New Margin- and Evidence-Based Approaches for EEG Signal Classification

N. Jeremy Hill and Jason Farquhar

Max Planck Institute for Biological Cybernetics, Tübingen

Brain-Computer Interface researchers at the
Department of Empirical Inference For Machine Learning And Perception:

Felix Bießmann, Cornelius Raths
Suzanne Martens
Jason Farquhar, Jeremy Hill
Bernhard Schölkopf

Goals

Develop systems which completely paralysed people (such as late-stage sufferers of Amyotrophic Lateral Sclerosis, ALS) can use to communicate, without relying on:

- muscles

Goals

Develop systems which completely paralysed people (such as late-stage sufferers of Amyotrophic Lateral Sclerosis, ALS) can use to communicate, without relying on:

- muscles
- peripheral nerves

Goals

Develop systems which completely paralysed people (such as late-stage sufferers of Amyotrophic Lateral Sclerosis, ALS) can use to communicate, without relying on:

- muscles
- peripheral nerves
- (vision)

Goals

Develop systems which completely paralysed people (such as late-stage sufferers of Amyotrophic Lateral Sclerosis, ALS) can use to communicate, without relying on:

- muscles
- peripheral nerves
- (vision)
- (motor cortex)

BCI projects

- Attention shifts to auditory stimuli.

BCI projects

- Attention shifts to auditory stimuli.
- Attention shifts to tactile stimuli.

  5-class paradigm in MEG (incl. NIC).

  50–85% correct, avg. 70% across 9 subjects.

  (Cornelius Raths, MSc awarded 2007)

BCI projects

- Attention shifts to auditory stimuli.
- Attention shifts to tactile stimuli.

  5-class paradigm in MEG (incl. NIC).

  50–85% correct, avg. 70% across 9 subjects.

  (Cornelius Raths, MSc awarded 2007)
- Improvement of visual "speller" paradigms
  - Manipulation of stimulus type.
  - Optimization of stimulus code according to information-theoretic and psychophysiological factors.

  (Felix Bießmann, MSc project in progress)

BCI projects

- Attention shifts to auditory stimuli.
- Attention shifts to tactile stimuli.

  5-class paradigm in MEG (incl. NIC).

  50–85% correct, avg. 70% across 9 subjects.

  (Cornelius Raths, MSc awarded 2007)
- Improvement of visual "speller" paradigms
  - Manipulation of stimulus type.
  - Optimization of stimulus code according to information-theoretic and psychophysiological factors.

  (Felix Bießmann, MSc project in progress)
- Ongoing work with Prof. Niels Birbaumer's group in Tübingen to analyse EEG/ECoG data from ALS patients.

BCI projects

- Attention shifts to auditory stimuli.
- Attention shifts to tactile stimuli.

  5-class paradigm in MEG (incl. NIC).

  50–85% correct, avg. 70% across 9 subjects.

  (Cornelius Raths, MSc awarded 2007)
- Improvement of visual "speller" paradigms
  - Manipulation of stimulus type.
  - Optimization of stimulus code according to information-theoretic and psychophysiological factors.

  (Felix Bießmann, MSc project in progress)
- Ongoing work with Prof. Niels Birbaumer's group in Tübingen to analyse EEG/ECoG data from ALS patients.
- *Algorithm development.*

Role of Machine Learning in BCI

- Get results quickly:

  Shift the burden of learning from the patient to the computer. Hours to recognize the relevant features, rather than weeks/months training a patient to modulate pre-specified features.

Role of Machine Learning in BCI

- Get results quickly:

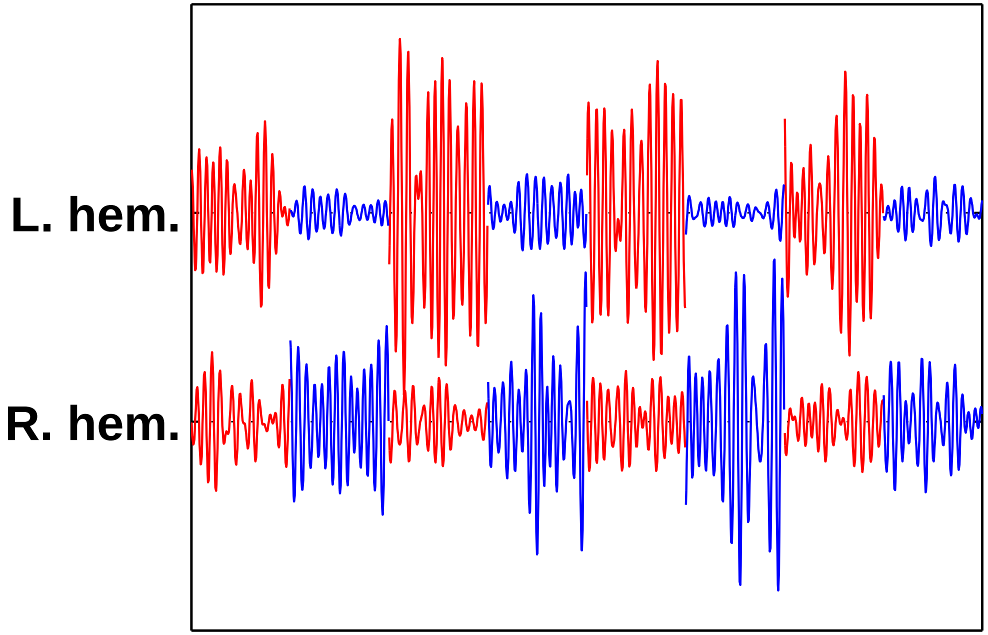
  Shift the burden of learning from the patient to the computer. Hours to recognize the relevant features, rather than weeks/months training a patient to modulate pre-specified features.

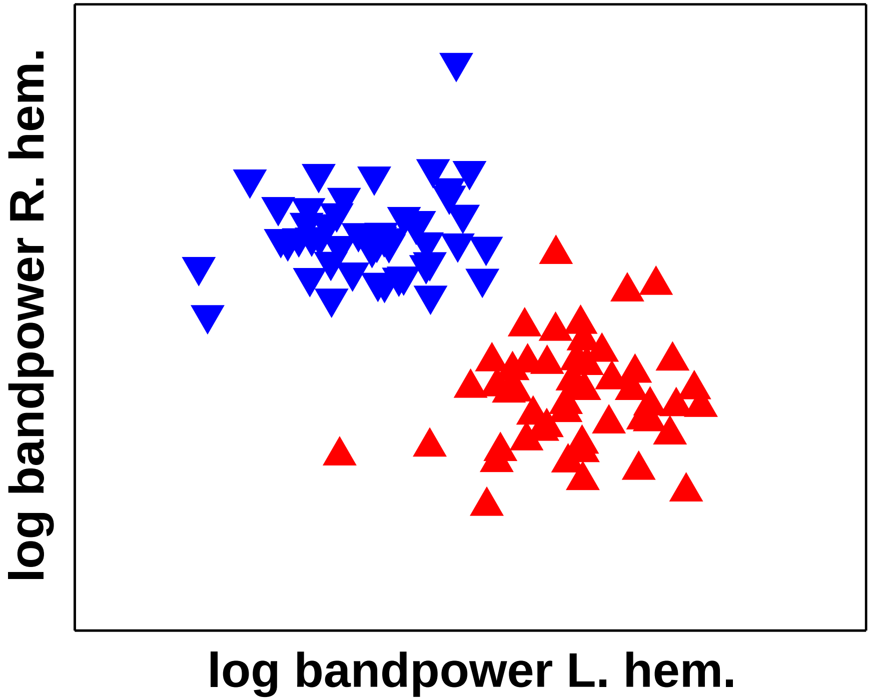
- Let the system run itself:

  No intervention from experts.

Example BCI setting

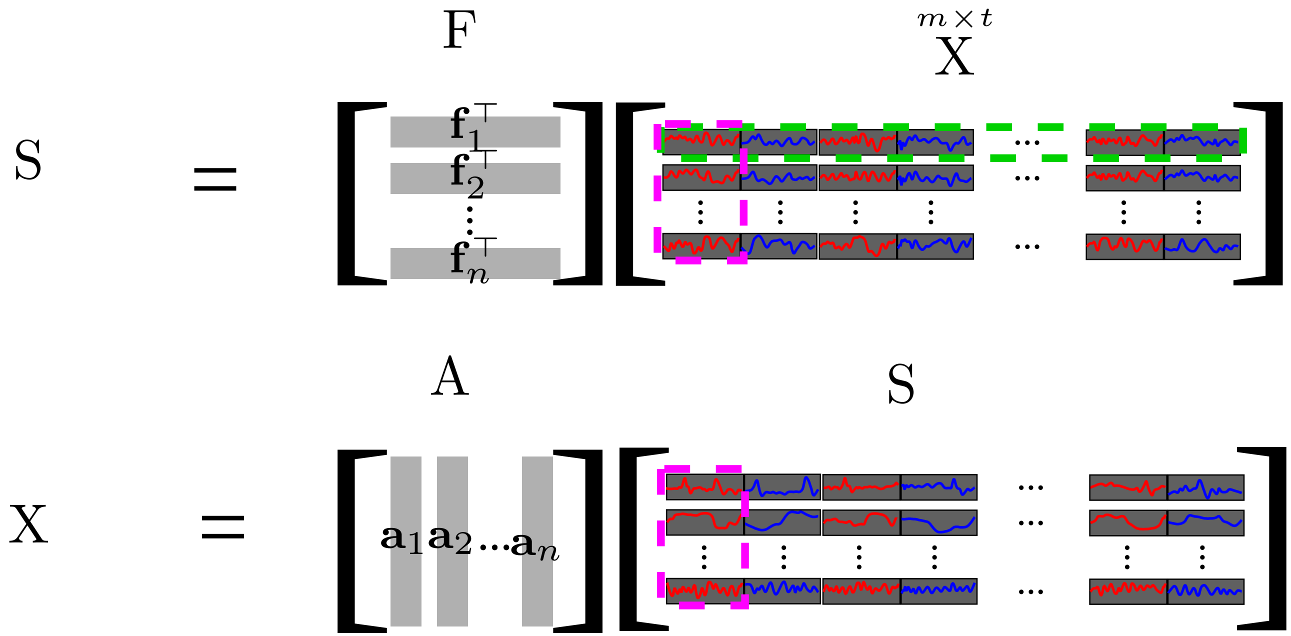
Event-Related Desynchronization in motor imagery: classify imagined left hand movement vs. imagined right hand movement based on α-band power of estimated pre-motor cortex sources in the left and right hemispheres.





Spatial Filtering

Given multichannel time-series X, we want appropriately spatially filtered time-series S = FX that contain only task-relevant information.

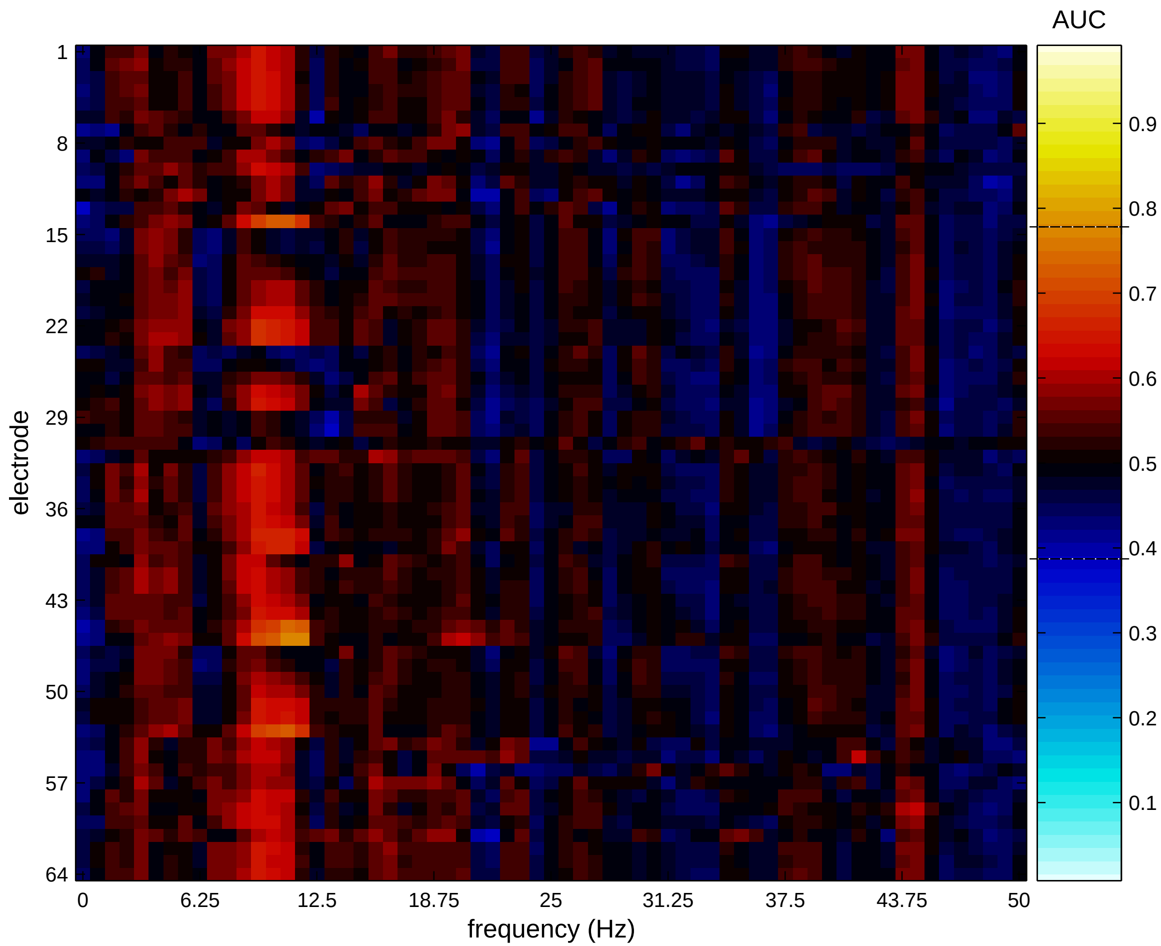


e.g.

- Independent Component Analysis (ICA)
- Common Spatial Pattern (CSP) — Koles 1991.

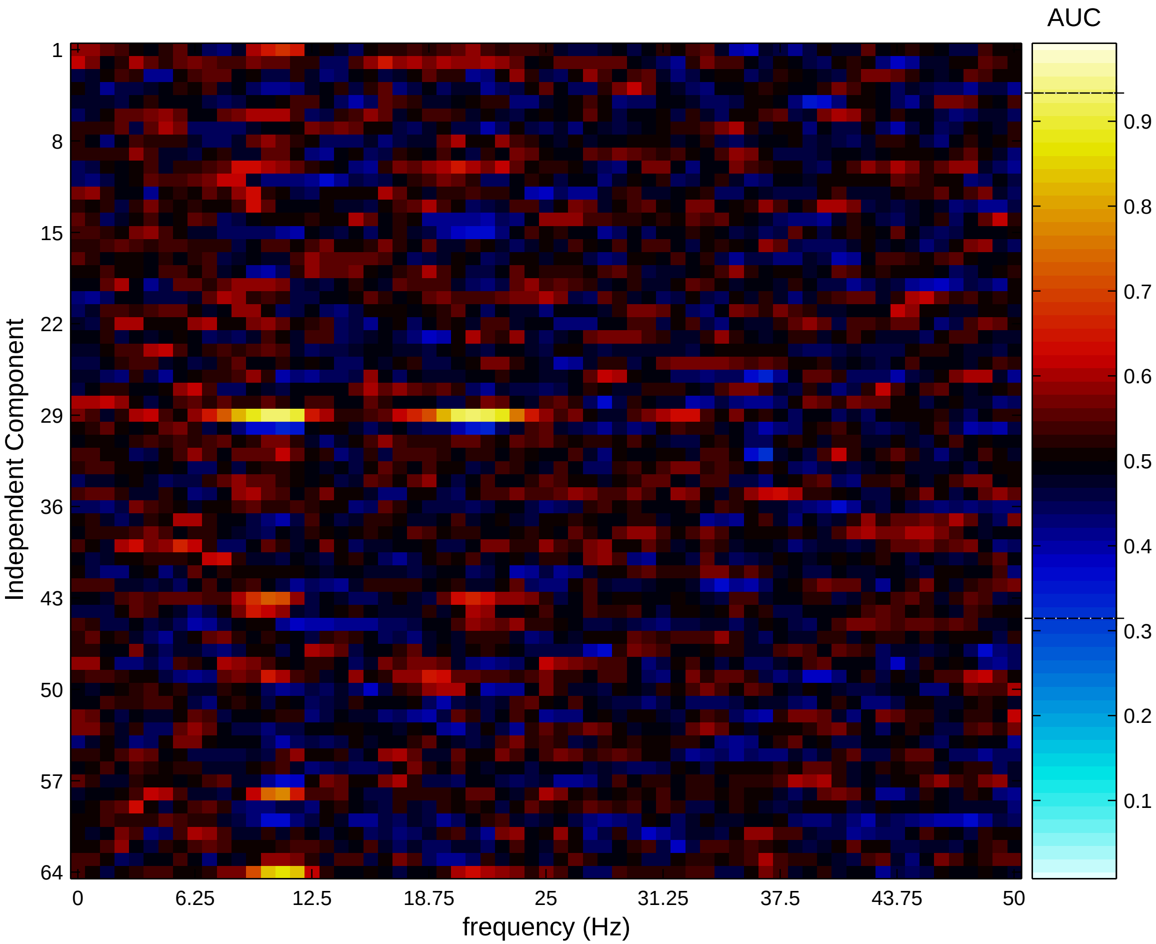
EEG Example

Amplitude spectra of raw EEG signals:

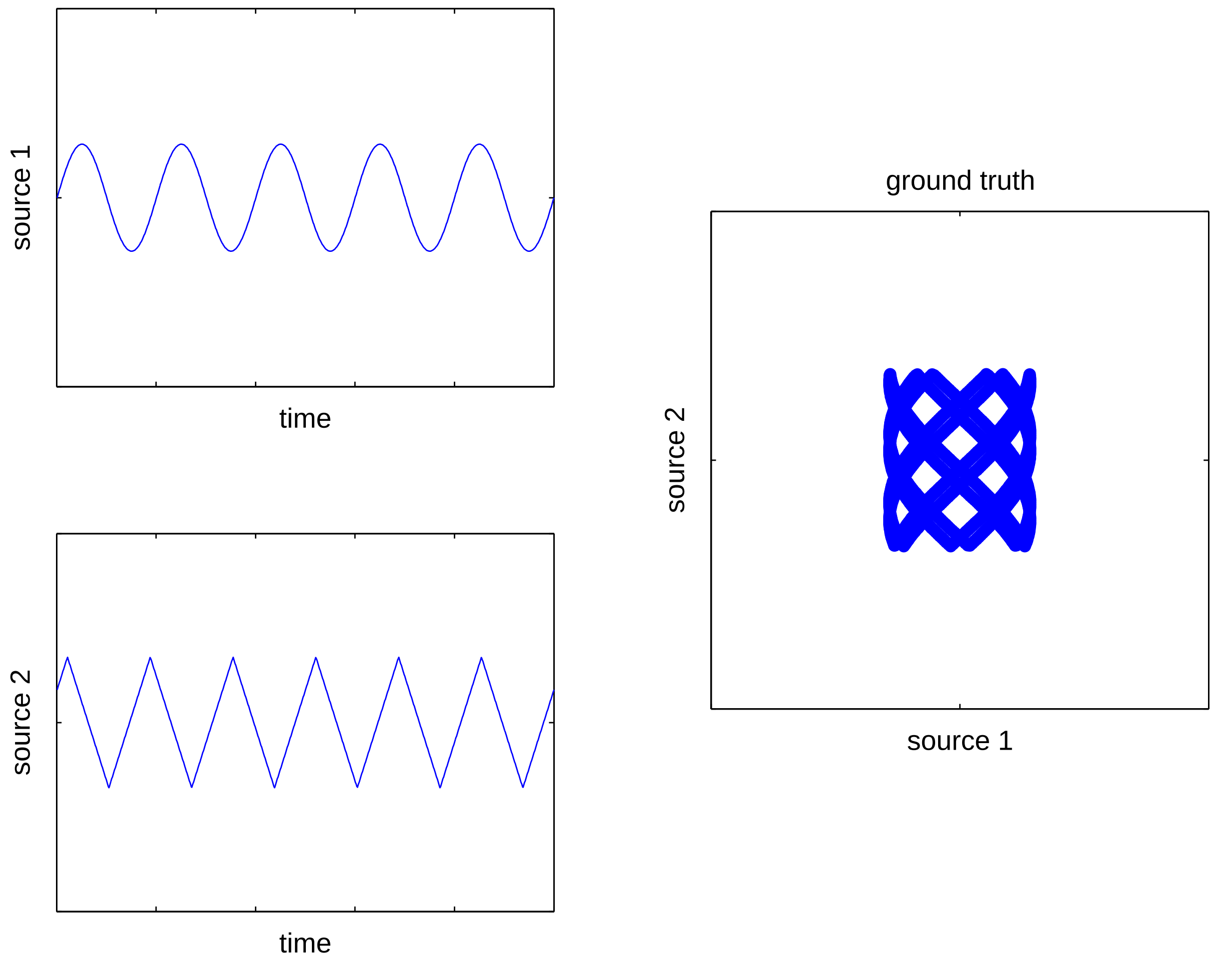


EEG Example

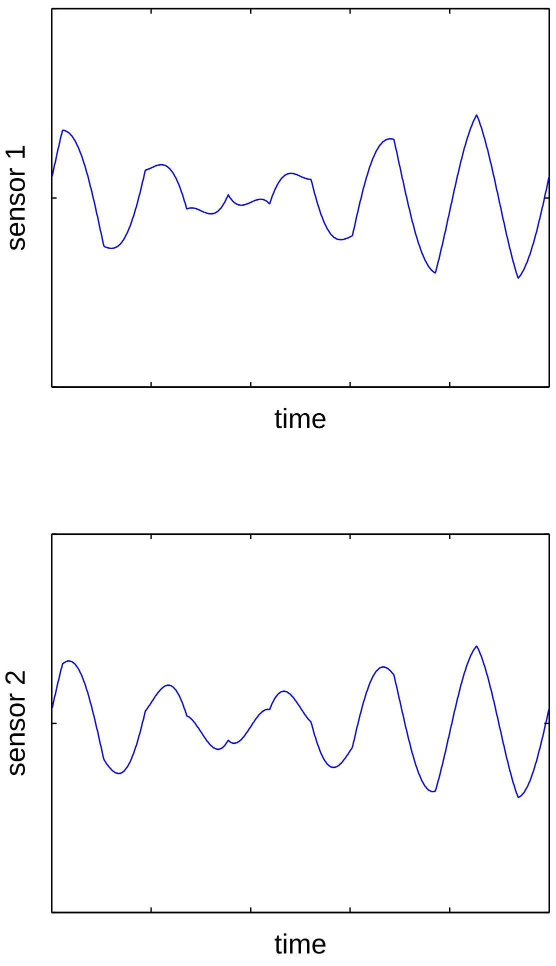
Amplitude spectra of sources estimated by Independent Component Analysis:

