NeuroImage 108 (2015) 328–342

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals

Denis A. Engemann [a,b,c,d,*], Alexandre Gramfort [a,e]

[a] CEA, DSV/I2BM, NeuroSpin, Bat. 145, 91191 Gif-sur-Yvette Cedex, France
[b] Neuropsychology & Neuroimaging Team, INSERM UMRS 975, ICM, Paris, France
[c] Cognitive Neuroscience (INM-3), Juelich Research Centre, Juelich, Germany
[d] Neuroimaging Group, Department of Psychiatry, University Hospital of Cologne, Cologne, Germany
[e] Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, France

ARTICLE INFO

Article history:
Accepted 4 December 2014
Available online 23 December 2014

Keywords:
Electroencephalography (EEG)
Magnetoencephalography (MEG)
Neuroimaging
Principal component analysis (PCA)
Factor analysis (FA)
Covariance estimation
Whitening
Model selection
Statistical learning

ABSTRACT

Magnetoencephalography and electroencephalography (M/EEG) measure non-invasively the weak electromagnetic fields induced by post-synaptic neural currents. The estimation of the spatial covariance of the signals recorded on M/EEG sensors is a building block of modern data analysis pipelines. Such covariance estimates are used in brain–computer interfaces (BCI) systems, in nearly all source localization methods for spatial whitening as well as for data covariance estimation in beamformers. The rationale for such models is that the signals can be modeled by a zero mean Gaussian distribution. While maximizing the Gaussian likelihood seems natural, it leads to a covariance estimate known as empirical covariance (EC). It turns out that the EC is a poor estimate of the true covariance when the number of samples is small. To address this issue the estimation needs to be regularized. The most common approach downweights off-diagonal coefficients, while more advanced regularization methods are based on shrinkage techniques or generative models with low rank assumptions: probabilistic PCA (PPCA) and factor analysis (FA). Using cross-validation all of these models can be tuned and compared based on Gaussian likelihood computed on unseen data.

We investigated these models on simulations, one electroencephalography (EEG) dataset as well as magnetoencephalography (MEG) datasets from the most common MEG systems. First, our results demonstrate that different models can be the best, depending on the number of samples, heterogeneity of sensor types and noise properties. Second, we show that the models tuned by cross-validation are superior to models with hand-selected regularization. Hence, we propose an automated solution to the often overlooked problem of covariance estimation of M/EEG signals. The relevance of the procedure is demonstrated here for spatial whitening and source localization of MEG signals.

Introduction

Magnetoencephalography and electroencephalography (M/EEG) measure non-invasively the weak electromagnetic fields induced by post-synaptic neural currents (Hämäläinen et al., 1993). At the current state-of-the-art, the use of magnetoencephalography and electroencephalography M/EEG data for neural engineering and neuroscience poses inherent mathematical and statistical signal processing challenges. A brain–computer interfaces (BCI) system uses M/EEG data to classify brain states and control a device (Lotte et al., 2007). It involves tedious preprocessing, extraction of predictive features and the design of dedicated classifiers. A modern M/EEG analysis workflow for brain imaging involves segmentation of anatomical MRI data, the computation of an electromagnetic forward model, multiple data-coregistration steps, extraction of signals of interest from the raw measurements and finally a numerical solution to the ill-posed biomagnetic inverse problem (Gramfort et al., 2014; Hämäläinen et al., 2010). In this work, we focus on one problem of such analysis pipelines which is the estimation of between-sensor covariance, also referred to as spatial covariance. The interest for such covariance estimates is motivated by the physics of the forward problem and Gaussian assumptions, which are commonly made by M/EEG methods. Due to the linearity of Maxwell's equations, M/EEG data are obtained by linear mixing of brain sources. The signals are then corrupted by some additive noise. Assuming the source amplitudes are Gaussian, the measured data are also Gaussian due to linear mixing. Assuming the additive noise to be also Gaussian, brain signals and noise can be fully characterized with a mean vector and a covariance matrix. In practice signals are high pass filtered or "baseline corrected", which allows us to assume the data to be zero mean. The only quantities to be estimated from the data are therefore the spatial covariances.

* Corresponding author at: CEA, DSV/I2BM, NeuroSpin, Bat. 145, 91191 Gif-sur-Yvette Cedex, France.
E-mail address: denis.engemann@cea.fr (D.A. Engemann).

http://dx.doi.org/10.1016/j.neuroimage.2014.12.040
1053-8119/

Data recorded with M/EEG can be used to localize the neural generators underlying the measurements. This procedure is known as the M/EEG inverse problem. Methods addressing this challenge are referred to as inverse solvers or inverse methods. While dipole models typically estimate the location of a few sources in the brain, so-called distributed source models consist of thousands of candidate dipolar sources which are positioned based on anatomical information. Methods from the minimum-norm estimates (MNE) family (weighted MNE (Lin et al., 2006), low resolution brain electromagnetic tomography (sLORETA) (Pascual-Marqui, 2002), dynamical statistical parametric mapping (dSPM) (Dale et al., 2000), mixed-norm estimates (MxNE) (Gramfort et al., 2012), time-frequency mixed-norm estimates (TF-MxNE) (Gramfort et al., 2013a)), as well as beamformers in the time domain, e.g. linear constrained minimum-variance (LCMV) (Veen et al., 1997), or the frequency domain, e.g. dynamic imaging of coherent sources (DICS) (Gross et al., 2001), all require the definition of a distributed source model and necessitate the estimation of spatial covariance matrices. Some of these solvers are non-linear: MxNE and TF-MxNE cannot be expressed as a single matrix multiplication applied to the data. Some are known as adaptive spatial filters: LCMV and DICS require the estimation of the covariance of the data. Yet, all these methods involve Euclidean ℓ_2 constraints which inherently assume Gaussian noise with zero mean and equal variances across sensors. To improve data with regard to these requirements a spatial whitening step is commonly implemented that allows to suppress between-sensor correlations related to noise. More specifically, a spatial covariance of the additive noise is estimated from data and subsequently used for whitening. This transforms data into independent white noise vectors characterized by an identical variance across channels.

The problem of estimating the covariance from multivariate samples is a problem that has been widely studied in statistics and for which various models have been proposed. In one such approach (Chen et al., 2010; Ledoit and Wolf, 2004) optimal coefficients are computed for the shrinkage of off-diagonal terms while other contributions propose structured models with reduced rank assumptions (Barber, 2012; Tipping and Bishop, 1999). In the context of M/EEG, noise can be biological (heart beat, eye blinks, muscle activity), environmental (line noise) and sensor-related. Purely sensor-related noise can be assumed to be independent across sensors. It can hence be modeled with a diagonal covariance matrix. In contrast, most sources of noise are structured and induce strong correlations between sensors. When estimating the spatial covariance from signal of interest as done for beamformers (Veen et al., 1997) or BCI for common spatial patterns (CSP) (Ramoser et al., 1998), strong between-sensor correlation occurs and can be explained by the following fact. If we assume one active source in the brain without the presence of noise, the linearity of the forward problem guarantees that the measured data span a subspace of dimension one. If we now assume that the source rotates, for a spherical head model, the subspace dimension is two in the case of MEG and three in the case of EEG (see for example (Mosher and Leahy, 1998) for discussions on this matter). Low rank hypotheses are also relevant for some MEG systems where the data are projected to a low rank signal subspace for denoising. This technique is known as signal space separation (SSS) (Taulu et al., 2005). Another peculiarity of modern MEG systems is the different sensor types used during recordings, e.g. magnetometers and planar gradiometers on Neuromag VectorView systems. These impose additional difficulties to the estimation because values differ by orders of magnitude between sensors while the sources captured only partially overlap.

We will therefore evaluate various strategies for the estimation of the spatial covariance of M/EEG data under Gaussian assumptions and develop a systematic approach of deciding between these alternatives. The study will focus on two particular kinds of approaches, shrinkage covariance estimators (Chen et al., 2010; Ledoit and Wolf, 2004) and on generative low rank models, also commonly referred to as latent variable models: probabilistic principal component analysis (PPCA) and factor analysis (FA) (Barber, 2012; Tipping and Bishop, 1999). In a first step, relevant statistical models and inference methods will be introduced and discussed in the context of M/EEG. Subsequently, implementation strategies will be detailed. Finally, we will present a comprehensive quantified evaluation of six approaches. This evaluation will be based on simulations and three M/EEG datasets. Impact of the proposed method on source localization results will be illustrated on a publicly available cognitive neuroscience dataset (Henson and Rugg, 2003).

Material and methods

Before detailing the covariance estimation models, we provide a motivating example: the problem of source reconstruction with ℓ_2 regularization, also known as minimum-norm estimates (MNE).

Minimum-norm estimates (MNE)

Minimum-norm estimates employ a distributed source model that consists of a large number of spatially fixed candidate dipoles whose amplitudes are estimated from the data (Gramfort et al., 2014; Hämäläinen et al., 2010). Let us denote by N the number of sensors, M the number of candidate dipoles and T the number of time samples in the data. Following the linearity of Maxwell's equation and the assumption of additive noise, the data matrix Y of size $N \times T$ is obtained by multiplication of the forward gain matrix G of size $N \times M$ by X, the unknown sources amplitudes of size $M \times T$, to which is added a noise term E of size $N \times T$. The model reads:

$$Y = GX + E.$$

The model can then be further specified by assuming that X and E have zero mean Gaussian distributions at each time sample t, i.e. $X_t \sim \mathcal{N}(0, R)$ and $E_t \sim \mathcal{N}(0, C)$. The matrices R and C, of size $M \times M$ and $N \times N$ respectively, refer to the source covariance and the noise covariance. Assuming C and R to be known, an estimate $\hat{X}$ of the amplitudes of the dipoles located on the cortical mantle is obtained by maximum a posteriori (MAP):

$$\hat{X} = \underset{X \in \mathbb{R}^{M \times T}}{\arg\min} \|Y - GX\|_C^2 + \|X\|_R^2$$

where $\|A\|_B^2 = \mathrm{Trace}(A^t B^{-1} A)$. This leads to:

$$\hat{X} = RG^t \left(GRG^t + C\right)^{-1} Y, \tag{1}$$

where G^t stands for the matrix transposition of G.

The noise is said to be white if the matrix C is the identity I. Let us denote by $C^{\frac{1}{2}}$ a square root matrix of C, such that $C^{\frac{1}{2}} C^{\frac{1}{2}} = C$. Note that there is no unique square root of a matrix. If C is invertible, so is $C^{\frac{1}{2}}$. If one denotes by $\tilde{Y} = C^{-\frac{1}{2}} Y$ and $\tilde{G} = C^{-\frac{1}{2}} G$ then Eq. (1) is equivalent to:

$$\hat{X} = R\tilde{G}^t \left(\tilde{G}R\tilde{G}^t + I\right)^{-1} \tilde{Y}. \tag{2}$$

In other words, after introducing $\tilde{Y}$ and $\tilde{G}$, the noise can be modeled as white. One can observe that Eq. (2) resembles Eq. (1) after replacing C by I. The process of computing $C^{\frac{1}{2}}$ and subsequently $\tilde{Y}$ and $\tilde{G}$ is called spatial whitening. The matrix $\tilde{Y}$ contains the whitened data, and $\tilde{G}$ is referred to as the whitened gain matrix.

In practice the square root $C^{\frac{1}{2}}$ is obtained from the eigenvalue decomposition under symmetry constraints of the estimated covariance $C = U_C \Lambda_C^2 U_C^t$ where U_C is an orthonormal matrix, $U_C U_C^t = I$, and Λ_C is a diagonal matrix with non negative entries. Assuming C to be full rank, it is

straightforward to verify that $C^{-\frac{1}{2}} = \Lambda_C^{-1} U_C^t$ is a valid inverse square root of C. An alternative is the symmetric matrix $C^{-\frac{1}{2}} = U_C \Lambda_C^{-1} U_C^t$, which we will use to visualize whitened data.

