAP to the sEMG example

claudio castellini

dlr - german aerospace center

## what is wrong in MLE?

MLE only considers the available data

MAP will take "more" information into account

be it from outside, from previous experiments, from reasonable or wishful thinking

MAP is a weighted/corrected/regularised version of MLE

let us apply it to the coin and sEMG examples!

## the coin, reloaded

remember: if we've seen  $s = [HTHHTHHHHTH]$ ,  $\theta^* = 0.3$  is all and only we can say (already quite stupid),

and if we get  $s = [HH]$  then we "know" that we will never get any Tails!

clear explanation: you need to do more flips!

- MLE only depends on  $\mathcal{D}$ , so

- MLE is flawed if  $\mathcal{D}$  is inadequate

and MAP can fix this

## the coin, reloaded

suppose we have flipped the coin 1000 times in the past, to check how fair that was.

we've got a sequence  $s_{BIG}$  with 475 Heads, 525 Tails

I bet  $s_{BIG}$  is more reliable than  $s...$  guess why? because it is *BIG* - it represents a "fair sampling" of your space (recall sEMG!)

now, of course every time we get a sequence we could calculate the likelihood anew, requiring that it be *BIG*... but isn't this slightly uncomfortable?

## the coin, reloaded

why not rather "store" the information gathered from  $s_{BIG}$  and reuse it ready-made later on, whenever we need it?

in probabilistic terms: let us turn this knowledge into a prior  $p_{BIG}(\theta)$ , plug it in Bayes's theorem, and then maximise the posterior whenever we need it, for any new sequence  $s$  (even the short ones!)

$$p(\theta|s) = \frac{p(s|\theta)p_{BIG}(\theta)}{\text{constant}}$$

## the coin, reloaded

the prior is  $p_{BIG}(\theta)$ , which we obtain as the Beta density for  $a = 525$  and  $b = 475$

$$p_{BIG}(\theta) = \text{Beta}(\theta|525, 475) = \frac{\Gamma(1000)}{\Gamma(525)\Gamma(475)} \theta^{524} (1 - \theta)^{474} = c \cdot \theta^{524} (1 - \theta)^{474}$$

it turns out that  $c$  is o.o.m.  $10^{300}$ , but as well  $\theta^{525}$ , for  $\theta \in [0, 1]$  is o.o.m.  $10^{-300}$ .

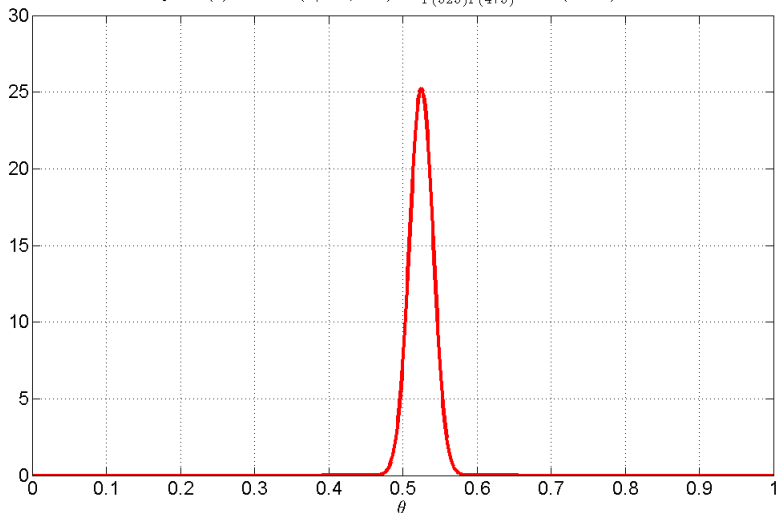
math freaks out there:  $\Gamma(n) = (n - 1)!$ , and now, how do I evaluate  $c$ ? cannot really evaluate  $999!$  ...

the constant is meant to normalise what actually is a probability density, and, well, the integral evaluates to one!

