NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 38 (2007) 501 – 510

Technical Note

Mixed-effect statistics for group analysis in fMRI: A nonparametric maximum likelihood approach

Alexis Roche,[a,c,*] Sébastien Mériaux,[a,c] Merlin Keller,[b,c] and Bertrand Thirion[b,c]

[a] CEA, Neurospin, Gif-sur-Yvette, France
[b] INRIA, Futurs, Orsay, France
[c] Institut d'Imagerie Neurofonctionnelle (IFR 49), Paris, France

Received 4 December 2006; revised 25 June 2007; accepted 29 June 2007
Available online 17 August 2007

This technical note describes a collection of test statistics accounting for estimation uncertainties at the within-subject level, that can be used as alternatives to the standard *t* statistic in one-sample random-effect analyses, *i.e.* when testing the mean effect of a population. We build such test statistics by estimating the across-subject distribution of the effects using maximum likelihood under a nonparametric mixed-effect model. For inference purposes, the statistics are calibrated using permutation tests to achieve exact false positive control under a symmetry assumption regarding the across-subject distribution. The new tests are implemented in a freely available toolbox for SPM called Distance.

Introduction

Conventional random-effect analysis in fMRI takes as input a set of BOLD contrast images, or effects estimated from a first-level, within-subject analysis, to produce a group statistical map which is further thresholded according to a desired significance level. Both parametric and nonparametric versions of that approach have been implemented in several software packages, generally in a massively univariate fashion. Such techniques have in common the fact that they reduce the experimental information at hand to a series of "first-order" summary statistics.

Over the past few years, mixed-effect (MFX) models have been proposed in order to account for the within-subject uncertainties, as represented in particular by the estimated variances of the effect estimates, therefore correcting the group statistical map for higher-order information (Holmes and Friston, 1998; Worsley et al., 2002; Beckmann et al., 2003; Neumann and Lohmann, 2003; Woolrich et al., 2004; Friston et al., 2005; Mériaux et al., 2006b). Roughly speaking, this correction acts as a reweighting of the subjects according to the reliability of their respective effect estimates. Statistics undergoing MFX correction may have more detection power than their uncorrected analogues, provided the within-subject uncertainties are not positively correlated with the effects (Mériaux et al., 2006a); in other words, first-level estimation performances should not be biased towards small effects.

To clarify the context, assume n subjects are selected randomly in a population of interest and submitted to the same fMRI experiment. The within-subject analysis of the scans then produces, in one particular voxel of the standardized space (usually, the MNI/Talairach space) and for each subject i, an estimate $\hat{\beta}_i$ of the BOLD effect relative to a given contrast of experimental conditions. For mathematical clarity, we will restrict ourselves to scalar (one-dimensional) effects. While the estimated effect $\hat{\beta}_i$ generally differs from the true but unobserved effect β_i, assume for now perfect within-subject estimation so that $\hat{\beta}_i = \beta_i$ for all i. We thus are given a sample $(\beta_1, \beta_2, \dots, \beta_n)$ drawn from an unknown probability density function (PDF) $f(\beta)$ that describes the distribution of the effects in the population.

This paper specifically addresses inferences about a location parameter (mean, median, mode, etc.). Assume for instance we wish to test the null hypothesis that the population mean is negative:

$$H_0: \quad \mu_f = \int \beta f(\beta) d\beta \leq 0$$

To that end, we may use the classical one-sample t test. We start with computing the t statistic,

$$t = \frac{\hat{\mu}_f}{\hat{\sigma}_f / \sqrt{n}}, \quad \text{with } \hat{\mu}_f = \frac{1}{n} \sum_i \beta_i,$$
$$\hat{\sigma}_f^2 = \frac{1}{n-1} \sum_i (\beta_i - \hat{\mu}_f)^2 \qquad (1)$$

Next, we reject H_0, hence accept the alternative H_1: $\mu_f > 0$, if the probability under H_0 of attaining the observed t value is lower than a given false positive rate. Under the assumption that $f(\beta)$ is normal, this probability is well known to be obtained from the

[*] Corresponding author. CEA, Neurospin, Gif-sur-Yvette, France.
E-mail address: alexis.roche@cea.fr (A. Roche).

Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.06.043

Student distribution with $n-1$ degrees of freedom. In this parametric context, the t statistic can be proved to be optimally sensitive (technically, in the sense of the uniformly most powerful unbiased test, see Good, 2005).

Non-Gaussian populations

If normality is not tenable, however, the Student distribution is valid only in the limit of large samples, and may thus lead to inexact control over the false positive rate in small samples. This problem can be worked around using non-parametric calibration schemes such as sign permutations (Holmes et al., 1996; Good, 2005), which allow exact inferences under a milder assumption of symmetry regarding $f(\beta)$. Although we strongly recommend permutation tests, they only provide an alternative strategy of thresholding a given statistic and, as such, address a *specificity* issue.

The fact that the sampling PDF $f(\beta)$ may not be normal also raises a *sensitivity* issue as the t statistic may no longer yield optimal power when normality does not hold. Without prior knowledge of the shape of $f(\beta)$, a reasonable default choice for the test statistic is one that maintains good detection performance over a wide range of PDFs. Such a statistic is robust, not quite in the classical sense of being resistant to outliers, but in the looser sense of being resistant to distributions that tend to produce outliers, such as heavy-tailed, skewed, or multimodal distributions. Standard examples of robust test statistics include Fisher's sign statistic, Wilcoxon's signed rank, the empirical likelihood ratio (Owen, 2001), or *M*-estimators (Rousseeuw and Leroy, 1987), which are fairly more sensitive than the t statistic in some non-Gaussian distributions. As a matter of fact, such statistics have been used previously in fMRI group analyses (Brammer et al., 1997; Wager et al., 2005; Mériaux et al., 2006c; Dehaene-Lambertz et al., 2006b; Rorden et al., 2007), most often combined with permutation tests.

Mixed effects

The problem of choosing an appropriate test statistic becomes more complex when mixed effects are taken into account. Up to now, we have assumed that the observations are exact in the sense that they are drawn from the very PDF $f(\beta)$ for which the inference is to be made. Assume more generally that, instead of β_i, we observe $\hat{\beta}_i=\beta_i+\varepsilon_i$, where ε_i is typically an additive Gaussian noise with known standard deviation σ_i. Intuitively, observations with high uncertainty are likely to be outliers, which reinforces the need for a robust test statistic. Hence, in a first approach, we could pick one of the above mentioned statistics and compute it from the (noisy) observations. While it is possible to set up reasonable assumptions under which this leads to a valid test, we suspect that further detection power can be gained by correcting the statistic for the uncertainty levels σ_i.

