L1-penalized linear mixed-effects models for high dimensional data with application to BCI

Siamac Fazli [a,b,*], Márton Danóczy [a], Jürg Schelldorfer [c], Klaus-Robert Müller [a,b,d]
a Berlin Institute of Technology, Franklinstr. 28/29, 10587 Berlin, Germany
b Bernstein Focus: Neurotechnology Berlin (BFNT-B), 10587 Berlin, Germany
c ETH Zürich,Rämistrasse 101, 8092 Zürich, Switzerland
d Institute for Pure and Applied Mathematics, UCLA, Los Angeles, CA 90095-7121 USA

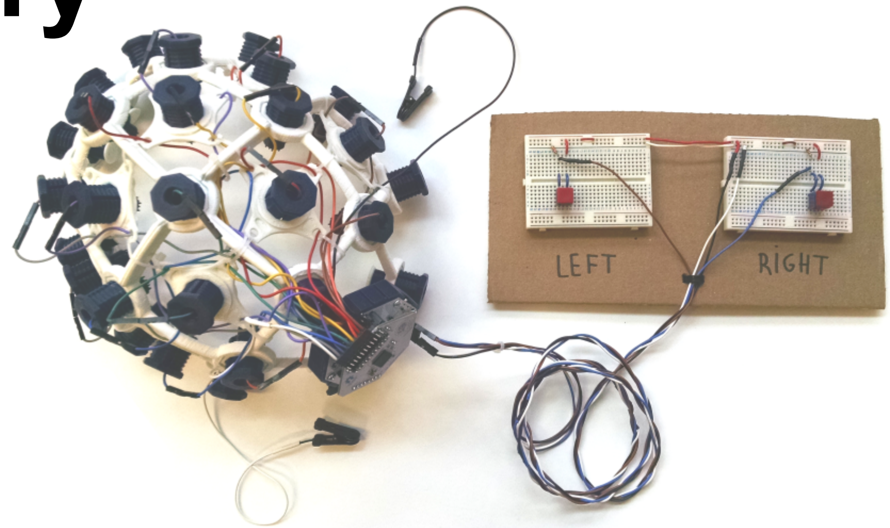
ELSEVIER – NeuroImage – Volume 56 – Issue 4 – 15 June 2011 – pages 2100-2108

Presented by Mehrnaz Motamed

1. **Introduction**
2. **Statistical Model:** model setup + L1-penalized maximum likelihood estimator + prediction of the random-effects + model selection + computational implementation
3. **Available data and experiments:** BCI setup and data basis: Generation of the ensemble + Temporal filters + Spatial filters and classifiers + Validation
4. **Results:** experimental results: Subject-to-subject transfer + Session-to-session transfer + Relation of baseline misclassification to σ2 and τ2 + Effective spatial filters and distances thereof
5. **Discussion and conclusions**

Motor Imagery

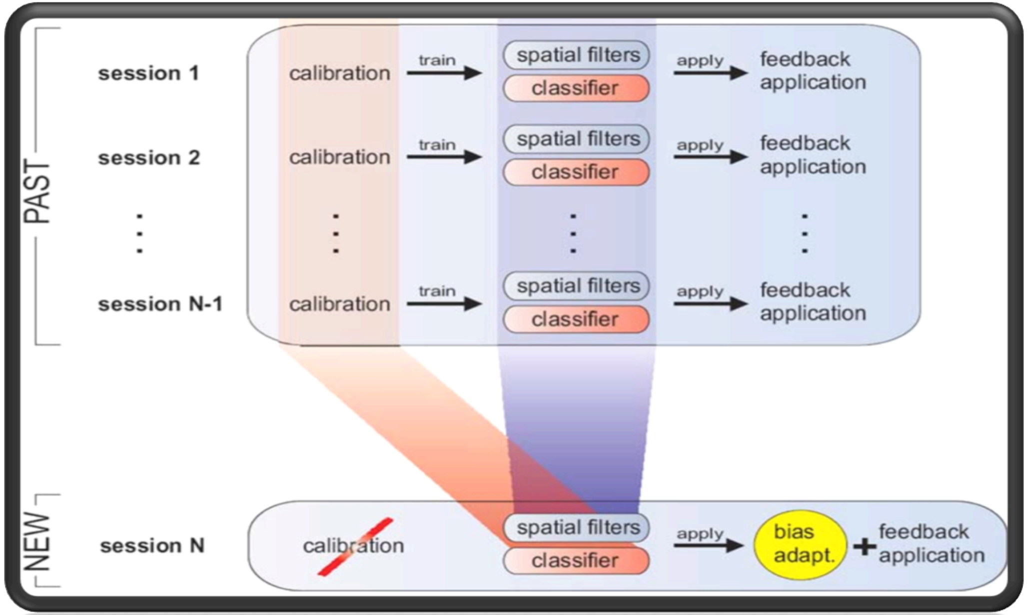
- Training Set



a mental process by which an individual rehearses or simulates a given action.

Zero Training!

- Depending on the number of past sessions, this procedure creates a larger set of filters.
- Once this set of historic filters is created, 6 so-called prototype filters are chosen



* Image from 'Towards Zero Training for Brain-Computer Interfacing'
Matthias Krauledat*, Michael Tangermann, Benjamin Blankertz, Klaus-Robert Mu¨ller

L1-penalized linear mixed-effects model

- Less calibration time (practically zero) (just bias)
- more accuracy
- within subject and across subject variation
- sparse
- better features

mixed effect: within subject and across subject variation
L1: Few data + high dimension (No need for calibration)

- mixed effect
- L1

Linear mixed-effects model

- inbuilt heterogeneity of data, distinct sources
  - Variance: Individual experiment and a fixed effect
- Grouping Structure
- Mixed Model: fixed and random effects

Problems:

- High Dimensional Data
- Computationally intractable for large covariates

Lasso method

- least absolute shrinkage and selection operator
- performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.
- Min Σe^2 s.t. $\Sigma|\beta j| \leq s$
- regression with an l1-norm penalty
- convex problem

L1 penalized

- assuming that the number of potential fixed effects is large and that the underlying true fixed-effects vector is sparse.
- splits high **variance** of BCI
- quantify both sources of variance + improved ensemble model
- Practically zero training (no calibration)
- BCI community goals: **reduce setup cost, increase Information Transfer Rates (ITR)**

How to implement Statistical Model?

Mixed effect model setup

- extensions of linear regression models for grouped data
- with coefficients that can vary with respect to one or more grouping variables
- Fixed and random effects
- individual experiments

Mixed effect model setup

- $y_i = X_i\beta + Z_i b_i + \varepsilon_i \qquad i = 1, \ldots, N$
- $i)\ b_i \sim \mathcal{N}_q(0, \tau^2 I_q),\ ii)\ \varepsilon_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i})$
- $y_i \sim \mathcal{N}_{n_i}(X_i\beta, \Lambda_i(\sigma^2, \tau^2)) \quad \text{with} \quad \Lambda_i(\sigma^2, \tau^2) = \sigma^2 I_{n_i} + \tau^2 Z_i Z_i^{\top}.$

N subjects
ni observations per subject
NT = Σni : total number of observations
yi: ni dimensional observation of each subject

Xi: ni*p fixed-effects covariates
Zi: ni*q random-effects covariates
β: p-dimensional fixed-effects vector
**bi: (i=1:N) q-dimensional random-effects vectors.
Assume: errors εi are mutually independent of the random effects bi.