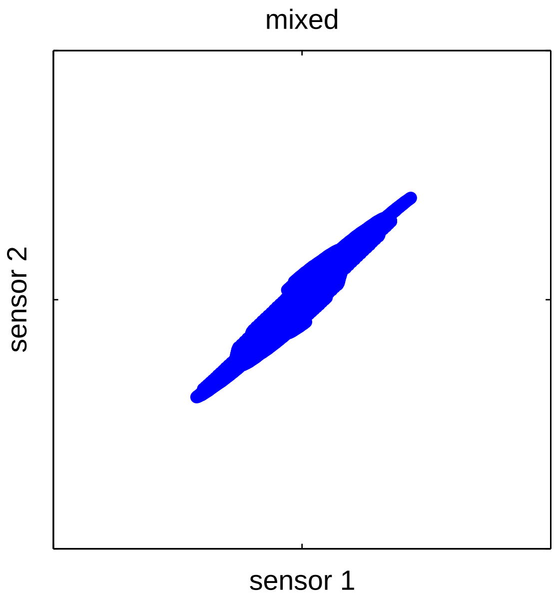


Whitening and rotation

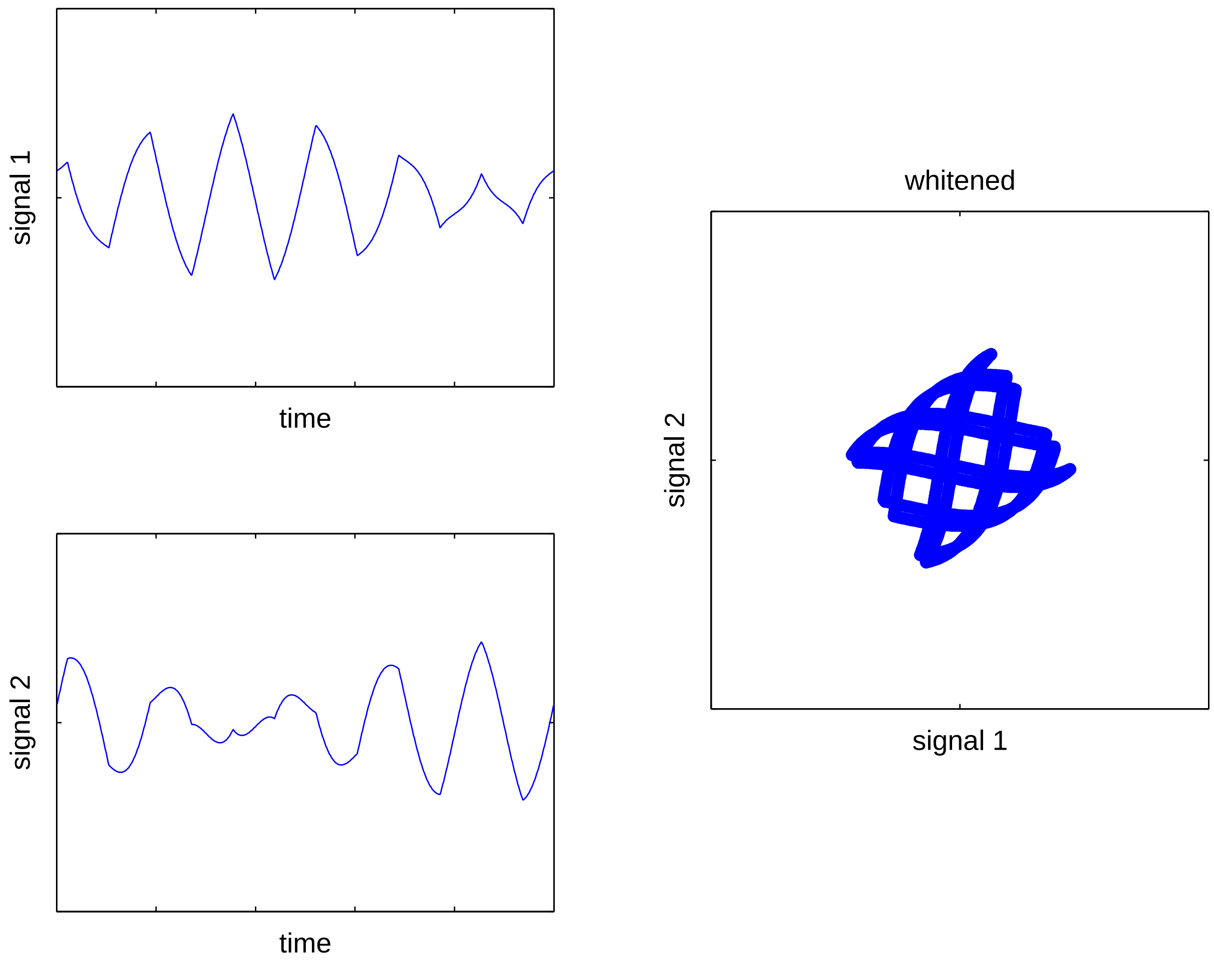


Whitening and rotation

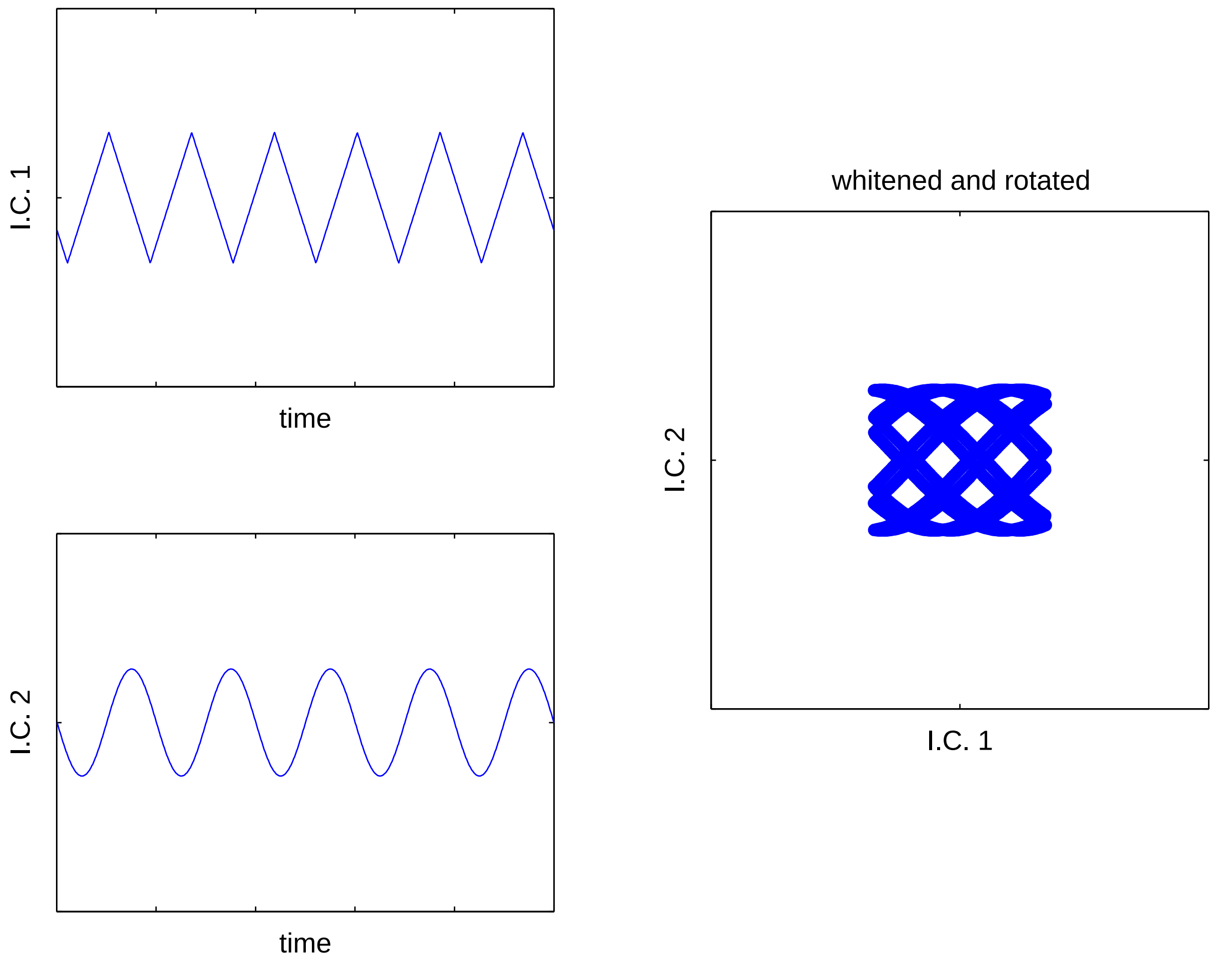




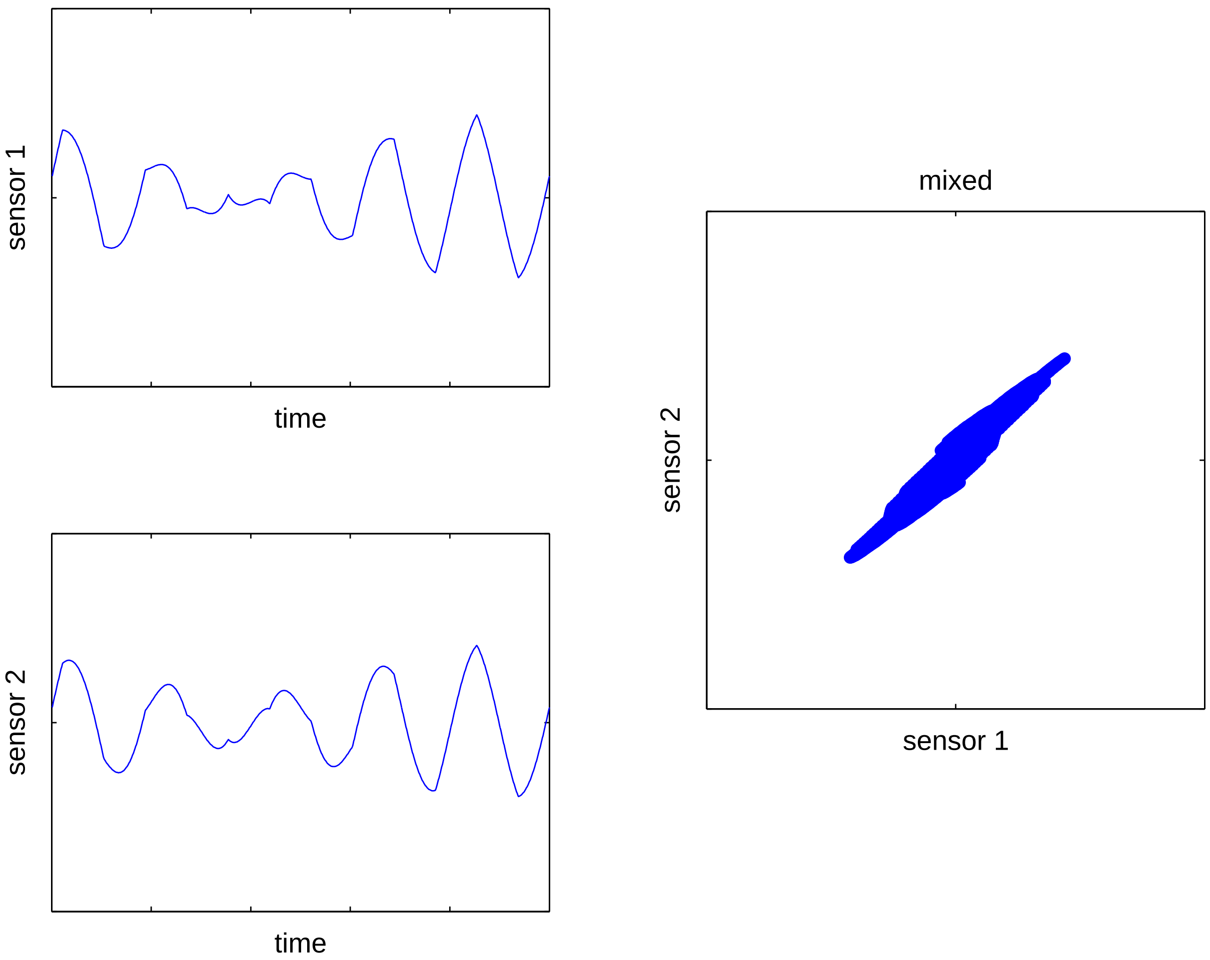
Whitening and rotation



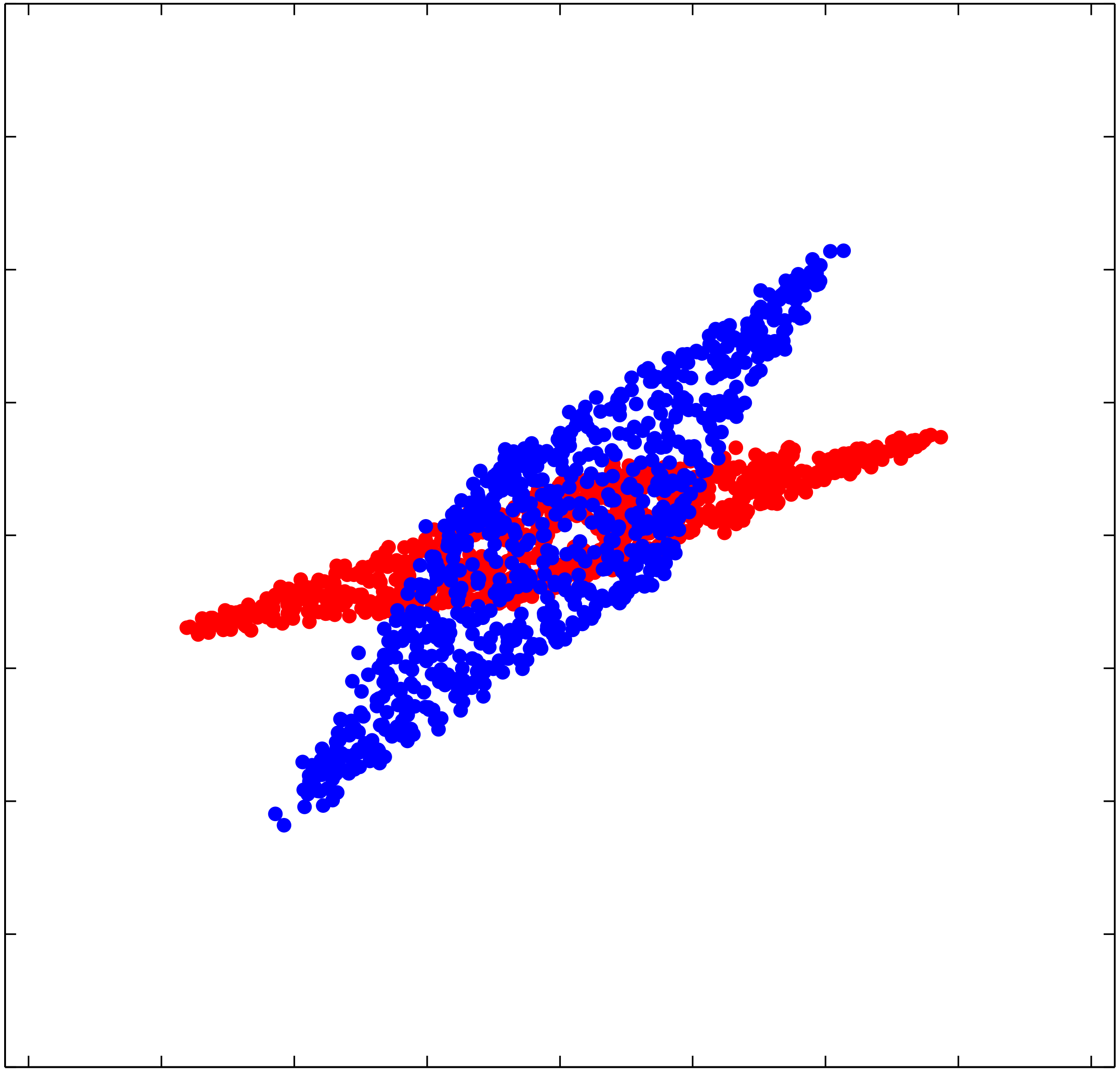
Whitening and rotation



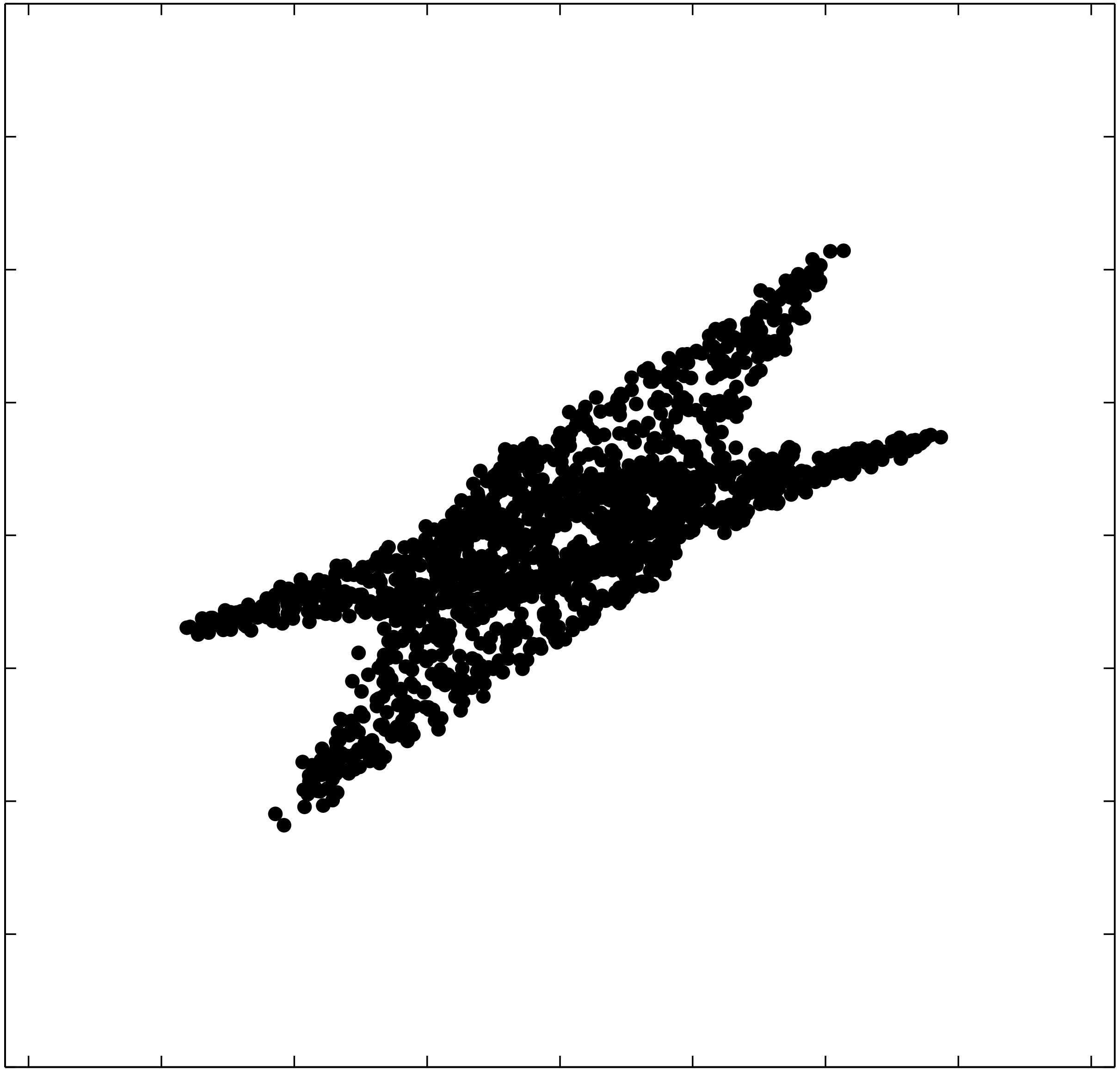
Whitening and rotation



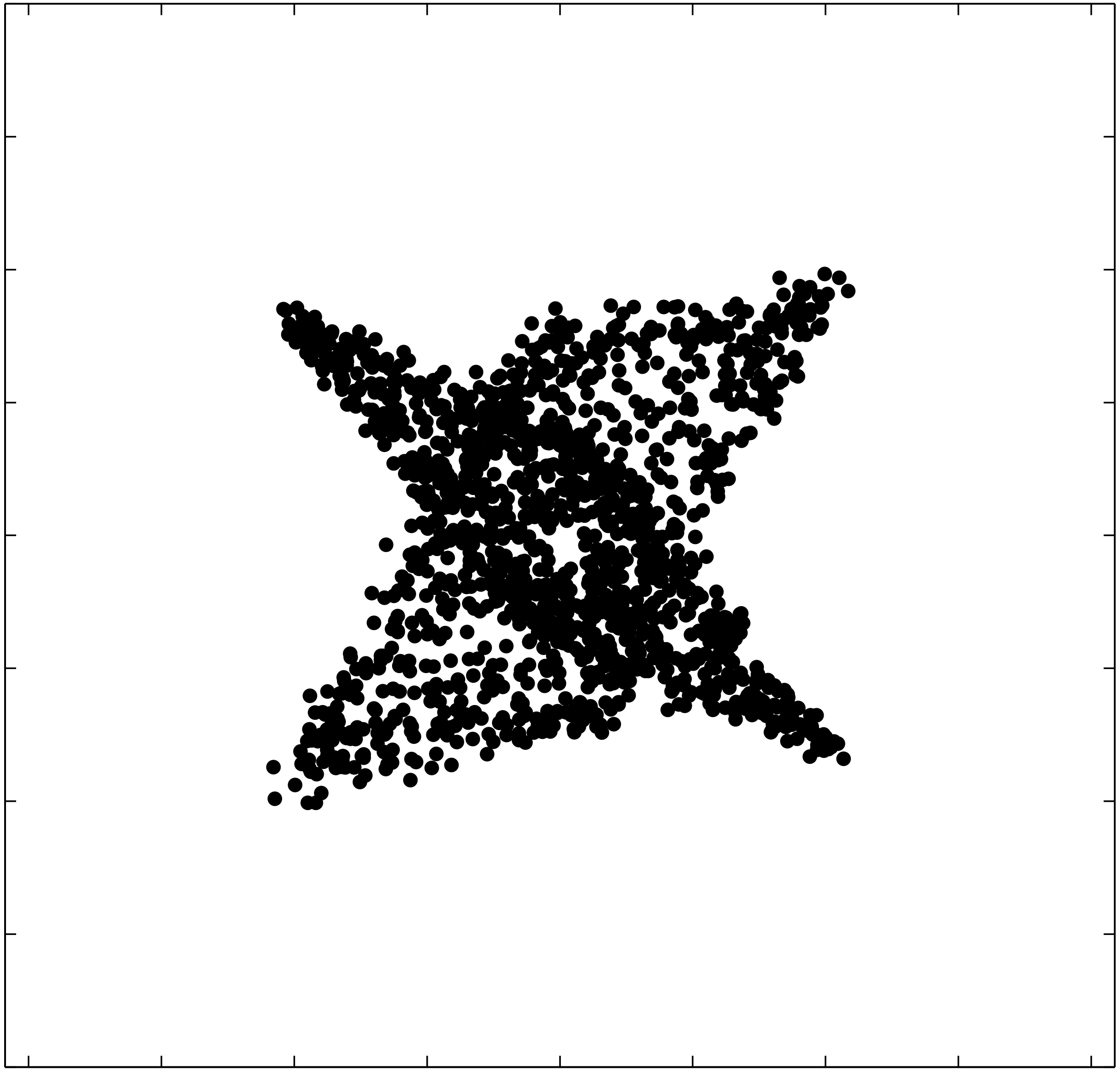
Cheap supervised rotation with CSP



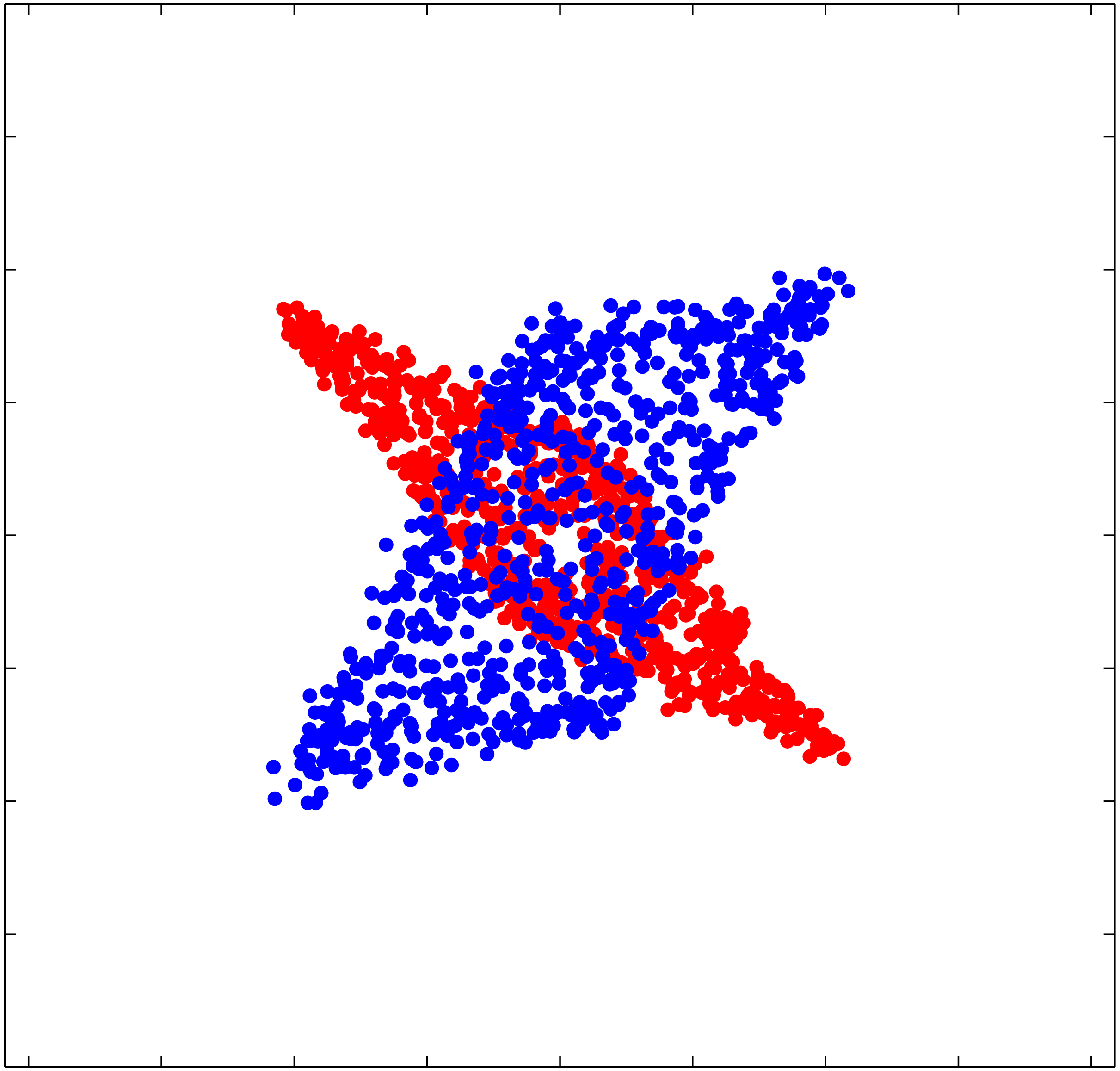
Cheap supervised rotation with CSP



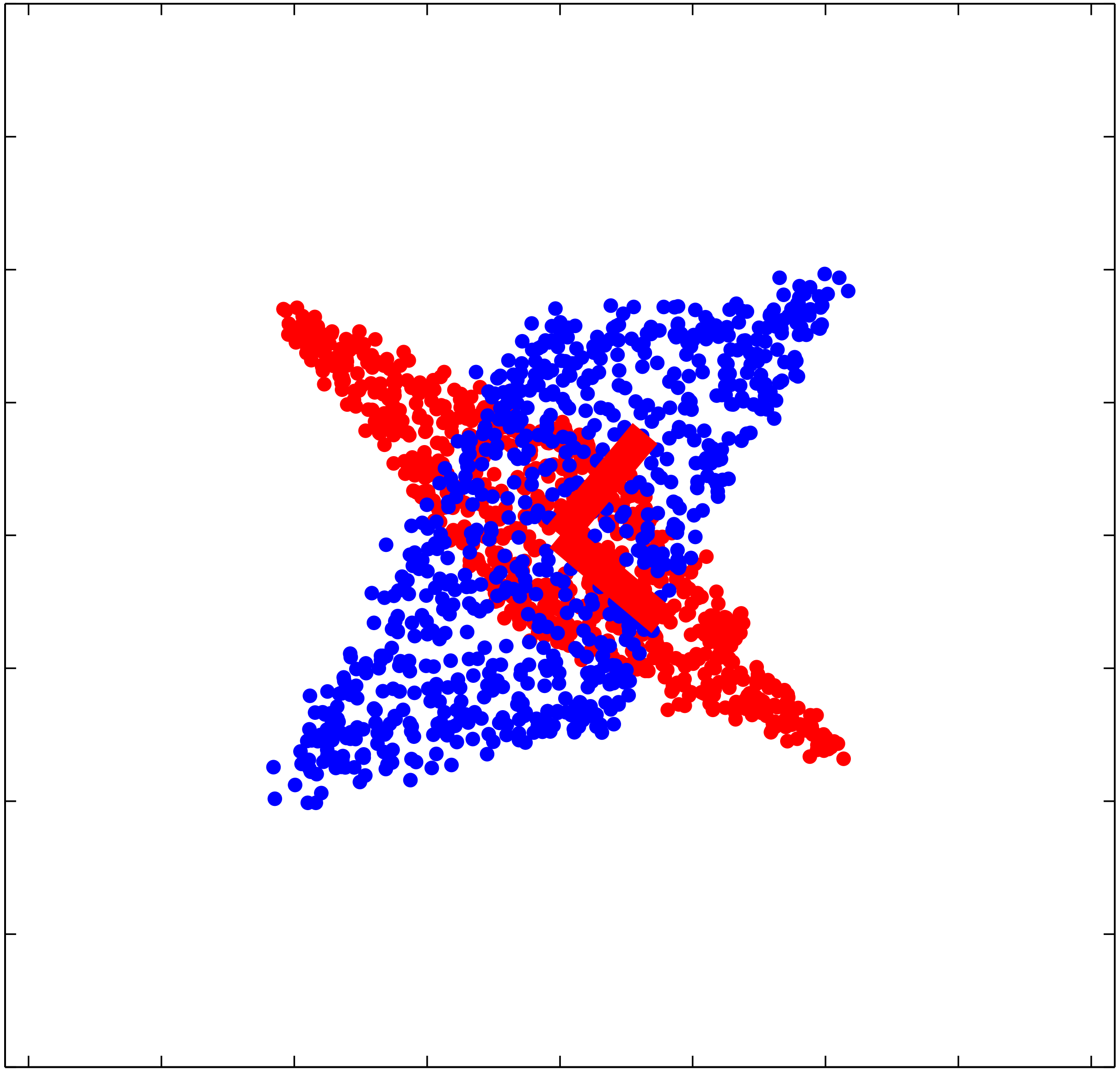
Cheap supervised rotation with CSP



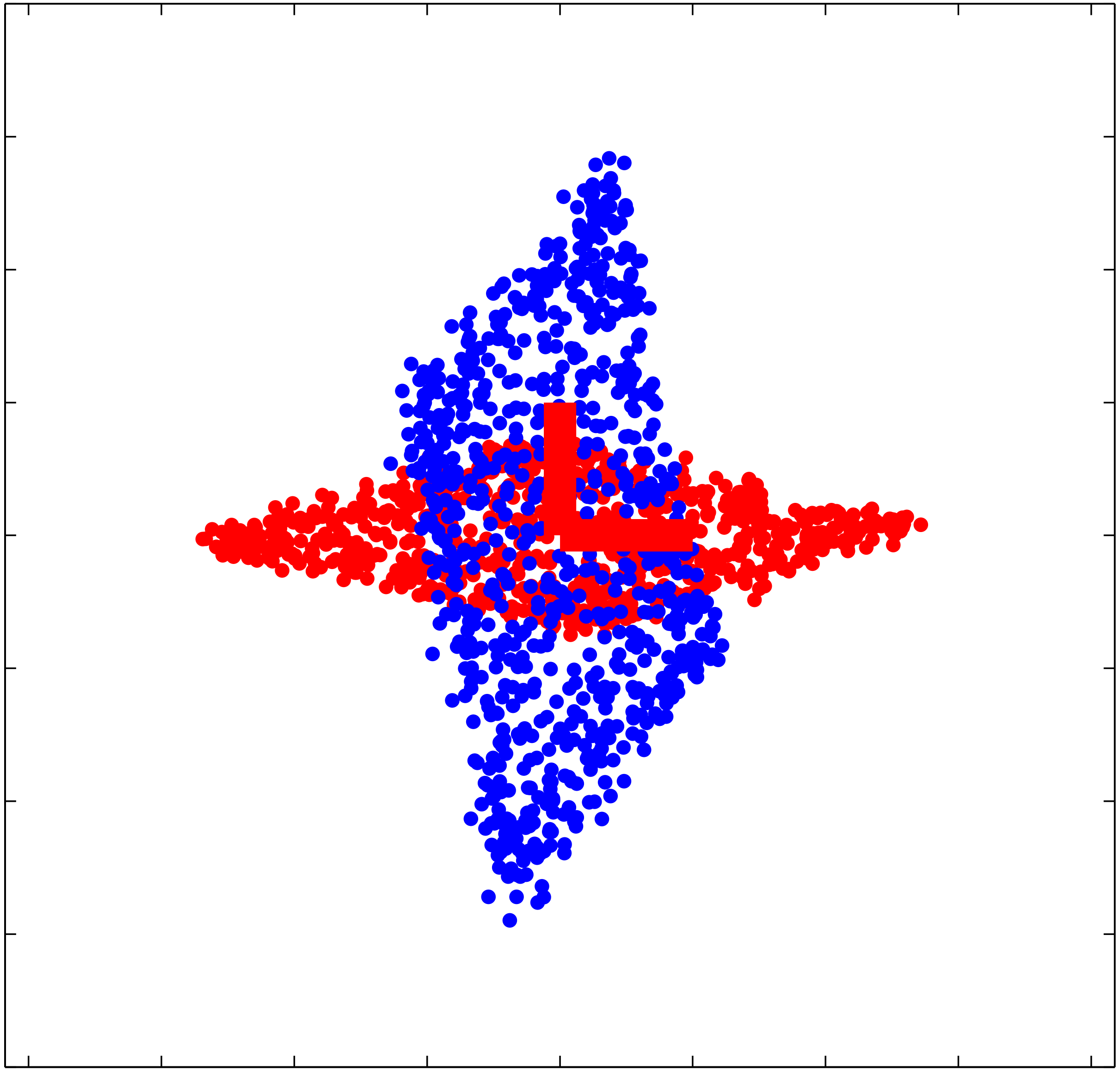
Cheap supervised rotation with CSP

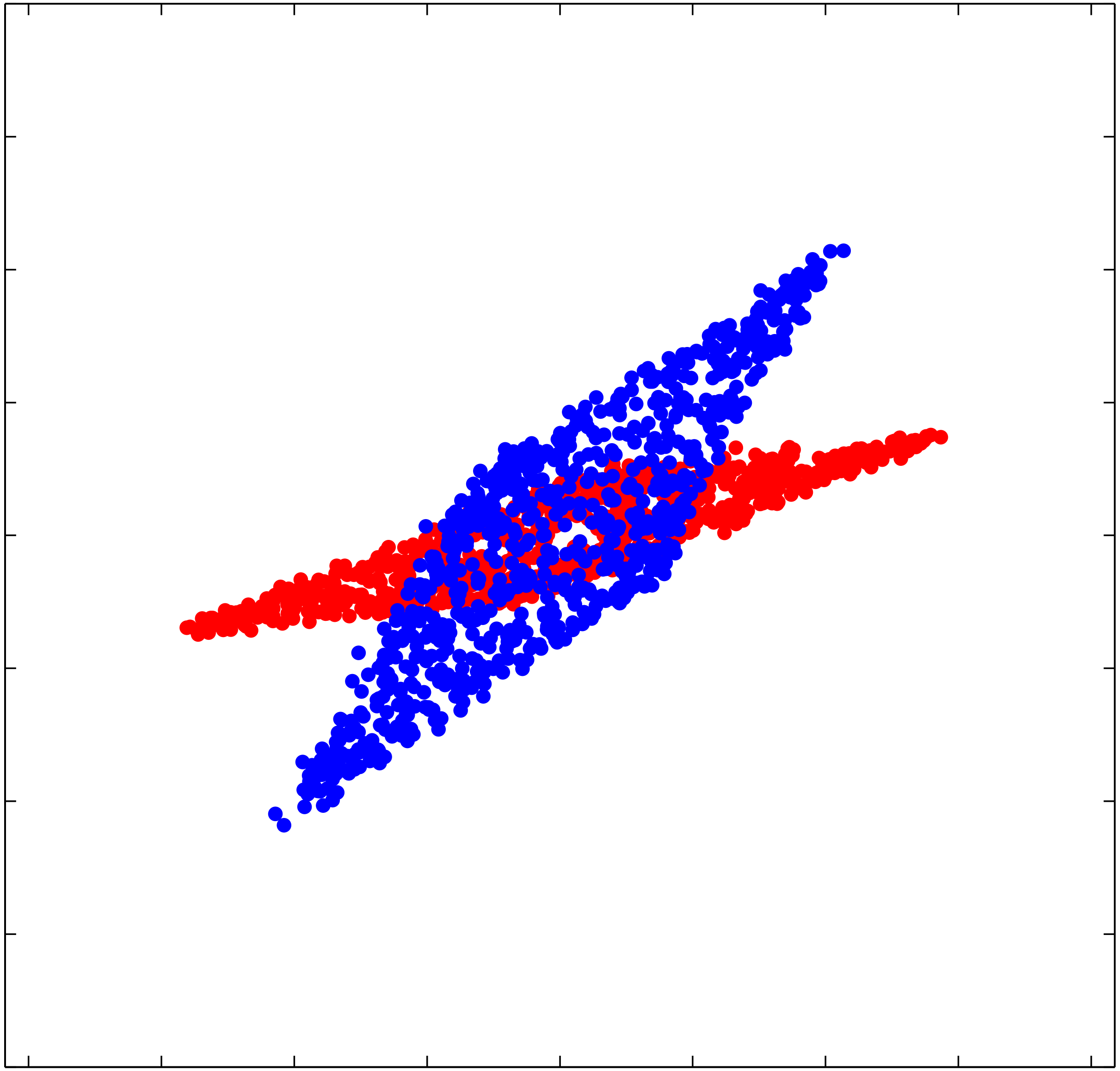


Cheap supervised rotation with CSP

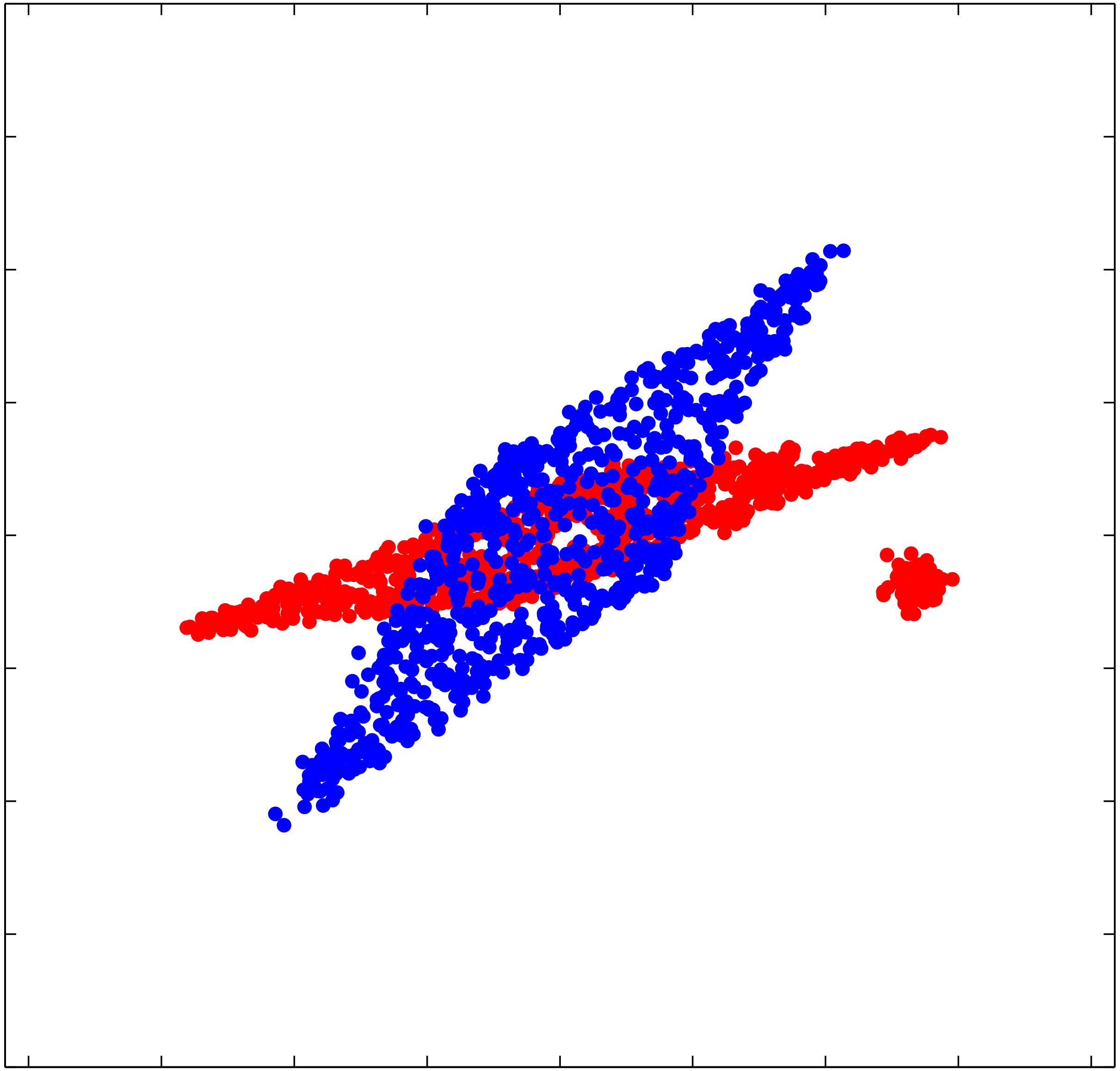


Cheap supervised rotation with CSP

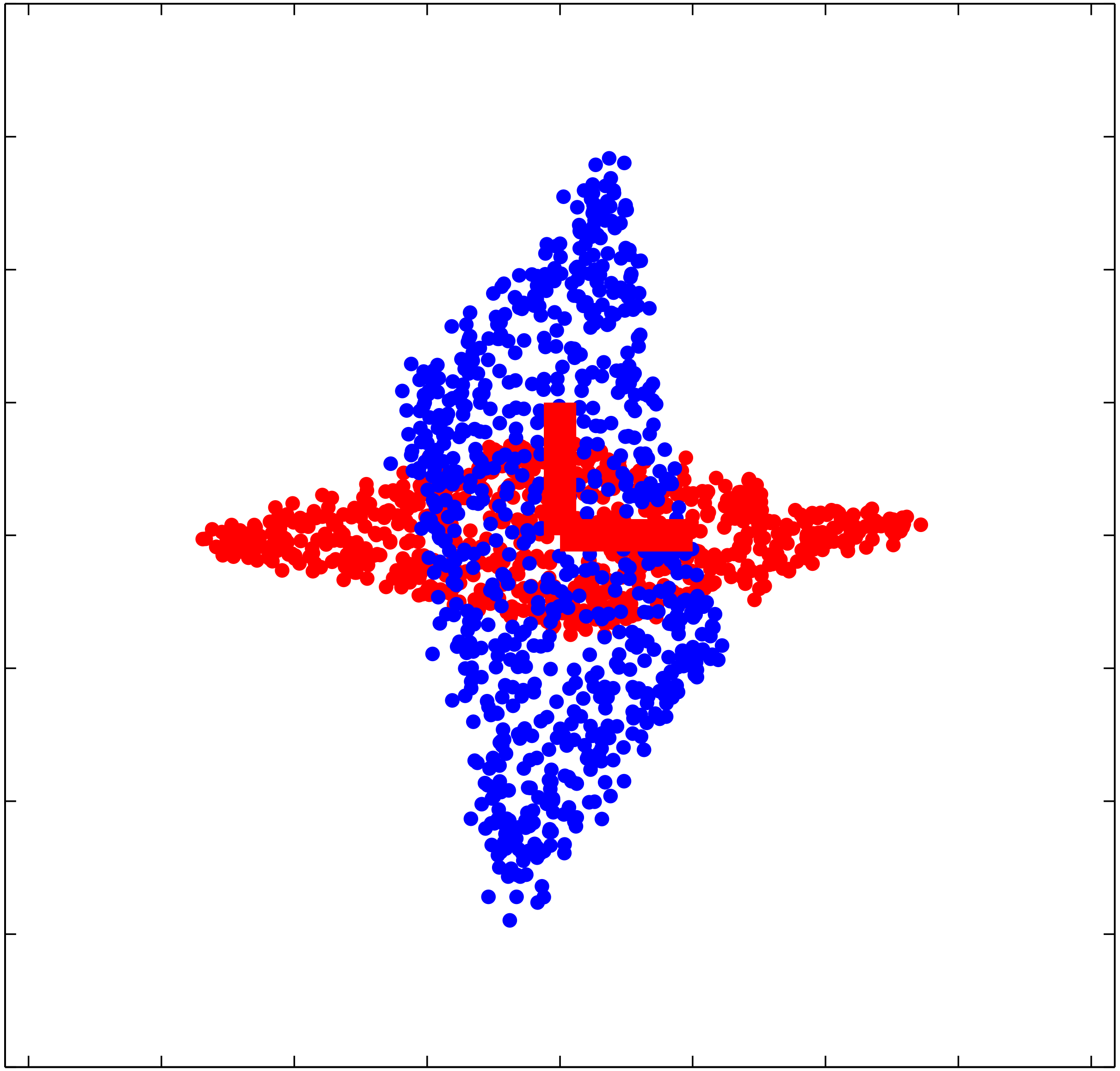




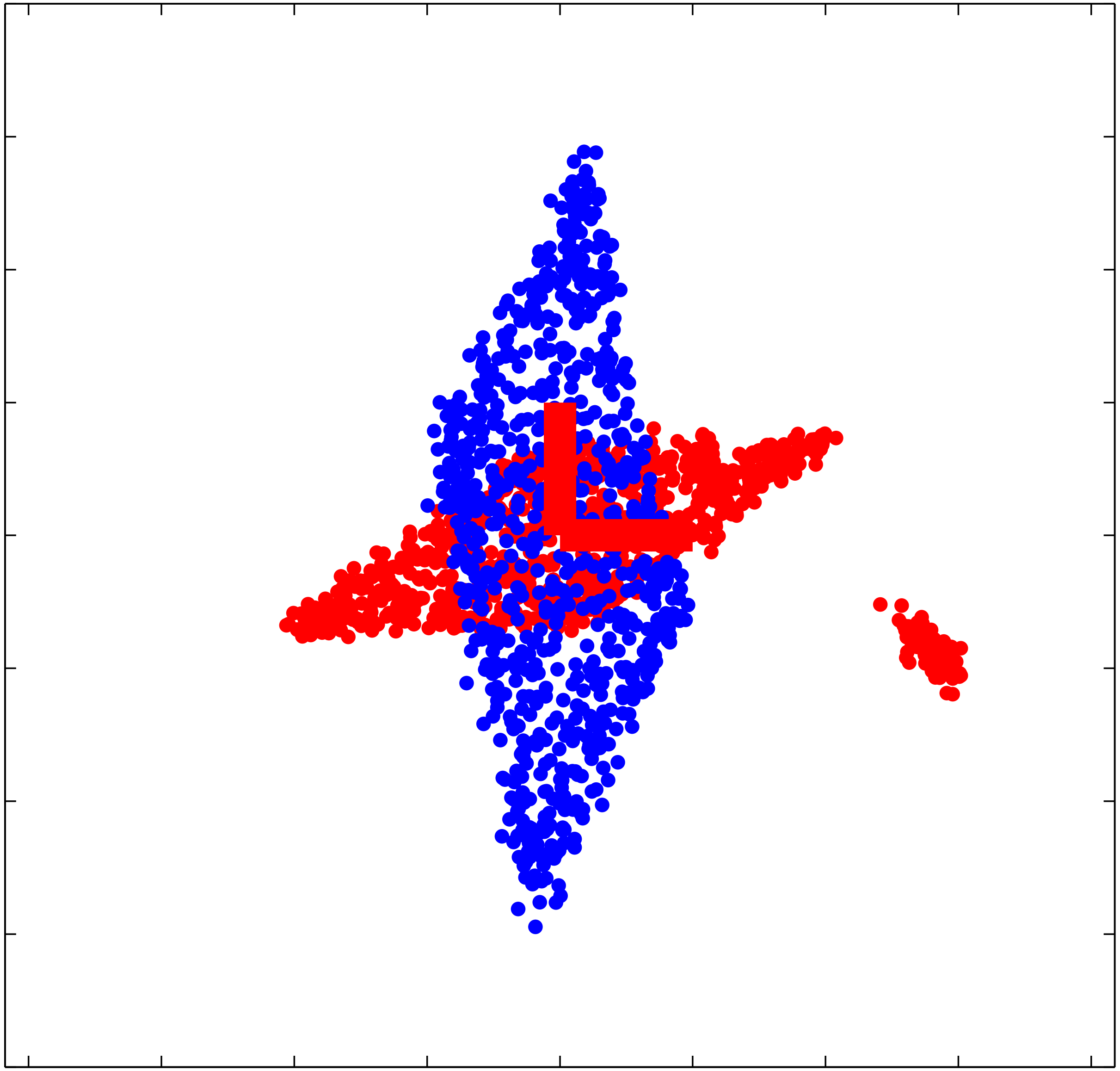
CSP: outlier- (artifact-) sensitivity



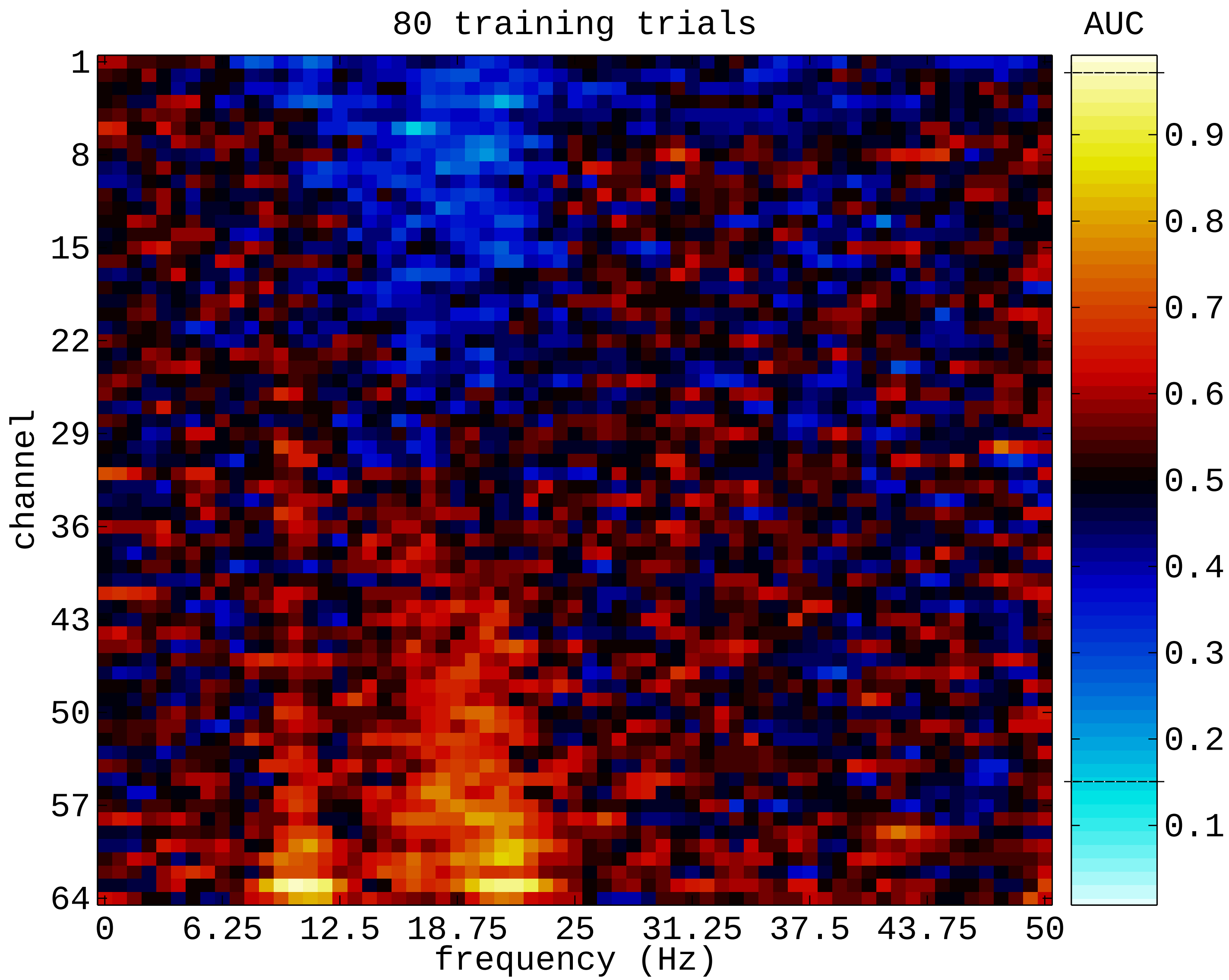
CSP: outlier- (artifact-) sensitivity



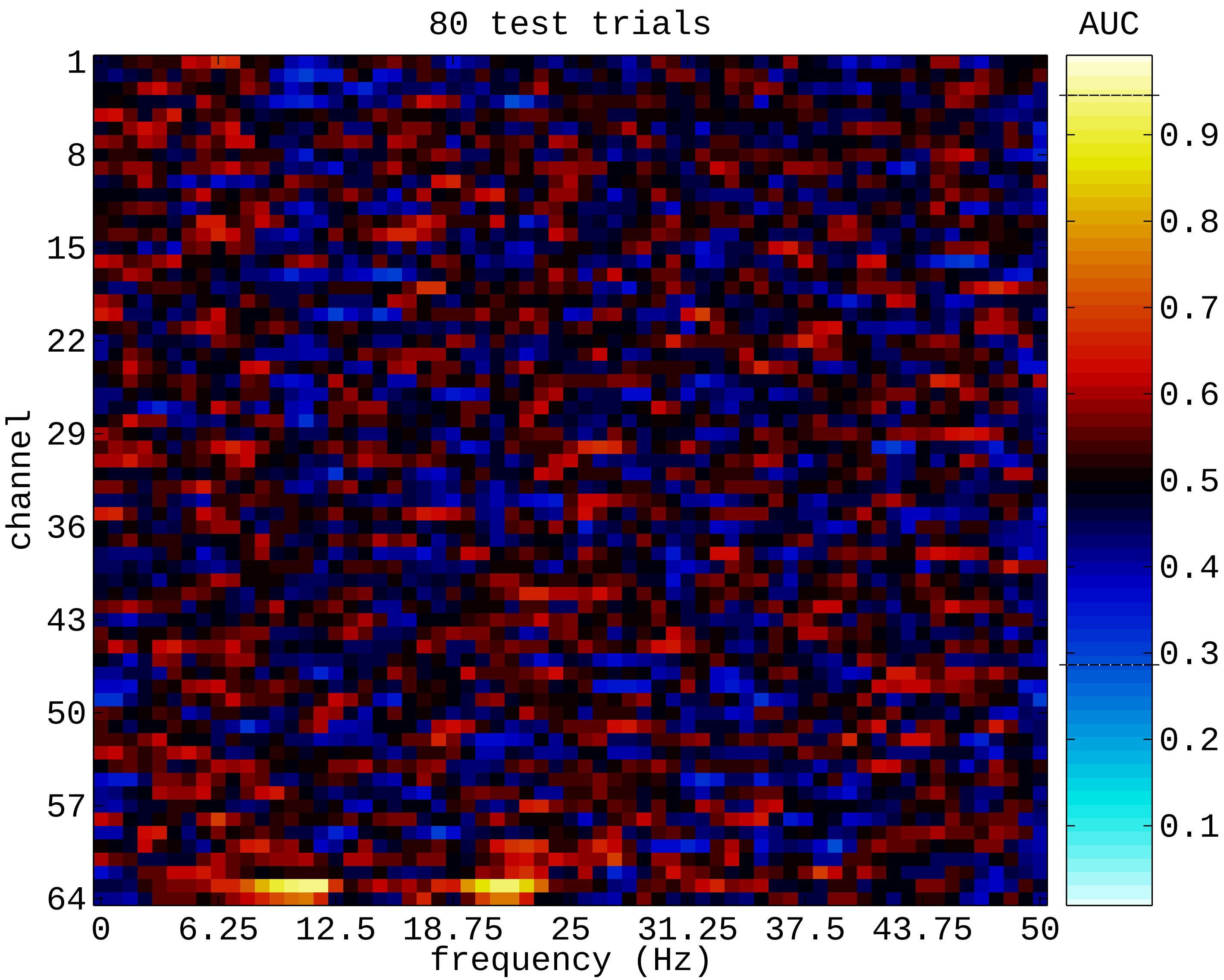
CSP: outlier- (artifact-) sensitivity



CSP: overfitting



CSP: overfitting



Problems with CSP

- Outlier-sensitivity, overfitting (due to poor objective)
- How to pick which components to use?

Problems with CSP

- Outlier-sensitivity, overfitting (due to poor objective)
- How to pick which components to use?
- Sensitivity to initial assumptions:
  - Which frequency band?
  - Which time window?

Problems with CSP

- Outlier-sensitivity, overfitting (due to poor objective)
- How to pick which components to use?
- Sensitivity to initial assumptions:
  - Which frequency band?
  - Which time window?

The exact frequency of sensory-motor rhythms varies between individuals.

In practice, component / band / time-window selection is often best performed by hand.