To reduce redundancy: Eq. (2) reveals that minimum-norm estimates (MNE) actually implements what is known as Tikhonov regularization (Tikhonov and Arsenin, 1977) or Ridge regression in the field of statistical learning (Hoerl and Kennard, 1970). As a consequence, if the gain matrix and the data are appropriately whitened, general conditions of statistical regression models apply to the magnetoencephalography and electroencephalography (M/EEG) inverse problem. Minimum-norm estimates, therefore, rely on the specification of the noise covariance matrix that needs to be estimated from the data. Or in other words, the quality of the inverse solution depends on the quality of the covariance estimate. This holds true for most other source localization in particular beamformers such as LCMV and DICS (Veen et al., 1997; Gross et al., 2001). However, this also applies to MNE variants such as dynamical statistics parametric mapping (dSPM) (Dale et al., 2000) or low resolution brain electromagnetic tomography (sLORETA) (Pascual-Marqui, 2002), as well as other distributed models such as minimum-current estimates (MCE) (Uutela et al., 1999) or mixed-norm estimates (MxNE) (Gramfort et al., 2012; Gramfort et al., 2013a). It therefore cannot be considered a local problem.

Covariance estimation

Model selection using cross-validation

The noise covariance estimator is typically applied to segments of (M/EEG) data that were not used to estimate the noise covariance and that typically include both, brain signals and noise. Its quality can hence be assessed by investigating how well the model describes new data. This idea of model quality assessment on unseen data is put into practice by aggregating results over random partitions of the data, and is referred to as cross-validation. Since data are assumed to follow a multivariate Gaussian distribution, parameterized by a covariance matrix C, the log-likelihood of some data Y reads:

$$\mathcal{L}(Y|C) = -\frac{1}{2T}\mathrm{Trace}\left(YY^t C^{-1}\right) - \frac{1}{2}\log\left((2\pi)^N det(C)\right). \quad (3)$$

The higher this quantity on unseen data, the more appropriate the estimated noise covariance C and the higher its success at spatially whitening the data. The log-likelihood, hence, allows us to select the best noise covariance estimators out of a given set of models using cross-validation with left out data. In the following we will discuss potentially relevant candidate strategies to estimate covariance matrices on M/EEG data.

Empirical covariance and regularization

The empirical covariance matrix can be computed by $C = \frac{1}{T} YY^t$, where Y contains the data of size $N \times T$. With a sufficient number of observations (T large), the sample covariance, which can be derived from maximum likelihood, is a good estimator of the true covariance. Typically, a noise covariance is computed on baseline segments preceding stimulation or for MEG on empty room measurements during which no subject is present. The latter is however not possible for electroencephalography (EEG) recordings for which the covariance estimation relies on data segments considered not relevant for the task, typically during baseline. Biological artifacts often contaminate the data leading to outlier samples, and sometimes the data statistics change over time, for example due to changes in environmental noise or changes in head position. If in such situations only a limited number of samples is available, the empirical covariance tends to to suffer from high variance. The estimate then is noisy and unreliable for further analysis.

One typical way to reduce the variance of the covariance estimator is to apply *diagonal loading*. It consists of amplifying the diagonal with a hand-selected constant which attenuates the off-diagonal elements that correspond to inter-sensor covariance:

$$C' = C + \alpha I, \quad \alpha > 0. \quad (4)$$

The value α is the regularization parameter. This diagonal weighting of the covariance stabilizes MNE-like estimates by reducing the variance. However, the introduced bias amounts to assuming a stronger uncorrelated noise level which leads to underestimated amplitudes in the source estimates. This especially applies to dSPM and sLORETA where the noise variance is used to rescale MNE estimates and convert them to statistical quantities such as F or T statistics. When used in beamformers, such a regularization of the data covariance matrices tends to increase the point spread function of the spatial filters and smear the estimates (Woolrich et al., 2011). In addition, hand-set regularization raises a new problem, which is how to choose the value of α.

Shrinkage models

An improvement of the hand-selected regularization or shrinkage approach introduced in the section called Covariance estimation is provided by the Ledoit–Wolf (LW) shrinkage model (Ledoit and Wolf, 2004). This covariance model constitutes an optimal weighted average of the invariant identity matrix and the empirical covariance matrix (Eq. (5)). The LW covariance estimates C_{LW} takes the form of:

$$C_{LW} = (1-\alpha)C + \alpha\mu I, \quad (5)$$

where I stands for the identity matrix, μ is the mean of the diagonal elements of C, and α is called the shrinkage parameter. The contribution of Ledoit and Wolf (2004) is to provide a formula to compute the optimal value for α. The solution is derived from the values of N, the number of dimensions, and T, the number of samples. It is provided in closed form and minimizes the mean squared error between the estimator and the population covariance. The underlying assumptions of the LW estimator are that the data are i.i.d. (independent identically distributed) which, as we will see below, is not a valid assumption for M/EEG data. However, Ledoit and Wolf (2004) have shown that the optimal shrinkage parameter guarantees C_{LW} to be well conditioned: matrix inversion is numerically stable, and more stable than with the empirical covariance.

A data-driven extension to the Ledoit–Wolf estimator can be motivated by Eq. (5). Instead of using the Ledoit–Wolf formula to compute α, cross-validation and likelihood estimation on unseen data can be compared over a range of α values to select the optimal regularization parameter. The optimal α can then be determined as the one yielding a covariance estimator with the maximum likelihood on unseen data. Throughout the manuscript, models with data-driven shrinkage coefficient as in Eq. (5) will be referred to as SC.

Probabilistic principal component analysis (PPCA)

M/EEG measurements are obtained by sensor recordings at various locations in space. They include signals from the brain but also artifacts. Such signals and artifacts yield spatially structured patterns on the sensor array. For example, a source in the brain that would be well modeled by an equivalent current dipole ECD produces a dipolar pattern on the sensors. If this dipole does not rotate, due to the physics of the forward problem, the signal space spanned by this ECD is of dimension one. The signal space is thus said to be of rank one. Both sources in the brain and artifacts share this property of generating low rank signals on the sensors. This is for example what justifies the use of signal space projection SSP (Uusitalo and Ilmoniemi, 1997). The idea behind SSP is that the noise subspace includes artifact-related sources of low dimensionality and that it is approximately orthogonal with the subspace spanned by the brain signals of interest. Therefore, projecting the data on the orthogonal of the noise subspace will remove artifacts and therefore denoise the data.

Principal component analysis (PCA) is a statistical method that is built on this idea of low rank signal space. When using classical PCA

one needs to pre-specify the number of components, which matches the rank of the subspace. While PCA was historically introduced as a method to reduce the dimension of data, or to approximate a matrix with one of lower rank, Tipping and Bishop (1999) have explained how it can be reframed as a generative probabilistic model and coined the term probabilistic PCA (PPCA). According to this perspective, PPCA corresponds to a multivariate Gaussian model where a random vector can be expressed as a random weighted linear combination of components added to some independent noise. The covariance can be decomposed as the sum of a low rank matrix and a scaled identity matrix. With this statistical model standard PCA is transformed into a latent variable model such as (FA).

To give a more formal description of the PPCA model, let K represent the number of components and y a sample generated by the model. The N-dimensional vector y is then obtained from a K-dimensional random vector w which is linearly transformed by K latent factors forming a matrix H of size $N \times K$, to which is added a fixed N-dimensional vector m and a random noise vector e:

$$y = Hw + m + e. \tag{6}$$

Both w and e are independent random vectors obtained from spherical[1] multivariate Gaussian distributions, respectively of size K and N. Without loss of generality, the covariance of w is the identity I_K and the covariance of e is $\sigma^2 I_N$:

$$e \sim \mathcal{N}\left(0, \sigma^2 I_N\right) \quad \text{and} \quad w \sim \mathcal{N}(0, I_K). \tag{7}$$

It naturally follows that given H, m and σ, the vector y is Gaussian:

$$y|H, m, \sigma \sim \mathcal{N}\left(m, HH^t + \sigma^2 I_N\right). \tag{8}$$

As a result, the covariance derived from the PPCA model is given by:

$$C_{PPCA} = HH^t + \sigma^2 I_N. \tag{9}$$

The natural question is then how to estimate m, H and σ from the data, and why the standard PCA method provides estimates of these quantities. Let us denote by $Y = \{y_1, \ldots, y_T\}$ the observed data. According to the PPCA model the likelihood of the data is expressed by:

$$p(Y|H, m, \sigma) = (2\pi)^{-\frac{TM}{2}} det\left(HH^t + \sigma^2 I_N\right)^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\text{Trace}\left(\left(HH^t + \sigma^2 I_N\right)^{-1} S\right)\right), \tag{10}$$

where

$$S = \sum_i (y_i - m)(y_i - m)^t. \tag{11}$$

The maximum-likelihood estimates of each parameter are given by Eq. (12) (Minka, 2000).

$$\hat{m} = \frac{1}{T}\sum_{i=1}^{T} y_i \qquad \hat{\sigma}^2 = \frac{\sum_{j=K+1}^{M} \lambda_j}{M-K} \qquad \hat{H} = U\left(\Lambda - \hat{\sigma}^2 I_K\right)^{\frac{1}{2}} Q, \tag{12}$$

where U is the matrix formed by the K top eigenvectors of S, the diagonal matrix Λ contains the corresponding eigenvalues λ_1 to λ_N while Q is a random orthogonal matrix. Importantly, to recover the principal components given by standard PCA, this matrix Q needs to be an identity matrix. From this it naturally follows how C_{PPCA} can be derived from standard PCA estimates.

[1] With identity covariance matrix.

The latter results are obtained assuming the number of components K to be known. In order to estimate this number from the data, various strategies have been developed. In Bishop (1999) Bayesian PCA has been proposed as an extension of PCA in which hyperparameters control the number of dimensions. This technique was used in Woolrich et al. (2011) in the context of linear constrained minimum-variance (LCMV) beamformers to estimate the spatial covariance of the data and its rank. In contrast, Minka (2000) proposed a Bayesian rank estimation technique based on Laplace approximation where inference is obtained from a variational Bayes approach. The resulting rank estimate will be referred to in the following as PCA Bayes. Finally, as detailed in Minka (2000), cross-validation can be used to obtain rank estimates based on PPCA without introducing additional hyperparameters as used in Bayesian PCA. With this approach, PPCA models are estimated on a fraction of the data over all possible numbers of components while the Gaussian likelihood of left out data is used as a principled quantitative measure to evaluate how well the model fits the data. The estimated number of components, K, is the value that maximizes the Gaussian likelihood of the left out data. In the course of the manuscript we will be focusing on the two latter approaches, Bayesian estimation with Laplace approximation and cross-validation.

Factor analysis (FA)

FA is another latent variable model that can be regarded as extension of PPCA (Tipping and Bishop, 1999; Barber, 2012). The crucial difference to PPCA is that instead of assuming a spherical noise, $e \sim \mathcal{N}(0, \sigma^2 I_N)$, it assumes a diagonal covariance, $w \sim \mathcal{N}(0, \Psi)$, where Ψ is diagonal with diagonal positive entries denoted by $\psi_1, \ldots, \psi_N$. PPCA is said to assume a homoscedastic noise: the noise variance is the same for all variables, here all sensors. Contrastingly, FA assumes a heteroscedastic noise: the noise variance differs between sensors.