This paper proposes a general method of correcting test statistics for mixed effects. The key to our approach is to interpret a variety of usual, "first-order" test statistics as maximum likelihood estimators of natural location parameters, or likelihood ratios. MFX correction then follows from developing maximum likelihood estimation under a nonparametric MFX model. Nonparametric, here, refers to the modeling assumptions regarding the random effect's PDF $f(\beta)$. To our best knowledge, the determination of tests for a location parameter under MFX models, has been restricted to date to the case where $f(\beta)$ is normal.

Throughout this paper, we will focus on inferences about a location parameter. Other group analysis problems, which are not covered here, include group comparison (two-sample tests) and more general tests of dependence. Most of the concepts presented below can be extended to those situations, albeit not necessarily in a straightforward way.

Method

In the following, we consider a particular voxel in the standardized space, therefore omitting voxel indexes in the notations. This is to say that we are adopting a massively univariate approach in which a test statistic is computed in each voxel independently from the others. This approach is chosen mainly for its simplicity and computational efficiency. The concept of voxel is not essential here as the same univariate approach can be used, *e.g.* within the parcel-based framework of Thirion et al. (2006) to handle localization uncertainties in the standardized space. Alternatively, using a multivariate extension of our formulation (which is beyond the scope of this paper), tests could be performed on spatial neighborhoods in a way similar to Friman et al. (2003), thus pooling the signal across voxels in an attempt to further improve detection power.

Mixed-effect model

Our goal is to make an inference about a location parameter such as the population mean or median, using the sample of estimated effects $(\hat{\beta}_1, \ldots, \hat{\beta}_n)$ as input data. In order to relate the unknown parameter with the observations, we adopt the following hierarchical sampling model:

First level (within-subject)

For each subject, the estimated effect relates to the true effect through a known conditional PDF, which is assumed Gaussian:

$$\forall i,\ \hat{\beta}_i|\beta_i \sim g_i(\hat{\beta}_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{\left(\hat{\beta}_i-\beta_i\right)^2}{2\sigma_i^2}} \tag{2}$$

Equivalently, $\hat{\beta}_i=\beta_i+\varepsilon_i$, where ε_i is an independent zero-mean Gaussian noise with standard deviation σ_i. We say that the noise is *homoscedastic* when σ_i is constant across subjects, and *heteroscedastic* in the converse situation. In practice, σ_i is usually estimated from the residuals of a multiple regression, suggesting that the Gaussian model neglects the precision on σ_i (equivalently, it assigns infinite degrees of freedom to each subject). Student distributions would be a more suitable choice than Gaussians, summarizing each subject by sufficient statistics, equivalently to the "full MFX" strategies advocated in Woolrich et al. (2004) and Friston et al. (2005). While we avoid that approach for computational reasons, we stress that assuming Gaussian noise does not hamper the test's specificity when calibrated using permutations (see section on Statistical calibration); the only potential impact is on the sensitivity.

Second level (between-subject)

The distribution of the true effects in the population is modeled as a PDF $f\in\mathcal{F}$, where $\mathcal{F}$ is some family:

$$\forall i,\quad \beta_i \sim f(\beta_i) \tag{3}$$

In the following, we mainly deal with the nonparametric case where $\mathcal{F}$ is the infinite-dimensional space of all PDFs on $\mathbb{R}$. If $\mathcal{F}$ is restricted to the Gaussian family, we fall back to the model considered *e.g.* in Worsley et al. (2002), Beckmann et al. (2003), and Mériaux et al. (2006b). Also notice, the linear model underlying the standard one-sample *t* statistic may be seen as a further restriction whereby the first-level noise is assumed homoscedastic.

Marginalizing out the true effects, one sees that each observation $\hat{\beta}_i$ has PDF $g_i \otimes f$ resulting from the convolution product of within-subject and between-subject distributions. This shows that the observations are non-identically distributed unless the noise is homoscedastic. The hierarchical model is summarized by its likelihood function which, under the usual independence sampling assumptions, reads:

$$L(f) = p\left(\hat{\beta}_1,\ldots,\hat{\beta}_n | f\right) = \prod_{i=1}^{n} g_i \otimes f\left(\hat{\beta}_i\right) \tag{4}$$

Test statistics

The general problem we tackle is to test a null hypothesis H_0: $\theta \in \Theta_0$ about some parameter of interest θ of the population, where $\Theta_0 \subset \mathbb{R}$ is some set of putative values for θ. Mathematically, $\theta = \phi(f)$ is defined as a real-valued function ϕ: $\mathcal{F} \to \mathbb{R}$ of the random effect's PDF *f*. A very common example of parameter is the population mean, which corresponds to the choice $\phi(f) = \int \beta f(\beta) d\beta$, the function known as the expectation, or mean.

Our approach is based on building any test statistic from a maximum likelihood estimate $\hat{f}$ of the random effect's PDF. While such an estimate is bound to be very noisy given the small number of subjects, some many-to-one functions of $\hat{f}$ may have sufficiently reduced statistical variability to enable powerful inferences. After developing Maximum likelihood PDF estimation, we will consider two families of test statistics: Parameter estimators and Likelihood ratios.

Maximum likelihood PDF

A maximum likelihood estimate $\hat{f}$ is one that maximizes $L(f)$ in Eq. (4) over the space $\mathcal{F}$ of admissible PDFs. In presence of nonzero first-level variances, such a maximization problem has typically no closed-form solution, which makes it necessary in practice to resort to numerical solvers such as expectation–maximization (EM) algorithms (Dempster et al., 1977). When the search space is restricted to the Gaussian family, one may use the EM algorithm proposed in Mériaux et al. (2006b) and reproduced in Appendix A.1.

We focus here on the case where $\mathcal{F}$ is the space of all PDFs on $\mathbb{R}$, so that we are searching for the nonparametric maximum likelihood estimate (NPMLE). From the general results given in Lindsay (1983), the NPMLE is necessarily a mixture of at most *n* Dirac masses, and may henceforth be tracked in a finite-dimensional space:

$$f(\beta) = \sum_{i=1}^{n} w_i \delta(\beta - z_i), \tag{5}$$

where w_i are unknown mixing proportions (nonnegative values that sum up to 1) and z_i are unknown support points. As seen in Fig. 1, the NPMLE may have fewer mixture components than *n*, if some support points coincide or if some mixing proportions vanish. This representation theorem generalizes the well-known property, central to nonparametric likelihood methods, that the NPMLE under exact observations is the so-called empirical distribution (Owen, 2001):

$$\hat{f}_e(\beta) = \frac{1}{n} \sum_{i=1}^{n} \delta\left(\beta - \hat{\beta}_i\right), \tag{6}$$

which corresponds to uniform mixing proportions, *i.e.* $\forall i$, $w_i = \frac{1}{n}$, and supports points coinciding with the observations.