Problems with CSP

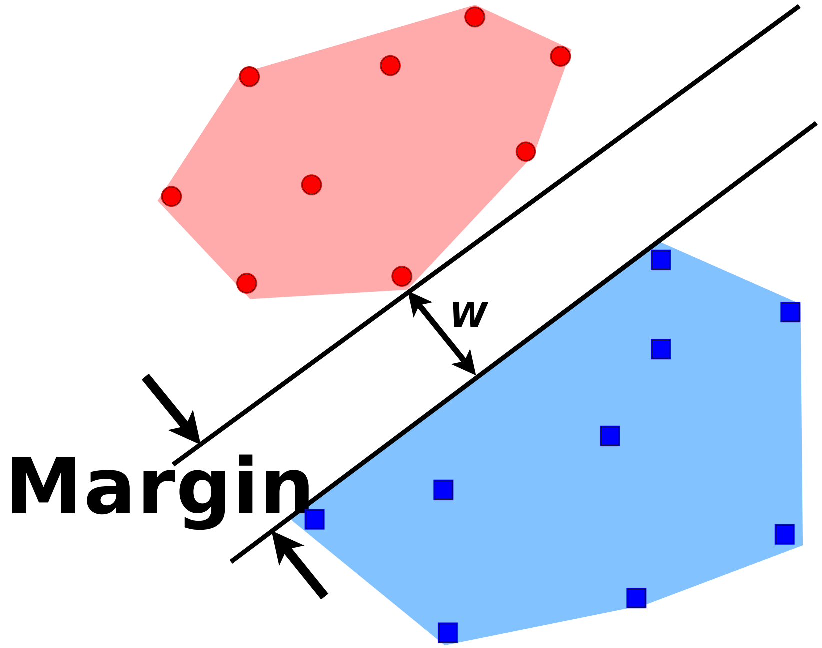
- Outlier-sensitivity, overfitting (due to poor objective)
- How to pick which components to use?
- Sensitivity to initial assumptions:
  - Which frequency band?
  - Which time window?

The exact frequency of sensory-motor rhythms varies between individuals.

In practice, component / band / time-window selection is often best performed by hand. The ideal BCI algorithm would be a “glass box” requiring no such intervention.

Get a better objective (I)

Approach #1: Margin Maximization (à la Support Vector Machine)



Maximize the margin in the space of log bandpower features $\psi(\mathrm{X}; \mathrm{F})$.

$$\psi(\mathrm{X}_i; \mathrm{F}) = \log \operatorname{diag} \left(\mathrm{F} \mathrm{X}_i \mathrm{X}_i^\top \mathrm{F}^\top \right)$$

Get a better objective (I)

Approach #1: Margin Maximization (à la Support Vector Machine)

Given time-series X_i and class labels y_i, simultaneously optimize

- spatial filtering coefficients F
- classifier weight-vector $\mathbf{w}$ in log-bandpower space
- classifier bias b in log-bandpower space

to minimize the SVM-like objective function:

$$\lambda \mathbf{w}^\top \mathbf{w} + \sum_i \max(0,\ 1 - y_i(\psi(\mathrm{X}_i; \mathrm{F})^\top \mathbf{w} + b))$$

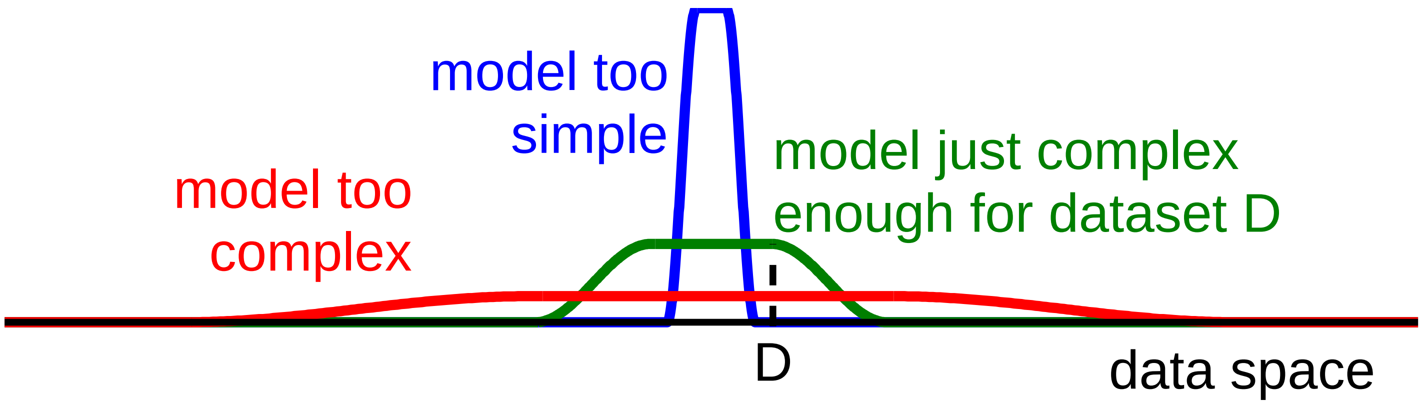
Regularization parameter λ can be found by cross-validation.

Approach #2: "Evidence" Maximization (using Gaussian Process classifiers)

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

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

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



Get a better objective (II)

Approach #2: "Evidence" Maximization (using Gaussian Process classifiers)

- Define a *covariance function* in the log-bandpower space, e.g. a linear covariance function

