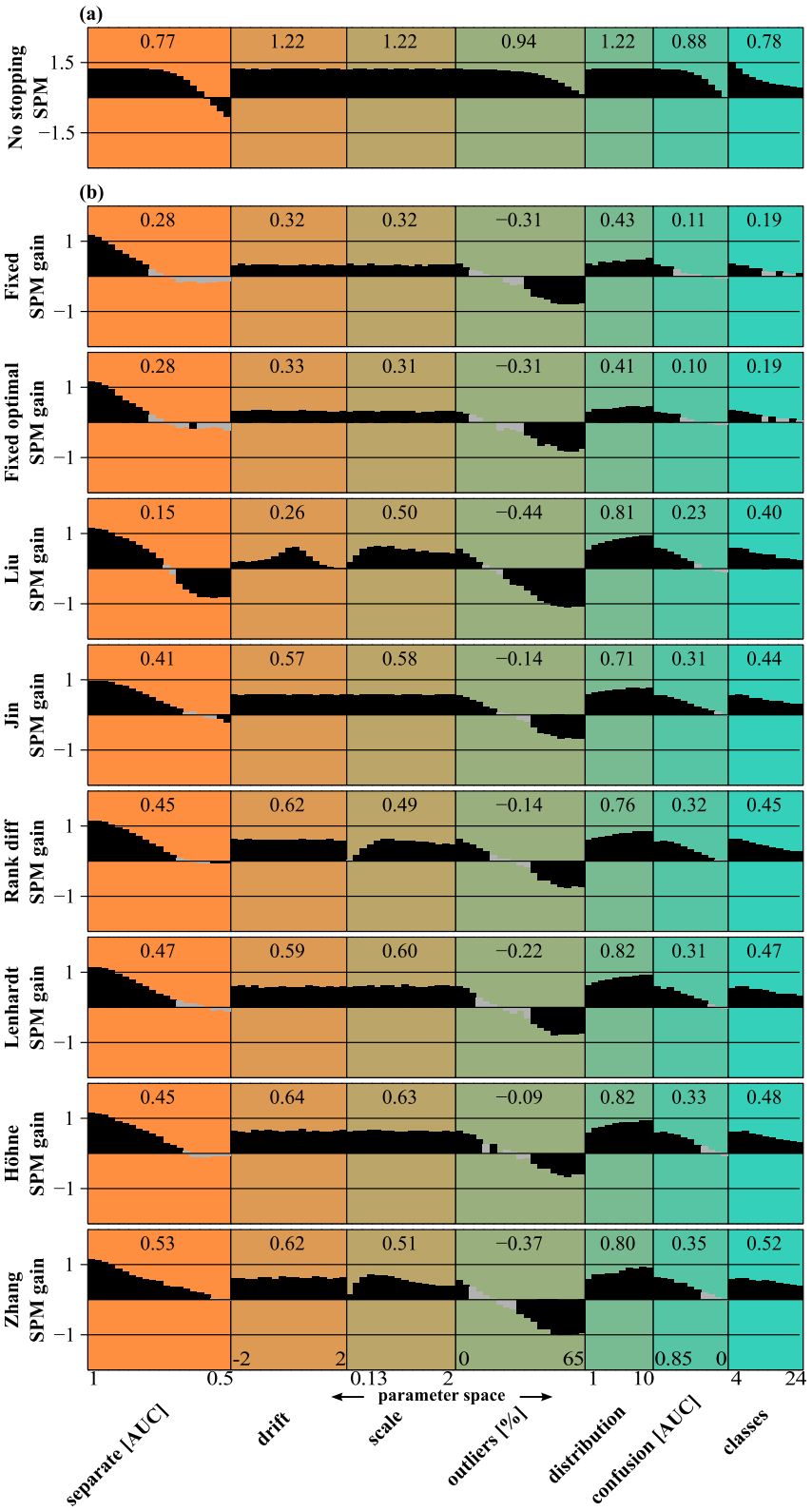
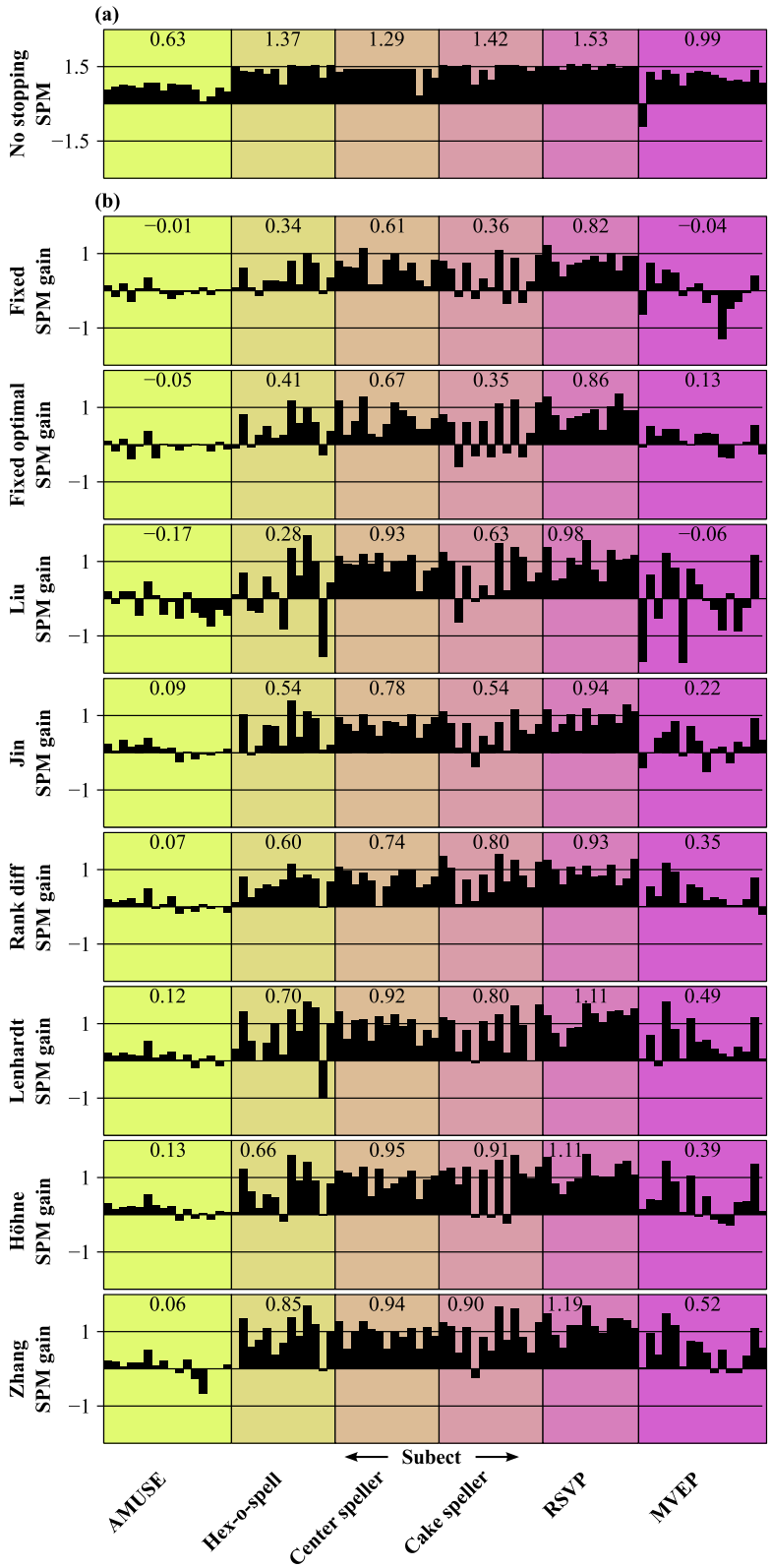