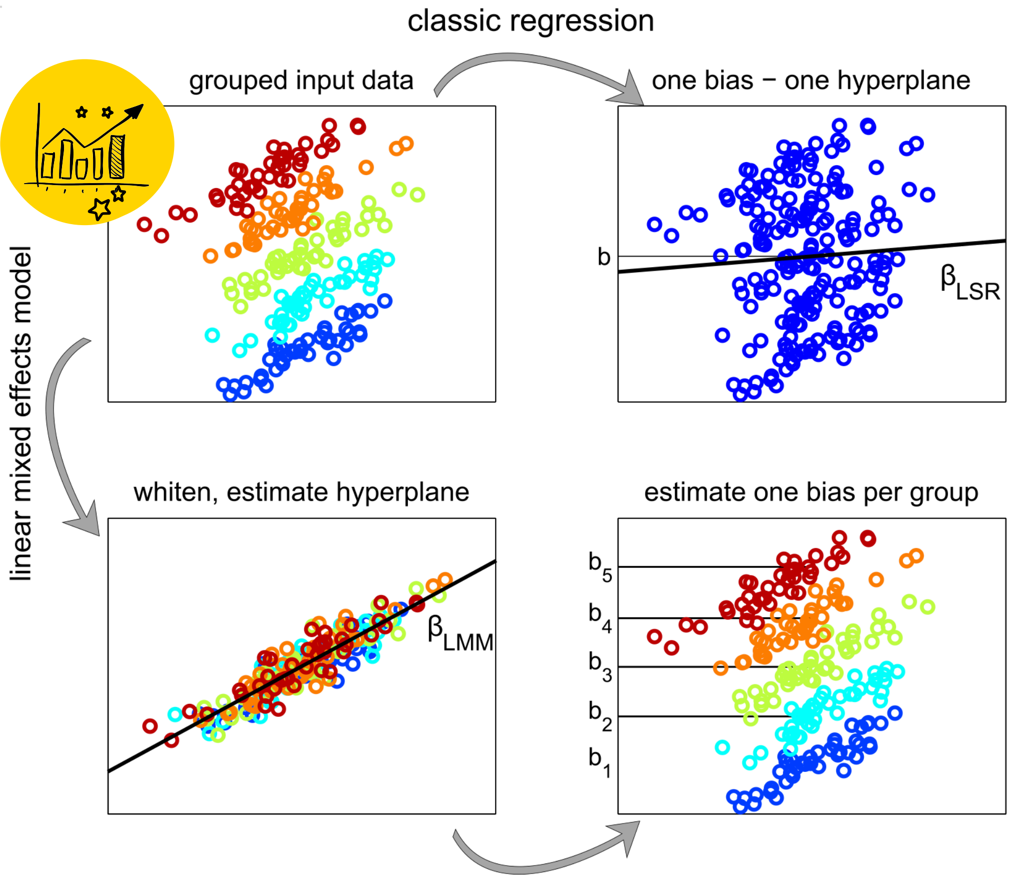


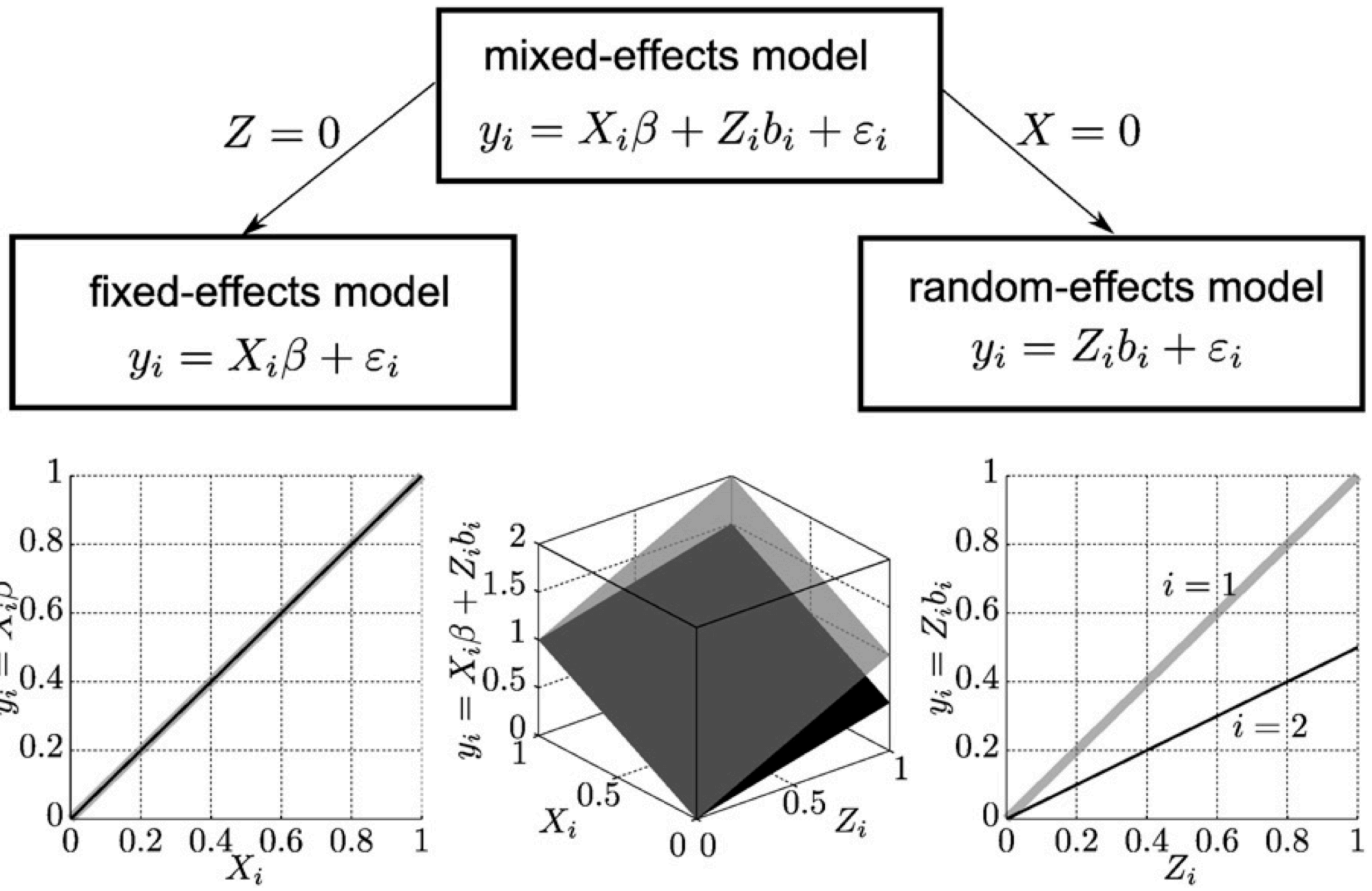
Figure 1

INTUITION

- $Z_i = I_{ni}$
  $\beta_{ORIG} = 0.5$
  $b_{ORIG} = [-2; -1; 0; 1; 2]$
  $\varepsilon_{ORIG} \sim N(0, 0.2)$