$$k(\mathrm{X}_i, \mathrm{X}_j) = 1 + \psi(\mathrm{X}_i; \mathrm{F})^\top \psi(\mathrm{X}_j; \mathrm{F})$$

  or some other function of ψ for non-linear classification.

- Plug this into a Gaussian Process Classifier

  (using Probit likelihood, and the Expectation-Propagation algorithm to approximate it—see Kuss & Rasmussen 2005, Journal of Machine Learning Research 6).

- The Gaussian Process framework yields an expression for the evidence, which is easily differentiable with respect to F.
- So optimize F by conjugate gradient descent.

Experiments

Both methods were tested on motor-imagery EEG data from 15 subjects:

- 9 from BCI competitions (Comp 2:IIa, Comp 3:IVa,IVc)
- 6 recorded at the MPI (Lal et al 2004, IEEE Trans. Biomed. Eng. 51)

Experiments

Both methods were tested on motor-imagery EEG data from 15 subjects:

- 9 from BCI competitions (Comp 2:IIa, Comp 3:IVa,IVc)
- 6 recorded at the MPI (Lal et al 2004, IEEE Trans. Biomed. Eng. 51)

Preprocessing:

- select time-windows 0.5–4 sec after stimulus presentation
- band-pass filtered in the broad 8–25Hz band.

Experiments

Both methods were tested on motor-imagery EEG data from 15 subjects:

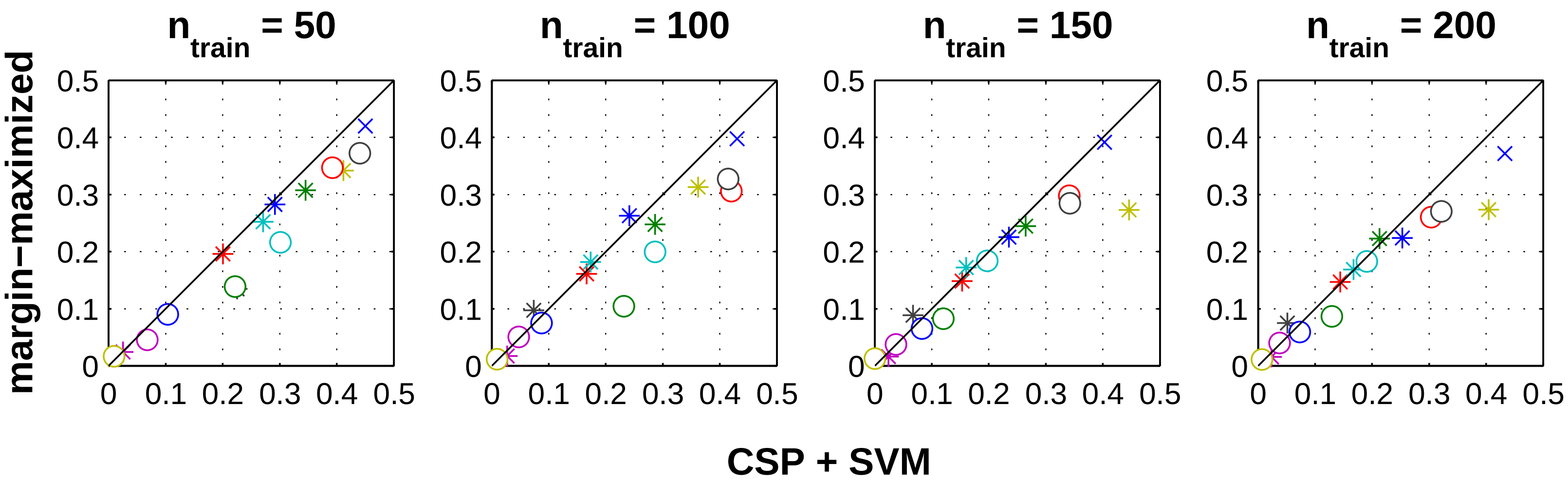
- 9 from BCI competitions (Comp 2:IIa, Comp 3:IVa,IVc)
- 6 recorded at the MPI (Lal et al 2004, IEEE Trans. Biomed. Eng. 51)

Preprocessing:

- select time-windows 0.5–4 sec after stimulus presentation
- band-pass filtered in the broad 8–25Hz band.

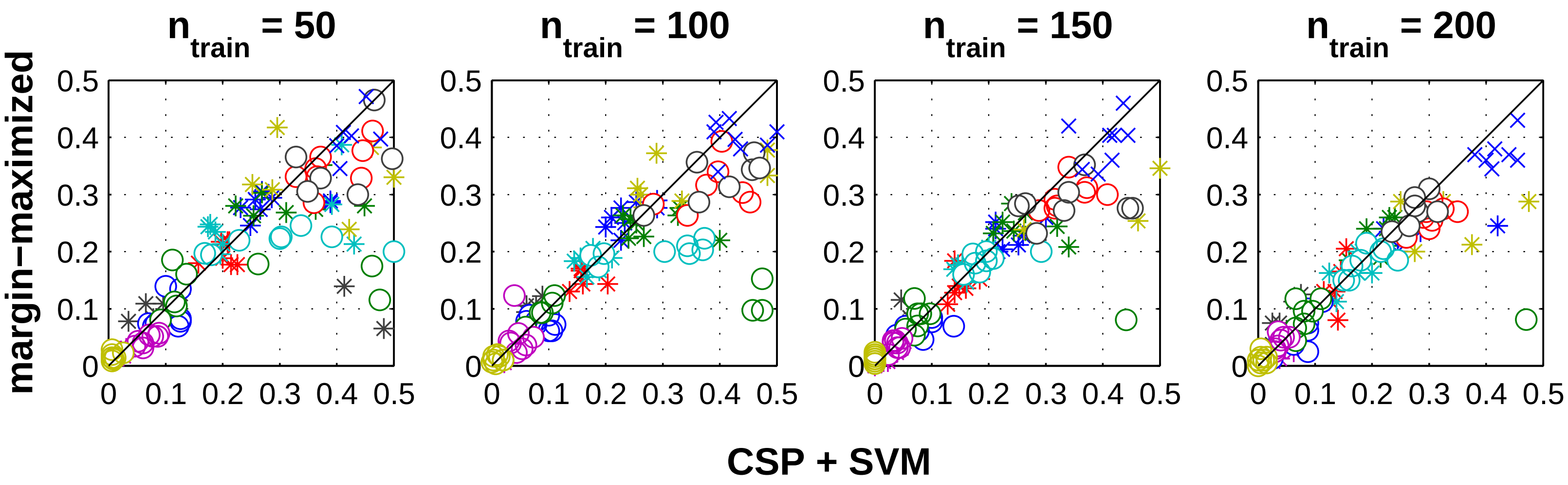
Design:

- Two spatial filters were optimized in each case.
- Performance was assessed as a function of training set size, $n_{\text{train}} \in \{50, 100, 150, 200\}$.
- Each assessment was repeated using 8 random training subsets.



Binary classification error rates:
new approach vs. traditional two-stage CSP + classifier approach.