Since no explicit expression is available in the general case where some observations are inexact, we propose in Appendix A.2 an EM algorithm to solve for the NPMLE iteratively. Like any EM, it is guaranteed to return an estimate with higher likelihood than its starting point, which we may choose as the empirical distribution $\hat{f}_e$. Note, however, that the EM algorithm may get trapped in a local maximum, a situation that seems to occur preferentially when the between-subject variability is swamped by the within-subject variability.

Alternative algorithms are given in Lindsay (1983) and Magder and Zeger (1996) but, being iterative and deterministic, they are also prone to local convergence. Robustness can be gained using stochastic EM variants, such as the SAEM algorithm (Delyon et al., 1999), at the expense of heavier computation time. In close spirit, the estimation of *f* may also be re-formulated in the nonparametric Bayesian framework described in Escobar and West

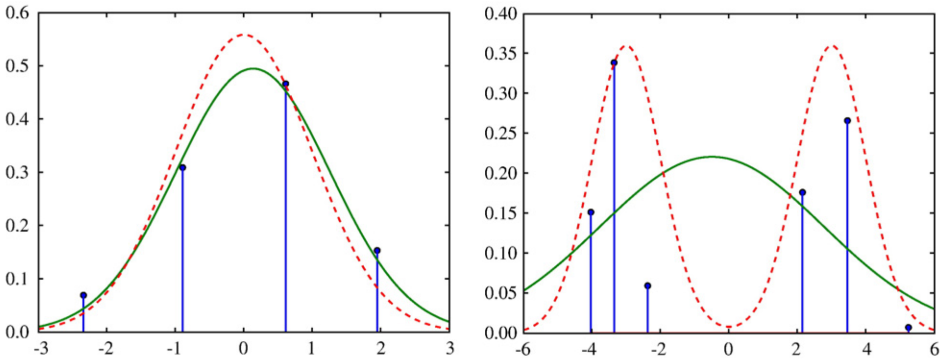

Fig. 1. Examples of parametric and nonparametric maximum likelihood PDF estimates in the MFX model, when the true PDF is respectively normal (left) and bimodal (right). In both cases, a sample size of 20 was drawn and contaminated with first-level Gaussian noise with heteroscedastic variances drawn in a Gamma distribution $\Gamma(3, \frac{1}{6})$. Each of the two panels shows the true PDF (dashed line), the parametric Gaussian fit (solid line), and the nonparametric fit (stems).

(1998), again leading to stochastic sampling. We are currently exploring, but it is important to realize that they are far more computationally intensive than the EM algorithm considered here, itself rather demanding (see Discussion).

Parameter estimators

An intuitive test statistic to consider for H_0: $\theta \in \Theta_0$ is a maximum likelihood estimator $\hat{\theta}$ of θ. A key remark is that, if $\hat{f}$ is a NPMLE of the PDF f, then $\hat{\theta}=\phi(\hat{f})$ is a NPMLE of $\theta=\phi(f)$ in the sense of the profile likelihood:

$$\tilde{L}(\theta) = \max_{f \in \mathcal{F}_\theta} L(f), \quad \text{with} \qquad \mathcal{F}_\theta = \{f \in \mathcal{F}, \varphi(f) = \theta\} \tag{7}$$

From Eq. (5), we know that the NPMLE $\hat{f}$ is adjusted via mixing proportions ($\hat{w}_1$, $\hat{w}_2$, …, $\hat{w}_n$) and support points ($\hat{z}_1$, $\hat{z}_2$, …, $\hat{z}_n$), and simplifies to the empirical distribution $\hat{f}_\mathrm{e}$ given by Eq. (6) in the special case of exact observations. The fact that $\phi(\hat{f})$ is optimal in the sense of profile likelihood suggests that it is asymptotically a more efficient estimator than $\phi(\hat{f}_\mathrm{e})$, thus presumably a more sensitive test statistic. This leads us to a general MFX correction principle.

Note that this argument relies on the availability of the global likelihood maximizer which, as already pointed out, is not guaranteed by the EM algorithm. In practice, MFX correction is performed without warranty of increasing profile likelihood, which may reduce the sensitivity of the ensuing test statistics as compared to the ideal situation where global maximization is guaranteed.

Location estimators. The population mean is defined by $\phi(f)=\int \beta f(\beta) \mathrm{d}\beta$, and thus has the following statistic as its NPMLE:

$$\hat{\mu} = \varphi(\hat{f}) = \sum_i \hat{w}_i \hat{z}_i, \tag{8}$$

which coincides with the classical sample mean when the observations are exact (then $\forall i$, $\hat{z}_i=\hat{\beta}_i$ and $\hat{w}_i=1/n$). While we note that the sample mean is unbiased under our modeling assumptions, the MFX-corrected estimator is more robust to those outliers that come with high first-level uncertainties. Such outliers are artefactual in the sense that they result from inaccuracies in the within-subject processing pipeline, due to poor distortion and/or motion correction, poor spatial normalization, resampling artifacts, unaccounted spikes in the time series, inappropriate BOLD response models, etc.

Also, if strongly non-Gaussian, the actual random effect's PDF f may have a tendency to bear behavioral outliers, *i.e.* subjects with activation patterns genuinely atypical even though they meet the recruitment criteria set up by the experimenter. Such a distribution may be better characterized by a "robust" location parameter. For instance, the NPMLE of the median is found by solving: $\hat{F}(z)=1/2$, where $\hat{F}$ is the estimated cumulative distribution function (CDF):

$$\hat{F}(\beta) = \sum_i \hat{w}_i \chi(\beta - \hat{z}_i), \tag{9}$$

$\chi(\cdot)$ being the indicator function of $\mathbb{R}^+$. The root of this equation is a weighted median.

A general difficulty with location estimators is that they are not invariant under multiplication of the sample by a positive factor. Since group effects have typically different scales across brain regions, location estimators tend to produce statistical maps that are highly non-stationary, thus requiring non-uniform thresholding strategies for good detection performance (see section of Statistical calibration).

Mixed-effect Fisher's sign statistic. The sign statistic t_s is the number of positive values in a sample (using the convention that zero counts half). If the observations are exact, t_s provides an efficient test of the population median: under the null hypothesis that the median is zero, t_s follows a binomial law $\mathcal{B}_{n,\frac{1}{2}}$ whatever the shape of f. An alternative interpretation is that $t_\mathrm{s}/n=1-\hat{F}_\mathrm{e}(0)$, where $\hat{F}_\mathrm{e}$ is the empirical CDF corresponding to Eq. (6), therefore t_s is, up to a constant factor, the NPMLE of $\phi(f)=1-F(0)$, which is the probability of a positive effect. Clearly, the median vanishes if and only if $\phi(f)=1/2$. This interpretation provides the key to an MFX-corrected Fisher's sign statistic:

$$t_\mathrm{s}^\mathrm{MFX} = n\left[1 - \hat{F}(0)\right] = n \sum_i \hat{w}_i \chi(\hat{z}_i),$$

where $\hat{F}$ is the NPMLE of the CDF given by Eq. (9).