## large numbers, still good results

matlab can still manage the precision, and this is the prior:

$$p_{BIG}(\theta) = \text{Beta}(\theta|525, 475) = \frac{\Gamma(1000)}{\Gamma(525)\Gamma(475)} \cdot \theta^{524}(1-\theta)^{474}$$



## the posterior

now it is clear that our old sequence  $s = [HTHHTHHHTH]$  cannot change things that much...

the likelihood, is, again,  $\theta^3(1 - \theta)^7$ ; the posterior is then

$$p(\theta|s) = \frac{\theta^3(1 - \theta)^7 \cdot c \cdot \theta^{524}(1 - \theta)^{474}}{\text{constant}} =$$
$$\text{new\_constant} \cdot \theta^{527}(1 - \theta)^{481}$$

still very similar to the prior:  $\theta^* = \frac{527}{481+527} \approx \frac{525}{475+525}$ . almost a fair coin, and it now pays *slightly* to bet on Tails

conclusion: maximising the posterior (MAP) gives us a much more "robust", "regular" result: new data cannot change things much (unless they are of the same size of the data used for the posterior, in which case we are already happy with the likelihood...)

## conjugate priors

hold on a sec... did you just watch the magic happen?

$$p(\theta|s) = \frac{\text{binomial}(3, 7) \cdot \text{Beta}(\theta|525, 475)}{\text{constant}} = \text{Beta}(\theta|527, 481)$$

1. the coin sequence likelihood is a *binomial (Bernoulli) distribution*;
2. the prior is a Beta density;
3. and the posterior is still a Beta density.

if the prior yields a posterior belonging to the same family of functions, it is called "conjugate prior" of its likelihood

which means that all the machinery of the MLE thing can be reused to evaluate the MAP. good news!

but hold on... there is an even more interesting consequence

## conjugate priors

operatively,

1. we have previous knowledge "distilled" from  $s_{BIG}$  in a prior;
2. we get a new  $s'$  in the form of a likelihood;
3. and we derive a posterior  $p_{(s_{BIG}, s')}$  which puts them together.

and now we can employ this information (gathered from  $s_{BIG}$  and  $s'$  together) by using the posterior  $p_{(s_{BIG}, s')}$  as a *prior* for the next MAP!

when a new sequence  $s''$  comes in, we do

$$p(\theta|s'') = \frac{p(s''|\theta)p_{(s_{BIG}, s')}(\theta)}{\text{constant}}$$

every time a new sequence comes in, we "adjust" the estimate for  $\theta^*$ , shifting it a little here, a little there, according to the new data coming in

basically we use the prior to "accumulate" useful knowledge, flip after flip. the use of conjugate priors enable us to enforce a sort of online learning: the prior "learns" more and more about  $\theta$ , as new sequences come in

# MAP applied to sEMG

consider the co-contraction set again. what is a prior in this case?

a prior will be a distribution over  $(\mu^*, \Sigma^*)$  gathered from having seen a lot of co-contraction samples in the past

**simplest possible idea:** use a normal distribution for  $(\mu^*, \Sigma^*)$  as well

one more piece of good news: the normal distribution is the conjugate prior... of itself!

$$\mathcal{N}(\mu_1, \Sigma_1) \cdot \mathcal{N}(\mu_2, \Sigma_2) = \mathcal{N}(\mu_3, \Sigma_3), \quad \text{where}$$

$$\mu_3 = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} \cdot (\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2)$$

$$\Sigma_3 = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$$

or, in the one-variable case,

$$\mathcal{N}(\mu_1, \sigma_1^2) \cdot \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\mu_3, \sigma_3^2), \quad \text{where} \quad \mu_3 = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \quad \sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

# MAP applied to sEMG

situation: consider our old friend,  $S_{co}$ , the set of 121 co-contraction samples

1. let us assume they are "good" samples, so
2. let us build a prior out of them
3. let the prior be exactly the pdf we evaluated beforehand:

$$p(\mu_{co}, \Sigma_{co}) = \mathcal{N}(\mu_{co}, \Sigma_{co})$$

...remember, this is easy to evaluate, once we have  $S_{co}$

given a dataset, it is easy to build a (normal) prior out of it

now six new samples come in, call the set  $S_{nco}$ . what are the related  $(\mu^*, \Sigma^*)$  gathered by MAP?