Results (I)

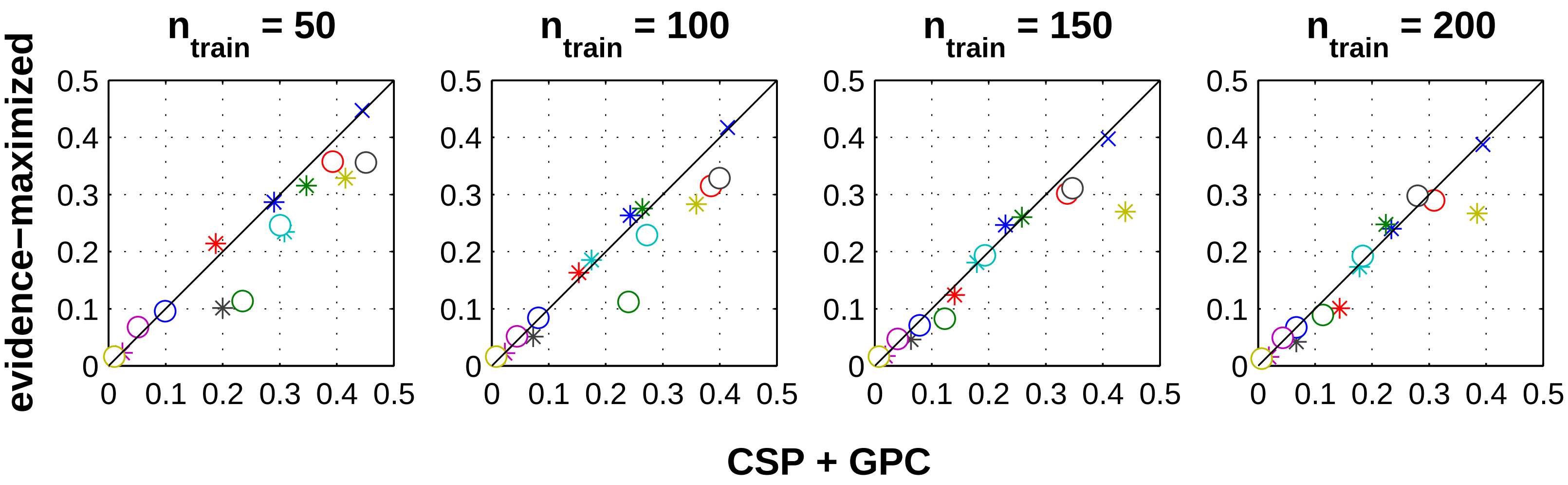


Note the consistent improvement, most markedly when we have:

- poor subject performance;
- small numbers of training trials.

This is encouraging from a clinical viewpoint.

Results (II)

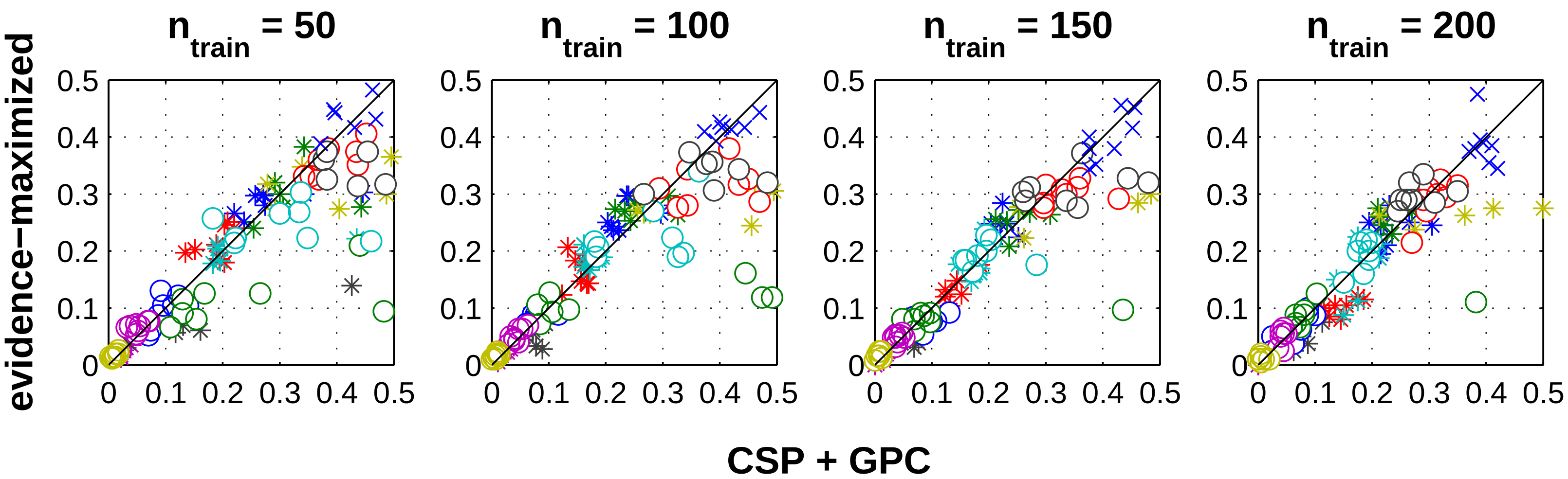


Note the consistent improvement, most markedly when we have:

- poor subject performance;
- small numbers of training trials.

This is encouraging from a clinical viewpoint.

Results (II)

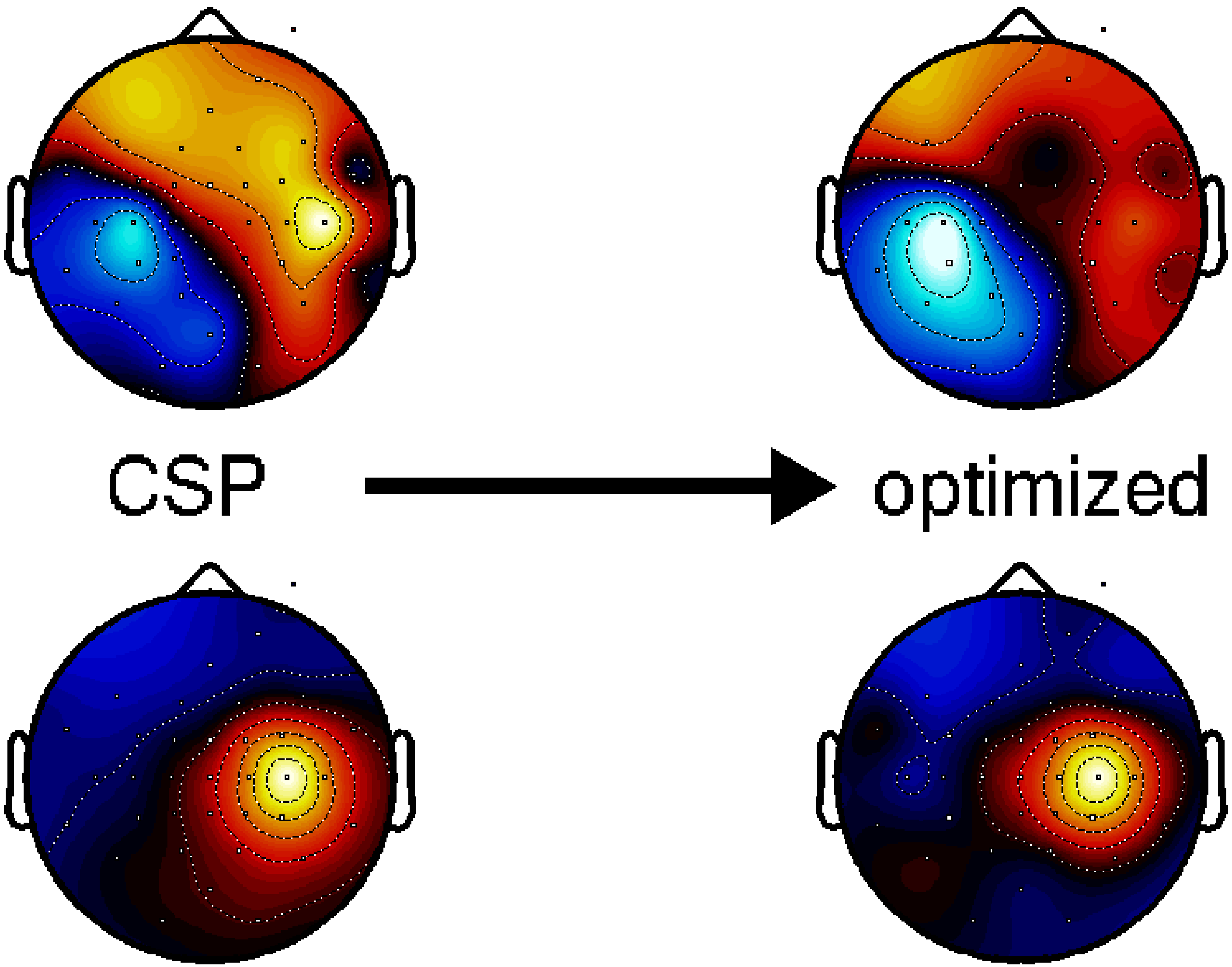


Note the consistent improvement, most markedly when we have:

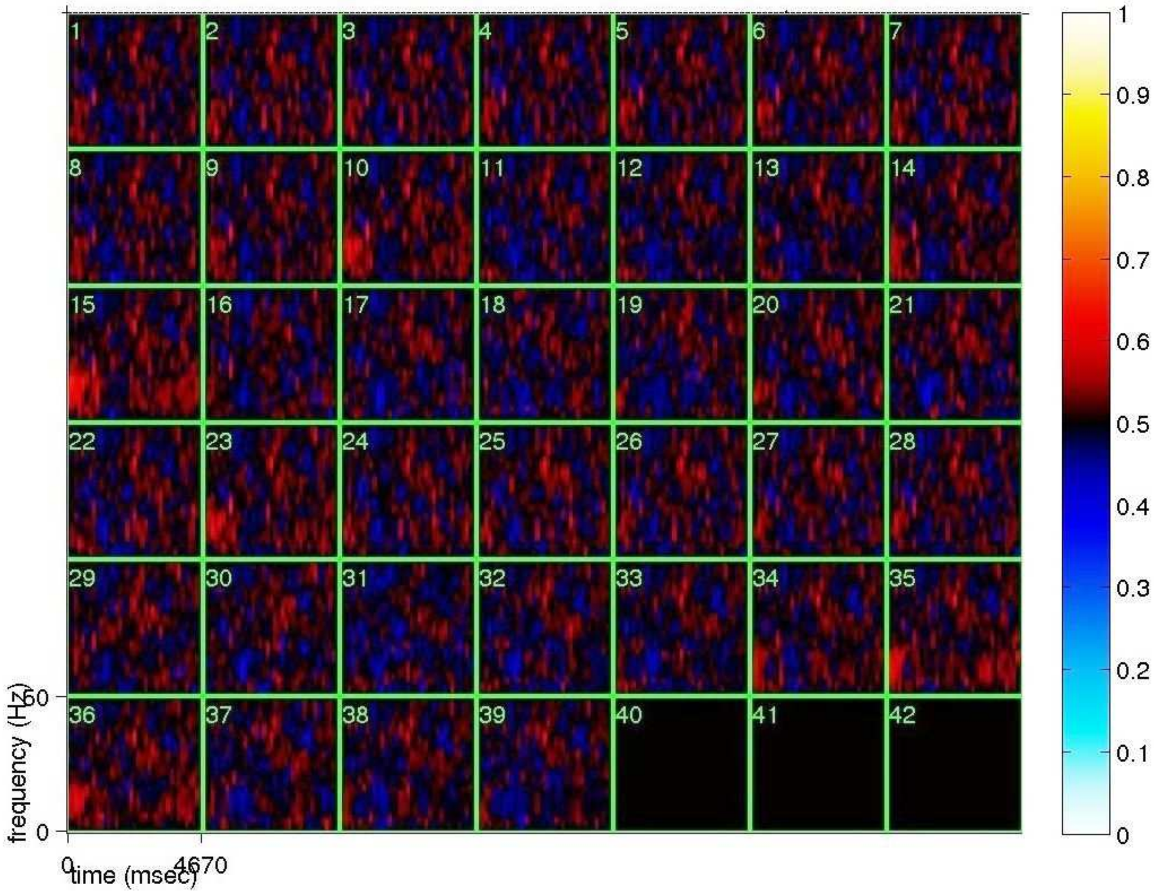
- poor subject performance;
- small numbers of training trials.

This is encouraging from a clinical viewpoint.

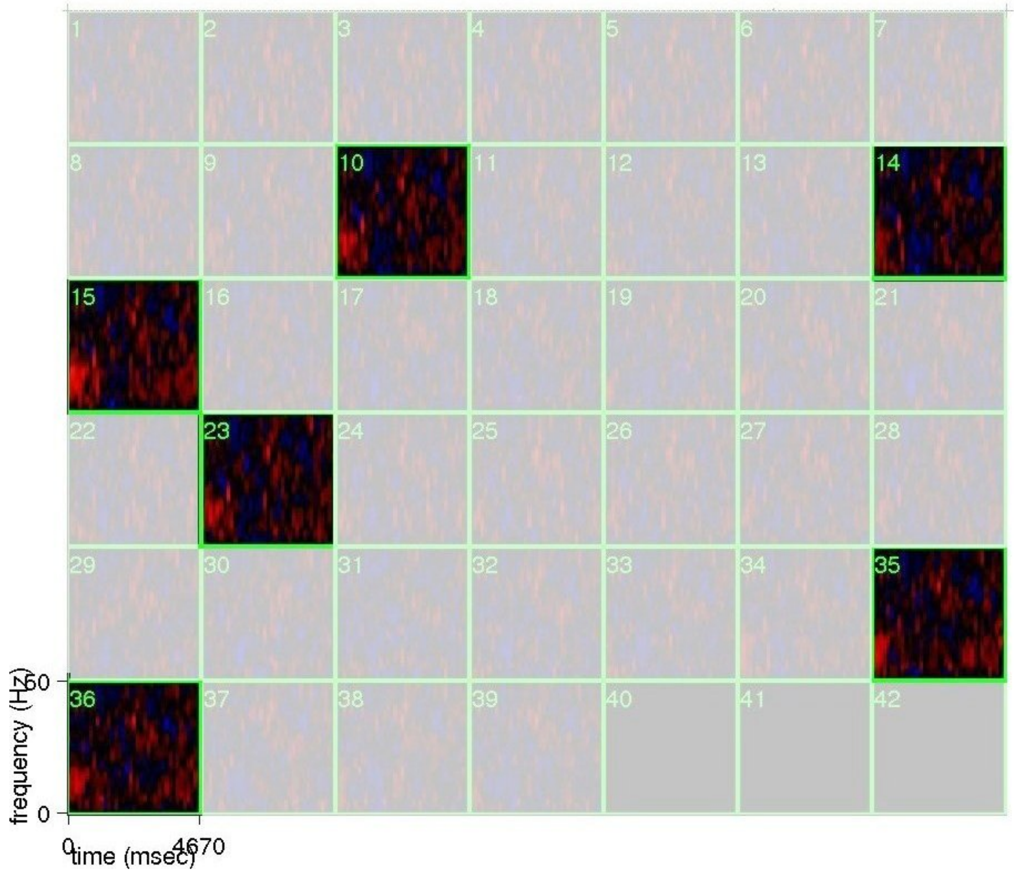
Example of spatial patterns “fixed” by evidence-maximization:



Spatial, temporal, spectral. . .

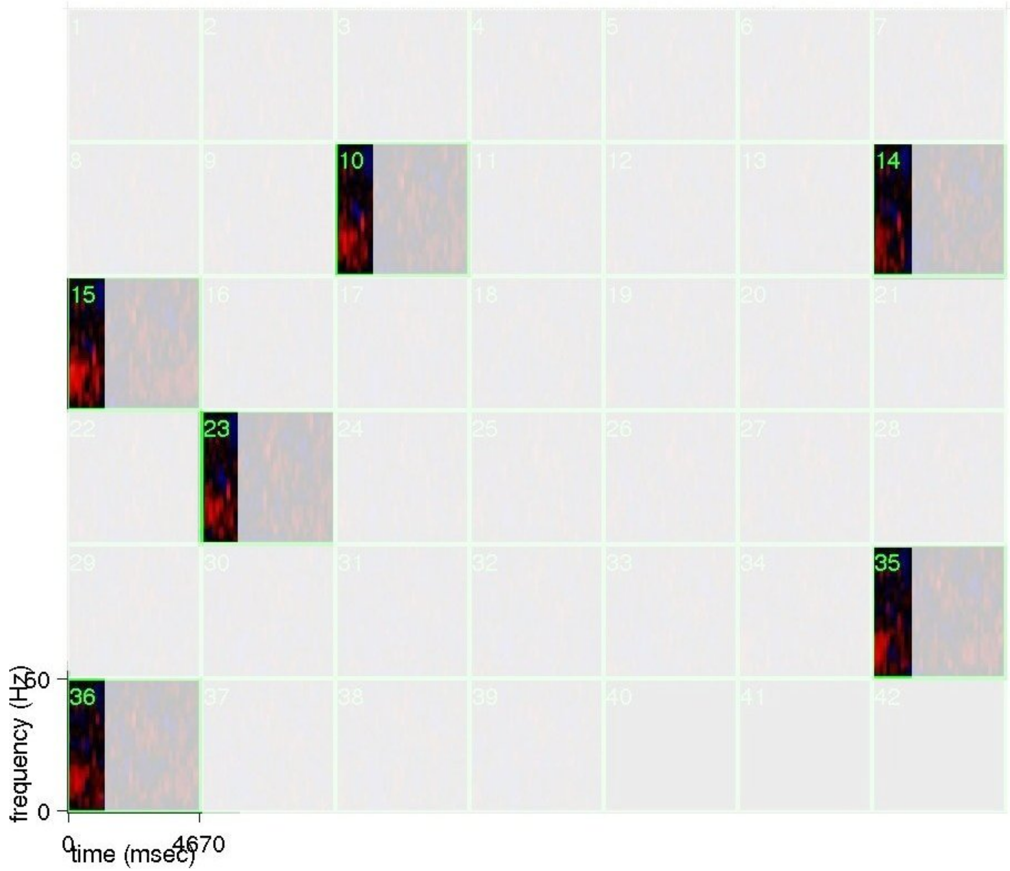


Spatial, temporal, spectral. . .



Ideally we want to optimize automatically over space

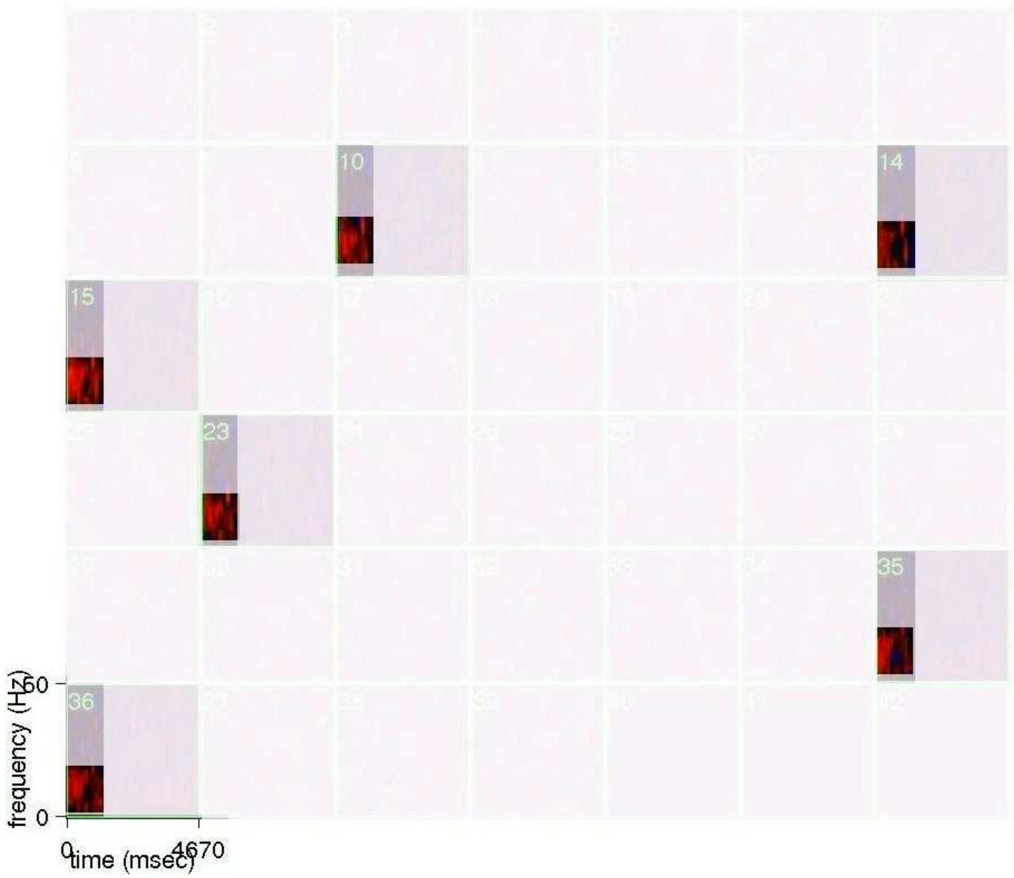
Spatial, temporal, spectral. . .



Ideally we want to optimize automatically over space

time

Spatial, temporal, spectral. . .



Ideally we want to optimize automatically over space
time
frequency

Spatial, temporal, spectral...

Weightings over time or frequency can be incorporated into our feature mapping:

$$\psi(\mathrm{X};\ \mathrm{F}) = \log \operatorname{diag}\left(\mathrm{F}\,\mathrm{X}\,\mathrm{X}^\top\,\mathrm{F}^\top \right)$$

Spatial, temporal, spectral. . .

