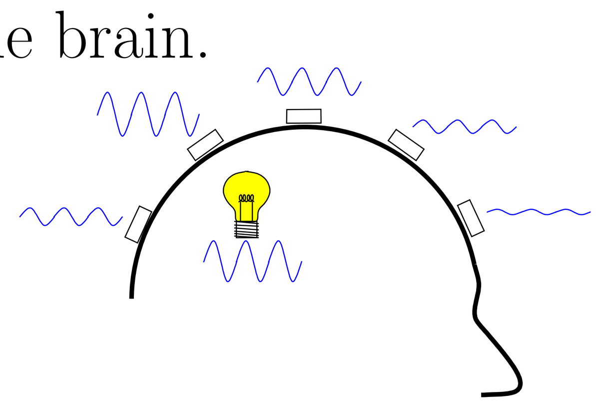
Optimizing Spatial Filters for BCI: Margin- and Evidence-Maximization Approaches

Jason Farquhar, N. Jeremy Hill, Bernhard Schölkopf

Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Spatial Filtering

Volume conduction means each EEG sensor picks up a superposition of signals from all over the brain.



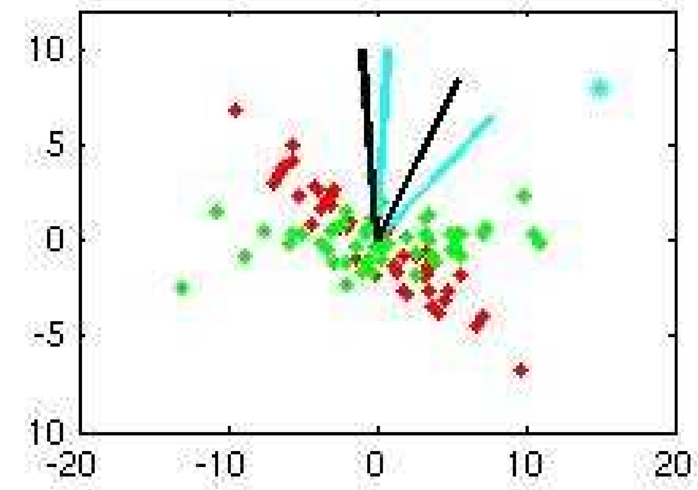
Our goal is to undo this superposition by spatial filtering, to re-focus on discriminative signals: a source separation problem. The most popular method for ERD-based BCI is the Common Spatial Pattern (CSP) algorithm which simply finds a spatial filter $\boldsymbol{s}$ to maximize:

$$\frac{\text{filtered var in class } c}{\text{total filtered var}} = \frac{\boldsymbol{s}^\top \left(\sum_{i:y_i=c} X_i X_i^\top \right) \boldsymbol{s}}{\boldsymbol{s}^\top \left(\sum_i X_i X_i^\top \right) \boldsymbol{s}}$$

Unfortunately:

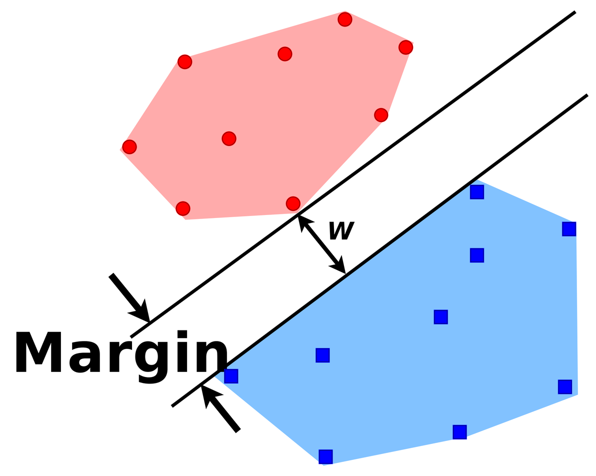
> CSP's objective is a poor predictor of classifier generalization performance.

The CSP objective's outlier-sensitivity (see figure below) leads to the problem of over-fitting.



Fix 1: Max-Margin

The maximum margin criterion (as used in SVMs) is a proven lower-bound on generalization performance.