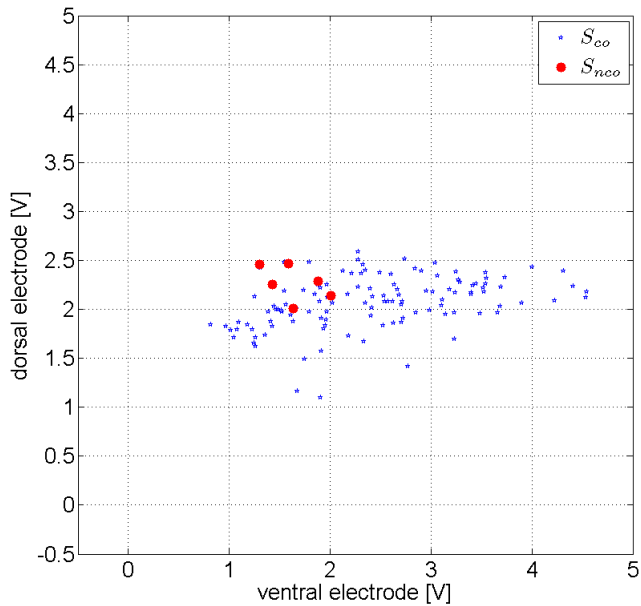
# MAP applied to sEMG

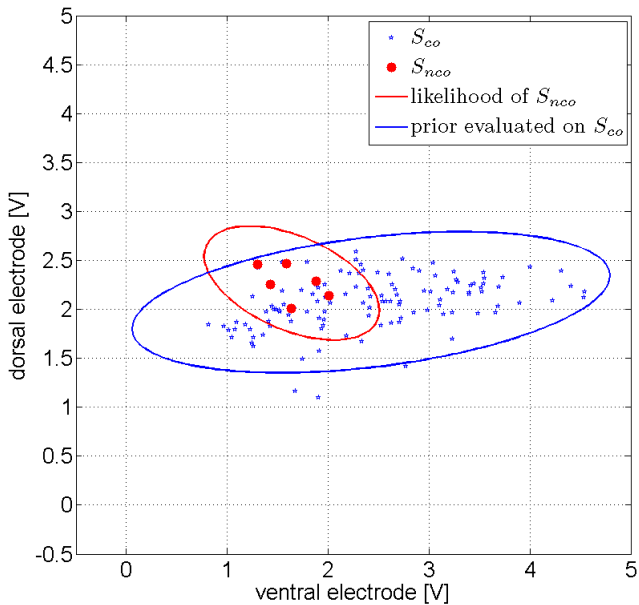
wait a sec... we can work it out in terms of product of two normals!

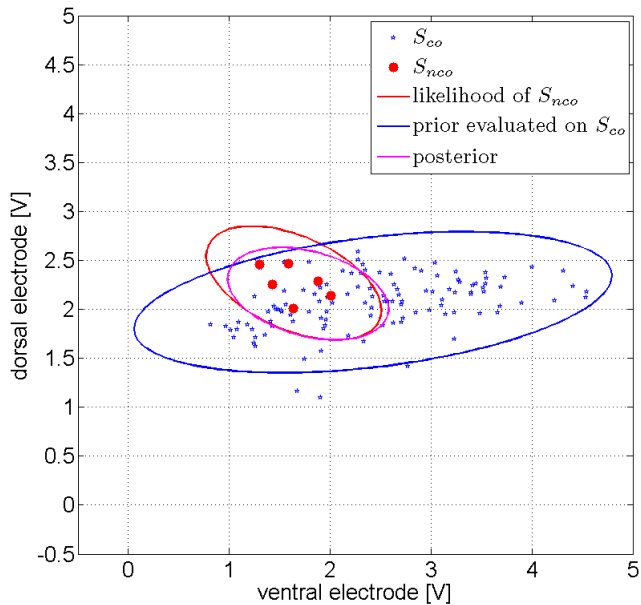
$$\begin{aligned}(\mu^*, \Sigma^*) &= \arg \max_{\mu, \Sigma} (\text{posterior}) = \\ \arg \max_{\mu, \Sigma} \left( \frac{\text{likelihood} \cdot \text{prior}}{\text{norm\_const}} \right) &= \arg \max_{\mu, \Sigma} (\text{likelihood} \cdot \text{prior}) = \\ \arg \max_{\mu, \Sigma} \left[ \mathcal{N}(\mu_{nco}, \Sigma_{nco}) \cdot \mathcal{N}(\mu_{co}, \Sigma_{co}) \right]\end{aligned}$$