- $\beta_{LSR} = 0.048$
  $b_{LSR} = 0.075$

- $\beta_{LMM} = 0.504$
  $b_{LMM} = [-1.96; -1.015; -0.014; 0.973; 2.013]$



mixed-effects model
$y_i = X_i\beta + Z_i b_i + \varepsilon_i$
$Z = 0$
$X = 0$
fixed-effects model
$y_i = X_i\beta + \varepsilon_i$
random-effects model
$y_i = Z_i b_i + \varepsilon_i$
$y_i = X_i\beta$
X_i
$y_i = X_i\beta + Z_i b_i$
X_i
Z_i
$y_i = Z_i b_i$
Z_i
$i = 1$
$i = 2$

foreach $(\sigma^2, \tau^2, \lambda)$ **do**

  foreach i **do**

    (Whiten data and labels)

$$\Lambda_i = \sigma^2 I_{n_i} + \tau^2 Z_i Z_i^\top$$

$$\bar{X}_i = \Lambda_i^{-1/2} X_i, \quad \bar{y}_i = \Lambda_i^{-1/2} y_i$$

  end

  (Fit ℓ_1-penalized least-squares to concatenated data)

$$\hat{\beta} = \operatorname{argmin} \|\bar{X}\beta - \bar{y}\|_2^2 + 2\lambda \textstyle\sum_{k=2}^{p} |\beta_k|$$

  foreach i **do**

    (Find random effects)

$$\hat{b}_i = [Z_i^\top Z_i + \sigma^2/\tau^2 I_q]^{-1} Z_i^\top (y_i - X_i \hat{\beta})$$

  end

end

Algorithm 1. Algorithm for fitting the mixed effects model.

L1-penalized maximum likelihood estimator

- $S_\lambda(\beta;\sigma^2;\tau^2) := (-1/2) \sum \{\log|\Lambda_i| + (y_i - X_i\beta)^T \Lambda_i^{-1} (y_i - X_i\beta)\} - \lambda \sum |\beta_k|$

λ: nonnegative regularization parameter.

- $\hat{\beta};\hat{\sigma}^2;\hat{\tau}^2 = \arg\max S_\lambda(\beta;\sigma^2;\tau^2)$

- *Prediction of the random-effects (b_i)(using MAP β, σ^2, τ^2)*

$$b_i = [\, Z^T{}_i \, Z_i + \sigma^2 / \tau^2 I_q]^{-1} \; Z^T{}_i \, (y_i - X_i \beta)$$

- *Model selection*

$$-2L(\beta\hat{}\,;\, \sigma\hat{}\,^2 ; \tau\hat{}^2) + \log N_T \cdot d\hat{}f_\lambda$$

- *Computational implementation*

$$\beta\hat{} = \arg\min \sum ||\Lambda_i^{-12} X_i \beta - y_i||_2{}^2 + 2\lambda \sum |\beta_k|$$

Data & Experiment

Data & Experiment

83 BCI experiments (sessions) * 150 trials

83 individual subjects

90 sessions*60-600 trials

44 subjects.

Balanced Data Set

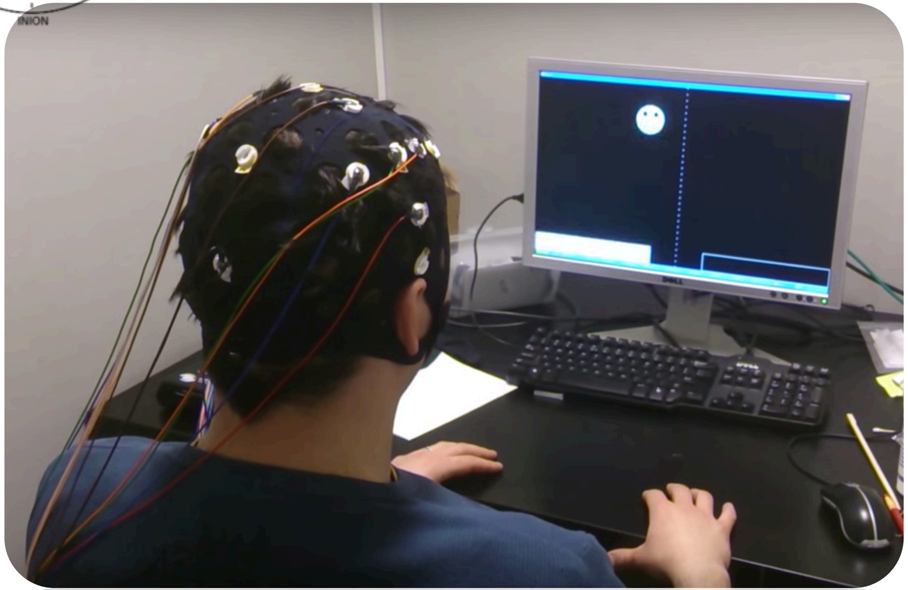
Unbalanced Data Set

'clean' model

individual sessions, underlying processes.

Experiment

I.



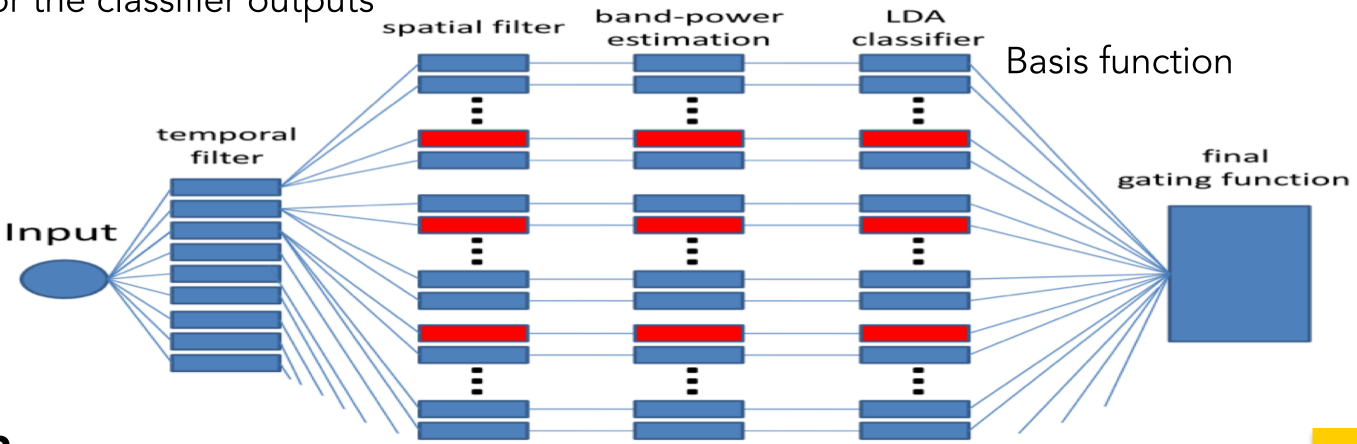
II.