Spatial filtering is introduced into the generalization objective as an explicit non-linear mapping to band-power features:

$$\psi(X_i; S) = \ln(\mathtt{diag}(S^\top X_i X_i^\top S))$$

where S is the spatial filter matrix $[\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots]$. Maximizing the margin in the space of these features yields the objective:

$$\lambda \boldsymbol{w}^\top \boldsymbol{w} + \sum_i \max(0,\ 1 - y_i(\psi(X_i; S)^\top \boldsymbol{w} + b))$$

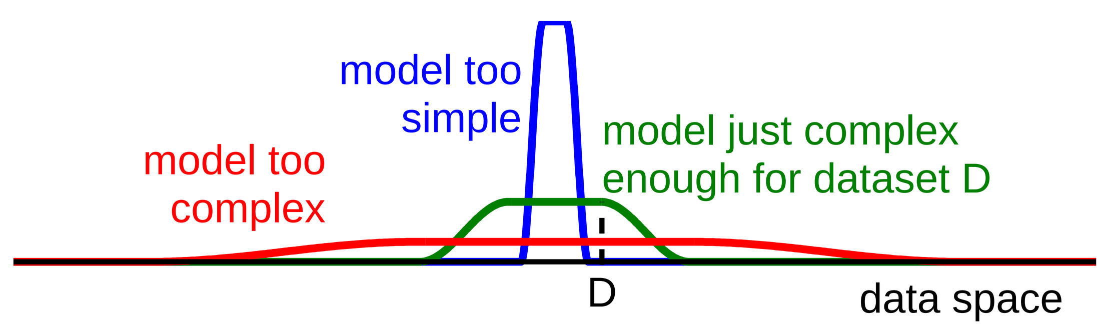
This is an unconstrained optimization. We minimize the objective with respect to $\boldsymbol{w}, b, S$, using conjugate gradient (seeded with CSP solutions to avoid local minima). Regularization hyperparameter λ is found by cross-validation.

Fix 2: Max-Evidence

The marginal likelihood or evidence of a probabilistic model with hyperparameters $\boldsymbol{\theta}$ is given by integrating the parameters (e.g. a classifier's weight vector $\boldsymbol{w}$) out of the likelihood for data D:

$$P(D|\boldsymbol{\theta}) = \int \Pr(D|\boldsymbol{w}, \boldsymbol{\theta}) \Pr(w|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{w}$$

As it is a probability density function, the evidence normalizes over the space of possible datasets. Maximizing it can be an effective means of complexity control and hence model selection:

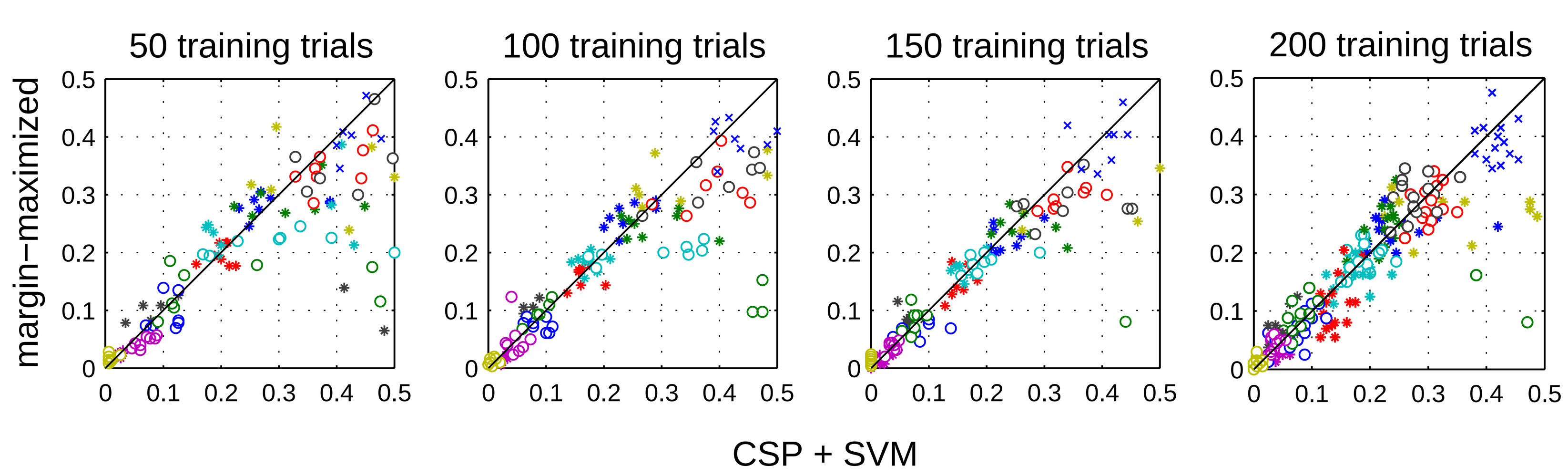


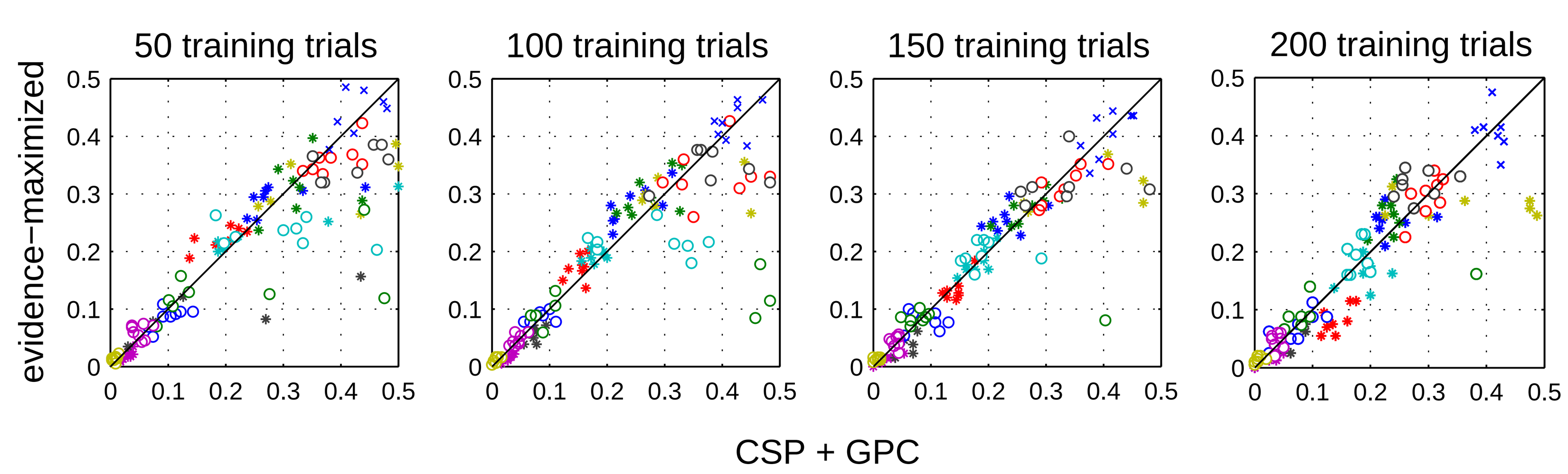
We treat spatial filter coefficients S as covariance function hyper-parameters in a Gaussian Process Classifier (GPC). As in our max-margin approach, we use a linear function of log filtered variances:

$$k(X_i, X_j) = 1 + \psi(X_i; S)^\top \psi(X_j; S)$$

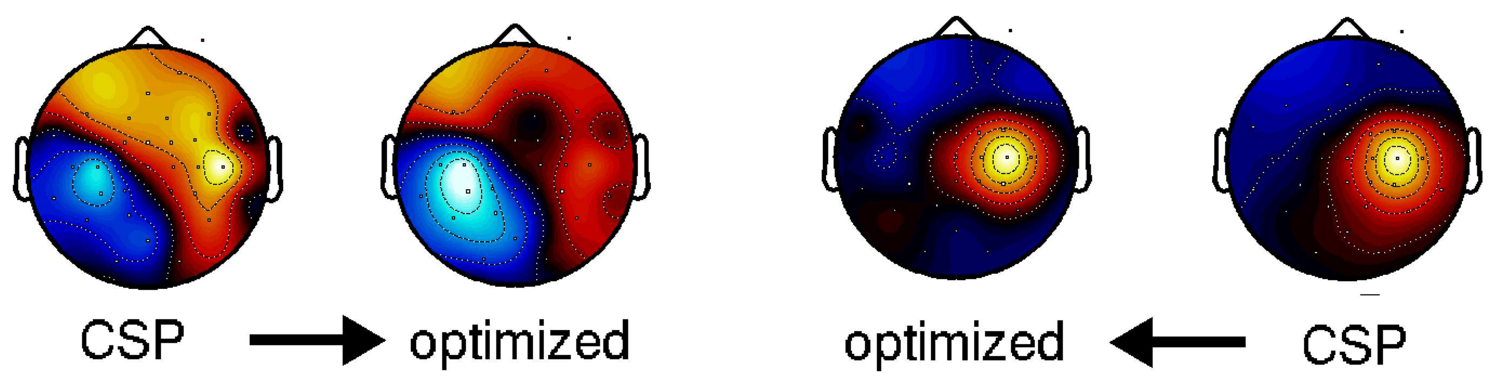
Since we can compute $\partial k / \partial S$, the Gaussian Process framework allows us to maximize $P(D|S)$ by a conjugate gradient method. Again, we use CSP as the seed.

Results





We show binary classification error from 15 imagined movement subjects: 9 from BCI competitions (Comp 2:IIa, Comp 3:IVa,IVc) and 6 from the MPI. These were pre-processed to select 0.5–4s after stimulus presentation and band-pass filtered to 8–25Hz.



> The two methods (margin and evidence maximization) perform similarly. Both show consistent improvements over ordinary CSP, most noticeably when few training trials are available or when initial performance is poor.