Weightings over time or frequency can be incorporated into our feature mapping:

$$\psi(\mathrm{X};\ \mathrm{F},\mathrm{G}) = \log\operatorname{diag}\begin{pmatrix} & & \ddots & & \\ \mathrm{F} & \mathrm{X} & \mathrm{G} & \mathrm{X}^\top & \mathrm{F}^\top \\ & & & & \end{pmatrix}$$

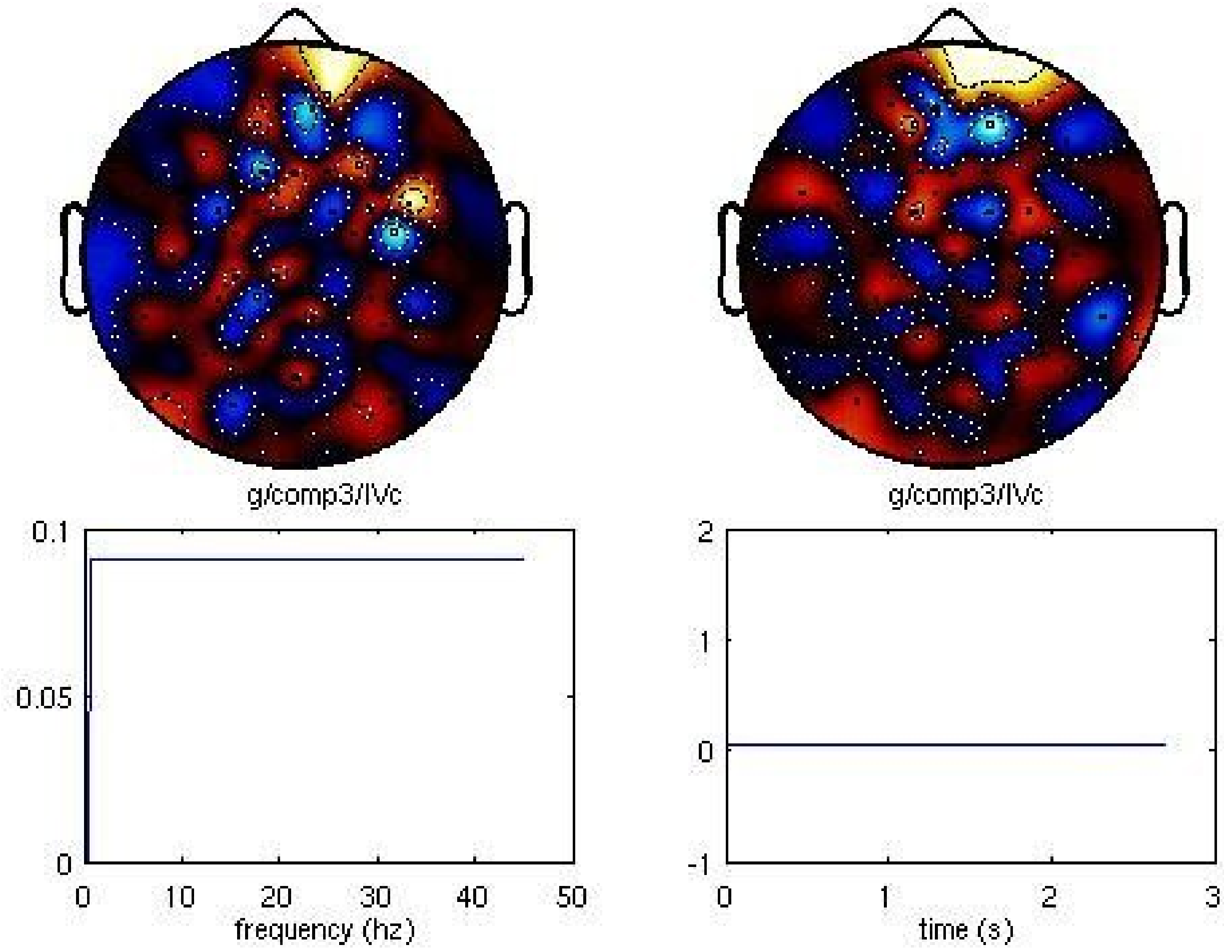
Spatial, temporal, spectral...

Weightings over time or frequency can be incorporated into our feature mapping:

$$\psi(\mathrm{X};\ \mathrm{F}, \mathrm{H}) = \log \operatorname{diag}\left(\mathrm{F}\ \tilde{\mathrm{X}}\ \begin{matrix} \ddots \\ \mathrm{H} \\ \end{matrix}\ \tilde{\mathrm{X}}^{\dagger}\ \mathrm{F}^{\top} \right)$$

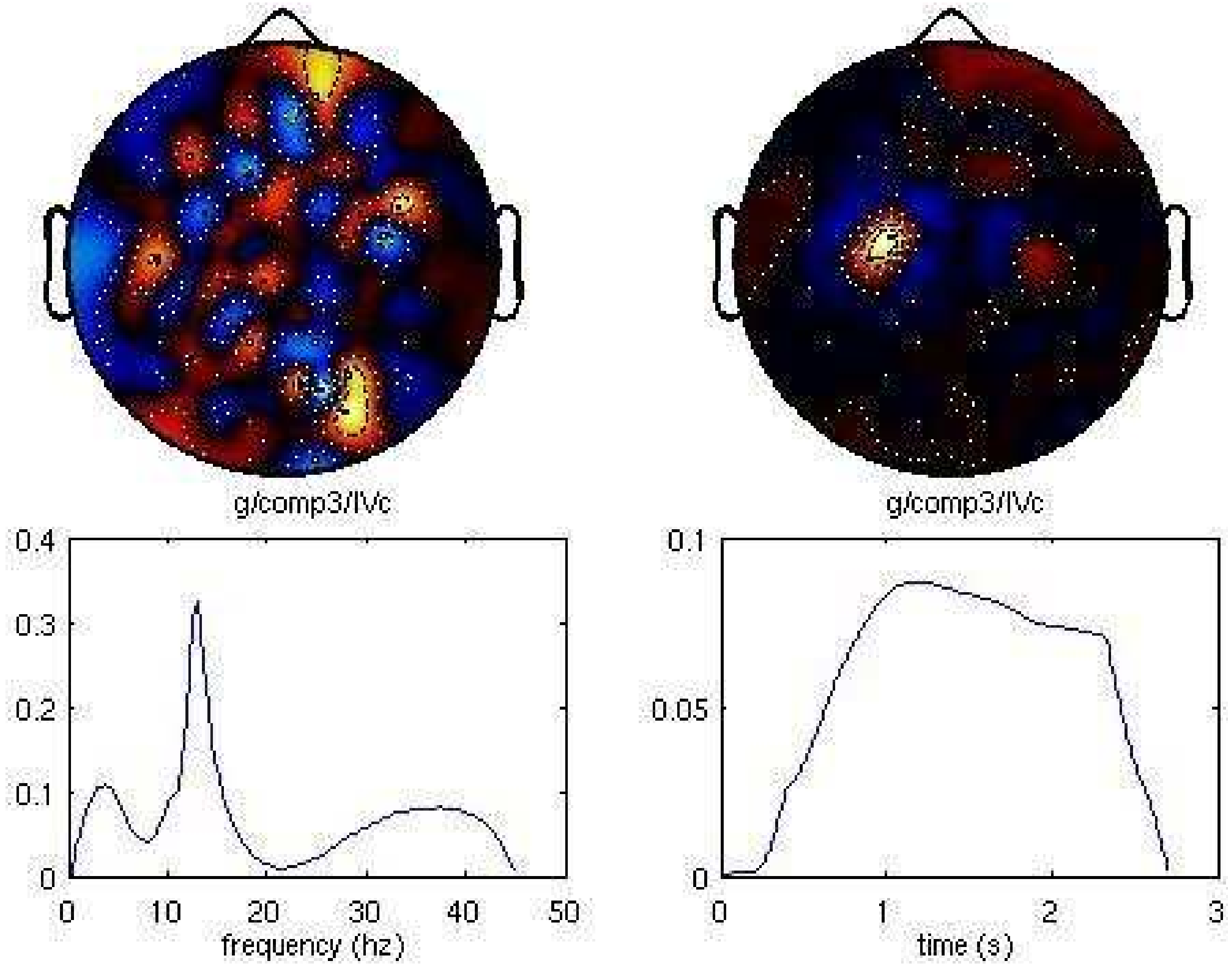
Spatial, temporal, spectral. . .

Preliminary experiments by Jason Farquhar show that iterated optimization of F, then G, then H. . . can yield sensible results with flat initialization over time and frequency, i.e. *without* requiring domain knowledge.



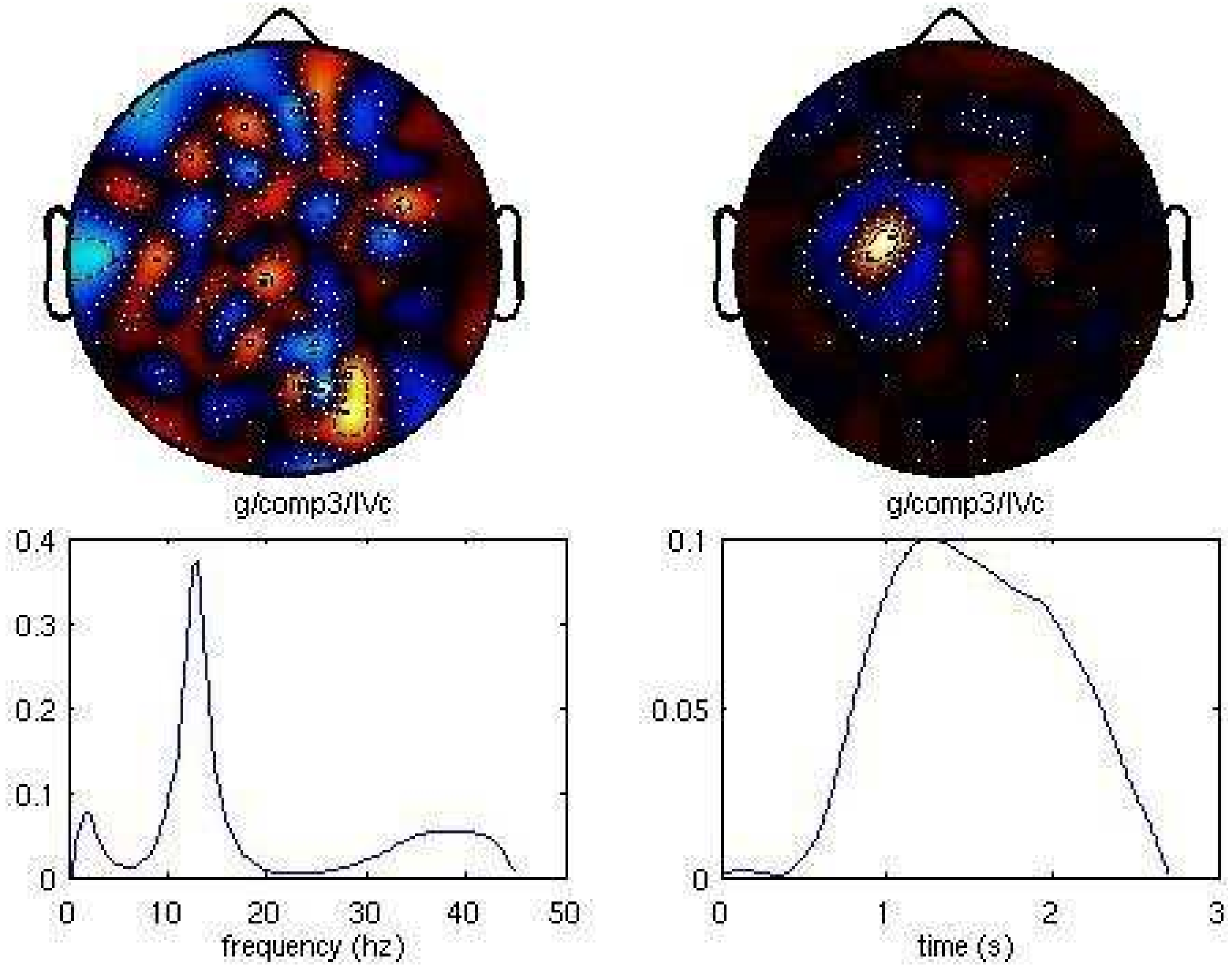
Spatial, temporal, spectral. . .

Preliminary experiments by Jason Farquhar show that iterated optimization of F, then G, then H. . . can yield sensible results with flat initialization over time and frequency, i.e. *without* requiring domain knowledge.



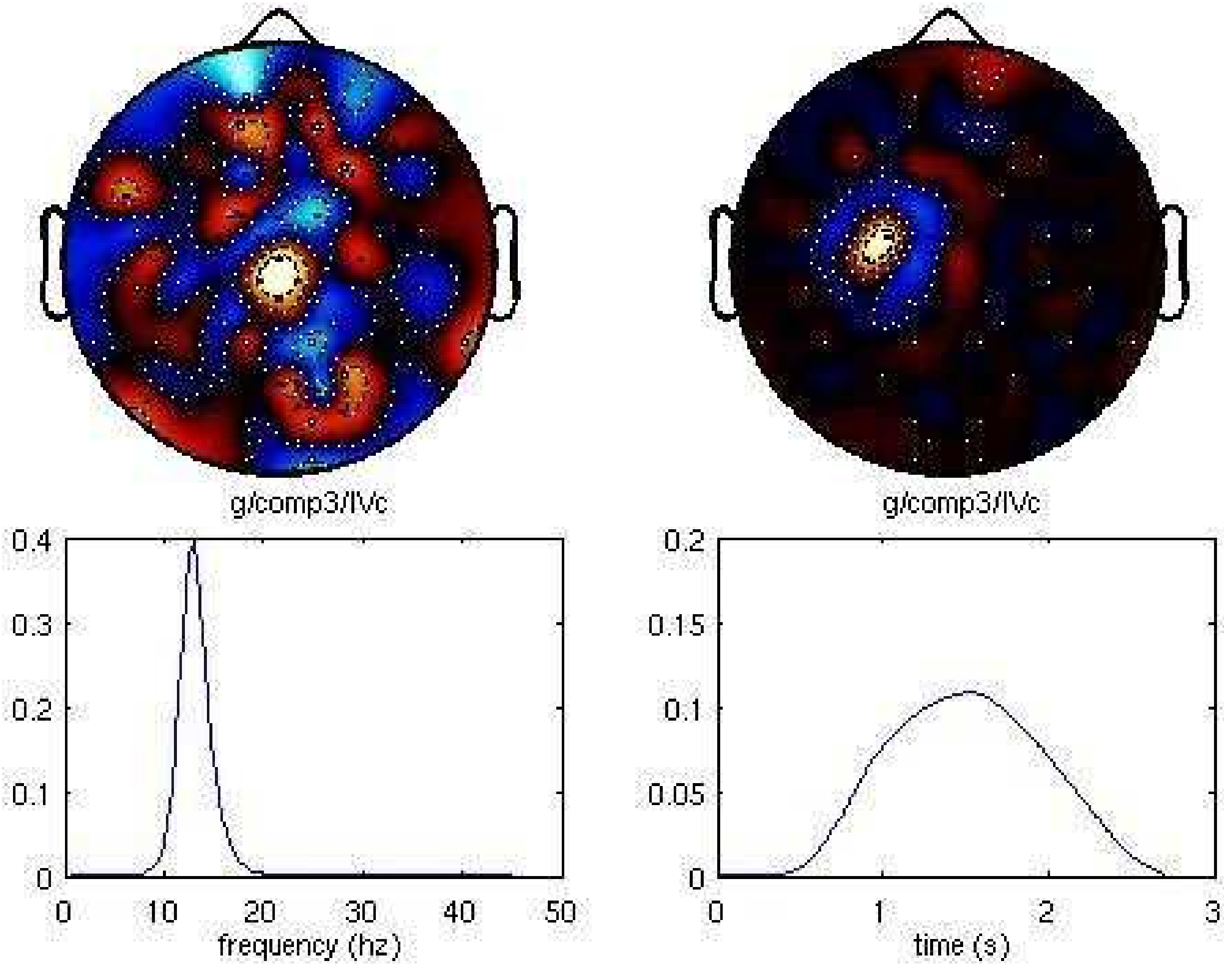
Spatial, temporal, spectral. . .

Preliminary experiments by Jason Farquhar show that iterated optimization of F, then G, then H. . . can yield sensible results with flat initialization over time and frequency, i.e. *without* requiring domain knowledge.



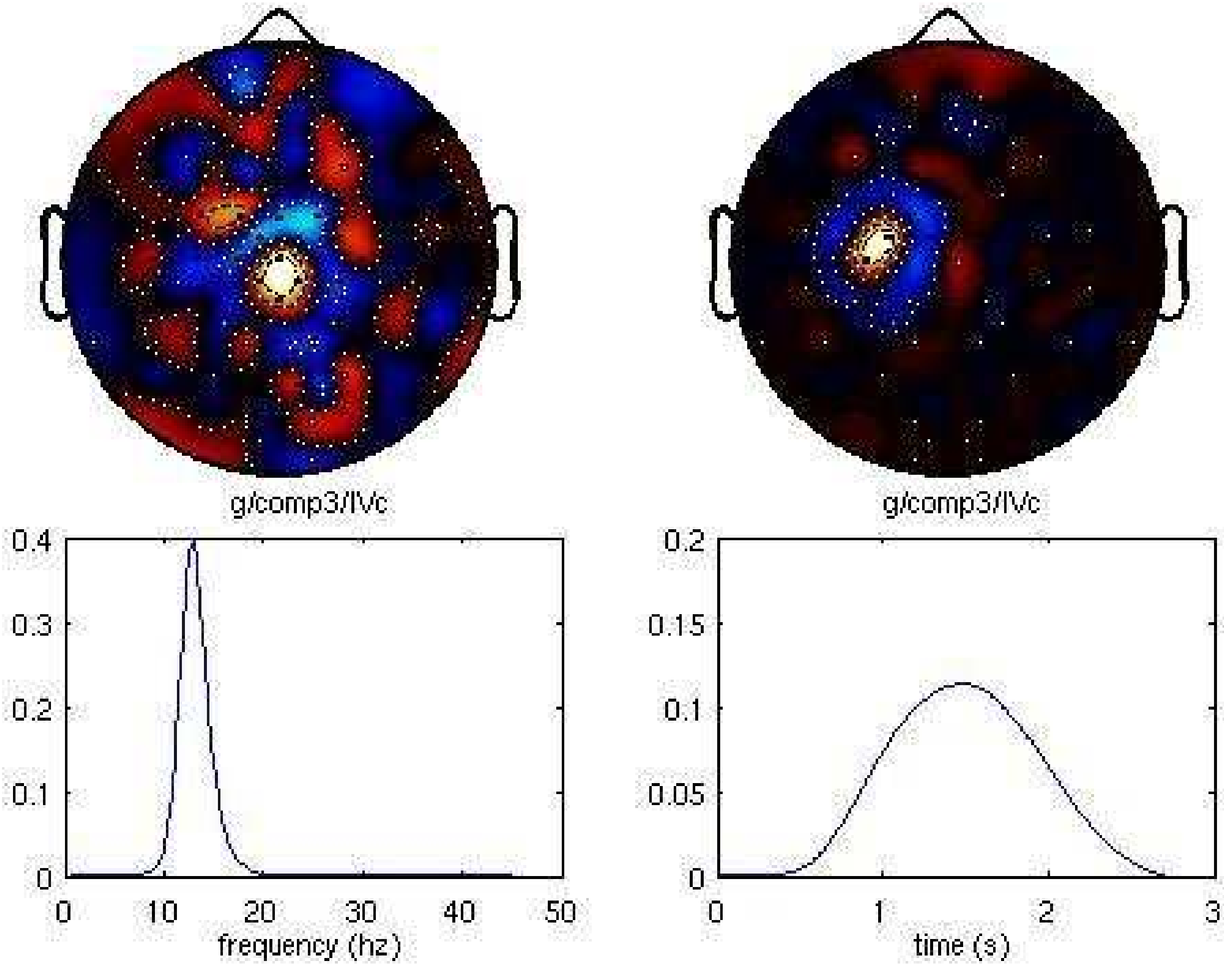
Spatial, temporal, spectral. . .

Preliminary experiments by Jason Farquhar show that iterated optimization of F, then G, then H. . . can yield sensible results with flat initialization over time and frequency, i.e. *without* requiring domain knowledge.



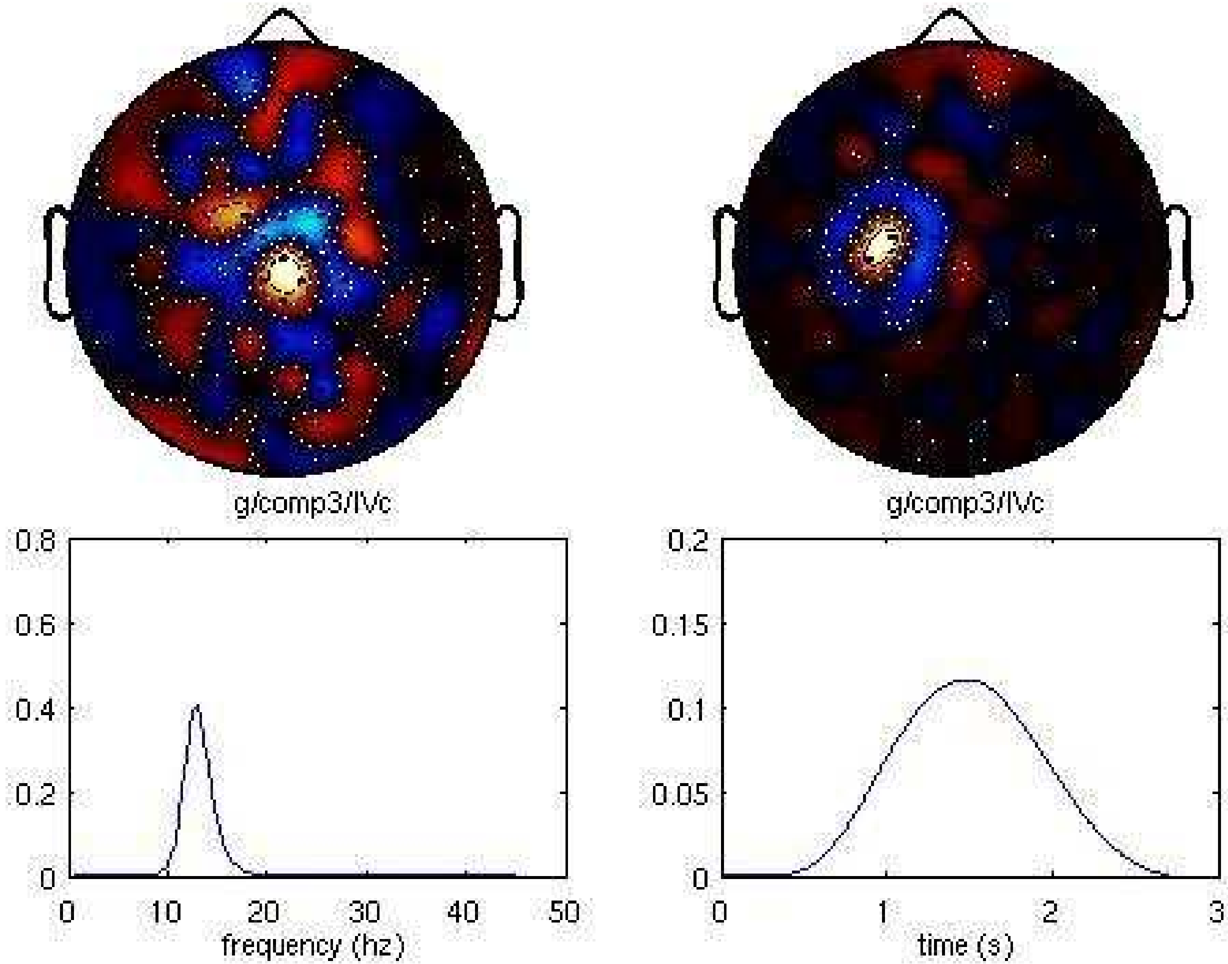
Spatial, temporal, spectral...

Preliminary experiments by Jason Farquhar show that iterated optimization of F, then G, then H. . . can yield sensible results with flat initialization over time and frequency, i.e. *without* requiring domain knowledge.