III. L R

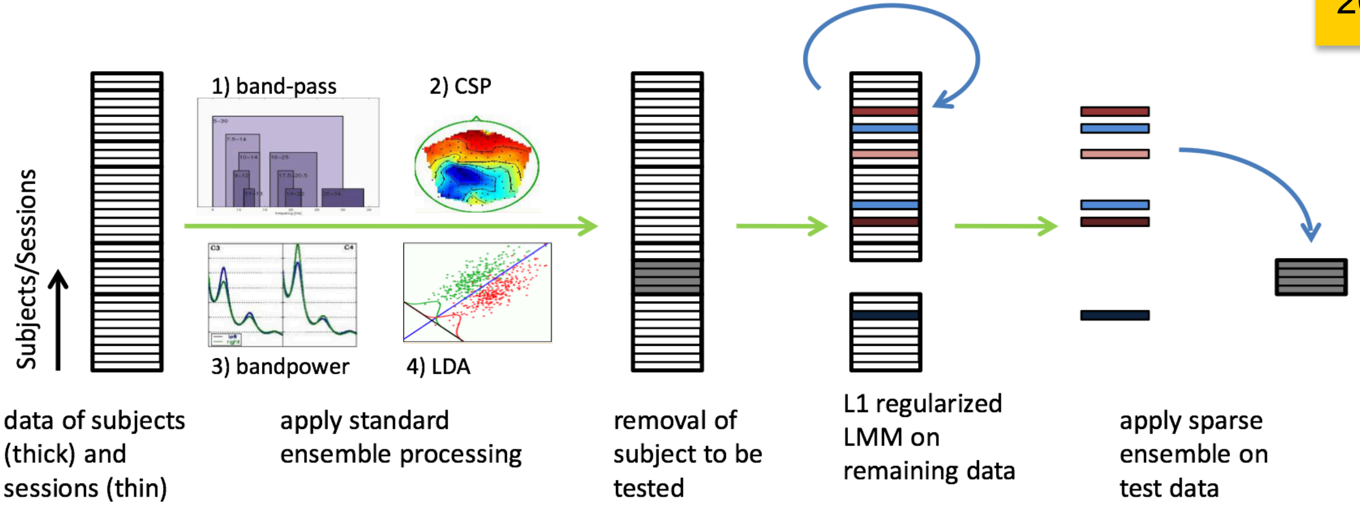
EMG, EOG

- *Generation of the ensemble*
- *Temporal filters*
- *Spatial filters and classifiers*
- *Validation*

appropriate weighting for the classifier outputs



- ***Generation of the ensemble***

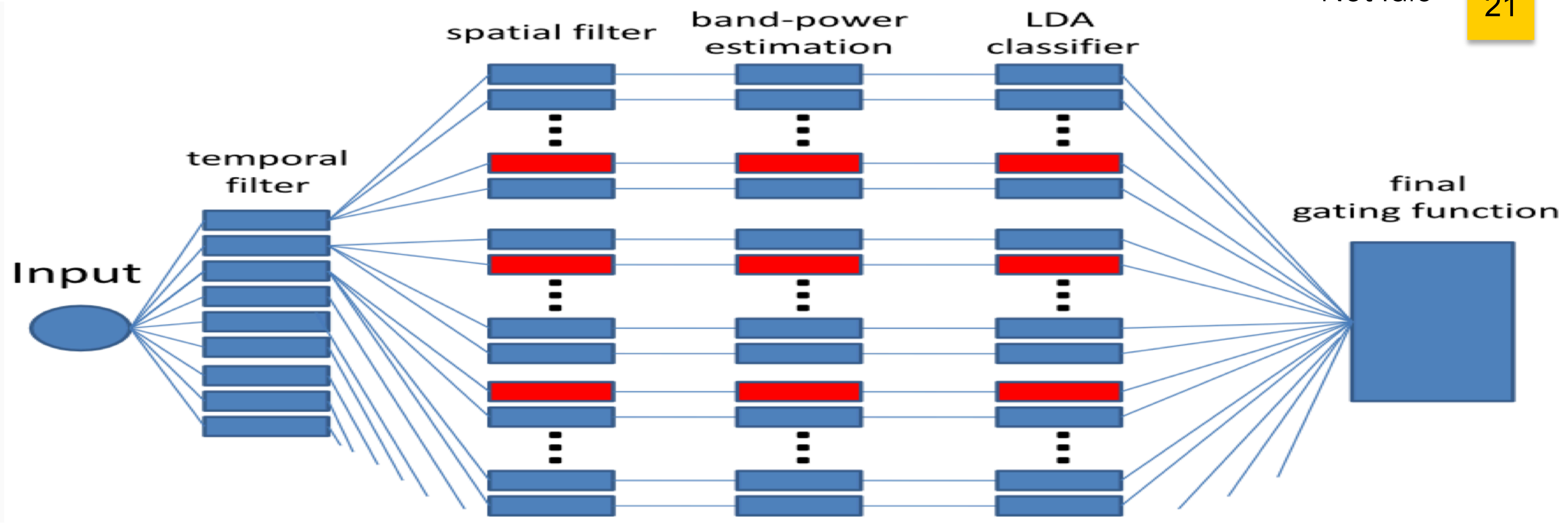


Remove frequencies within the raw signal, that are not of interest.

Temporal filters

Not Idle

21



Spatial filters and classifiers:

Σ_1 , Σ_2 → **CSP** → W , D → **LDA**

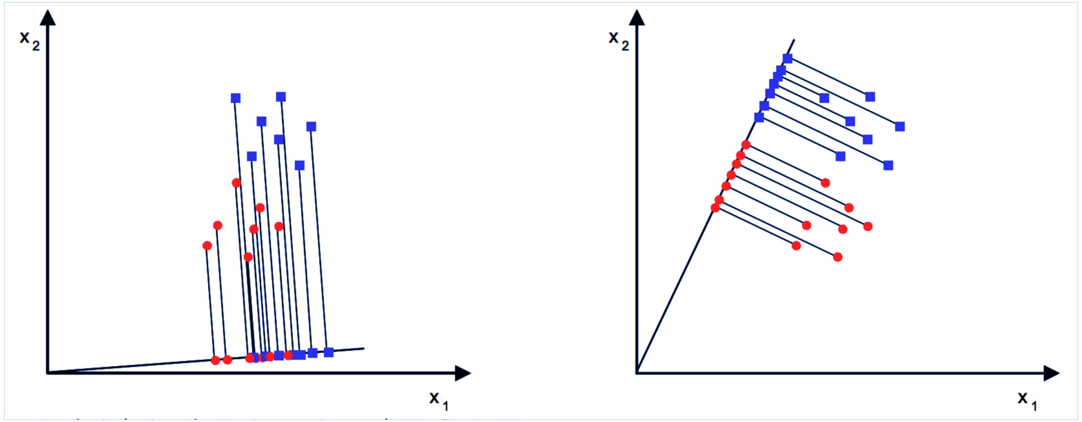
a matrix of projections
Diagonal

$W \Sigma_1 W^T = D$ and $W \Sigma_2 W^T = I - D$

separating a multivariate signal into additive subcomponents which have maximum differences in variance