Mixed-effect Wilcoxon's signed rank statistic. The Wilcoxon's signed rank statistic is a classical alternative to the sign statistic that works by sorting the absolute effects in ascending order, then summing up the ranks multiplied by the corresponding effect's signs, yielding:

$$t_\mathrm{w} = \sum_{i=1}^{n} \mathrm{sign}\left(\hat{\beta}_i\right) \mathrm{rank}\left(|\hat{\beta}_i|\right)$$

It is meaningful to interpret Wilcoxon's statistic as a measure of symmetry about zero. More specifically, if the observations are exact and if the population median is zero, we easily prove that t_w is n times the NPMLE of the covariance $\phi(f)=\mathrm{Cov}(\mathrm{sign}(Z), F_+(|Z|))$ where F_+ is the CDF of $|Z|$, that is: $F_+(u)=F(u)-F(-u)$ for $u\geq 0$. Clearly, $\phi(f)=0$ if the effect's sign and the effect's absolute value are statistically independent, a situation that occurs in particular if the PDF f is symmetric about zero.

Therefore, under the assumption that f is symmetric, rejecting $\phi(f)=0$ implies that the (unique) location parameter of f is different from zero. Note that the reasoning is valid whatever the function applied to the absolute effects, yet the fact that $F_+(|Z|)$ is uniformly distributed in [0, 1] ensures that t_w is scale-invariant. The MFX generalization follows straightforwardly from this interpretation:

$$t_\mathrm{w}^\mathrm{MFX} = \sum_{i=1}^{n} \hat{w}_i \mathrm{sign}(\hat{z}_i) \left[\hat{F}(|\hat{z}_i|) - \hat{F}(-|\hat{z}_i|)\right],$$

with $\hat{F}$ specified as in Eq. (9).

Likelihood ratio

Rather than working in the parameter space and picking a parameter estimate that maximizes likelihood, we may work in the likelihood space and consider the maximum likelihood ratio (LR) as a test statistic for H_0: $\theta \in \Theta_0$,

$$r = \frac{\max_{f \in \mathcal{F}_0} L(f)}{\max_{f \in \mathcal{F}} L(f)} = \frac{\max_{\theta \in \Theta_0} \tilde{L}(\theta)}{\max_{\theta \in \mathbb{R}} \tilde{L}(\theta)},$$

where $\mathcal{F}_0 \subset \mathbb{R}$ is the subspace of PDFs for which $\phi(f) \in \Theta_0$. Clearly, r takes on values between 0 and 1, small values indicating that the random effect's PDF is unlikely to lie in $\mathcal{F}_0$. The LR is partly justified by the Neyman–Pearson lemma, which says that the most sensitive statistic for the test of two *simple* hypotheses is the ratio of their respective likelihoods.

For the sake of clarity, we now specify $\phi(f)$ as the mean. Computing the LR involves performing both constrained and unconstrained likelihood maximizations. Consider first the simple null hypothesis H_0: $\theta=0$ so that $\Theta_0=\{0\}$. Practical constrained maximization is then performed using an adaptation of the above discussed EM algorithm for the unconstrained problem (see Appendix A.2). Scale invariance is guaranteed here by the fact that each of the two maximization sets contains all scalings of any of its PDFs. By construction, however, this LR statistic is well suited for a two-sided test as it is blind to the mean effect's sign.

In order to perform a one-sided test, hence testing H_0: $\theta\leq 0$, one should ideally solve the constrained maximization problem on $\Theta_0=\mathbb{R}^-$ but this is practically intractable as the profile likelihood may not be bell-shaped. To work around this problem, we postulate a property of the above maximum likelihood ratio defined for the simple null H_0: $\theta=0$, which holds for a variety of likelihood ratios and is known as Wilks' phenomenon: if $f\in\mathcal{F}_0$, then $-2\log r$ converges in distribution towards a χ^2 as the sample size goes to infinity, the χ^2 having as many degrees of freedom as the dimension of the parameter of interest, one in this case. This result is proved in Owen (2001) in the case of exact observations; its generalization to mixed effects will be omitted here as the proof is rather technical and out of the scope of this paper.

According to Wilks' phenomenon, the following one-sided LR variant:

$$t_r^{\mathrm{MFX}} = sign(\hat{\mu})\sqrt{-2\log r}, \tag{10}$$

where $\hat{\mu}$ is the mean NPMLE as in Eq. (8), is asymptotically distributed like a normalized Gaussian. Simulations suggest that this is generally an anticonservative approximation (see Fig. 2). We will refer to $t_{\mathrm{r}}^{\mathrm{MFX}}$ as the *MFX empirical LR* statistic (MFX–ELR), following Owen's preference for the word "empirical" rather than "nonparametric" (Owen, 2001).

Parametric statistics

Similar maximum likelihood estimators or likelihood ratios can be derived in parametric context, that is, by restricting the search space $\mathcal{F}$ to a finite-dimensional family. For instance, substituting the NPMLE for the PDF that maximizes likelihood over the Gaussian family, we obtain parametric maximum likelihood location estimators that generally differ from their nonparametric versions discussed in the section on Parameter estimators. Likewise, we define a Gaussian version of the one-sided likelihood ratio (see section on Likelihood ratio), hereafter referenced as the *MFX Gaussian LR* statistic (MFX–GLR), which is shown to consistently generalize the standard one-sample t statistic (Mériaux et al., 2006b).

In the fMRI group analysis, however, we tend to prefer nonparametric statistics because prior information about the random effect's PDF is barely available as it depends on the cognitive task under study. Admittedly, nonparametric estimation results in overfitting the PDF owing to the so-called "curse of dimensionality", however low-dimensional parameters computed from that nonparametric fit may actually be efficient estimators. Simulations confirm this intuition, showing that in non-Gaussian distributions, the MFX–ELR can substantially outperform the MFX–GLR in terms of power, while, surprisingly enough, both statistics turn out to perform comparably in Gaussian distributions, the case most favorable to the MFX–GLR (see Fig. 3).

Statistical calibration

Having chosen a test statistic t for H_0: $\theta\in\Theta_0$, we now turn to the problem of designing a proper statistical test, which involves defining the PDF $p(t|H_0)$ of the statistic given H_0. In a pure frequentist approach, however, this PDF is not uniquely defined since it is *a priori* dependent on the unknown random effect's PDF f satisfying H_0. For likelihood ratio statistics, this dependence may vanish in the limit of large samples as a consequence of Wilks' phenomenon (see section on Likelihood ratio), but such asymptotic properties only provide crude significance assessments in small samples. An alternative is then to use resampling techniques (Nichols and Hayasaka, 2003), a particularly useful example of which are, in our context, sign permutations (Holmes et al., 1996).