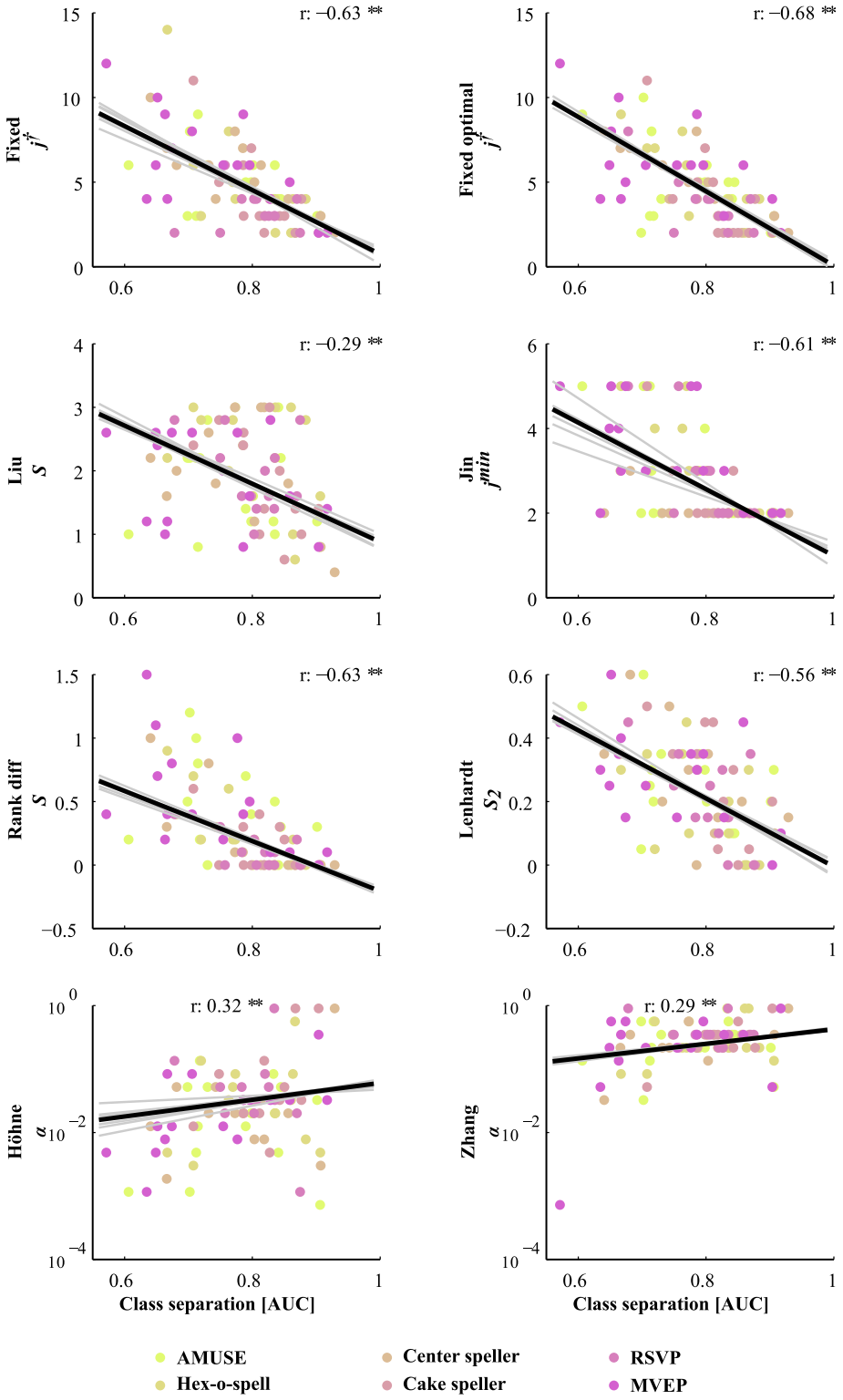