The covariance as delivered by FA is given by:

$$C_{FA} = HH^t + diag(\psi_1, \ldots, \psi_D). \tag{13}$$

Factor analysis therefore covers a richer class of models and can be more suitable for data such as M/EEG where the noise varies between sensors, for example, due to undetected bad channels, or when combining different sensor types, e.g. magnetometers and gradiometers. The consequence of this difference between PPCA and FA models, is that the component matrix in FA differs from the principal components, also referred to as principal axes of the data (Tipping and Bishop, 1999). This implies in practice that the FA model parameters cannot be inferred as easily as with PPCA. Indeed, no closed form solution is available for FA. Inference for FA hence relies on an iterative algorithm. Due to its diagonal noise term, FA can cope with more complex noise structures in which noise variance varies across channels. It suggests that FA can describe data with fewer dimensions than PPCA but also that it can cope with more datasets. However, this flexibility has its price: estimating more complex models requires more samples.

The estimation of the FA model parameters is performed using Expectation Maximization (EM) as described in Barber (2012). In practice each iteration consists of a spatial whitening of the data using the present estimate of the data covariance followed by an update of the components. This later step is performed with a singular value decomposition SVD, which is also used to compute the standard PCA solution. Usually a minimum of 20 iterations is necessary to reach convergence of the FA estimation on M/EEG data. FA is therefore significantly slower to compute than a PCA. However, thanks to randomized numerical linear algebra (Martinsson et al., 2011), FA computation can be significantly sped up making covariance estimation based on FA and cross-validation tractable, even when combining MEG and EEG. Such an efficient implementation is provided in the scikit-learn machine learning library (Abraham et al., 2014; Pedregosa et al., 2011).

Evaluation metrics: whitened data and global field power

We now detail model quality evaluation metrics that we will use in the experiments. The whitened evoked response is a sensor space metric which is obtained by multiplying the array of sensor measurements by the symmetric whitener:

$$C^{-1/2}Y. \tag{14}$$

The resulting signals should follow a standard-normal distribution. The amplitudes are expected to be situated between -1.96 and 1.96 for baseline segments from which the covariance was estimated. This result follows from the 2.5% and 97.5% quantiles of the standard-normal distribution. In other words, 95% of the data should be in that range of values.

The whitened global field power GFP is a second sensor space metric that quantifies variability over the full sensor array at a given time sample. We define this GFP, or more precisely the whitened rank-adjusted GFP, as:

$$\frac{\sum_{i=1}^{N} \chi_i^2}{P} \tag{15}$$

where P is the rank of the data and N is the number of sensors. Should the dimensionality of the data have been previously reduced then $P < N$. This typically happens when an independent component analysis (ICA), SSP or signal space separation (SSS) has been applied to the data. If no rank reduction has been applied then $P = N$. When computing the GFP on the whitened evoked data, it appears that the numerator in Eq. (15) is a χ^2 random variable with P degrees of freedom so that expected value of the GFP as defined here is 1. On actual whitened data, deviation of the GFP from 1 will indicate an improper whitening.

General data analysis and software

All covariance estimators and the cross-validation were computed using the Python machine learning package scikit-learn (Pedregosa et al., 2011). The empirical covariance and the regularization were computed using the MNE software (Gramfort et al., 2013b; Gramfort et al., 2014). The FA implementation was based on algorithm 21.1 from Barber (2012). Estimation of FA parameters is iterative with expensive SVDs, one at each iteration. To improve suitability for cross-validation and extensive rank estimation, we contributed a modified implementation of factor analysis to the scikit-learn package, based on the randomized SVD algorithm (Halko et al., 2011; Martinsson et al., 2011). While producing results equivalent to a full SVD, the randomized SVD uses significantly less memory and allowed to cut computation times by up to a factor of seven.[2]

The MNE software (Gramfort et al., 2013b; Gramfort et al., 2014) was used to process and analyze all magnetoencephalography (MEG) and EEG data. For the source space analyses, the FreeSurfer[3] software was used to obtain cortical surface reconstructions.

Simulated data

To compare the behavior of the covariance estimators across a varying numbers of samples, four different data scenarios were simulated. They can be represented on a 2 (homoscedastic VS heteroscedastic noise) by 2 (low VS high rank) grid. For each scenario, covariance estimates and rank estimates were computed for PPCA, the PCA (Bayes) and FA with a continuously increasing number of samples. In addition, model likelihood was computed for the Ledoit–Wolf and the shrunk covariance (SC) estimator as well as for PPCA and FA. To reduce data variability, results were averaged over 50 runs using different random seeds. The data were simulated as follows: to obtain low rank data, a random $N \times N$ square matrix was computed. Number of dimensions was set to $N = 50$. In a second step, the rank of the matrix was reduced by applying a truncated SVD. The K singular vectors with highest singular values were kept to form a matrix H of size $N \times K$ as in Eq. (6). An arbitrary orthogonal matrix of size $K \times T$ was then used to form T independent samples that were projected using H into the N dimensional space. The outcome is a $N \times T$ dataset living in a subspace of dimension K. Finally, either homoscedastic or heteroscedastic Gaussian noise was added to the data. This was achieved by adding a $N \times T$ random matrix formed by T samples drawn from Gaussian distributions with diagonal covariances. In the heteroscedastic case the entries on the diagonal are all positive but different (each feature, sensor, is corrupted with a different noise level), while in the homoscedastic case all the entries on the diagonal are positive and equal. The rank was set to either $K = 10$ (low rank) or to $K = 40$ (high rank). T was varied between 200 and 2000 in steps of 50.

To determine the optimal SC estimator with cross-validation, estimators were computed with α varying on a logarithmic grid of 30 values between 0.01 and 1. Each estimator was then evaluated with a three-fold Monte Carlo cross-validation procedure. The optimal shrinkage was then determined based on the highest likelihoods on left out data. To determine the hyperparameter K of the low rank models, PPCA and FA were computed on a grid of rank values. K varied between one to 49 in steps of three. For the sake of completeness, 50 (the number of observed dimensions) was included in this range. Each value was used to select the number of dimensions directly. At each step, the models obtained were evaluated with the same cross-validation procedure. The estimated rank was then determined by the K parameter of the model with the highest log-likelihood.

M/EEG datasets

The covariance and rank estimation procedures were subsequently tested using MEG data recorded by three commercial and widely used meg systems: 1) a 4D-Neuroimaging whole-head magnetometer system with 248 channels (MAGNES-3600WH MEG), 2) a VSM MedTech Inc. whole-head axial gradiometer system with 275 channels (CTF/VSM) using second-order axial gradiometers and synthetic third gradient for denoising and 3) a Neuromag VectorView whole-head system with 306 channels (Elekta Neuromag, Finland), which are formed from 102 sensor triplets, each comprising two orthogonal planar gradiometers and one magnetometer.

The Neuromag dataset is shipped with the MNE software (Gramfort et al., 2013b; Gramfort et al., 2014) and includes combined M/EEG recordings conducted at the Martinos Center of Massachusetts General Hospital. EEG was recorded simultaneously using an MEG-compatible cap with 60 electrodes. Data were sampled at 600 Hz. In the experiment, auditory stimuli (delivered monaurally to the left or right ear), and visual stimuli (shown in the left or right visual hemifield) were presented in a random sequence with a stimulus onset asynchrony (SOA) of 750 ms.

The CTF/VSM data-set includes MEG recordings conducted by the Functional Imaging Laboratory, London. It is available on the SPM webpage[4] (Litvak et al., 2011) and can also be downloaded using the MNE software.[5] Data were sampled at 480 Hz. In this experiment, faces and scrambled faces were presented to the participant. The paradigm is detailed in Henson and Rugg (2003).

The 4D-Neuroimaging dataset was kindly provided by Breuer et al. (2013). Recordings were conducted at the Institute of Neuroscience and Medicine (INM-4), Forschungszentrum Jülich, Germany and sampled at 1017.25 Hz. In the experiment, auditory stimuli (simple

[2] cf. https://github.com/scikit-learn/scikit-learn/pull/2406.
[3] http://surfer.nmr.mgh.harvard.edu/.
[4] http://www.fil.ion.ucl.ac.uk/spm/data/mmfaces/.
[5] http://martinos.org/mne/auto_examples/datasets/plot_spm_faces_dataset.html.

sinusoidal tones at 1000 Hz and 2000 Hz) were presented to the participant in a random sequence with a SOA of 1000–2000 ms.

All data were bandpass filtered between 1 and 45 Hz using a zero-phase 4th order Butterworth filter. The low pass at 45 Hz excluded the power line frequencies at 50 Hz and 60 Hz for data recorded in Europe and the USA. The high pass at 1 Hz removed low-frequency drifts as well as baseline offsets from the data. To allow the comparison of the results obtained with the different datasets, all epochs were resampled at 150 Hz. Segments contaminated by biological artifacts were detected based on peak-to-peak amplitude and ignored during estimation to avoid distorted covariance estimates due to outliers. Note that this led to slightly different sample sizes when comparing datasets.

MEG data expressed in T or T/m are very small, and close to machine precision. To improve numerical stability, data were scaled by the order of magnitude corresponding to the measurement unit, for example by in the case of magnetometers. For datasets combining gradiometers and magnetometers, the latter were upweighted as recommended by MaxFilter software (Elekta-Neuromag). A factor of 4 was chosen as this value turned out to improve numerical stability. The estimated covariances were then rescaled to the squared measurement unit. Epochs were defined from −200 ms to 500 ms with respect to the stimulus onset. To estimate the noise covariance, baseline segments (−200 to 0 ms) were extracted and concatenated to form a two-dimensional matrix comprising channels and time samples.

Sensor space validation

The same protocol was applied to the M/EEG datasets as for the simulation. For each dataset, covariances and their log-likelihoods were computed based on each estimator. The PPCA and FA parameters were evaluated using cross-validation over a range of different values for rank parameter K from five and to the multiple of five that was closest to the actual number of channels, advancing in steps of five. Subsequently, the log-likelihood, the whitened evoked response and the corresponding GFP were computed for each estimator. The log-likelihood scores were then used to inform model selection. Graphical displays of log-likelihood scores were computed to illustrate the data on which model selection was based. The procedure was executed separately for each channel type as well as for magnetometers and gradiometers combined. For the combined-sensors runs, whitening effects related to either gradiometers or magnetometers are presented separately. The acronyms used to refer to the different datasets or to the views on datasets are summarized in Table 1.

This procedure was conducted at two discrete sample sizes, one including the first 15 epochs encompassing 465 samples, and a second one including the first 50 epochs of 1550 samples. These values reflect arbitrary choices. However, both levels approximate the lower and the upper bounds for the number of samples used for the simulation.

For each dataset, whitened-evoked responses were then computed based on the covariance estimator with the highest model likelihood. The ensuing display is informative in two ways. First, assuming a correct whitening, for baseline whitened evoked responses, 95% of the amplitudes are expected to have a value between −1.96 and 1.96 (cf. Evaluation metrics: whitened data and global field power). Second, the post-baseline segments are important to evaluate as appropriate whitening should leave intact the 'butterfly' shape which is typical for visualizations of evoked responses in which time courses from multiple channels are superimposed.

A second graphical monitoring technique was implemented by computing the whitened GFP for each estimator and super-imposing the results separately for each estimator. The estimators tested on M/EEG data are presented in Table 2 with their corresponding abbreviations. Assuming correct whitening, the whitened GFP should produce baseline scores around one (cf. Evaluation metrics: whitened data and global field power). Importantly, post-baseline segments easily reveal incorrect scaling, i.e., if GFP scores do not return to the baseline where it would be expected. Moreover, this approach allows to compare the impact of different estimators more directly, as the sensor signals are reduced to a single time course. It is noteworthy that both techniques allow to compare baseline and post-baseline segments to visually assess the impact of a given estimator on the signal-to-noise ratio. For comparison, the raw, non-whitened evoked responses are displayed in Fig. 1.