Sign permutations

Sign permutations consist of tabulating the reference PDF $p(t|H_0)$ by resampling the estimated effects across all possible flips

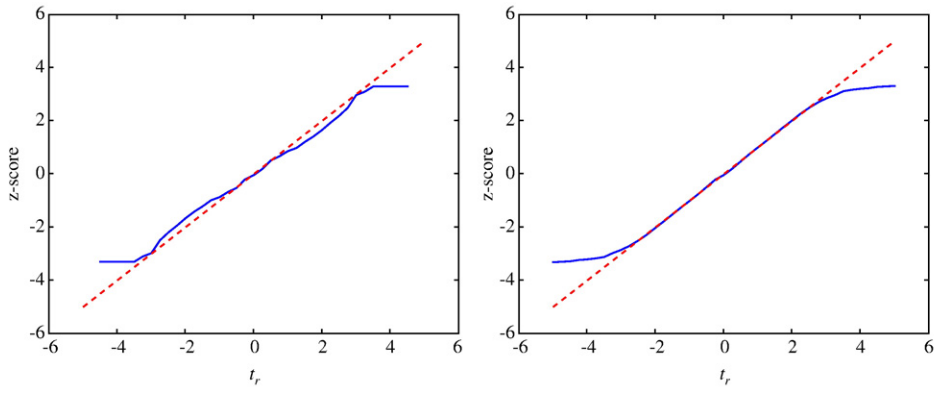

Fig. 2. QQ plot of permutation resampled values of $t_{\mathrm{r}}^{\mathrm{MFX}}$ plotted against normal quantiles, for sample sizes of 10 (left) and 30 (right) drawn in a normal population $N(0, 1)$ with first level variances drawn in a Gamma distribution $\Gamma(3, 1/6)$. The solid and dashed lines show, respectively, the z-score computed using sign permutations and Wilks' asymptotic z-score.

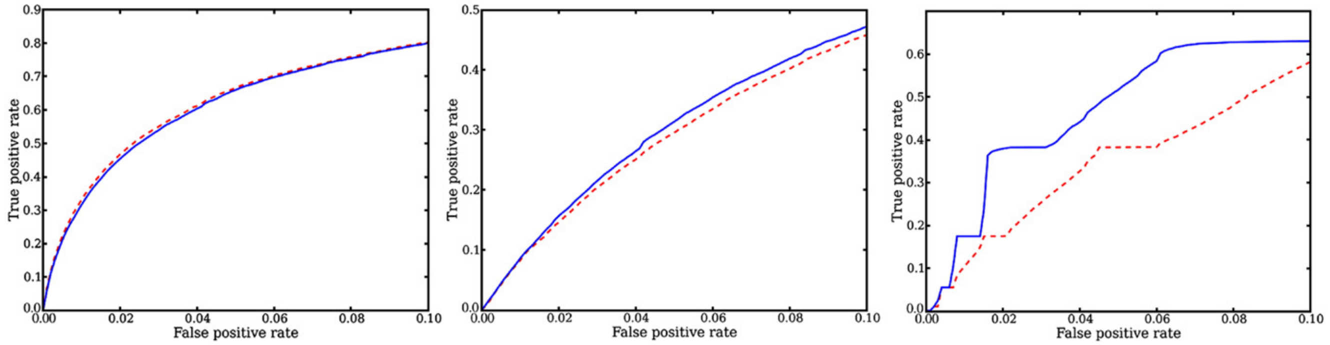

Fig. 3. Monte Carlo-simulated ROC curves comparing the detection power of the permutation tests based respectively on the MFX–ELR (solid curve) and the MFX–GLR (dotted curve). In all three panels, 10,000 samples of size 10 were drawn from a fixed PDF $f(\beta)$ with mean 1, and corrupted with homoscedastic Gaussian noise with standard deviation 1. Left, $f(\beta)=N(1, 1)$ is Gaussian. Middle, $f(\beta)=1/2[N(-1, 1)+N(3, 1)]$ is a symmetric mixture of two Gaussians. Right, $f(\beta)=1/2[N(-1, 0.1)+N(3, 0.1)]$ is a mixture with sharper modes. Note that the MFX–GLR statistic is equivalent here to the standard t statistic because the noise is homoscedastic.

of signs, the number of which is 2^n. They are primarily intended to test a simple null hypothesis H_0: $\theta=0$ about a location parameter under the restrictive assumption that f is symmetric, in which case all location parameters are identical. Therefore, testing the mean is equivalent to testing any other location parameter, although the test statistics derived according to different location parameters (see section on Test statistics) may not be equally sensitive.

In MFX context, Mériaux et al. (2006b) formulate two additional conditions on which sign permutations are applicable: (i) the subjects are drawn independently; (ii) first-level estimators are location equivariant and scale-invariant. These conditions imply that the distributions of a well-behaved test statistic are stochastically increasing w.r.t. the population mean effect, which validates the use of sign permutations to perform one-sided tests, *i.e.* testing the composite hypothesis H_0: $\theta \leq 0$. Notice that those two conditions are more general than those underlying the MFX model stated in the section on Mixed-effect model to guide the test statistic's derivation. In particular, sign permutations do not require the first-level variability to be genuinely Gaussian, nor do they assume a parametric form for the random effect's PDF provided it is symmetric.

The permutation-based distribution of t is conditional on the effects' absolute values and the first-level variances when present, meaning that the test is conditionally exact, hence unconditionally exact, up to the discretization induced by the finite number of permutations; P-values are conservatively precise at 2^{-n}. This is the same permutation mechanism as that classically used to calibrate both Fisher's sign statistic and Wilcoxon's signed rank (see section on Parameter estimators), which are special cases for which the permutation-based distribution is data-independent. Another intriguing example is the t statistic, whose permutation-based distribution converges towards the Student distribution in the limit of large samples (Fieller, 2005).

Multiple testing

In image analysis context, a test is performed at each and every voxel in a search volume. It is computationally efficient to threshold the statistical map using a uniform value, which can be tuned, *e.g.*, so as to control the false positive rate below a desired level. While spatially variable thresholds are conceivable and, to some extent, more natural, uniform thresholding yields good detection power provided the test statistic's distribution is reasonably stationary. As already pointed out, this constraint calls for scale-invariant statistics.

A major advantage of permutation tests is that they readily solve the multiple comparison problem under multivariate exchangeability in a way that circumvents both the parametric assumptions and the approximations underlying random field theory (Worsley, 1994). Like in Holmes et al. (1996), Bullmore et al. (1999), Nichols and Holmes (2002), and Hayasaka and Nichols (2003), we compute voxel-level P-values corrected for the familywise error rate by calibrating the maximum statistic over the search volume. Similarly, we compute corrected cluster-level P-values from the permutation-based distribution of the maximum cluster extent statistic after thresholding.