Supplementary Figure 1: Artificial datasets. A) To indicate the effect size, the baseline performance of *no stopping* is shown. B) Shown is the performance gain over the *no stopping* condition for all methods, on several artificially created datasets. For any method to be robust, it should not have negative gains, meaning it should not reduce performance with respect to *no stopping*. With many outliers in the online data, all methods do reduce the performance. All methods show a similar dependency on data separation, and apart from *Liu* and, to a much lesser extend, *Jin* and *Rank diff*, all methods reduce to baseline for inseparable data. *Liu* as the only method is susceptible to all sorts of data distortion, including drifts and scaling.



Supplementary Figure 2: Individual subject profiles. A) To indicate the effect size, the baseline performance of *no stopping* is shown. B) Shown are for each subject and each stopping methods the achieved gains over the *no stopping* condition. Numbers indicate the average gain per method per paradigm (see main manuscript, Figure 6). Negative gains mean that a method reduced the performance for a subject (negative bar) or on average over an entire paradigm (negative number).



Supplementary Figure 3: Estimating $S^{\dagger}$. The individually optimized scaling parameter S is plotted against the AUC of the training data. High linear correlations are found between these, which leads to a set of coefficients for each methods that allow the direct estimate of $S^{\dagger}$ from the training data. The black line indicates the linear fit of all data, using an iteratively reweighted least squares fit. The gray lines indicate the fit of all but one dataset, showing high resemblance with the overall fit. Note that the α values for *Höhne* and *Zhang* are plotted on a log scale, given the nature of the parameter.