Spatial filters and classifiers:

LDA



Linear Discriminate analysis

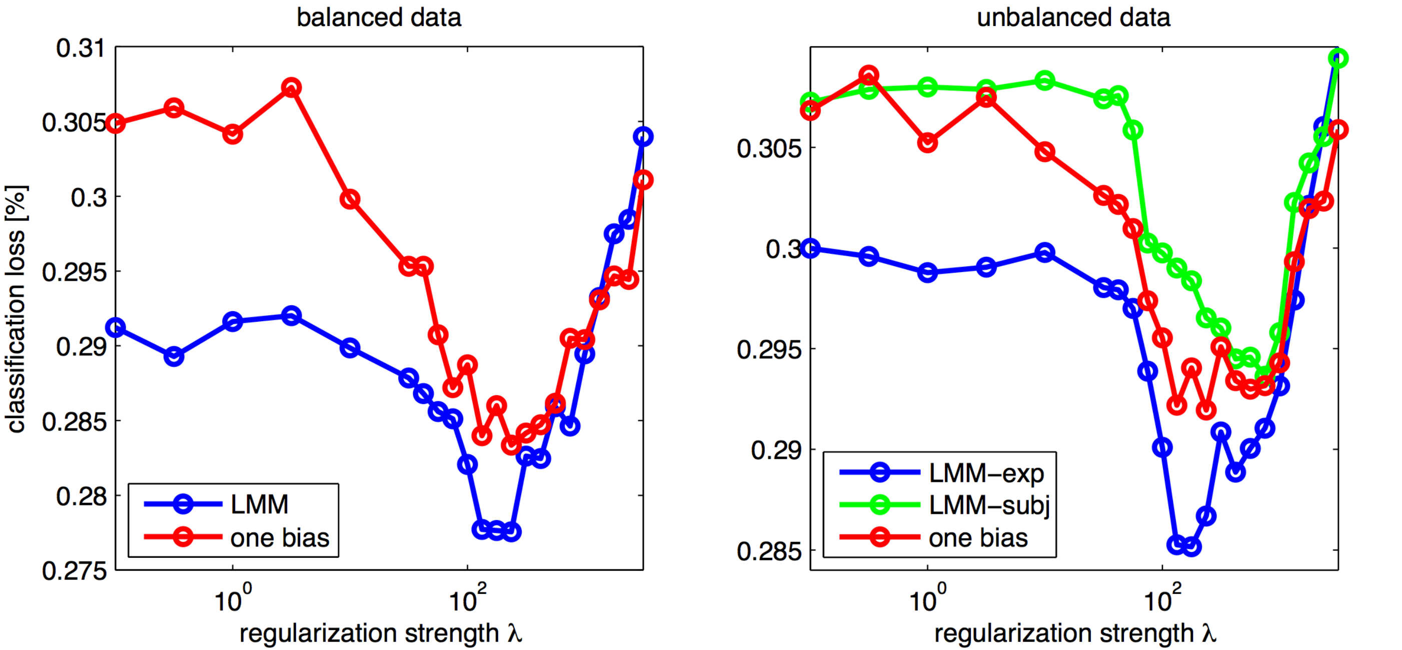
*Figure from Introduction to Common Spatial Pattern Filters for EEG Motor Imagery Classification
Tatsuya Yokota
Tokyo Institute of Technology
July 17, 2012

Validation

8-fold cross-validation,
splitting the data chronologically.

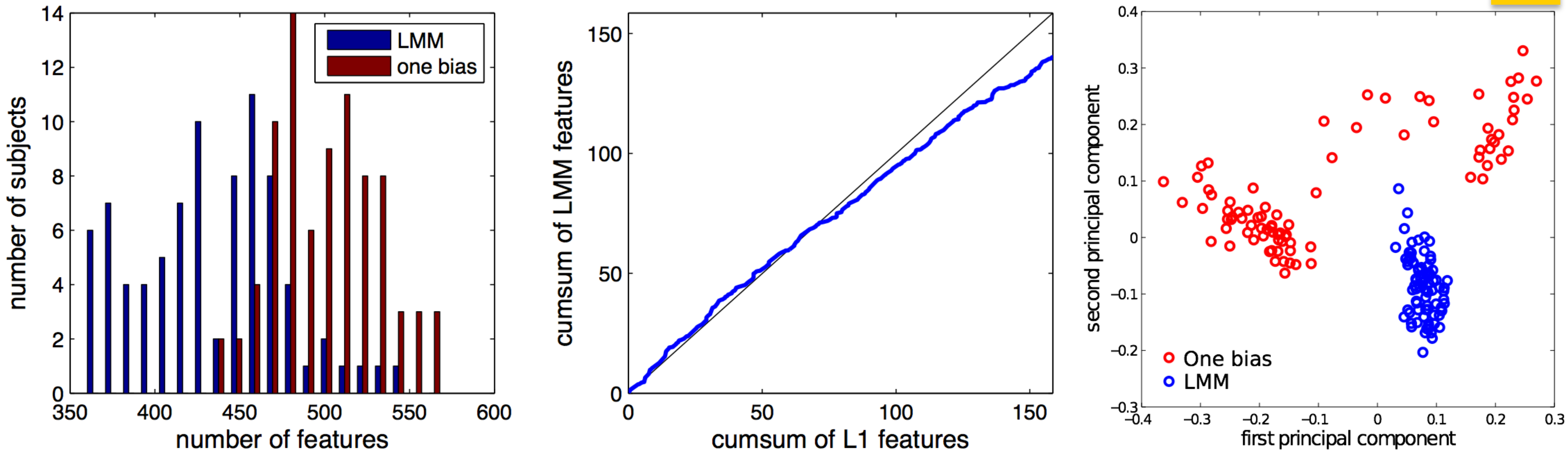
chronological | since the non- stationarity of the data is thus preserved