Source space validation

To demonstrate the practical impact of estimator quality on source localization in applied contexts, the single subject SPM-faces dataset described above was analyzed at the source level using the above covariance estimation and selection procedure. This dataset was chosen because it implicates experimental contrasts relevant to cognitive and social–cognitive neuroscience. Data were not resampled and a cutoff frequency of 30 Hz was used. Apart from this, data were preprocessed as for the sensor space validation. MNE source estimates were then computed separately for the faces and the scrambled faces condition. Resulting maps of cortical activity maps were then subtracted to form a paired contrast. Except for the covariance parameter, MNE estimates were computed using the default parameters proposed by the MNE-software. The regularization-parameter was set to $1.0/\mathrm{SNR}^2$ where SNR refers to the signal-to-noise ratio parameter which defaults to 3. A depth-weighting of 0.8 was used in combination with a loose-constraint of 0.2 and free orientation. The dSPM procedure was used for noise normalization. This resulted in unsigned dSPM source estimates reflecting normalized current magnitude. Positive values resulting from a paired contrast of the form $dSPM_{faces} - dSPM_{scrambled}$ reflect activity specific to the faces condition.

To assess the differential impact of covariance estimation, this analysis was conducted over varying numbers of epochs using both best and the worst noise covariance estimator (cf. MNE computation in Eq. (1)). To quantify the different statistical properties of the resulting dSPMS, means and standard deviations were computed across source locations at any given time sample. This summarizes the spatial variability of the dSPM maps.

Table 1
Overview on datasets used and corresponding legend keys.

Key	Dataset and channel type	Number of channels used
bti	4D Magnes 3600 WH magnetometers	248
ctf	CTF-275 axial gradiometers	274
vv-eeg	VectorView EEG electrodes	59 (1 bad)
vv-grad	VectorView planar gradiometers	203 (1 bad)
vv-mag	VectorView magnetometers	102
vv-meg-grad	VectorView planar gradiometers, combined estimation	203 (1 bad)
vv-meg-mag	VectorView magnetometers, combined estimation	102

Table 2
Overview on covariance estimators used in concert with M/EEG data.

Key	Estimator
Raw	Empirical covariance computed from restricted number of epochs
Reg	Regularized covariance with $\alpha = 0.1$ (default regularization parameter in the MNE software)
LW	Ledoit–Wolf estimator
SC	Shrunk covariance with cross-validation
PPCA	Probabilistic PCA with cross-validation to set K
FA	Factor analysis with cross-validation to set K

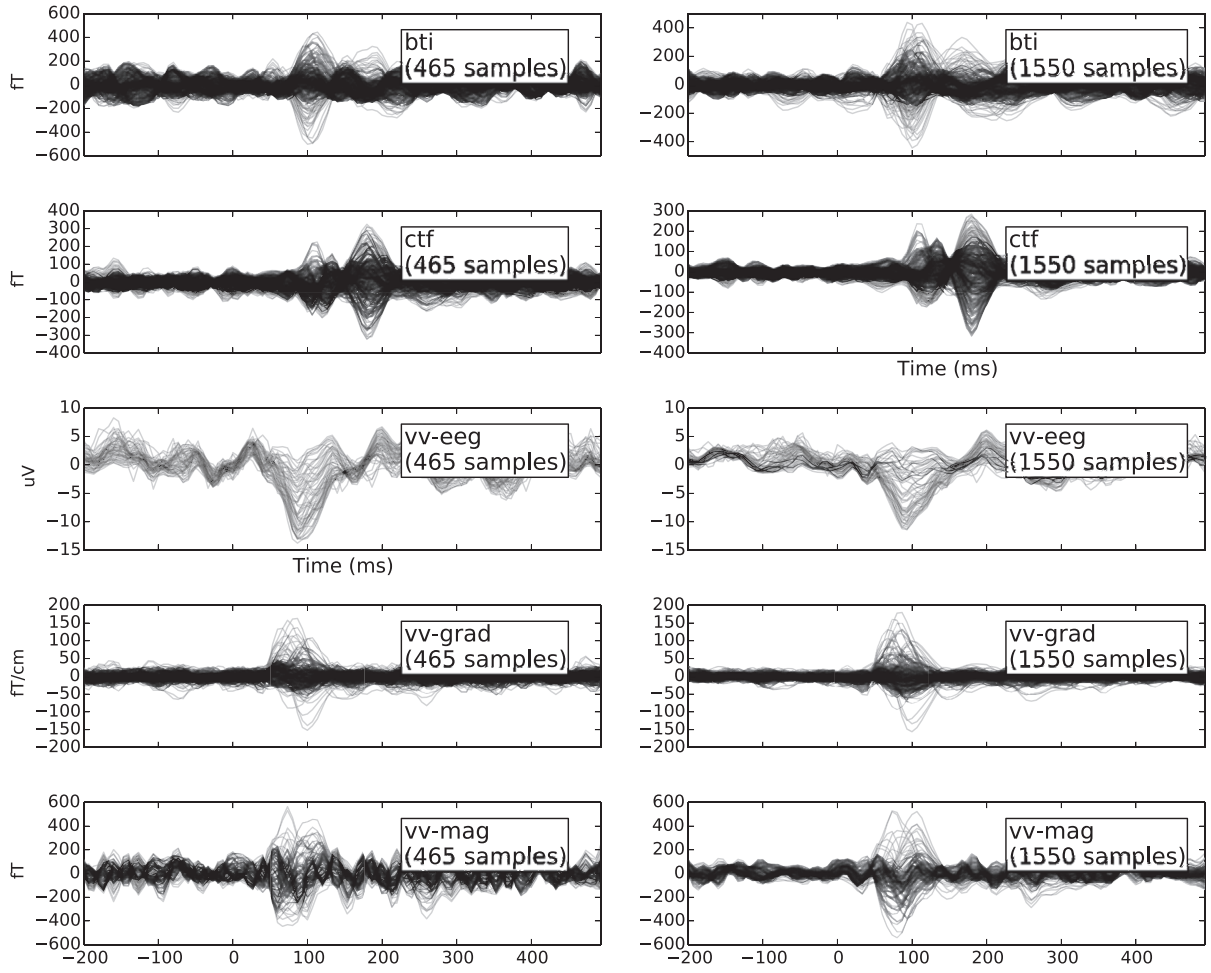

Fig. 1. Non-whitened evoked responses of all datasets for 15 epochs (450 samples) and 50 epochs (1500 samples).

Results

Simulated data

Rank estimation

Fig. 2 presents the rank estimation results based on PPCA, Minka's Bayes PCA (Minka, 2000), and FA. All three estimators recovered the true rank of the data when noise was homoscedastic. When heteroscedastic noise was present, only FA was able to recover the true rank, irrespective of the true rank (10 or 40). When the noise was heteroscedastic, dramatic overestimations of the true rank occurred for PPCA. Furthermore, it can be observed that PPCA and FA only produced stable results if the sample size exceeded a minimum of roughly 350 samples.

Model likelihood

This is further illustrated in Fig. 3 which shows the model likelihoods of the covariance estimators. For all conditions, the model likelihood increased with the number of samples, and more steeply in the range where the rank estimates exhibited high instability. In the low rank scenario, the latent variable models were unequivocally more appropriate than the "unstructured" shrunk covariance models. For homoscedastic noise and a rank of 10, both PPCA and FA performed equal. When noise was heteroscedastic, FA had the highest model likelihood across the entire sample range, followed by the shrunk covariance models and PPCA. However, differences between the other estimator's performance disappeared with increasing number of samples. In contrast, the high rank scenario was governed by a different regime. Independent of the noise structure, a clear performance pattern emerged where SC exhibited the best results at a low number of samples while the probabilistic latent variable models only gradually improved with increasing numbers of samples, ultimately reaching comparable model likelihoods.

Sensor space validation

Rank estimation

The results on magnetoencephalography and electroencephalography (M/EEG) data are presented in Fig. 4. Probabilistic PCA and FA both indicated a low rank structure for the data, except for the EEG scenario with the larger sample size where PPCA suggested full rank. On average, the FA rank estimate ($M_K = 35.833$, $SD_K = 14.410$) was lower than the PPCA rank estimate ($M_K = 42.083$, $SD_K = 17.376$).

Second, the estimated rank was generally higher for the high number of samples ($M_K = 26.667$, $SD_K = 1.952$) as compared to the low number of samples scenario ($M_K = 51.250$, $SD_K = 0.203$).

Model likelihood

The model likelihoods of the covariance estimators are reported in Fig. 5. Three main observations can be made. First, the automatically selected covariance estimator were consistently more appropriate than the empirical covariance. Second, FA and SC consistently delivered better models than PCA and Ledoit–Wolf (LW) respectively. The SC estimator prevailed where the number of samples was lower while FA produces the most appropriate fit when applied to multi-sensor datasets or when the number of samples was high.

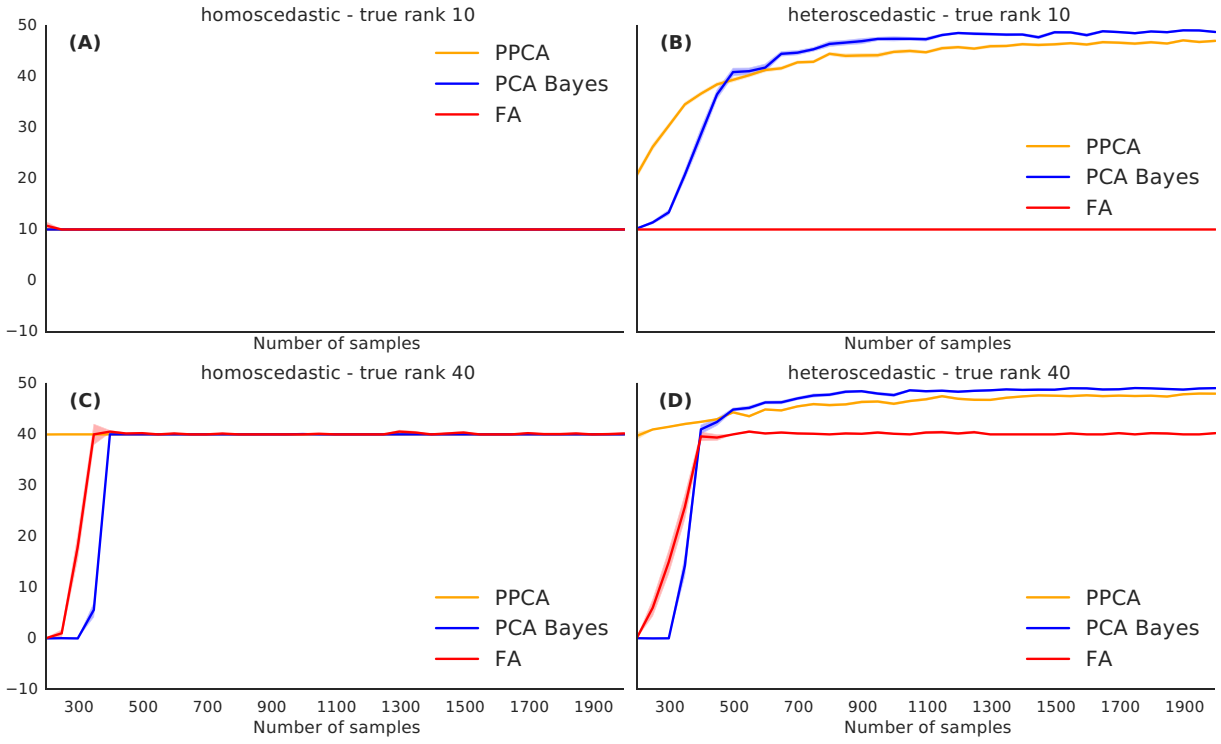

Fig. 2. Comparison between rank estimators on simulated data. Areas represent the standard error of the mean. Panel (A). Homoscedastic noise, ground truth rank of 10. All rank estimators converge. Panel (B). Same ground truth as in (A) but heteroscedastic noise. Only FA recovered the true rank. Panel (C). Homoscedastic noise, ground truth rank of 40. The low rank estimators converge, however, the sample size must be sufficiently large for FA and PCA Bayes rank estimation procedures. Panel (D). Same rank as in (C), heteroscedastic noise. Note that FA consistently recovered the true rank of the data but only if the number of samples exceeded a minimum.