Appendix A. Höhne method in depth

The method is based on the Welch's t-test, which can be used to assess whether two normally distributed samples with individual variance have equal mean. The test statistic can be calculated as

$$t(X_1, X_2) = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}. \tag{A.1}$$

where $\overline{X}_i$, s_i^2, and N_i are sample mean, sample variance and sample size, respectively.

In the BCI scenario, the two normally distributed samples are classifier outputs of class c vs. the joint set of classifier outputs of all other classes $\tilde{c}$, thus $t(\boldsymbol{D}_{c,1\ldots j,t}, \boldsymbol{D}_{\tilde{c},1\ldots j,t})$. The test statistic is transformed into a p-value in order to obtain a measure which is independent of sample size.

In the early stopping problem, we are finally testing, if there is a class c that fulfills the stopping criterion,

$$M(\boldsymbol{D}^{test}, t, j) < \alpha \tag{A.2}$$

Sub-function $M(\boldsymbol{D}^{test}, t, j)$ is defined by

$$M(\boldsymbol{D}, t, j) = \min_c p(t(\boldsymbol{D}_{c,1\ldots j,t}, \boldsymbol{D}_{\tilde{c},1\ldots j,t}), \nu) \tag{A.3}$$

As one can see in Equation A.6, the transformation of the t-value into the p-value requires an additional parameter ν, which can be directly computed from the data as well. The transformation from the t-value into the corresponding p-value is implemented in numerous standard statistical toolboxes. Nevertheless, a detailed analytic derivation is given below.

$$p(t, \nu) = \int_{-\infty}^{t} f(u)\, du \tag{A.4}$$

$$\nu = \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}} \tag{A.5}$$

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{A.6}$$

Here, Γ denotes the Gamma-function, in contrast to its previous usage as the training function.

Appendix B. Classifier output distributions in Zhang

In practice, the classifier output distributions f_{a} and f_{u} are modeled as univariate Gaussians

$$f_{\mathrm{a/u}}(d) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathrm{a/u}}}} \exp\left(-\frac{\left(d-\mu_{\mathrm{a/u}}\right)^2}{2\sigma^2_{\mathrm{a/u}}}\right) , \tag{B.1}$$

where the means and variances are estimated from the training data by

$$\hat{\mu}_{\mathrm{a}} = \frac{1}{JT}\sum_{t=1}^{T}\sum_{j=1}^{J} \boldsymbol{D}^{train}_{\boldsymbol{l}_t,j,t} \tag{B.2}$$

$$\hat{\mu}_{\mathrm{u}} = \frac{1}{JT(C-1)}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{c\neq\boldsymbol{l}_t} \boldsymbol{D}^{train}_{c,j,t} \tag{B.3}$$

and

$$\hat{\sigma}^2_{\mathrm{a}} = \frac{1}{JT-1}\sum_{t=1}^{T}\sum_{j=1}^{J}(\boldsymbol{D}^{train}_{\boldsymbol{l}_t,j,t} - \hat{\mu}_{\mathrm{a}})^2 \tag{B.4}$$

$$\hat{\sigma}^2_{\mathrm{u}} = \frac{1}{JT(C-1)-1}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{c\neq\boldsymbol{l}_t}(\boldsymbol{D}^{train}_{c,j,t} - \hat{\mu}_{\mathrm{u}})^2 . \tag{B.5}$$

Such modeling is in line with the LDA classifier assuming the features of the two classes to be normally distributed as well. In fact, $\hat{\mu}_{\mathrm{a/u}}$ and $\hat{\sigma}^2_{\mathrm{a/u}}$ can be obtained from the feature means and covariance matrix $\boldsymbol{m}_{\mathrm{a/u}}$ and $\boldsymbol{\Sigma}$ estimated by LDA via

$$\hat{\mu}_{\mathrm{a/u}} = \boldsymbol{w}^\top \boldsymbol{m}_{\mathrm{a/u}} \quad \text{and} \quad \hat{\sigma}^2_{\mathrm{a}} = \hat{\sigma}^2_{\mathrm{u}} = \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} , \tag{B.6}$$

where $\boldsymbol{w}$ is the LDA projection vector.