Discussion

The one-sample tests described in this paper have been implemented in C language within the NiPy library project (http://neuromimaging.scipy.org). A Matlab™ interface exists (Distance toolbox; http://www.madic.org) and is released as a plug-in for the Statistical Parametric Mapping software (SPM; http://www.fil.ion.ucl.ac.uk/spm).

Practical use

Our initial motivation was to avoid a naive approach in which users would perform outlier diagnosis, drop some subjects and run the parametric t test (or any other test) on the remaining subjects. Technically, that approach boils down to a MFX analysis where outliers are assigned infinite uncertainty, but it breaks statistical independence whenever outliers are detected from a joint analysis of the subjects. Unless intrinsic evidence supports excluding some subjects, this "drop and test" approach is bound to produce critically anti-conservative inferences. Instead, we suggest using MFX corrected test statistics, which are robust to artefactual outliers, yet assessing significance levels through permutations from the *complete* original dataset.

In practice, MFX statistics may fail to increase detection power over statistics that ignore within-subject variances, if there exists a positive correlation between the within-subject variances and the corresponding effect estimates (Mériaux et al., 2006a). This situation typically occurs when first-level analyses use an inaccurate

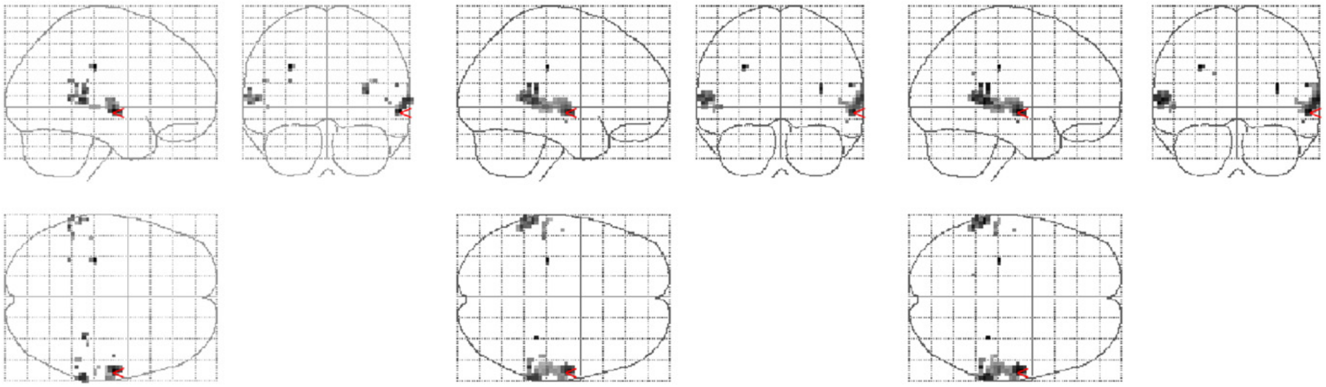

Fig. 4. Sentence × Speaker interaction (FIAC'05 dataset). Comparison of three permutation tests using different statistics, from left to right: the Student statistic (parametric thresholding produces an almost identical map), the MFX–GLR, the MFX–ELR. Maximum intensity projections are displayed after thresholding at $P = 0.01$ uncorrected.

BOLD impulse response model, or do not account for nonstationarities such as habituation effects. The first-level errors then capture large amounts of signal and are artificially inflated. We shall therefore stress the importance of accurate within-subject models.

Finally, some guidelines for choosing a test statistic are useful, given that only one test is authorized in theory. The method presented in the Method section determines a unique test statistic according to the following three criteria: (*i*) the model for the random effect's PDF: parametric or nonparametric; (*ii*) the location parameter of interest: mean, median, mode, etc. (although we noted earlier that all location parameters are equivalent under the symmetry assumption underlying the sign permutation mechanism); (*iii*) the type of test statistic: parameter estimator or likelihood ratio. Regarding (*i*), we prefer nonparametric models for their potential to accommodate non-Gaussian populations. Regarding (*ii*), we choose the mean as it is clearly the most commonly used location parameter. Finally, as discussed above, the scale invariance property of the likelihood ratio makes it suitable for spatially uniform thresholding. We therefore recommend the MFX–ELR (10) as a default test statistic.

Case study: FIAC'05 dataset

The Functional Imaging Analysis Contest (FIAC) dataset (Dehaene-Lambertz et al., 2006a) served as a comparative benchmark for various fMRI analysis methods in a workshop held in Toronto, 2005, at the 11th Meeting of the Organization for Human Brain Mapping. Results were reported in a special issue of the Human Brain Mapping journal (Poline et al., 2006). The experiment exploited sentence repetition and involved 15 participants who were asked to listen carefully to sentences from "The Three Little Pigs" read by different speakers. The experimental design was 2 × 2 factorial, with sentence type as the first factor (same vs. different) and speaker type as the second factor (same vs. different), and included two block-related sessions. Data were acquired on a Bruker 3 T scanner.

We performed one-sided permutation tests combined with each of the following four statistics: the *t* statistic, the standard empirical LR statistic (ELR) and their respective MFX-corrected versions. Note that the results of the test using the MFX–GLR statistic were reported in the same issue (Mériaux et al., 2006b). Analyses were performed from first-level summary statistics produced by SPM2 (after $5 \times 5 \times 5$ mm^3 FWHM spatial Gaussian smoothing), and were restricted to a manually segmented mask of 2920 voxels surrounding the perisylvian areas. In order to save computation time, all tests were performed using $N = 10{,}000$ random independent permutations rather than $2^{15} = 32{,}768$ exhaustive permutations. The randomization introduces a standard error on any *P*-value (corrected or uncorrected) of $\sqrt{(P - P^2)/N} \leq 1/2\sqrt{N} = 0.005$. Finally, each test was thresholded for a 1% false positive rate ($P = 0.01$ uncorrected) in order to obtain comparable activation maps.

When comparing the standard statistics with their MFX-corrected counterparts, we observe that the local maxima have similar locations in the respective maps, but the clusters detected under MFX correction are significantly bigger in the sense that they have smaller corrected cluster-level *P*-values. A contrast of special interest is the Sentence × Speaker interaction, for which both MFX statistics detect a significant cluster in the right superior temporal sulcus (STS), as shown in Fig. 4 and Table 1. The same cluster falls short of significance using the tests based

Table 1
Sentence×Speaker interaction (FIAC'05 dataset)

Statistical test	Cluster-level $P_{\text{FWE-corr}}$	Cluster size (voxels)	Peak voxel-level $P_{\text{FWE-corr}}$	Peak position (mm)			Time/perm (s)
				x	*y*	*z*	
Parametric *t* test (SPM)	0.13	24	0.70	60	−12	−3	–
Permutation *t* test	0.09	25	0.28	60	−12	−3	0.01
Permutation ELR	0.13	20	0.44	60	−9	−3	0.01
Permutation MFX–GLR	0.02	97	0.04	60	−12	−3	0.03
Permutation MFX–ELR	0.01	111	0.04	60	−12	−3	2.97

Comparison of *P*-values corrected for the family-wise error rate, for the biggest cluster found in the right superior temporal sulcus after thresholding each test at $P = 0.01$ uncorrected. The computation times are given for a standard PC, 2.80 GHz single processor running Linux.