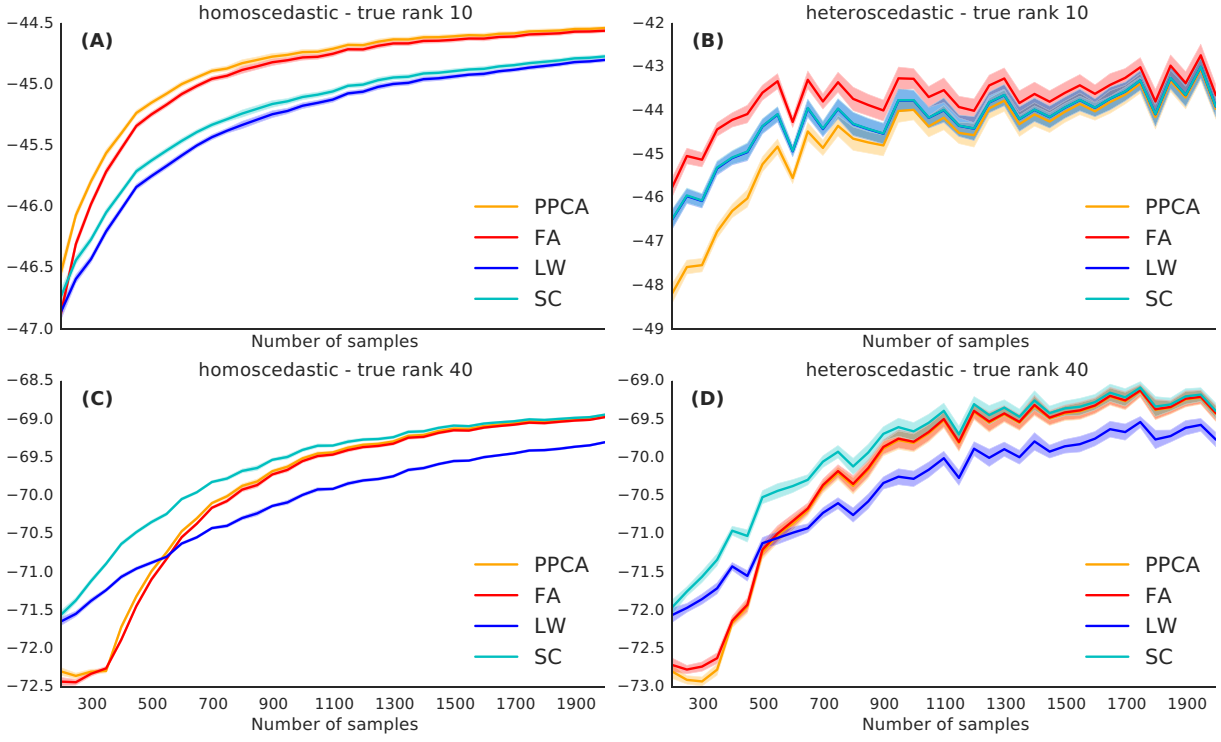

Fig. 3. Comparison between different covariance estimators on simulated data. Panel (A). Homoscedastic noise, ground truth rank of 10. Low rank models exhibit a higher model likelihood than shrunk covariance models. Areas represent the standard error of the mean. Panel (B). Same ground truth as in (A) but heteroscedastic noise. FA exhibits the highest likelihood across different numbers of samples. Panel (C). Homoscedastic noise, ground truth rank of 40. The SC covariance had the highest likelihood with few samples available. The likelihood for FA and PCA model improves with increasing numbers of samples and finally reached the level of SC. Panel (D). Same rank as in (C) with heteroscedastic noise. The same pattern emerged as in (C).

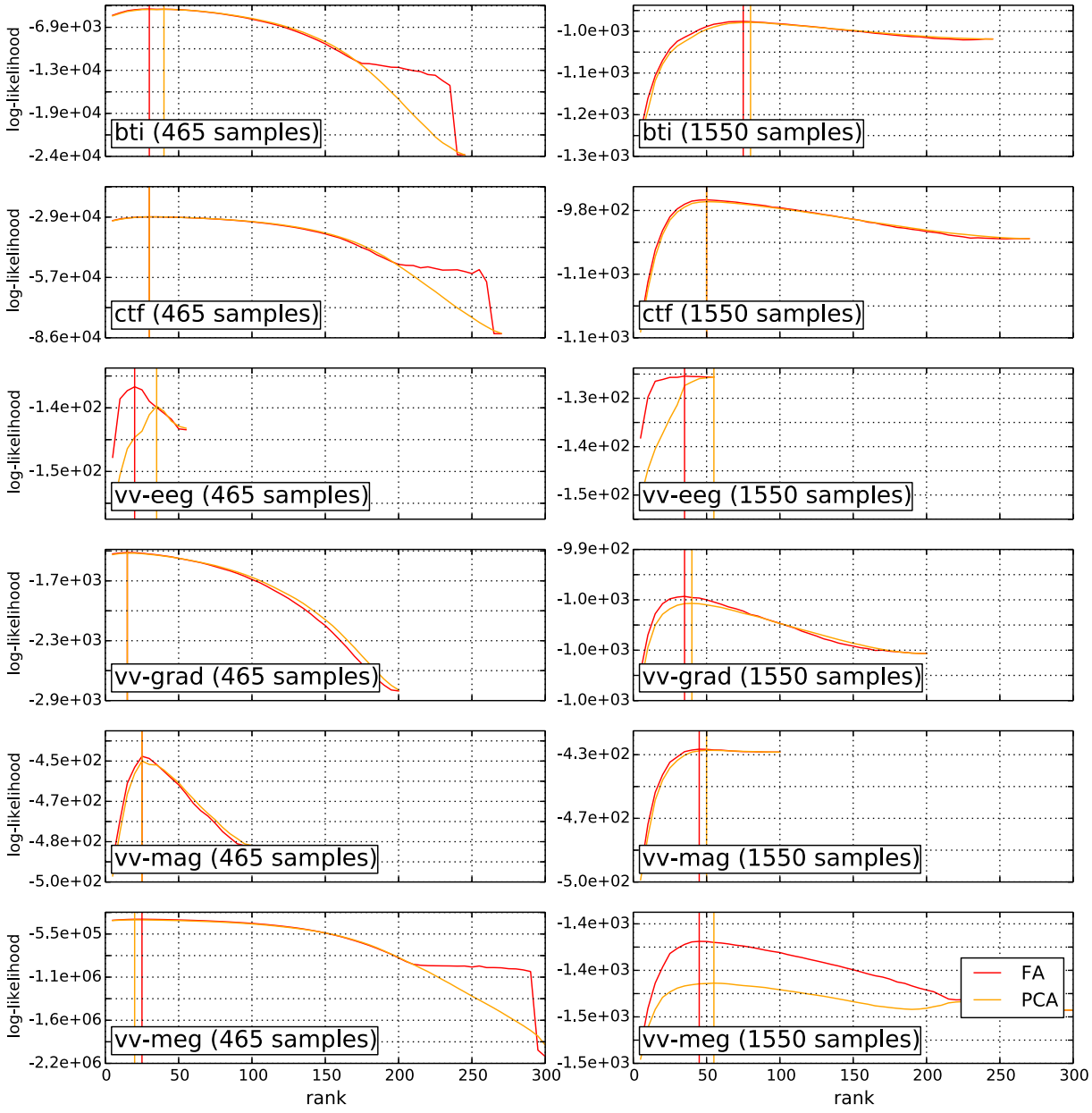

Fig. 4. Rank estimates for low and high numbers of baseline samples computed on the different datasets. The estimated rank was higher when increasing the number of samples. FA always outperformed PCA suggesting that M/EEG noise is heteroscedastic, not homoscedastic. One also observes that PPCA rank estimates are almost always higher than equivalents estimated with FA.

Whitened global field power

In Fig. 6 whitened global field power (GFP) plots are presented for each estimator. The GFP dynamics exemplify respective under- and overestimation tendencies. The black dotted horizontal line indicates the expected value for white Gaussian data. GFP values below and above this line correspond to overestimation and underestimation of the noise level, respectively. When the noise is underestimated, that is when normalized GFPs are below one, the procedure is said to underfit. In contrast, it is said to overfit if the noise is greater than one during the baseline periods (between −200 and 0 ms). The huge deflections in the post-baseline window represent time-locked brain responses.

In four out of seven datasets the empirical covariance produced clearly visible overfitting while the regularized covariance tended to underfit the noise. However, in almost any other case, differences seem hard to distinguish by mere visual inspection.

Whitened evoked response

Fig. 7 shows whitened evoked responses for each dataset where the whitener was computed from the best fitting covariance estimator determined by its log-likelihood on unseen data. Except one case, where hand-set regularization was most-appropriate, either SC or FA performed best.

Source estimates

The impact of covariance estimation was practically examined by computing signal contrasts that reflect cortical activity related to face perception over a range of different numbers of epochs for both the worst and the best estimators. Fig. 8 shows contrast-results for 20, 40, and 60 input epochs, respectively. The empirical covariance and SC were the worst and the best estimator across the entire range of epochs, respectively. When comparing signal dynamics, i.e., the spatial standard