Results

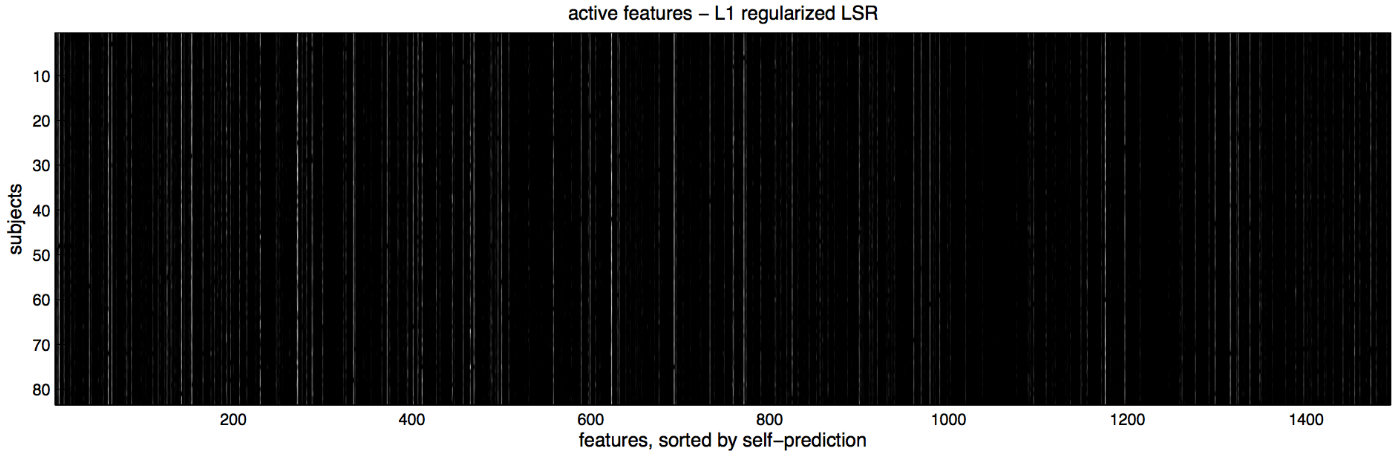


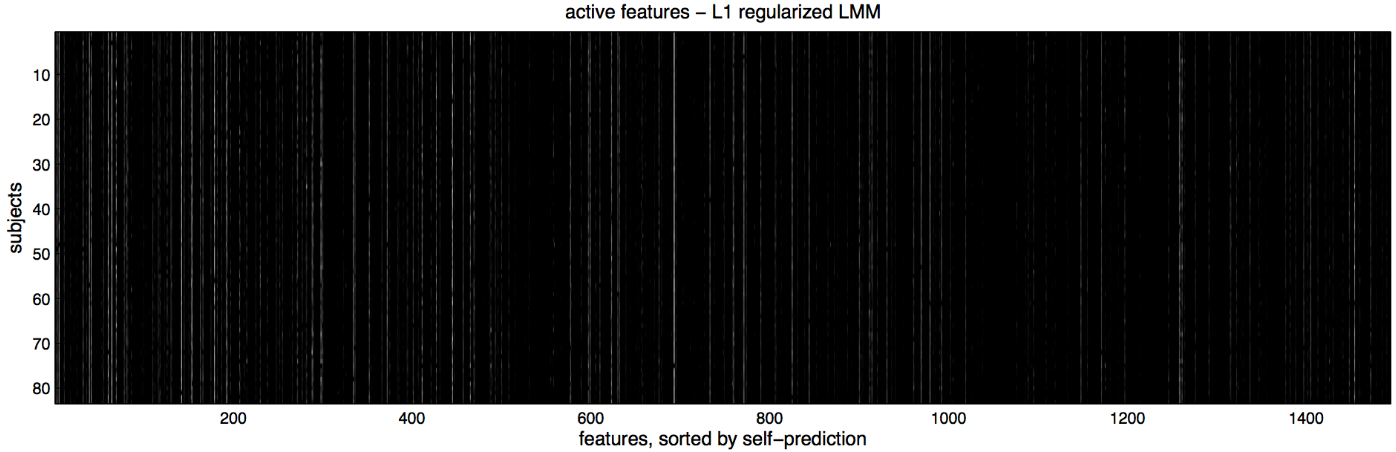
LMM-exp one bias per experiment

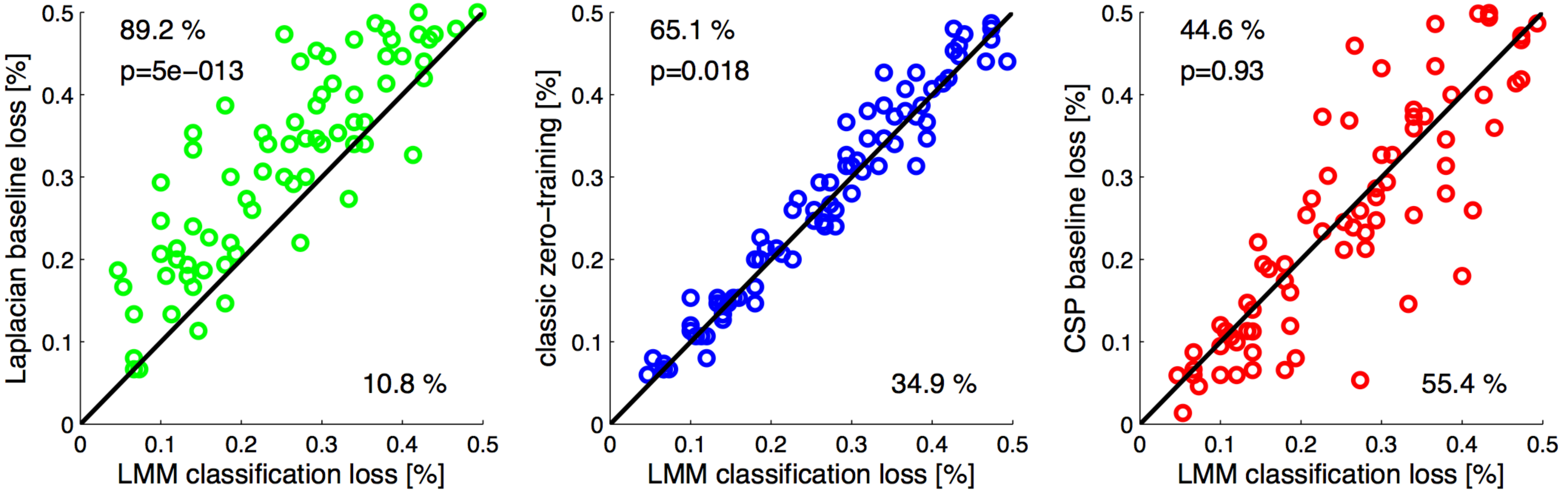
LMM-subj estimates one bias per subject



LMM: higher fraction of features with low self-prediction errors
sorted by increasing self-prediction accuracy.


active features – L1 regularized LSR
subjects
10
20
30
40
50
60
70
80
200
400
600
800
1000
1200
1400
features, sorted by self–prediction