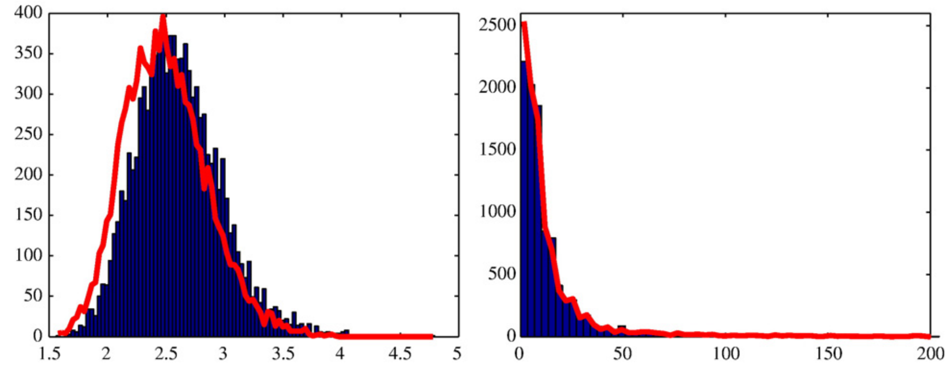

Fig. 5. Sentence × Speaker interaction (FIAC'05 dataset). Histograms of the maximum statistic over the search volume, $T_{\max}$ (left) and the maximum cluster size, $S_{\max}$ (right), respectively for the MFX–ELR (bars) and the MFX–GLR (solid lines).

on either the standard t statistic, or the ELR, which confirms the improved detection power achieved by MFX correction on the FIAC'05 dataset.

These experiments also suggest that the MFX–ELR does not sacrifice the detection performances of its parametric version, the MFX–GLR, which is consistent with the simulations (see Fig. 3). On each contrast we tested, the activation maps were similar, both qualitatively and in terms of significance; see the histograms of maximum statistics over the search volume in Fig. 5. While this trend has no reason to generalize to all datasets, it suggests that the across-subject distributions of the contrasts are fairly normal in this study. Yet we anticipate that substantial differences between the two MFX statistics could be seen in strongly inhomogeneous populations.

In our implementation, the MFX–ELR is about 100 times slower than the MFX–GLR, which reflects the respective computational costs of the associated EM algorithms. In this example, the permutation test using the MFX–ELR took about 8 h, compared to 6 min using the MFX–GLR. We do not consider this computation time prohibitive as it remains negligible compared to the time required to design an fMRI experiment and acquire a complete group dataset. Also note that permutation tests are easy to parallelize.

Conclusion

This work is an attempt to boost the detection power of massively univariate random-effect analyses. Random-effect analyses in fMRI call for MFX models, since BOLD effects cannot be observed exactly from the scans. Standard t (and F) statistics may be interpreted in MFX context as corresponding to a homoscedastic two-level Gaussian model. We have investigated a more flexible model accounting for heteroscedasticity and relaxing normality at the between-subject level, from which new test statistics maybe derived that are potentially more sensitive at the expense of heavier computations. Those statistics may be combined with permutation tests to enable exact specificity control under a symmetry assumption regarding the across-subject distribution of the effects.

Acknowledgments

This work was partly funded by the French Ministry of Research (ACI Neurosciences Intégrées et Computationnelles). Many thanks to Ghislaine Dehaene, Christophe Pallier, Nick Fieller, Philippe Ciuciu and Jean-Baptiste Poline for thoughtful discussion.

Appendix A. A practical likelihood maximization

We detail here two EM algorithms used to maximize the likelihood function given by Eq. (4), respectively on the Gaussian family and on the space of all PDFs on $\mathbb{R}$. EM algorithms are a class of iterative techniques for likelihood maximization that are guaranteed to increase the likelihood value on each iteration (Dempster et al., 1977). These two algorithms may be seen as special cases of a more general EM algorithm that models the random effect's PDF as a mixture of Gaussians and considers the *true* effects and the class labels as missing data.

A.1. EM algorithm for a Gaussian population

In this case, $\mathcal{F}$ is the Gaussian family so that the random effect's PDF f is parameterized by its mean μ_{f} and standard deviation σ_{f}. The algorithm iteratively refines initial estimates $\hat{\mu}_{\mathrm{f}}$ and $\hat{\sigma}_{\mathrm{f}}$ by alternating two steps, the E-step (expectation) and the M-step (maximization). In our implementation, the initial estimates are respectively taken as the classical sample mean and sample standard deviation of the observed effects ($\hat{\beta}_1$, $\hat{\beta}_2$, …, $\hat{\beta}_n$).

E-step. Assume current estimates are exact, and compute the posterior joint PDF of all subject's true effects. Since the subjects are conditionally independent, this reduces to computing each subject's posterior, $p(\beta_i|\hat{\beta}_i, \sigma_i, \mu_{\mathrm{f}}, \sigma_{\mathrm{f}})$ which is a Gaussian with parameters (m_i, s_i):

$$m_i = \frac{\hat{\sigma}_{\mathrm{f}}^2}{\sigma_i^2 + \hat{\sigma}_{\mathrm{f}}^2}\,\hat{\beta}_i + \frac{\sigma_i^2}{\sigma_i^2 + \hat{\sigma}_{\mathrm{f}}^2}\,\hat{\mu}_{\mathrm{f}}, \qquad s_i = \frac{\sigma_i\,\hat{\sigma}_{\mathrm{f}}}{\sqrt{\sigma_i^2 + \hat{\sigma}_{\mathrm{f}}^2}}$$

M-step. Update ($\hat{\mu}_{\mathrm{f}}$, $\hat{\sigma}_{\mathrm{f}}$) by maximizing the expected log-likelihood of the complete data:

$$\mathcal{Q}(\mu_{\mathrm{f}},\sigma_{\mathrm{f}}) = n\log\sqrt{2\pi}\sigma_{\mathrm{f}} + \frac{1}{2\sigma_{\mathrm{f}}^2}\sum_i \left[s_i^2 + (\mu_{\mathrm{f}} - m_i)^2\right],$$

yielding:

$$\hat{\mu}_{\mathrm{f}} = \frac{1}{n}\sum_i m_i, \quad \hat{\sigma}_{\mathrm{f}}^2 = \frac{1}{n}\sum_i \left[s_i^2 + (\hat{\mu}_{\mathrm{f}} - m_i)^2\right]$$

In the constrained problem of maximizing likelihood subject to $\mu_f = 0$, the algorithm is identical except that $\hat{\mu}_f$ is frozen to zero in the M-step. In practice, we perform five EM iterations, which proves generally sufficient to achieve good precision on the ensuing test statistics.