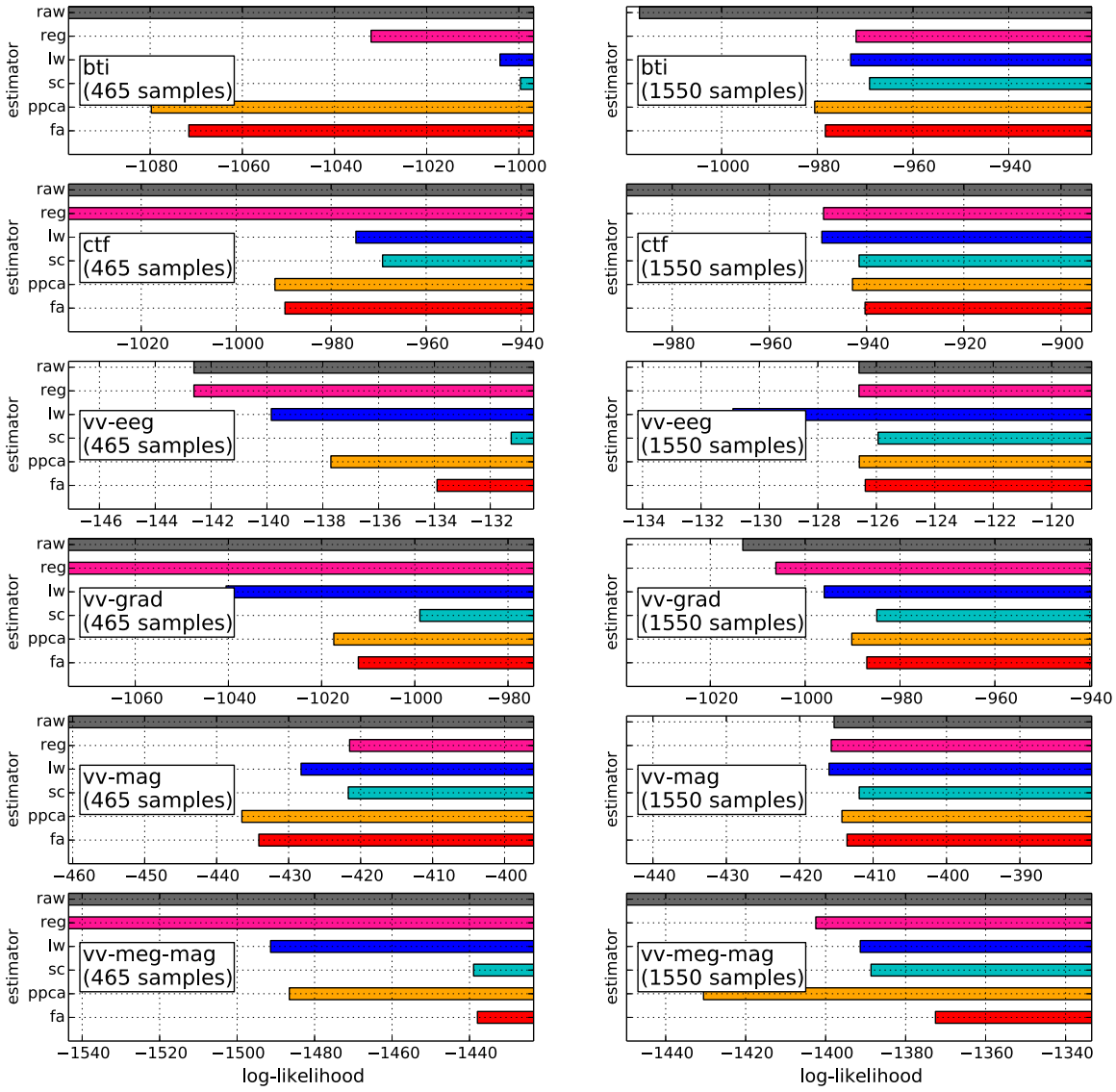

Fig. 5. Log-likelihoods of covariance models for low and high numbers of baseline samples obtained on the different datasets. Either cross-validated SC or FA turned out to the best model. FA was always more appropriate than PCA. Except for the multi-sensor dataset, FA outperformed SC only when the number of samples was high. In most cases, standard regularization either under- or overestimated the baseline noise.

deviation, across epochs, the best covariance estimator showed virtually no variability whereas the worst estimator led to different source amplitudes, depending on the number of epochs. This higher consistency for source amplitudes based on the best estimator was also reflected in more consistent spatial extents of the activity maps at a fixed threshold. With only 20 epochs of input data, the activation maps based on the best estimator suggest a pronounced ventral–temporal center of activity in the mid-fusiform gyrus, a brain region commonly associated with processing faces (Haxby et al., 2002; Yovel and Freiwald, 2013). In contrast, the worst estimator produced activation maps that emphasize nearly the entire ventral part of the temporal lobe. Importantly, for the best estimator, spatial maps look visually identical when varying numbers of epochs. This is consistent with the spatial standard deviation depicted in the second row of the figure. While the red area shrinks with increasing numbers of epochs, the blue area remains constant across epochs. Taken together, with the worst estimator, source amplitudes more strongly depended on the number of samples whereas differential dynamics in spatial standard deviation also indicated a more variable spatial extent of cortical activation. In general, differences between the best and the worst estimator decreased with increasing numbers of epochs.

Discussion

The present study addressed the problem of data-driven regularization of spatial covariance estimates computed on magnetoencephalography and electroencephalography M/EEG data. Such covariance estimates are a building block of most (M/EEG) data analysis pipelines. They are particularly useful for spatial whitening of data which is required by most distributed source localization methods. This problem was approached by employing model selection with cross-validation. In detail, the log-likelihood of the covariance was proposed as a metric to select the best model out of a set of alternative covariance estimators. In addition to empirical and regularized covariance estimates which reflect common standard choices, covariance models with shrinkage estimators and latent variable models were subjected to model selection. Data were validated by simulations, sensor space metrics and, practically, by source localization of an experimental contrast from a face processing (MEG) experiment.

Both, simulation and sensor space results unequivocally demonstrated that there was not one single model that fitted all scenarios and datasets. Different global parameters, such as sample size, the true

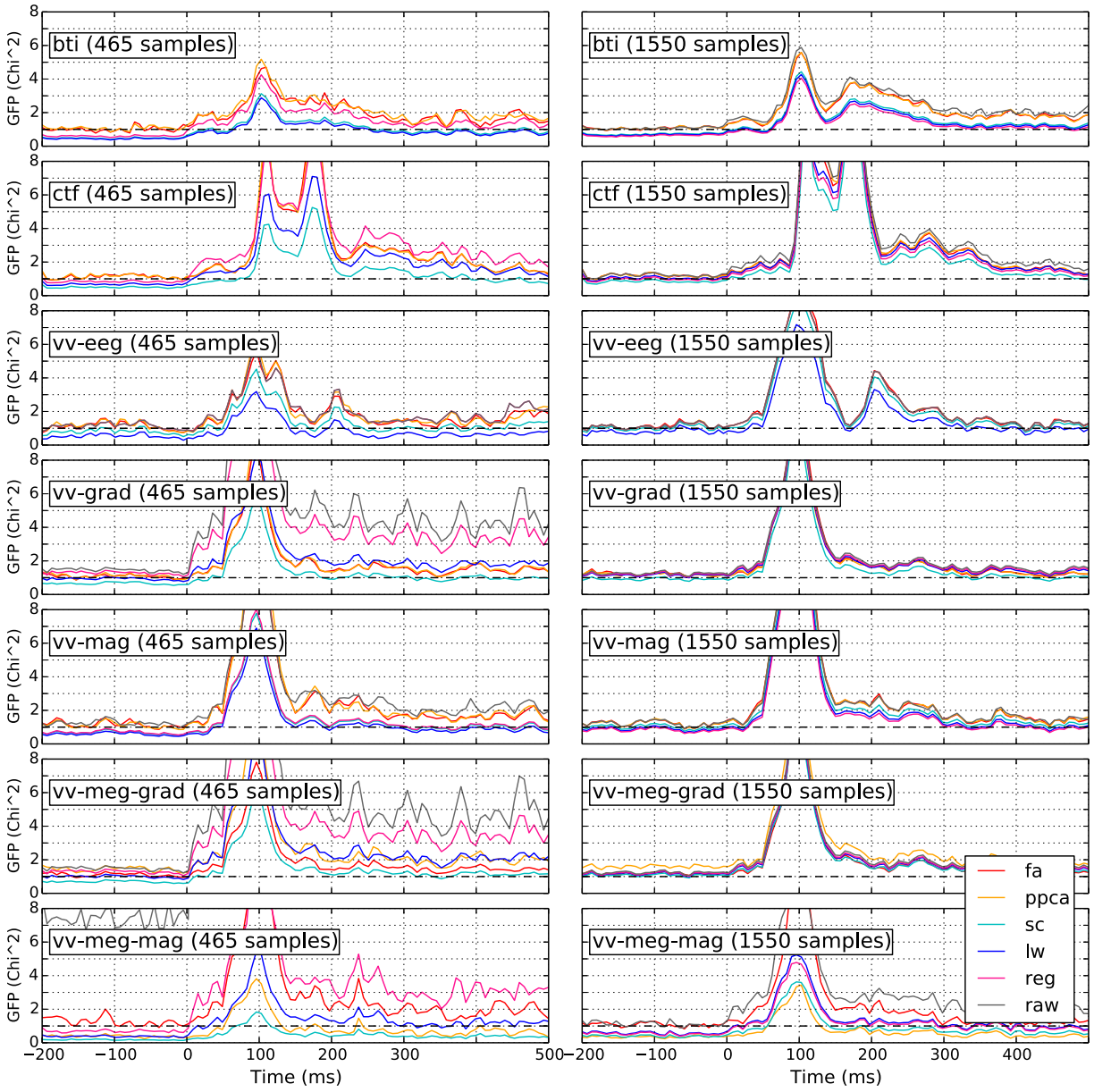

Fig. 6. Global field power (χ^2 statistic) of whitened evoked data for low and high numbers of baseline samples and different datasets. The dotted vertical black represents the expected baseline amplitude of one, given Gaussian baseline data.

number of dimensions, the noise structure and the sensor type shaped the log-likelihood of the covariance estimators. The simulation suggested that covariance models based on latent variable models delivered a more appropriate covariance estimate if the true rank of the data was low and the number of samples was sufficient. In detail, given a sufficient sample size, factor analysis (FA) was the best model when the structure of noise was heteroscedastic while probabilistic principal component analysis (PCA) performed best with homoscedastic noise. This stands in contrast to the M/EEG study were probabilistic principal component analysis (PPCA) never achieved the highest model probability on any of the M/EEG datasets. However, the sensor space validation suggests that on M/EEG data two solutions are likely to be selected, either favoring FA or shrunk covariance (SC) models. In this context it is helpful to recapitulate differences between FA and PCA. Due to its diagonal noise term, the former can cope with varying noise levels across channels. FA can therefore describe the data with fewer dimensions. This is because in such a model, variance that is not captured by the components is captured in the diagonal noise term. However, such complex models require more samples than simple models to be properly estimated. Simulation and sensor space findings are consistent with these characteristics suggesting that FA leads to a lower rank estimates than PPCA but was only preferred when the number of samples was sufficiently high. Second, in the multi-sensor dataset only FA produced appropriate noise estimates for combined sensor types. Also, in other cases where FA was selected the number of samples was higher, not lower. In contrast, if analysis was constrained to one sensor type, in most cases SC was selected irrespective of the sample size.

Taken together, these findings indicate that the model likelihood may depend on the system type and the recordings themselves in ways which are not sufficiently understood. In this sense, M/EEG data problems are subject to the “no free lunch theorem” (Wolpert, 1996) which characterizes problems that do not permit finding short cuts. In practice, these findings suggest evaluating at least FA, SC in addition to the default regularization and to then choose the best model using cross-validation with unseen data.