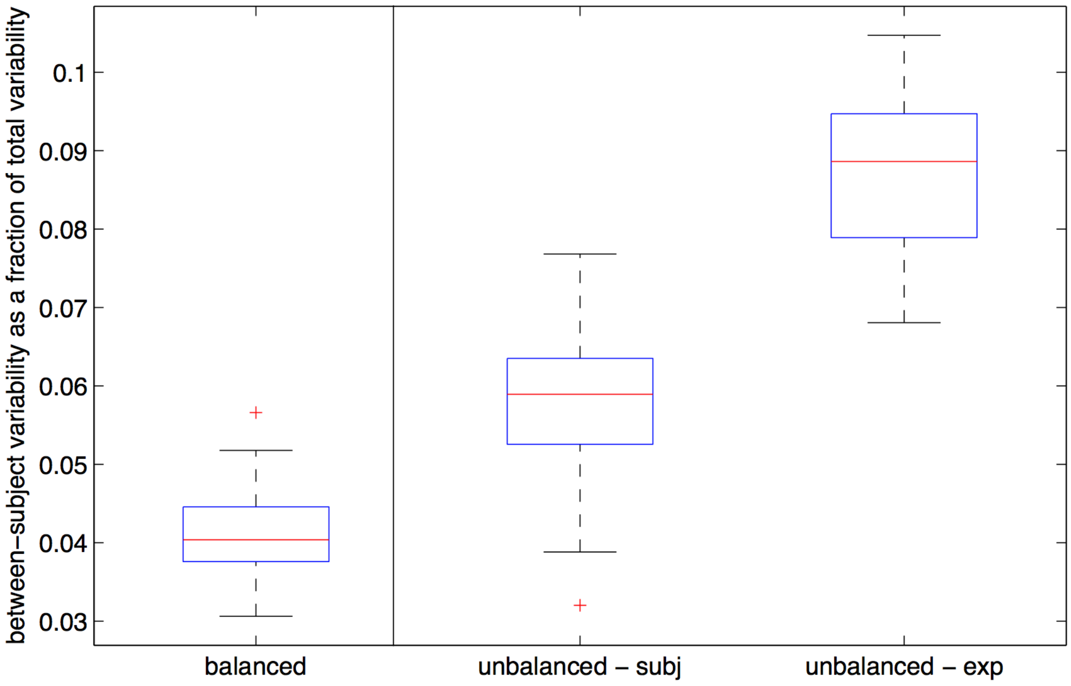

active features – L1 regularized LMM
subjects
10
20
30
40
50
60
70
80
200
400
600
800
1000
1200
1400
features, sorted by self–prediction


89.2 %
p=5e−013
10.8 %
Laplacian baseline loss [%]
LMM classification loss [%]
65.1 %
p=0.018
34.9 %
classic zero−training [%]
LMM classification loss [%]
44.6 %
p=0.93
55.4 %
CSP baseline loss [%]
LMM classification loss [%]

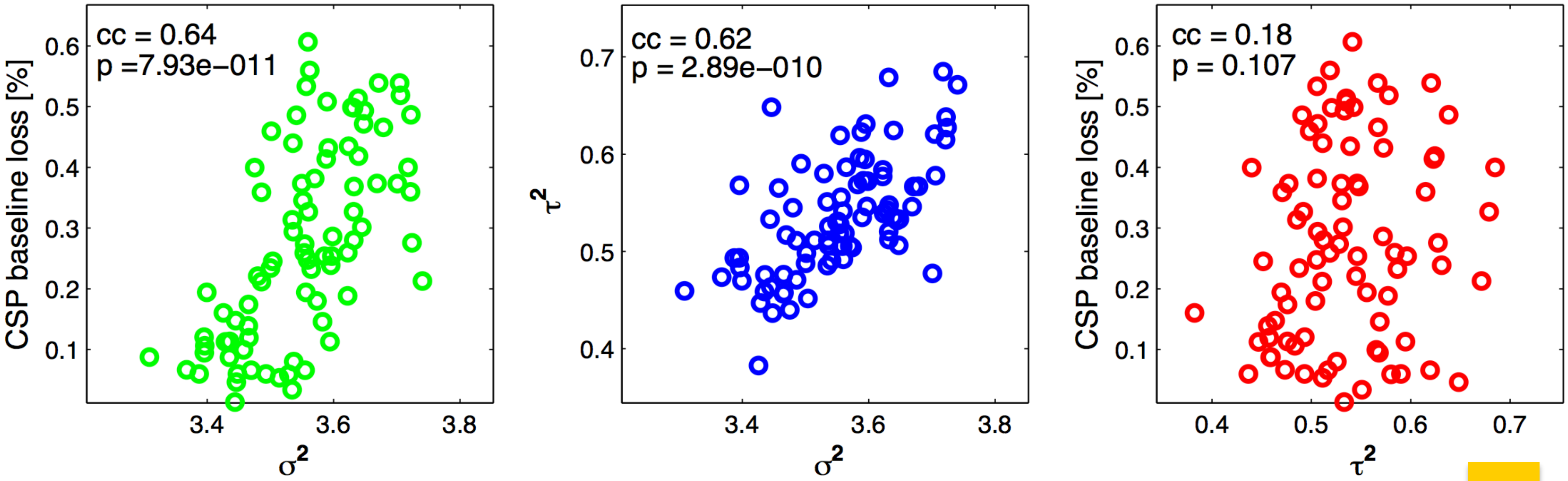
30

the **transfer of classifiers from sessions to sessions** required a **bias correction**

Valid: we are able to **capture a meaningful part of the variability** which would otherwise **be ignored as noise**.



It is not possible to draw conclusions about the quality of a subject's data by the variance of its assigned biases.



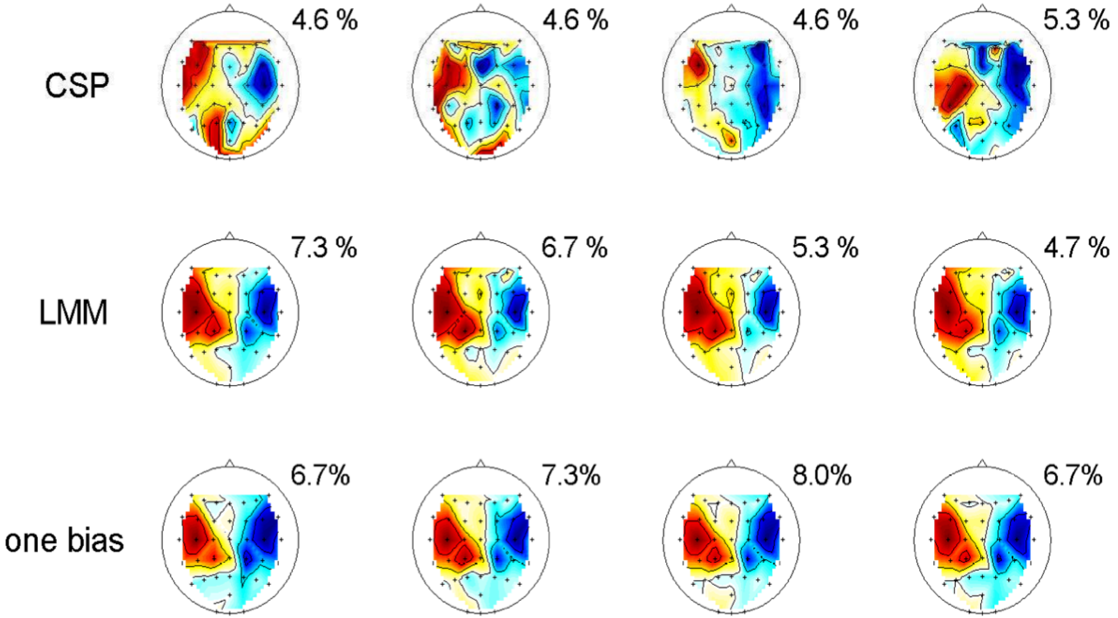
σ2 have **a strong positive correlation**