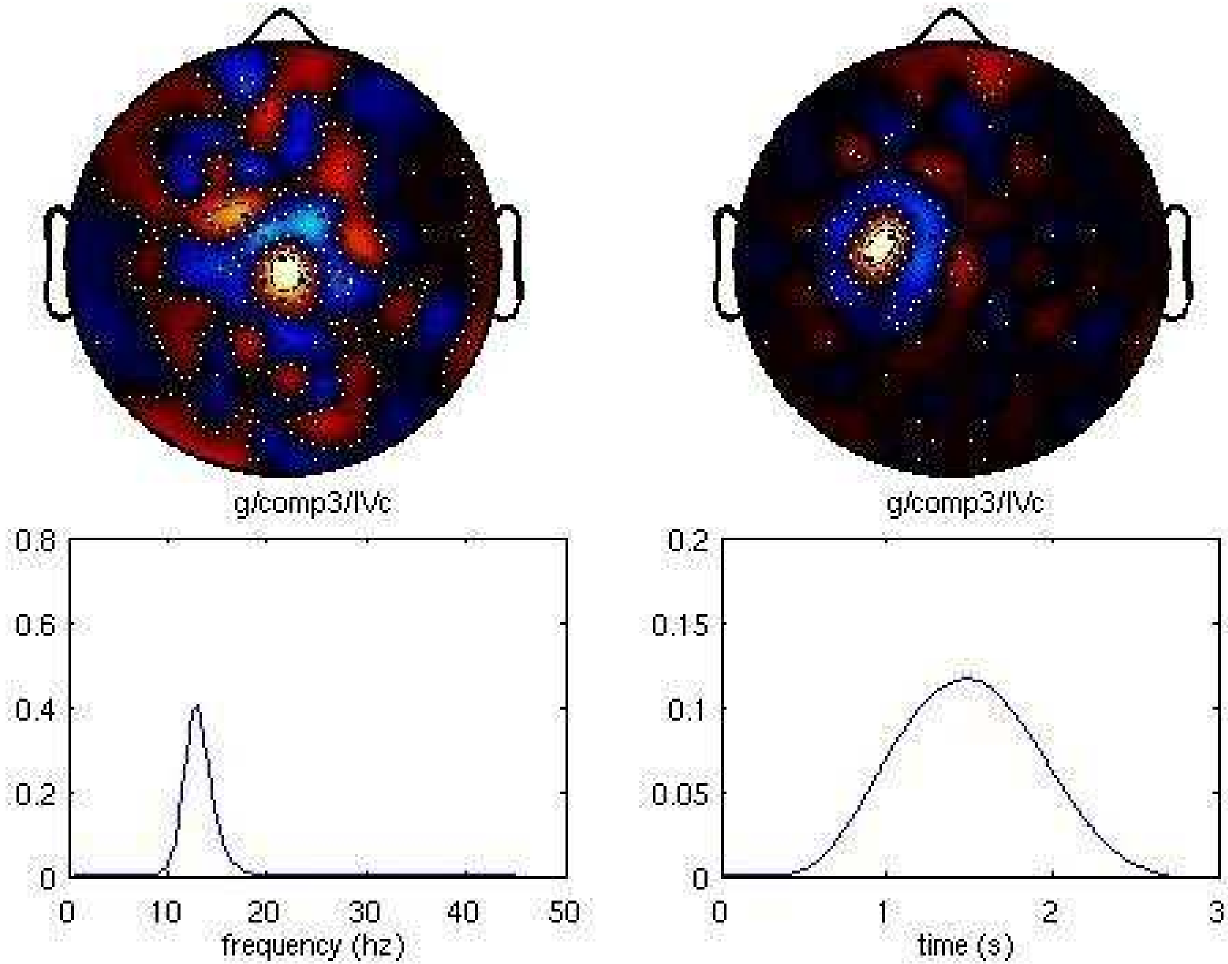
Spatial, temporal, spectral...

Preliminary experiments by Jason Farquhar show that iterated optimization of F, then G, then H... can yield sensible results with flat initialization over time and frequency, i.e. *without* requiring domain knowledge.



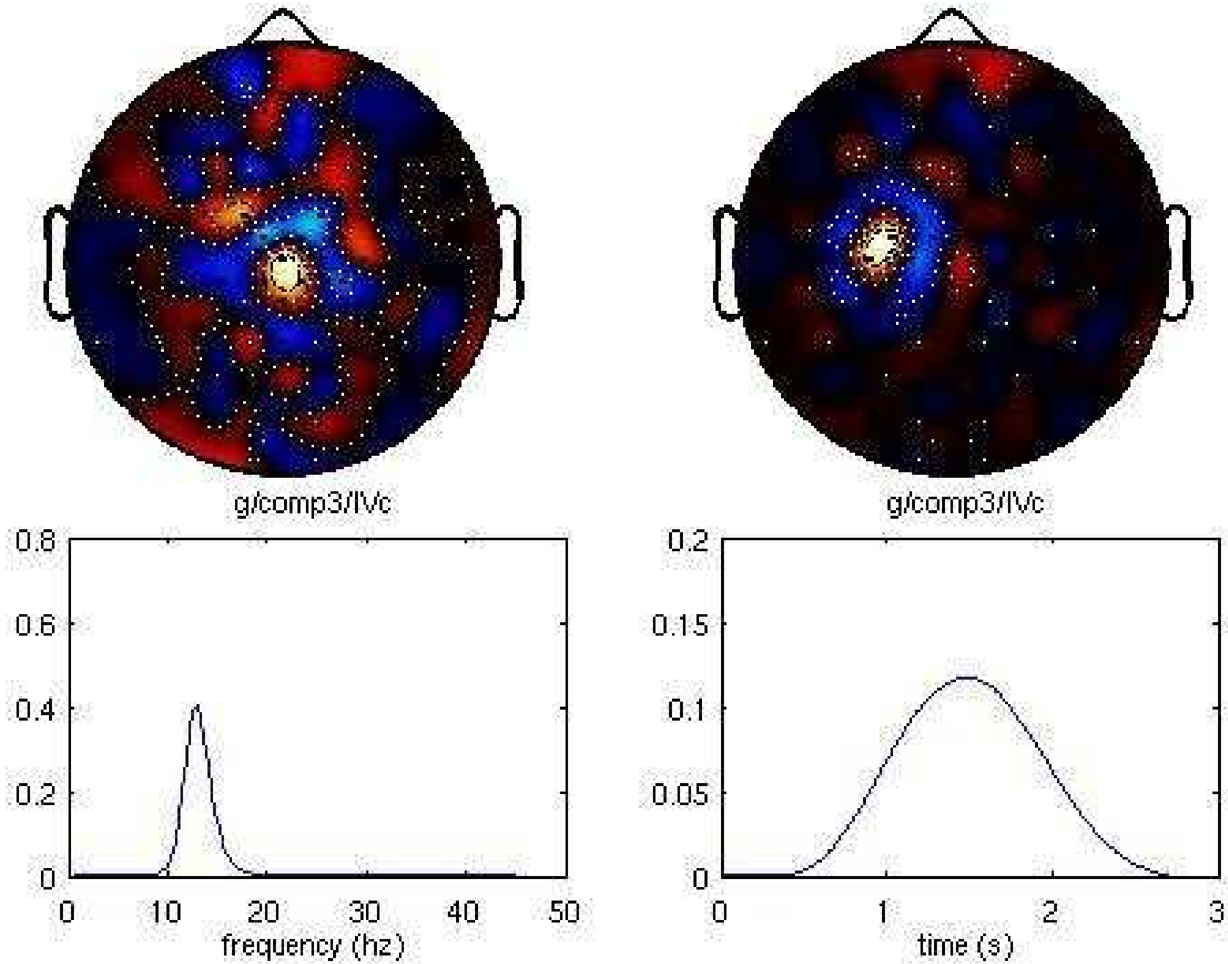
Spatial, temporal, spectral...

Preliminary experiments by Jason Farquhar show that iterated optimization of F, then G, then H... can yield sensible results with flat initialization over time and frequency, i.e. *without* requiring domain knowledge.

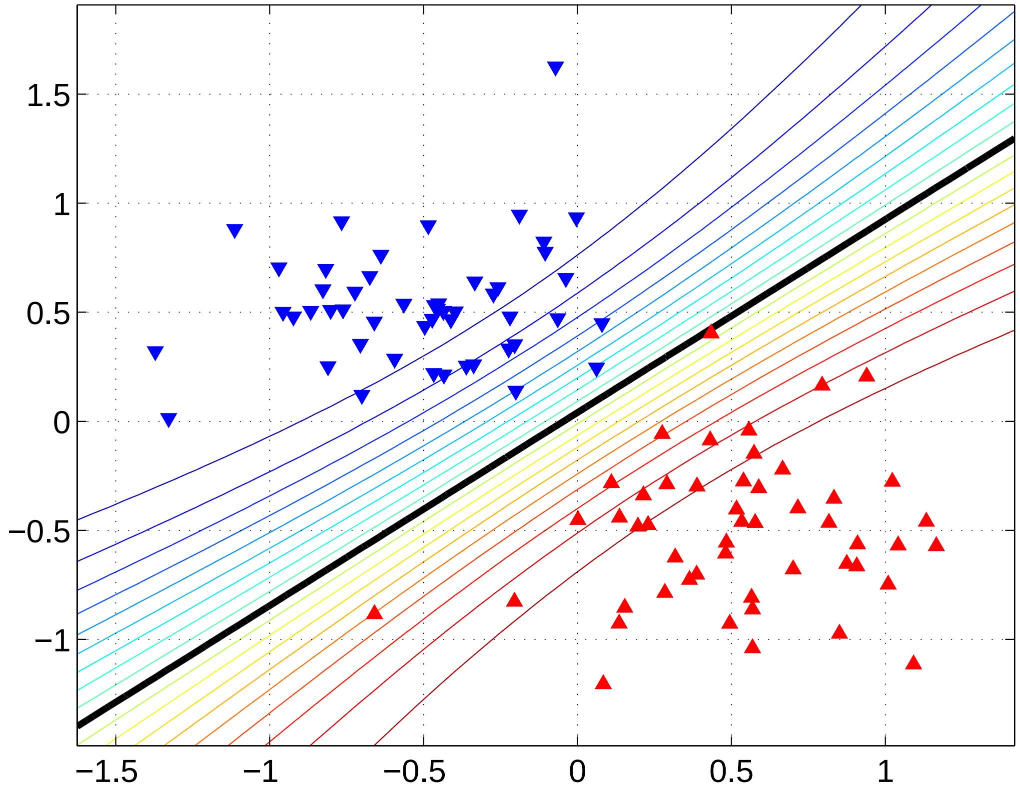


Spatial, temporal, spectral...

Preliminary experiments by Jason Farquhar show that iterated optimization of F, then G, then H... can yield sensible results with flat initialization over time and frequency, i.e. *without* requiring domain knowledge.

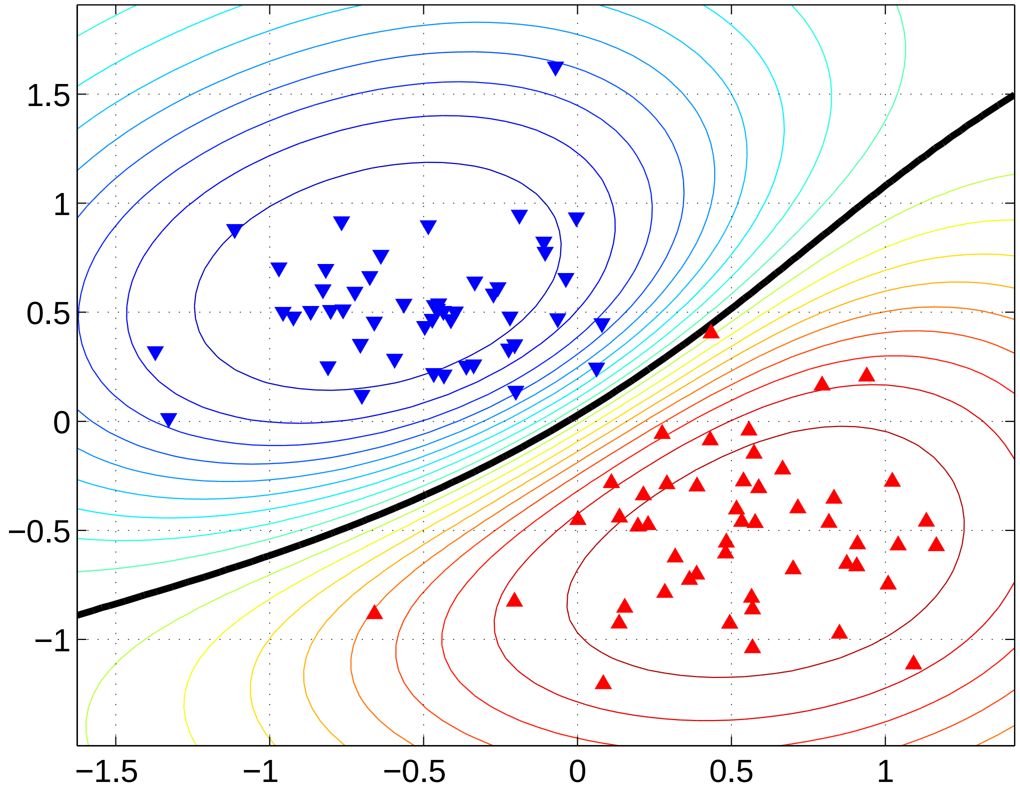


Linear or non-linear?



$$k(\mathrm{X}_i, \mathrm{X}_j) = v \left[1 + \sum_{k=1}^{m} \sigma_k^{-2} \log\left(\mathbf{f}_k^\top \mathrm{X}_i \mathrm{X}_i^\top \mathbf{f}_k\right) \; \log\left(\mathbf{f}_k^\top \mathrm{X}_j \mathrm{X}_j^\top \mathbf{f}_k\right) \right]$$

Linear or non-linear?

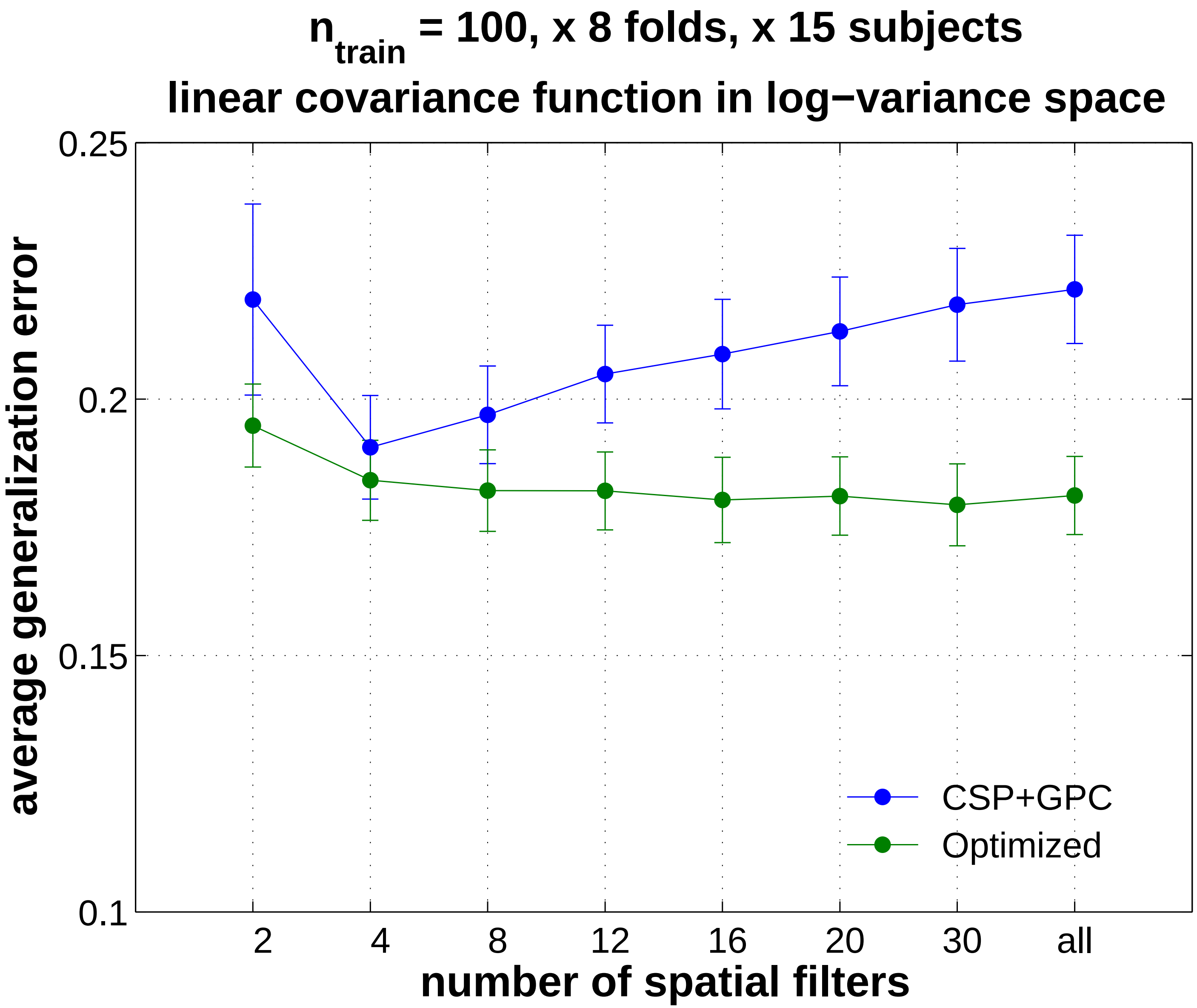


$$k(\mathrm{X}_i, \mathrm{X}_j) = v \exp\left(-\frac{1}{2}\sum_{k=1}^{m} \sigma_k^{-2} d_k^2\right)$$

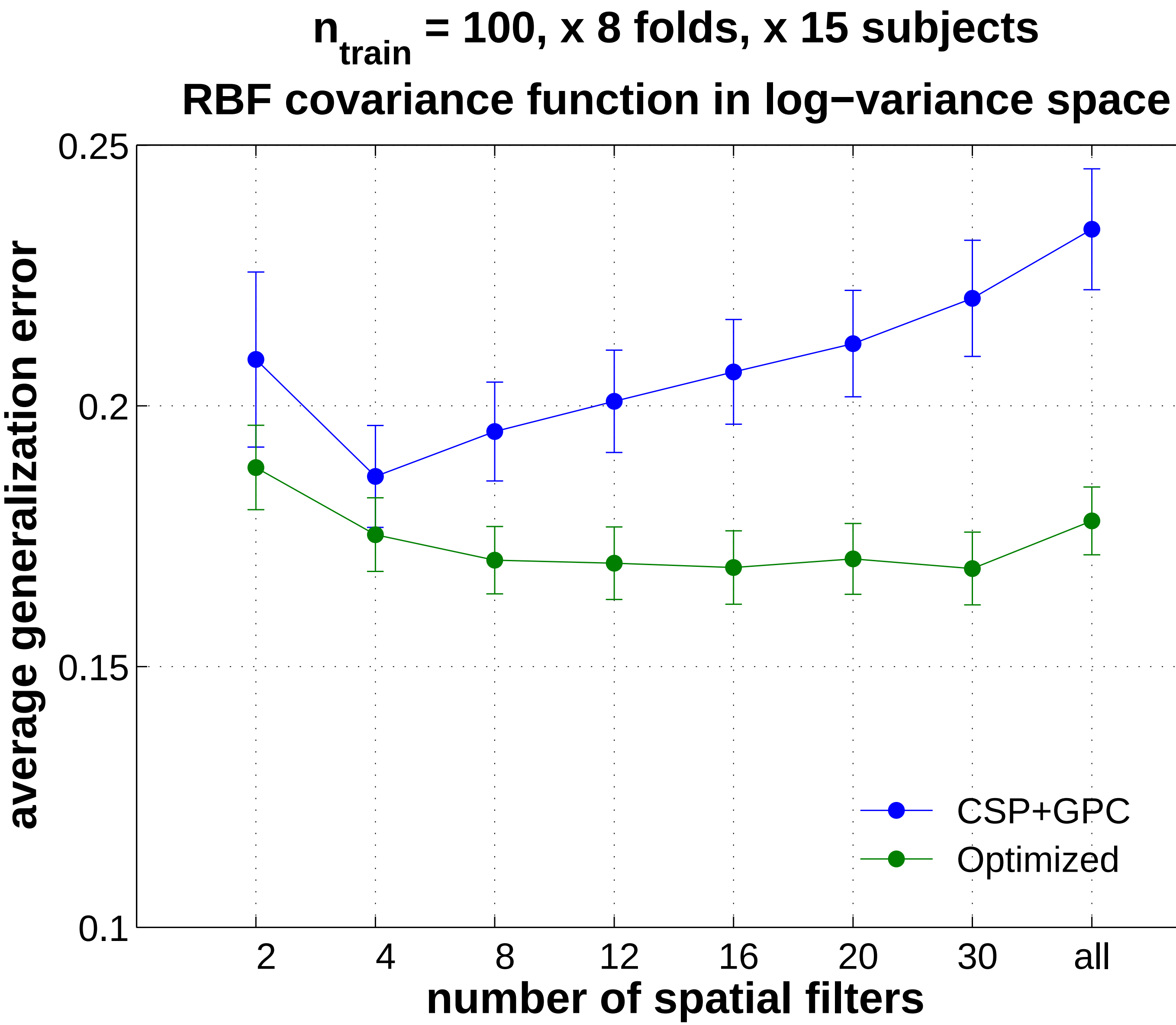
where

$$d_k = \log\left(\mathbf{f}_k^\top \mathrm{X}_i \mathrm{X}_i^\top \mathbf{f}_k\right) - \log\left(\mathbf{f}_k^\top \mathrm{X}_j \mathrm{X}_j^\top \mathbf{f}_k\right)$$

Linear or non-linear?



Linear or non-linear?



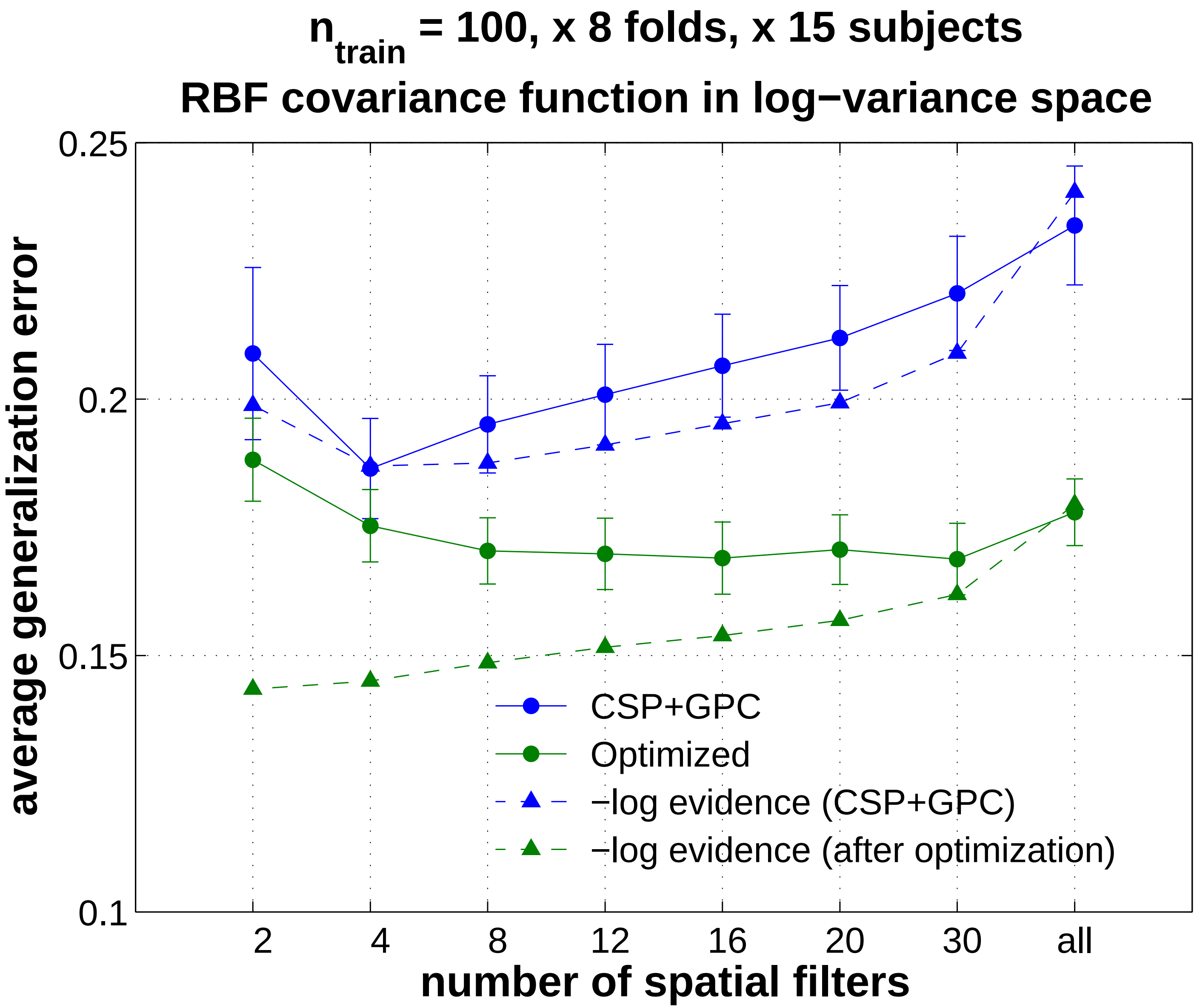
Linear or non-linear?