and this (recall  $\mu_1$ ,  $\mu_2$  and  $\mu_3$ ) leads to a normal posterior, whose parameters are

$$\begin{aligned}\mu_p^* &= (\Sigma_{nco}^{-1} + \Sigma_{co}^{-1})^{-1} \cdot (\Sigma_{nco}^{-1} \mu_{nco} + \Sigma_{co}^{-1} \mu_{co}) \\ \Sigma_p^* &= (\Sigma_{nco}^{-1} + \Sigma_{co}^{-1})^{-1}\end{aligned}$$







are we happy? no we are not.

now hold on a sec... that is wrong! or, isn't it?

actually, the posterior looks much more like the likelihood rather than like the prior... which is exactly what we hoped would *not* happen

consider the expression for  $\mu_p^*$  again: small  $\sum_{n_{CO}}$  means that  $\mu_{n_{CO}}$  prevails over  $\mu_{CO}$  !!

in other words, "small variance wins"

convince yourself by looking at the 1D case again...

## so what is wrong?

it's as easy as that: we fitted a prior... as if it were a likelihood, and that is wrong!

that way, we were giving the same dignity to six samples, and to 121 samples

the "simplest possible idea" (slide 11) was not the right one

recall Albert Einstein: *Make things as simple as possible, but not simpler.*  
[paraphrase from a lecture at Oxford, 1933]

## how to patch it

recall Christian's lecture on multivariate Gaussians (0300, slides 23/24)

first of all we assume that the covariance of the data set is known,  $\Sigma_{co}$ .  
the covariance estimated from the observed  $S_{co}$  is a good guess!

then we build a prior for its mean value,  $\mu_{co}$ , as if we were inferring a vector from a noisy measurement (which is exactly what we are doing!)

the prior  $p(\mu|S_{co}, \Sigma_{co})$  is then a Normal density with

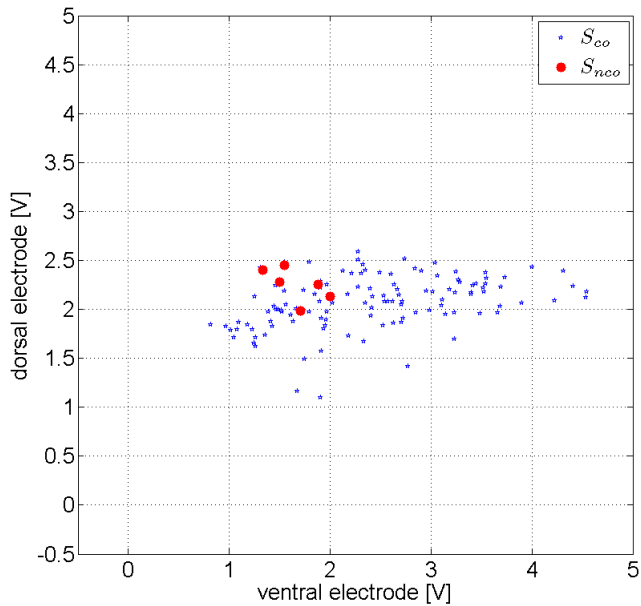
$$\mu_p = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \mu_{co} \quad \text{and} \quad \Sigma_p = \frac{1}{n} \Sigma_{co}$$