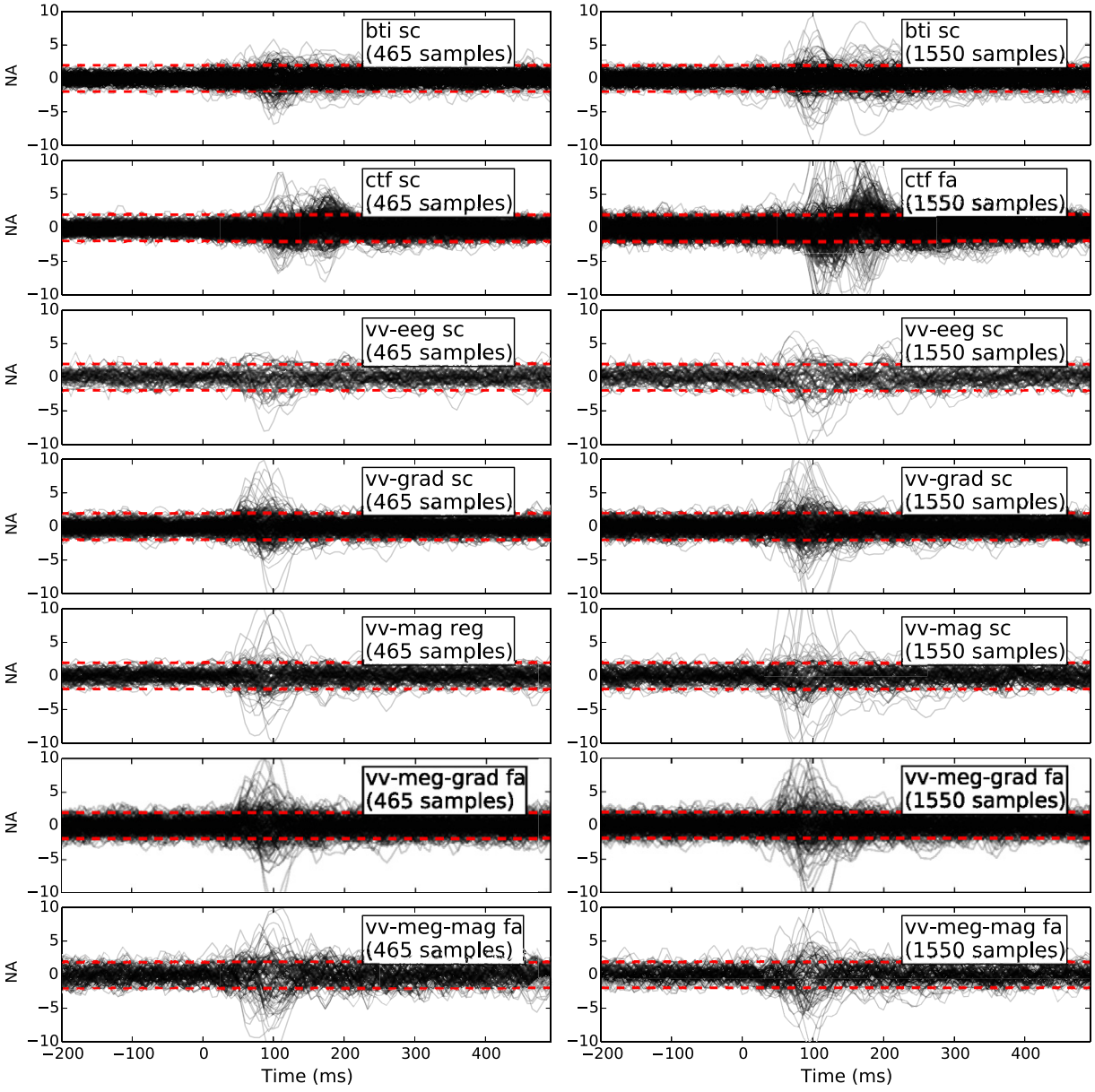

Fig. 7. Time-locked whitened evoked responses with the optimal covariance model for low and high numbers of baseline samples and different datasets. Selection based on results depicted in Fig. 5.

The simulation results on heteroscedastic noise are not surprising per se, given the model assumptions of FA. However, combining simulation and sensor space results reveals a basic M/EEG signal-characteristic. Sensor noise is heteroscedastic, and not homoscedastic. This is reflected by the fact that on MEG data PPCA models never prevailed, whereas such models were preferred under certain simulated conditions. M/EEG methods should, therefore, focus on models that take into account the variable sensor noise in M/EEG. This aspect seems only partially covered by the literature. Examples are given by research on independent component analysis (ICA) in neuroimaging (Beckmann and Smith, 2004; Dammers et al., 2008; Hyvärinen et al., 2004) and approaches which leverage FA for source imaging and artifact rejection (Nagarajan et al., 2007; Zumer et al., 2007; Zumer et al., 2008).

The sensor space findings demonstrate another important implication of this study. In the context of model selection, mere visualization is insufficient to assess the quality of the whitening step. Two visual inspection methods have been demonstrated, namely whitened global field power (GFP) plots and whitened evoked plots. Both provide with basic diagnostics for spatial whitening but only provide limited guidance on how to rank the different models. To go beyond graphical data exploration, the multivariate Gaussian log-likelihood score evaluated on left out data can be regarded as an unbiased quantified measure of estimator performance. Assuming that the whitened noise covariance is an identity matrix, this quantity measures how close the covariance of the whitened data approximates the identity matrix. Or, put differently, to which degree the data will be whitened by a whitening operator computed from the covariance estimate. In other words, it is a quantified measure of the success of the spatial whitening procedure.

Importantly, the log-likelihood procedure is closely linked to commonly neglected aspect of M/EEG data analysis. Since the log-likelihood was evaluated on data that were not used for parameter estimation it measures what is called in machine learning the “out of sample performance” (Breiman and Spector, 1992). This is relevant, since, in fact virtually all whitening operations on M/EEG data are applied to time intervals which were not used for the estimation of the noise model, i.e., post-baseline time-locked signals which reflect brain activity in

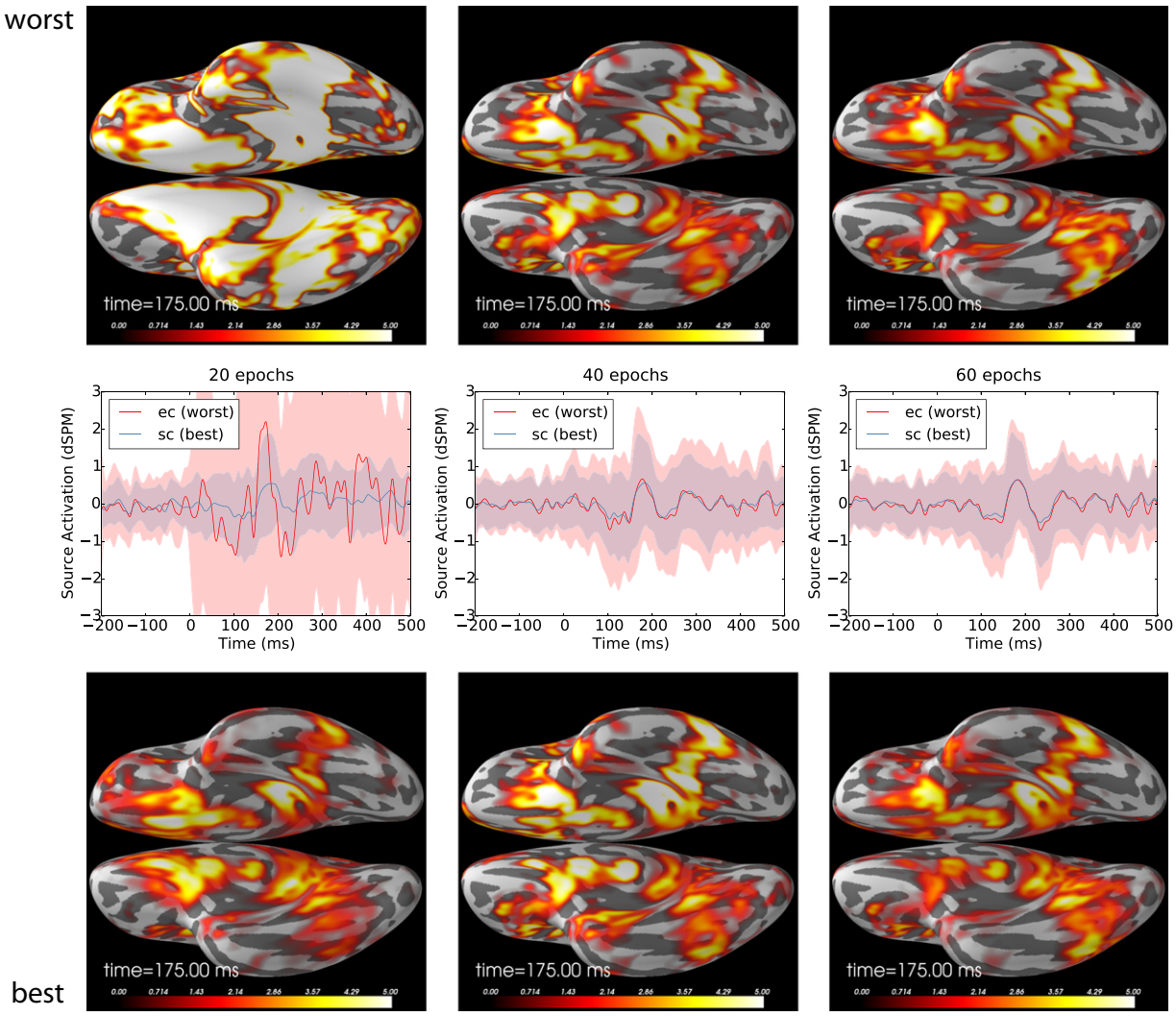

Fig. 8. Worst and best covariance estimators for faces > scrambled contrast. The top row represents dSPM maps around the maximum amplitude at 175 milliseconds for the worst covariance estimators. The mid-row represents average temporal dynamics for worst and best estimators superimposed. The lines refer to the average signal across vertices, the areas depict the standard deviation across vertices. The bottom-row represents dSPM maps around the maximum amplitude at 175 milliseconds for the best covariance estimators. The columns refer to results for 20, 40 and 60 epochs of input data. Statistical maps are thresholded at the 99th percentile of the maximum amplitude at 60 epochs. For comparability to other studies, results are shown on the FreeSurfer average brain's inflated surface in ventral view. The curvature of the cortical surface is indicated by light and dark gray colors for gyri and sulci, respectively.

addition to noise. Likewise, often the meg noise covariance is not even computed from the dataset analyzed but from so called empty room recordings, which correspond to measurements during which no subject is present. Such empty room recordings are commonly used when running time frequency analyses and are regarded as mandatory for resting state analyses (Gramfort et al., 2013b; Hämäläinen et al., 2010; Lin et al., 2004). As a consequence, goodness of fit measures computed on unseen data are generally desirable. Based on formula (3), the procedure that was developed and evaluated in this study can be easily generalized to any covariance estimate, even beyond the estimators that were investigated in the present context. Importantly, this new method can be used with any kind of inverse solution, either of minimum-norm estimates (MNE) type or beamforming type. In this context it is important to note that the impact of the method is expected to depend on the inverse solver and its exact use case. For non-normalized minimum-norm estimates, i.e. plain vanilla MNE, the covariance matrix only matters for large regularization parameters, i.e. low SNRs. For perfect (noise-free) data, the covariance matrix would not be required at all. For noise-normalized estimates (dSPM and sLORETA), covariance matrices are more essential. Beamformers are a different case, as their estimation typically relies on two covariance estimates, one that aims to describe the spatial structure of noise, and one that is concerned with the spatial structure of the data. An accurate estimate of data covariance matrix is particularly crucial in the context of beamformers (Hauk and Stenroos, 2014; Woolrich et al., 2011). Our findings are therefore expected to be even more relevant for methods based on beamformers.

The practical impact assessment on a publicly available dataset demonstrated at least two critical implications. Across varying amounts of trials, the best estimator led to more stable source estimates. Around 20 epochs this effect was dramatic. The worst estimator led to massive increase of cortical source amplitudes. The best estimator still estimated the same contrast-amplitudes with only 20 trials of exposure compared to three times as many trials. These results demonstrate that small datasets, where a covariance estimate from baseline segments is preferred and which only consist of few epochs (see, for example Lu et al., 2014), are compatible with source analysis, should the noise covariance be carefully estimated. More importantly, this procedure can be expected to reduce overall variability in source estimates across subjects. The source localization results suggest an asymptotic trend towards convergence between the worst and the best estimator. Consistent with the simulation findings, with increasing numbers of epochs their differences became increasingly smaller. But, practically, stability differences are still visible when comparing results between 40 epochs and 60 epochs exposure, which is a more common scenario than the

previously mentioned analysis of sparse events. This implies that the worst covariance estimator which, in this case, turned out to be the default empirical sample covariance, will lead to increased variance as a function of different numbers of epochs. This case is practically relevant if one assumes that, for a group of subjects different numbers of trials will be selected, based on behavioral and artifact-related exclusion criteria. It is then easily conceivable that different epoch counts can lead to a ramping-up of variance which will be prevented by a robust estimator that exerts a stabilizing impact on the amplitudes of source estimates.