Use of the classifier's criterion to optimize preprocessing parameters means

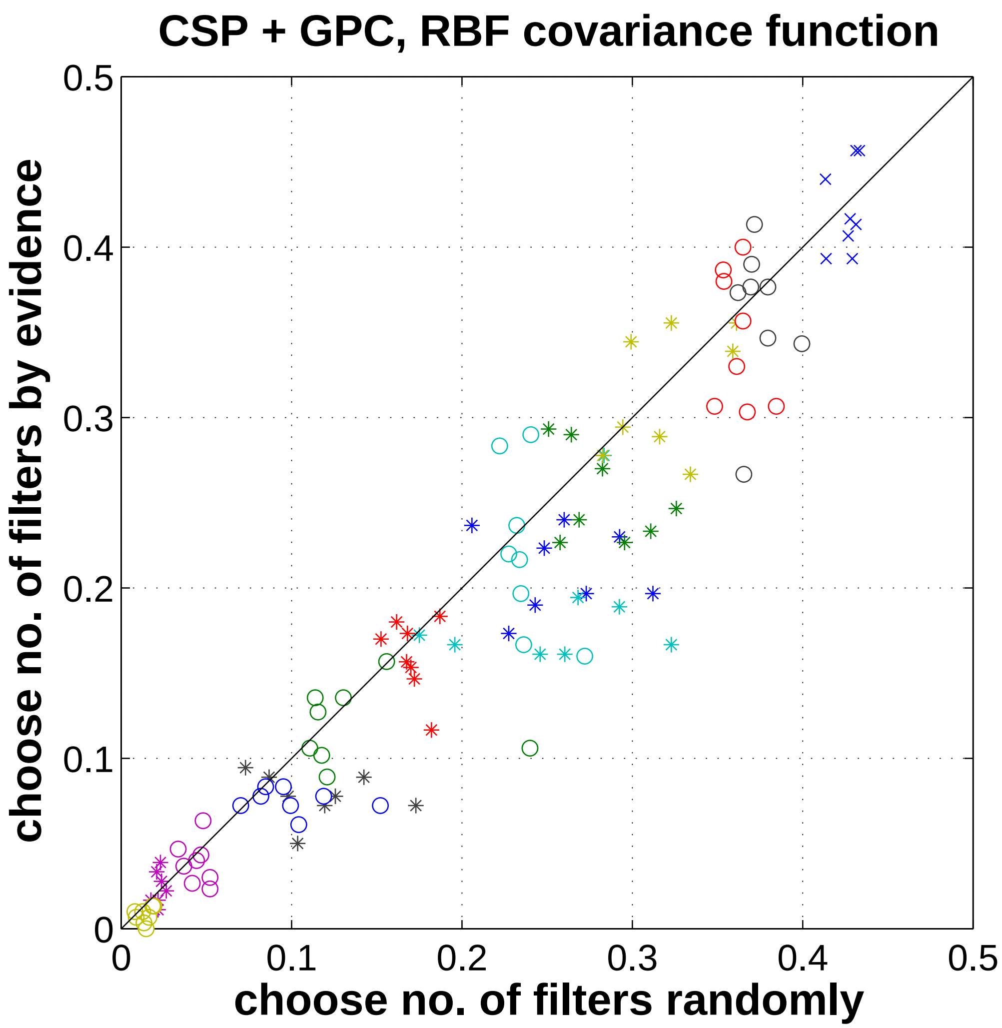
- projection into higher-dimensional feature spaces via a non-linear kernel can help;
- not just "any classifier will do."

See also: Tomioka et al. (NIPS 2006) - logistic regression on (non-logged) variance features.

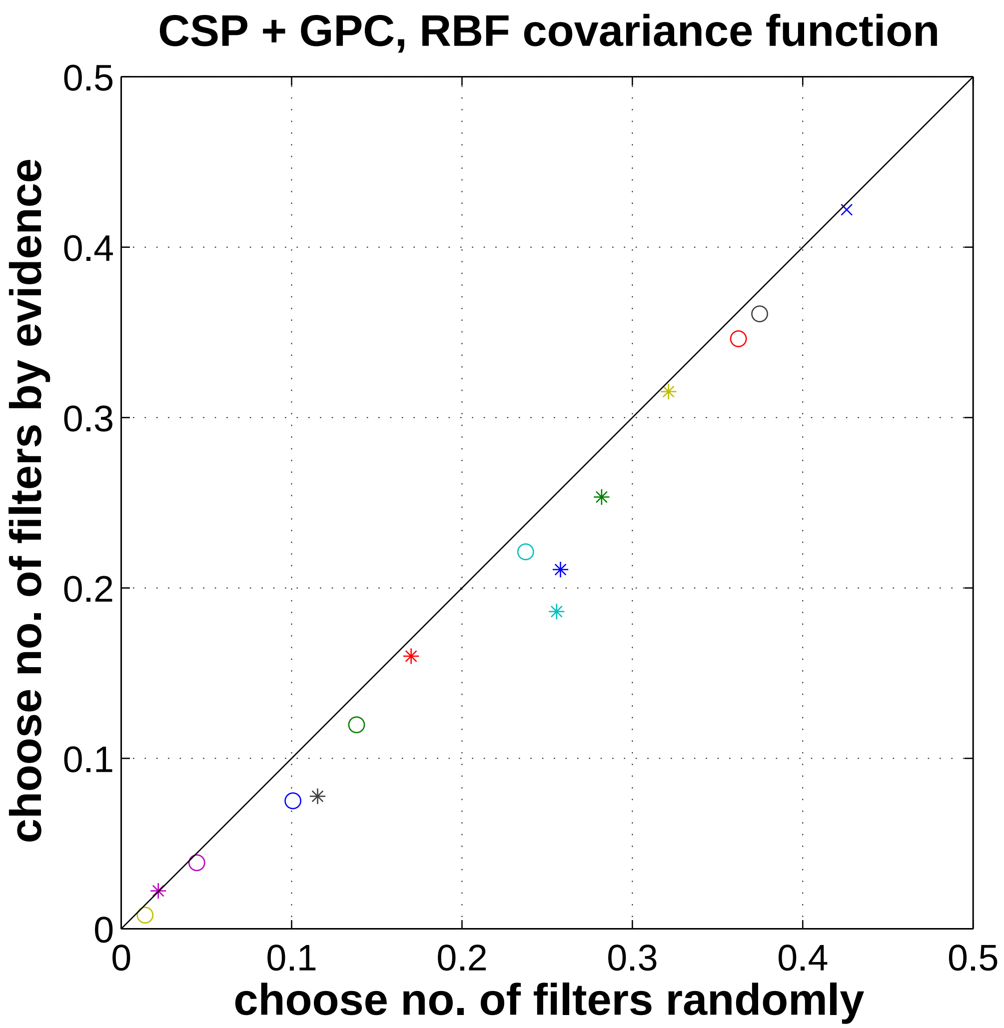
Model selection: how many filters?



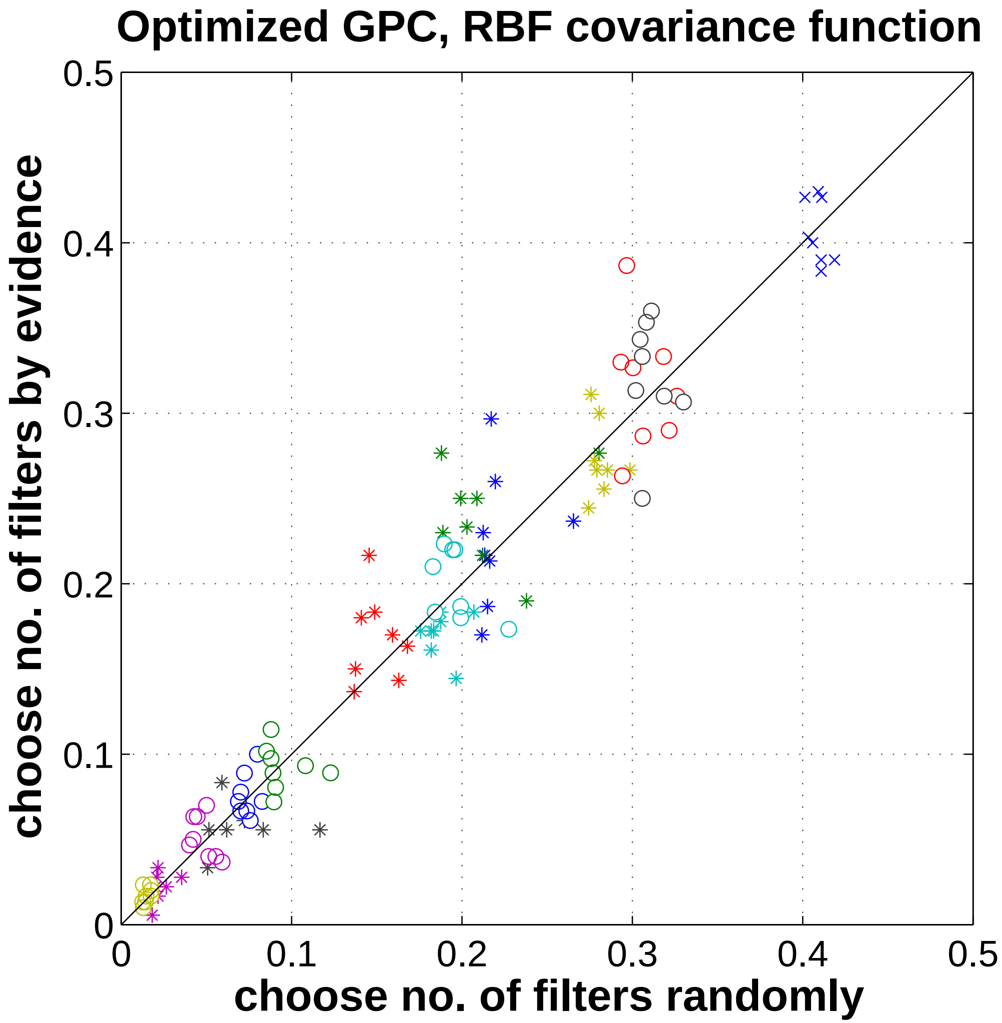
Model selection: how many filters?



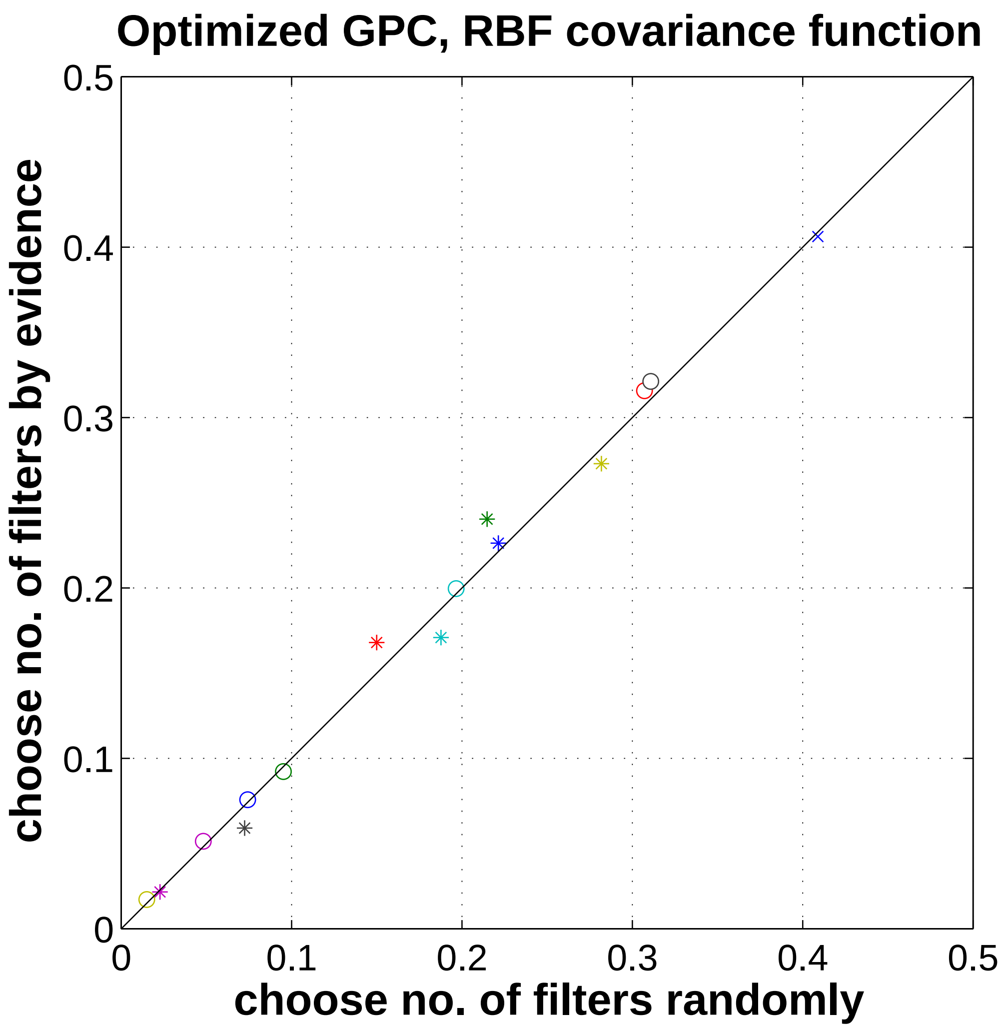
Model selection: how many filters?



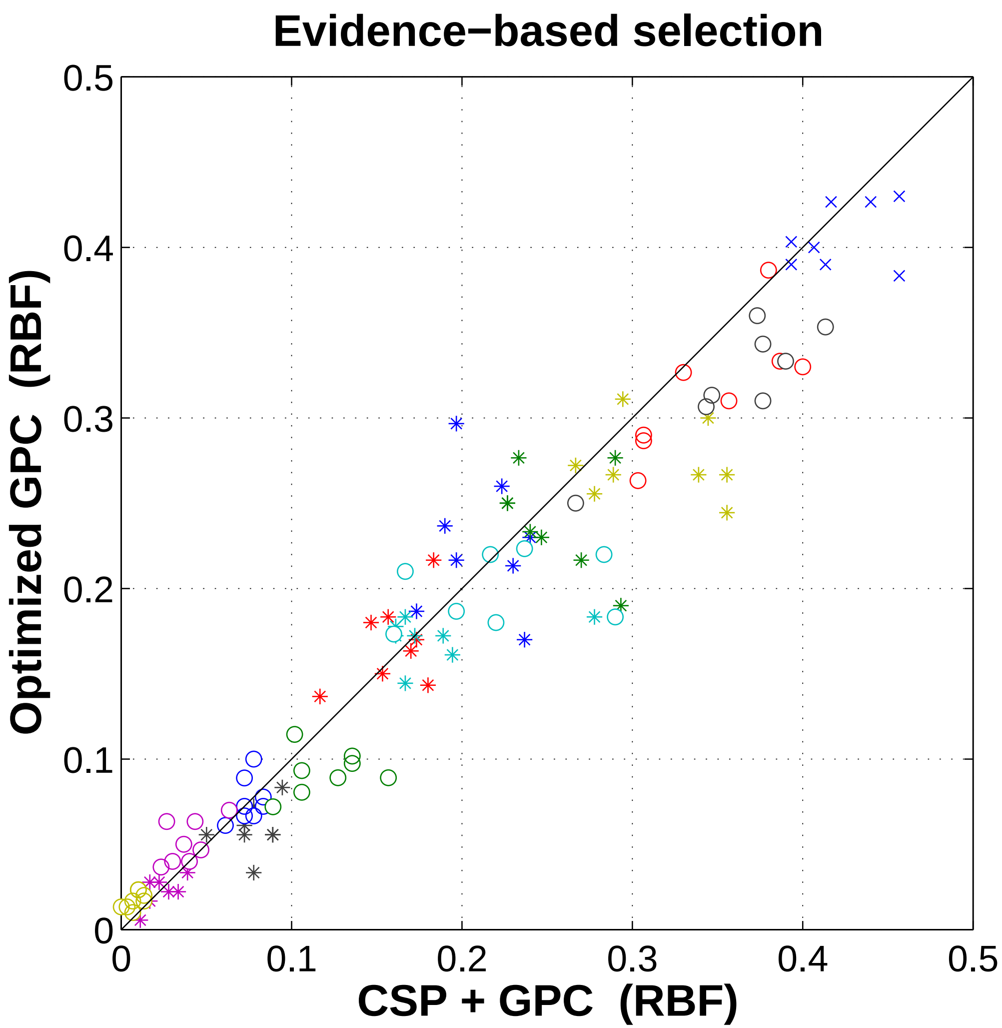
Model selection: how many filters?



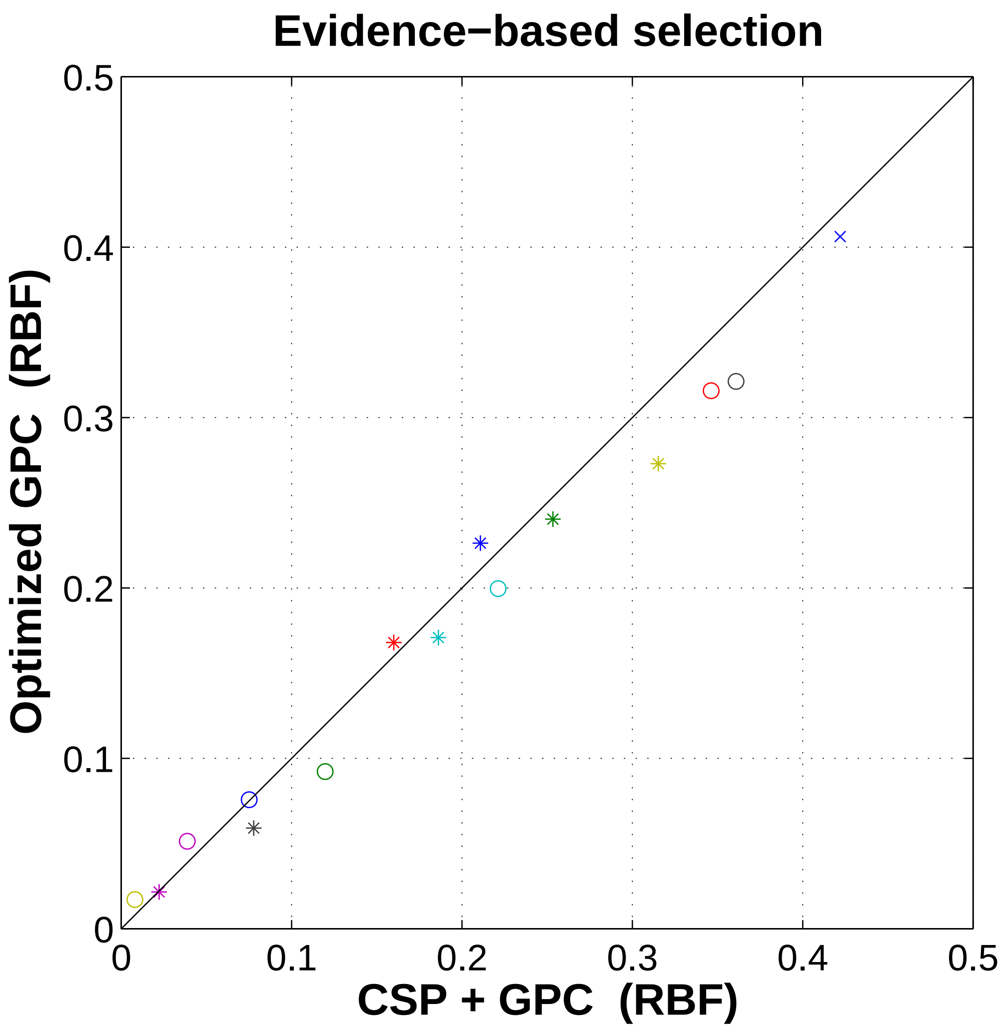
Model selection: how many filters?



Model selection: how many filters?



Model selection: how many filters?



Further goals

- Beyond optimization: MCMC sampling and prediction averaging within the GPC framework.

Further goals

- Beyond optimization: MCMC sampling and prediction averaging within the GPC framework.
- Adapting the system to cope with shifts in background activity between training and test sessions.
  - Hill, Farquhar & Schölkopf 2006 (Proc. 3rd Intl. BCI Workshop)
  - Tomioka, Hill, Blankertz & Aihara 2006 (IBIS, Osaka.)

Further goals

- Beyond optimization: MCMC sampling and prediction averaging within the GPC framework.
- Adapting the system to cope with shifts in background activity between training and test sessions.
  - Hill, Farquhar & Schölkopf 2006 (Proc. 3rd Intl. BCI Workshop)
  - Tomioka, Hill, Blankertz & Aihara 2006 (IBIS, Osaka.)
- Constraining spatial filters to use a small number of electrodes (for convenience of setup).
  - Farquhar, Hill & Schölkopf 2006 (Proc. 3rd Intl. BCI Workshop)

Further goals

- Beyond optimization: MCMC sampling and prediction averaging within the GPC framework.
- Adapting the system to cope with shifts in background activity between training and test sessions.
  - Hill, Farquhar & Schölkopf 2006 (Proc. 3rd Intl. BCI Workshop)
  - Tomioka, Hill, Blankertz & Aihara 2006 (IBIS, Osaka.)
- Constraining spatial filters to use a small number of electrodes (for convenience of setup).
  - Farquhar, Hill & Schölkopf 2006 (Proc. 3rd Intl. BCI Workshop)
- Application of the same approach to features in the time domain (and automatic combination of time-domain & band-power features).

Conclusion

- Significant improvements can be made by applying the principles of margin- and evidence- maximization to the automatic extraction of bandpower features in EEG (and other signal sources..?)

Conclusion

- Significant improvements can be made by applying the principles of margin- and evidence- maximization to the automatic extraction of bandpower features in EEG (and other signal sources..?)
- Benefits are greatest in the difficult cases: high noise and/or small amounts of data.

Conclusion

- Significant improvements can be made by applying the principles of margin- and evidence- maximization to the automatic extraction of bandpower features in EEG (and other signal sources..?)
- Benefits are greatest in the difficult cases: high noise and/or small amounts of data.
- Simultaneous optimization of filters and classifier weights eliminates the need to select filters by hand.

Conclusion

- Significant improvements can be made by applying the principles of margin- and evidence- maximization to the automatic extraction of bandpower features in EEG (and other signal sources..?)
- Benefits are greatest in the difficult cases: high noise and/or small amounts of data.
- Simultaneous optimization of filters and classifier weights eliminates the need to select filters by hand.
- Early indications are that interpretable, optimal weightings across space, time and frequency can be obtained
  - simultaneously;
  - without being very sensitive to prior assumptions.

Thank you for your attention.