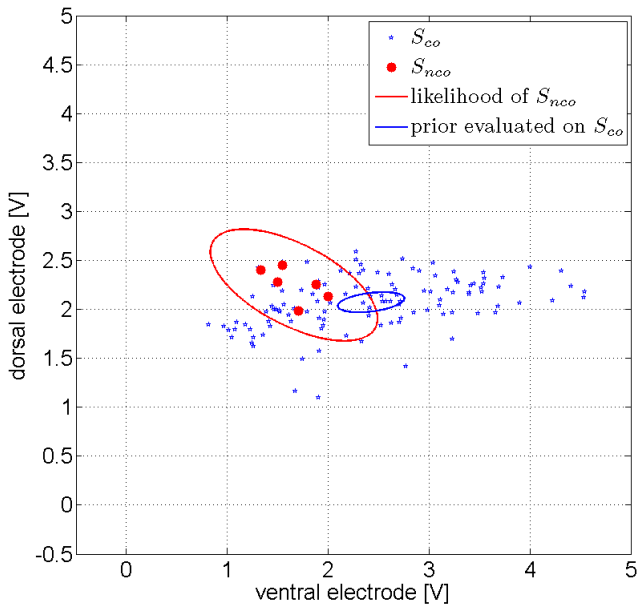
where  $n = |S_{co}|$

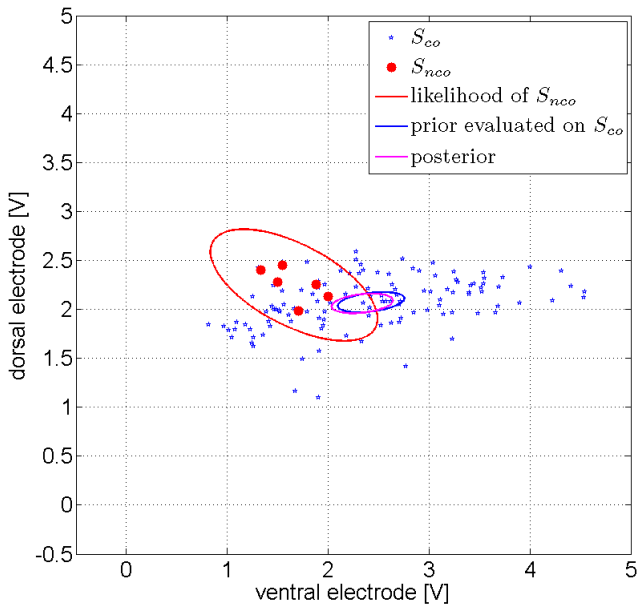
## how to patch it

now do as before but this time using the prior we just evaluated:

$$\begin{aligned}(\mu^*, \Sigma^*) &= \arg \max_{\mu, \Sigma} (\text{posterior}) = \\ \arg \max_{\mu, \Sigma} \left( \frac{\text{likelihood} \cdot \text{correct\_prior}}{\text{norm\_const}} \right) &= \arg \max_{\mu, \Sigma} (\text{likelihood} \cdot \text{correct\_prior}) = \\ \arg \max_{\mu, \Sigma} \left[ \mathcal{N}(\mu_{nco}, \Sigma_{nco}) \cdot \mathcal{N}\left(\mu_{co}, \frac{1}{n} \Sigma_{co}\right) \right]\end{aligned}$$







## conclusions

MAP is superior to MLE: given "prior knowledge", it will make our estimate more robust

can take into account *any* form of prior knowledge!

in the example here we have built a prior from previous, trustworthy data...  
...but you could also have had "external" information about the pdf

only thing: you need to be able to pack it into a prior...  
...and be able to work with it (conjugate priors are manna from heaven!)

but beware of solutions which look too easy: recall our failed attempt!