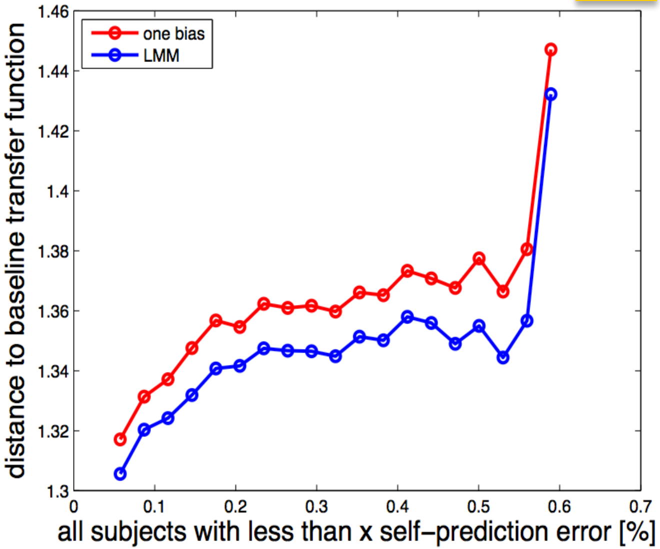
of **τ2** and **σ2** and find a **strong positive relation**.

within-subject variability σ^2, between-subject variability τ^2

sinusoid

LEFT: resulting response matrices





Discussion and conclusions

- Sparse modeling approach based on L1-penalized linear mixed-effects model and novel BCI zero- training model

- Better overall prediction (improved zero-training model)
- Attributing variability to differences between subjects, not noise.
- Same subject on two subsequent days, on the second day the classifier can often be reused without much retraining, only the bias needs to be adjusted
- More robust feature selection
- **Future study:** study online adaptation of penalized linear mixed-effects models in the context of medical diagnosis.

References

Alamgir, M., Grosse-Wentrup, M., Altun, Y., 2010. Multitask learning for brain–computer interfaces. In: Teh, Y.W., Titterington, M. (Eds.), Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010), pp. 17–24.

Andrew, G., Gao, J., 2007. Scalable training of L_1-regularized log-linear models. Proceedings of the 24th International Conference on Machine Learning, ACM, Corvalis, Oregon, pp. 33–40.

Blankertz, B., Curio, G., Müller, K., 2002. Classifying single trial EEG: towards brain computer interfacing. In: Diettrich, T.G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Inf. Proc. Systems (NIPS 01), pp. 157–164.

Blankertz, B., Dornhege, G., Schäfer, C., Krepki, R., Kohlmorgen, J., Müller, K., Kunzmann, V., Losch, F., Curio, G., 2003. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. IEEE Trans. Neural Syst. Rehabil. Eng. 11, 127–131.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K., Kunzmann, V., Losch, F., Curio, G., 2006a. The Berlin brain–computer interface: EEG-based communication without subject training. IEEE Trans. Neural Syst. Rehabil. Eng. 14, 147–152.

Blankertz, B., Dornhege, G., Lemm, S., Krauledat, M., Curio, G., Müller, K., 2006b. The Berlin brain–computer interface: machine learning based detection of user specific brain states. J. Universal Comput. Sci. 12, 2006.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K., Curio, G., 2007. The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. Neuroimage 37, 539–550.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K., 2008. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Process. Mag. 25, 41–56.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K., 2010a. Single-trial analysis and classification of ERP components — a tutorial. Neuroimage.

Blankertz, B., Sannelli, C., Halder, S., Hammer, E.M., Kubler, A., Müller, K., Curio, G., Dickhaus, T., 2010b. Neurophysiological predictor of SMR-based BCI performance. Neuroimage 51, 1303–1309.

Bondell, H.D., Krishna, A., Ghosh, S.K., 2010. Joint variable selection of fixed and random effects in linear mixed-effects models. Biometrics 66, 1069–1077.

Cheng, M., Gao, X., Gao, S., Xu, D., 2002. Design and implementation of a brain–computer interface with high transfer rates. IEEE Trans. Biomed. Eng. 49, 1181–1186.

Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., Müller, K., 2006. Combined optimization of spatial and temporal filters for improving brain–computer interfacing. IEEE Trans. Biomed. Eng. 53, 2274–2281.

Dornhege, G., Millán, J.R., Hinterberger, T., McFarland, D., Müller, K. (Eds.), 2007. Toward Brain–Computer Interfacing. MIT Press, Cambridge, MA.

Fazli, S., Grozea, C., Danoczy, M., Blankertz, B., Popescu, F., Müller, K., 2009a. Subject independent EEG-based BCI decoding. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems, 22. MIT Press, pp. 513–521.

Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K.R., Grozea, C., 2009b. Subject-independent mental state classification in single trials. Neural Netw. 22, 1305–1312.

Ibrahim, J.G., Zhu, H., Garcia, R.I., Guo, R., 2010. Fixed and random effects selection in mixed effects models. Biometrics. doi:10.1111/j.1541-0420.2010.01463.x.

Koles, Z., Soong, A.C.K., 1998. EEG source localization: implementing the spatio-temporal decomposition approach. Electroencephalogr. Clin. Neurophysiol. 107, 343–352.

Krauledat, M., Tangermann, M., Blankertz, B., Müller, K., 2008. Towards zero training for brain-computer interfacing. PLoS One 3, e2967.

Meier, L., van de Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. J. R. Statist. Soc. B 70, 53–71.

Müller, K., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Trans. Neural Netw. 12, 181–201.

Müller, K., Anderson, C.W., Birch, G.E., 2003. Linear and nonlinear methods for brain–computer interfaces. IEEE Trans. Neural Syst. Rehabil. Eng. 11, 165–169.

Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2005. Recipes for the linear analysis of EEG. Neuroimage 28, 326–341.

Parra, L., Christoforou, C., Gerson, A., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M., Sajda, P., 2008. Spatiotemporal linear decoding of brain state. IEEE Signal Process. Mag. 28, 107–115.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-Plus. Springer, New York.

Ramoser, H., Müller-Gerkin, J., Pfurtscheller, G., 2000. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans. Rehabil. Eng. 8, 441–446.

Schelldorfer, J., Bühlmann, P., van de Geer, S., 2010. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. arXiv:1002.3784 preprint.

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R., Müller, K., 2006a. Towards adaptive classification for BCI. J. Neural Eng. 3, R13–R23.

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R.P.N., Müller, K., 2006b. Towards adaptive classification for BCI. J. Neural Eng. 3, 13–23.

Thomas, K.P., Guan, C., Lau, C.T., Vinod, A., Ang, K.K., 2009. A new discriminative common spatial pattern method for motor imagery brain–computer interfaces. IEEE Trans. Biomed. Eng. 56, 2730–2733.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B 58, 267–288.

Tomioka, R., Müller, K., 2010. A regularized discriminative framework for EEG analysis with application to brain–computer interface. Neuroimage 49, 415–432.

Vidal, J.J., 1973. Toward direct brain–computer communication. Annu. Rev. Biophys. Bioeng. 2, 157–180.

Wang, Y., Wang, R., Gao, X., Hong, B., Gao, S., 2006. A practical vep-based brain–computer interface. IEEE Trans. Neural Syst. Rehabil. Eng. 14, 234–240.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Statist. Soc. B 68, 49–67.

Thanks!

You can find me at

- memotame@eng.ucsd.edu