To the best of our knowledge, this is the first time that an automated procedure based on cross-validation on unseen data has been employed for model selection in spatial whitening of M/EEG data and has been validated on source localization results which are relevant to the broader cognitive neuroscience community. To avoid an unbalanced view, related studies need to be mentioned though. One such approach has been recently proposed by Woolrich et al. (2011) who employed Bayesian PCA (Bishop, 1999) to estimate noise and data covariance matrices in the context of beamforming. Bayesian PCA is an alternative approach to infer the number of latent components in the PCA model. It solves this problem using a Bayesian inference approach, what is achieved here with PPCA and cross-validation (to avoid overfitting). Practically, as with the PPCA and FA estimators, the amount of regularization and the number of components are learned from the data. Also, the Bayesian PCA model is not specific to one M/EEG inverse problem. It can hence be plugged into any M/EEG imaging technique that is formulated as a constrained linear model. However, the method presented in the current study goes beyond the Bayesian PCA (Woolrich et al., 2011), as it quantifies the benefit of each modeling assumption and can select the best estimator over a richer class of models. This is an important consideration since latent variable models can be outperformed by shrinkage when few samples are available for estimation. Second, Bayesian PCA is a PCA model and hence assumes a homoscedastic noise which has been shown to be a suboptimal assumption for M/EEG data. More advanced models have been proposed in which the baseline noise covariance is estimated jointly with the post baseline data covariance (Nagarajan et al., 2007; Zumer et al., 2007, 2008). However, these approaches are particularly tailored for beamformer methods and not for MNE-type inverse solvers which do not rely a post-baseline data covariance. Interestingly, spatiotemporal estimators of M/EEG covariance that consider both the spatial correlations between sensors and the temporal dependencies between time samples have been previously proposed (Bijma et al., 2005). This approach is promising, nevertheless it is based on pure maximum likelihood, hence, does not implement any type of regularization to reduce estimation variance. This warrants future investigations which combine the automated regularization we propose here and such spatiotemporal covariance models.

It is important to note that the proposed approach is subject to certain numerical constraints. The computation of the low rank estimators can result in numerical errors if the data is rank-deficient. As a consequence, at the current stage of development it is recommended to compute the PPCA and FA models before applying processing steps such as signal space separation (SSS), signal space projection (SSP) or ICA (Hyvärinen et al., 2004; Taulu et al., 2005; Uusitalo and Ilmoniemi, 1997) and then apply dimensionality reducing operators to both the data and the covariance estimators. Alternatively, when computing the noise covariance after application of dimensionality reducing operators, the PPCA and FA models should be used with care. Second, outlier samples may strongly distort model selection in certain estimators, especially FA. It is therefore recommended to remove heavily corrupted time segments before estimation.

To conclude, this study has developed an automated procedure to tune covariance estimates computed from M/EEG data. This method establishes a quality-preserving function, since it will lead to estimates that will not fall behind the default empirical covariance. Indeed the result of the automatic whitening performance was in almost all cases more accurate than whitening based on hand-set regularization. But it was always preferable to the empirical covariance. However, for the unlikely case that all other options fail, the empirical covariance would be selected as fall-back option. Automatic whitening, hence, constitutes a solution to the regularization problem and helps avoiding ad-hoc parameterization and other heuristics that are difficult to generalize across the variety of M/EEG data analysis pipelines. The impact demonstrated on face-related signal contrasts suggest that this study contributes one small but important element in a set of measures which help promote laboratory- and data-independent analysis pipelines which are so urgently needed to improve reproducibility of M/EEG research (Gramfort et al., 2013b, 2014; Gross et al., 2013).

Acknowledgment

We would like to thank the MNE-Python and scikit-learn developers for their peer review of our code. We would also like to thank John Mosher for bringing this problem of spatial whitening to our attention. We thank Danilo Bzdok, Michael Eickenberg and Teon Brooks for their feedback on this manuscript. We thank the two anonymous reviewers for their helpful suggestions.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G., 2014. Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. 8. http://dx.doi.org/10.3389/fninf.2014.00014.

Barber, D., 2012. Bayesian Reasoning and Machine Learning. Cambridge University Press.

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. Med. Imaging, IEEE Trans. 23, 137–152.

Bijma, F., De Munck, J.C., Heethaar, R.M., 2005. The spatiotemporal meg covariance matrix modeled as a sum of kronecker products. NeuroImage 27, 402–415.

Bishop, C.M., 1999. Bayesian PCA. Advances in Neural Information Processing Systemspp. 382–388.

Breiman, L., Spector, P., 1992. Submodel Selection and Evaluation in Regression. The x-Random Case. International statistical review/revue internationale de Statistique. pp. 291–319.

Breuer, L., Axer, M., Dammers, J., 2013. A new constrained ICA approach for optimal signal decomposition in polarized light imaging. J. Neurosci. Methods 220, 30–38.

Chen, Y., Wiesel, A., Eldar, Y.C., Hero, A.O., 2010. Shrinkage algorithms for MMSE covariance estimation. Signal Process. IEEE Trans. 58, 5016–5029.

Dale, A., Liu, A., Fischl, B., Buckner, R., 2000. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron 26, 55–67.

Dammers, J., Schiek, M., Boers, F., Silex, C., Zvyagintsev, M., Pietrzyk, U., Mathiak, K., 2008. Integration of amplitude and phase statistics for complete artifact removal in independent components of neuromagnetic recordings. Biomed. Eng. IEEE Trans. 55, 2353–2362.

Gramfort, A., Kowalski, M., Hämäläinen, M., 2012. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. Phys. Med. Biol. 57, 1937–1961.

Gramfort, A., Strohmeier, D., Haueisen, J., Hämäläinen, M., Kowalski, M., 2013a. Time-frequency mixed-norm estimates: sparse M/EEG imaging with non-stationary source activations. Neuroimage 70, 410–422.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hämäläinen, M., 2013 ab. MEG and EEG data analysis with MNE-Python. Front. Neurosci. 7. http://dx.doi.org/10.3389/fnins.2013.00267.

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M., 2014. MNE software for processing MEG and EEG data. Neuroimage 86, 446–460. http://dx.doi.org/10.1016/j. neuroimage .2013.10.027.

Gross, J., Kujala, J., Hämäläinen, M., Timmermann, L., 2001. Dynamic imaging of coherent sources: studying neural interactions in the human brain. Proc. Natl. Acad. Sci. 98, 694–699.

Gross, J., Baillet, S., Barnes, G., Henson, R., Hillebrand, A., Jensen, O., Jerbi, K., Litvak, V., Maess, B., Oostenveld, R., Parkkonen, L., Taylor, J., van Wassenhove, V., Wibral, M., Schoffelen, J., 2013. Good practice for conducting and reporting MEG research. NeuroImage 65, 349–363.

Halko, N., Martinsson, P.G., Tropp, J.A., 2011. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. 53, 217–288.

Hämäläinen, M., Hari, R., Ilmoniemi, R., Knuutila, J., Lounasmaa, O., 1993. Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev. Mod. Phys. 65, 413–497.

Hämäläinen, M.S., Lin, F.H., Mosher, J.C., 2010. Anatomically and Functionally Constrained Minimum-Norm Estimates. MEG: An Introduction to Methods: An Introduction to Methods 186.

Hauk, O., Stenroos, M., 2014. A framework for the design of flexible cross-talk functions for spatial filtering of EEG/MEG data: deflect. Hum. Brain Mapp. 35, 1642–1653.

Haxby, J., A., H.E., Gobbini, M.I., 2002. Human neural systems for face recognition and social communication. Biol. Psychiatry 51.

Henson, R., Rugg, M., 2003. Neural response suppression, haemodynamic repetition effects, and behavioural priming. Neuropsychologia 41, 263–270.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.
Hyvärinen, A., Karhunen, J., Oja, E., 2004. Independent Component Analysis. 46. John Wiley & Sons.
Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. 88, 365–411. http://dx.doi.org/10.1016/S0047-259X(03)00096-4.
Lin, F.H., Witzel, T., Hämäläinen, M.S., Dale, A.M., Belliveau, J.W., Stufflebeam, S.M., 2004. Spectral spatiotemporal imaging of cortical oscillations and interactions in the human brain. Neuroimage 23, 582–595.
Lin, F.H., Witzel, T., Ahlfors, S.P., Stufflebeam, S.M., Belliveau, J.W., Hamalainen, M.S., 2006. Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. Neuroimage 31, 160–171.
Litvak, V., Mattout, J., Kiebel, S., Phillips, C., Henson, R., Kilner, J., Barnes, G., Oostenveld, R., Daunizeau, J., Flandin, G., Penny, W., Friston, K., 2011. EEG and MEG data analysis in SPM8. Comput. Intell. Neurosci. 2011. http://dx.doi.org/10.1155/2011/852961.
Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain–computer interfaces. J. Neural Eng. 4, 1–13.
Lu, Q., Jiang, H., Bi, K., Liu, C., Yao, Z., 2014. Discriminative analysis with a limited number of MEG trials in depression. J. Affect. Disord. 167, 207–214. http://dx.doi.org/10.1016/j.jad.2014.06.007.
Martinsson, P.G., Rokhlin, V., Tygert, M., 2011. A randomized algorithm for the decomposition of matrices. Appl. Comput. Harmon. Anal. 30, 47–68. http://dx.doi.org/10.1016/j.acha.2010.02.003.
Minka, T.P., 2000. Automatic Choice of Dimensionality for PCA. NIPSpp. 598–604.
Mosher, J., Leahy, R., 1998. Recursive MUSIC: a framework for EEG and MEG source localization. Biomed. Eng. IEEE Trans. 45 (11).
Nagarajan, S.S., Attias, H.T., Hild, K.E., Sekihara, K., 2007. A probabilistic algorithm for robust interference suppression in bioelectromagnetic sensor data. Stat. Med. 26, 3886–3910.
Pascual-Marqui, R., 2002. Standardized low resolution brain electromagnetic tomography (sLORETA): technical details. Methods Find. Exp. Clin. Pharmacol. 24, 5–12.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Ramoser, H., Müller-Gerking, J., Pfurtscheller, G., 1998. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans. Rehab. Eng 8, 441–446.
Taulu, S., Simola, J., Kajola, M., 2005. Applications of the signal space separation method. IEEE Trans. Signal Proc. 53, 3359–3372.
Tikhonov, A., Arsenin, V.Y., 1977. Solutions of Ill-posed Problems. WH Winston, Washington, DC 330.
Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. J. R. Stat. Soc. Ser. B (Stat Methodol.) 61, 611–622.
Uusitalo, M., Ilmoniemi, R., 1997. Signal-space projection method for separating MEG or EEG into components. Med. Biol. Eng. Comput. 35, 135–140.
Uutela, K., Hämäläinen, M., Somersalo, E., 1999. Visualization of magnetoencephalographic data using minimum current estimates. Neuroimage 10, 173–180.
Veen, B.V., Drongelen, W.V., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. Biomed. Eng. IEEE Trans. 44, 867–880.
Wolpert, D.H., 1996. The lack of a priori distinctions between learning algorithms. Neural Comput. 8, 1341–1390.
Woolrich, M., Hunt, L., Groves, A., Barnes, G., 2011. MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization. Neuroimage 57, 1466–1479.
Yovel, G., Freiwald, W.A., 2013. Face Recognition Systems in Monkey and Human: Are They the Same Thing? F1000prime reports 5, 10
Zumer, J.M., Attias, H.T., Sekihara, K., Nagarajan, S.S., 2007. A probabilistic algorithm integrating source localization and noise suppression for MEG and EEG data. Neuroimage 37, 102–115.
Zumer, J.M., Attias, H.T., Sekihara, K., Nagarajan, S.S., 2008. Probabilistic algorithms for MEG/EEG source reconstruction using temporal basis functions learned from data. Neuroimage 41, 924–940.