A.2. EM algorithm for a general population

In this case, $\mathcal{F}$ is the space of all PDFs on $\mathbb{R}$. It is then proved in Lindsay (1983) that the maximum likelihood PDF is necessarily a mixture of n or less Dirac masses:

$$f(\beta) = \sum_{k=1}^{n} w_k \delta(\beta - z_k),$$

which amounts to saying that each observation $\hat{\beta}_i$ is drawn from an unobserved class $k(i)$. In practice, we initialize the EM algorithm with uniform mixing proportions and support points coinciding with the observations (corresponding to the maximum likelihood PDF under exact observations).

E-step. Given the PDF parameters, we compute the posterior probability q_{ik} of class label k for subject i, yielding:

$$q_{ik} = \frac{\hat{w}_k gi\left(\hat{\beta}_i - \hat{z}_k\right)}{\sum_{k'} \hat{w}_{k'} gi} \left(\hat{\beta}_i - \hat{z}_{k'}\right)$$

M-step. Given the posterior probabilities q_{ik}, we form the negated expected complete-data log-likelihood:

$$\mathcal{Q}(w,z) = \sum_{i,k} q_{ik} \left[\log\sqrt{2\pi}\sigma_i + \frac{\left(\hat{\beta}_i - z_k\right)^2}{2\sigma_i^2} - \log w_k \right]$$

This criterion is to be minimized subject to the constraint that the mixing proportions sum up to one, and possibly that the mean population effect vanishes. We therefore consider the Lagrangian:

$$\mathcal{L}(w, z, \lambda_0, \lambda_1) = \mathcal{Q}(w,z) + \lambda_0 \left(\sum_k w_k - 1 \right) + \lambda_1 \left(\sum_k w_k z_k \right),$$

whose derivatives read:

$$\frac{\partial \mathcal{L}}{\partial w_k} = -\frac{1}{w_k} \sum_i q_{ik} + \lambda_0 + \lambda_1 z_k,$$

$$\frac{\partial \mathcal{L}}{\partial z_k} = \sum_i \frac{q_{ik}}{\sigma_i^2} \left(z_k - \hat{\beta}_i \right) + \lambda_1 w_k$$

When no mean constraint is applied, so that $\lambda_1 = 0$, the M-step yields an explicit updating rule. We easily get $\lambda_0 = n$, then:

$$\hat{w}_k = \frac{1}{n} \sum_i q_{ik}, \quad \hat{z}_k = \frac{1}{S_k} \sum_i \frac{q_{ik}}{\sigma_i^2} \hat{\beta}_i \quad \text{with} \qquad S_k = \sum_i \frac{q_{ik}}{\sigma_i^2}$$

Constrained M-step. When maximizing likelihood subject to the zero mean constraint, the Lagrange multiplier λ_1 becomes a free parameter. In this case, there is no explicit solution to the M-step. A possibility is to recast the joint constrained minimization w.r.t. w and z as sequential constrained minimizations:

- Along w (at fixed z). We, again, find that $\lambda_0 = n$, and:

  $$\hat{w}_k = \frac{\frac{1}{n} \sum_i q_{ik}}{1 + \lambda_1 z_k}$$

  We then solve for λ_1 by writing the zero-mean constraint, leading to a weighted version of the standard empirical likelihood equation (Owen, 2001), whose solution generally exists uniquely, and can be found using a Newton algorithm. The only exception is when all the support points have the same sign, in which case no solution exists and we simply leave the mixing proportions unchanged until the next iteration.
- Along z (at fixed w). We easily get the following implicit equation:

  $$\hat{z}_k = \frac{1}{S_k} \left(\sum_i \frac{q_{ik}}{\sigma_i^2} \hat{\beta}_i - \lambda_1 w_k \right),$$

  (with S_k like in the unconstrained case) which becomes explicit after expressing the zero-mean constraint and solving the resulting linear equation for λ_1.

In practice, we do not iterate the alternate minimization, meaning that, in the constrained case, our algorithm is actually an EM variant known as the expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993), which maintains the essential property that the likelihood value increases on each iteration.

References

Beckmann, C., Jenkinson, M., Smith, S., 2003. General multi-level linear modelling for group analysis in fMRI. NeuroImage 20, 1052–1063.

Brammer, M., Bullmore, E., Simmons, A., Grasby, P., Howard, R., Woodruff, P., Rabe-Hesketh, S., 1997. Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. Magn. Reson. Imaging 15 (7), 763–770.

Bullmore, E., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M., 1999. Global, voxel, and cluster tests, by theory and permutation, for difference between two groups of structural MR images of the brain. IEEE Trans. Med. Imag. 18, 32–42.

Dehaene-Lambertz, G., Dehaene, S., Anton, J.-L., Cam-pagne, A., Ciuciu, P., Dehaene, G.P., Denghien, I., Jobert, A., Le Bihan, D., Sigman, M., Pallier, C., Poline, J.-B., 2006a. Functional segregation of cortical language areas by sentence repetition. Hum. Brain Mapp. 27, 360–371.

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., 2006b. Functional organization of perisylvian activation during presentation of sentences in preverbal infants. Proc. Natl. Acad. Sci. U. S. A. 103, 14240–14245.

Delyon, B., Lavielle, M., Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. Ann. Stat. 27 (1), 94–128.

Dempster, A., Laird, A., Rubin, D., 1977. Maximum likelihood from incomplete data via the em algorithm (with discussion). J. R. Stat. Soc., Ser. B Stat. Methodol. 39, 1–38.

Escobar, M., West, M., 1998. Computing Bayesian nonparametric hierarchical models. Practical Nonparametric and Semiparametric Bayesian Statistics. Springer Verlag, New York, pp. 1–22.

Fieller, N., 2005. Statistical Modelling and Computing. Lecture notes published on the web.

Friman, O., Borga, M., Lundberg, P., Knutsson, H., 2003. Adaptive analysis of fMRI data. NeuroImage 19 (3), 837–845.

Friston, K., Stephan, K., Lund, T., Morcom, A., Kiebel, S., 2005. Mixed-effects and fMRI studies. NeuroImage 24, 244–252.

Good, P., 2005. Permutation, Parametric, and Bootstrap Tests of Hypotheses, 3rd edition. Springer.

Hayasaka, S., Nichols, T., 2003. Validating cluster size inference: random field and permutation methods. NeuroImage 20 (4), 2343–2356.

Holmes, A., Friston, K., 1998. Generalisability, Random Effects and Population Inference. NeuroImage, vol. 7, p. S754.

Holmes, A., Blair, R., Watson, J., Ford, I., 1996. Nonparametric analysis of statistic images from functional mapping experiments. J. Cereb. Blood Flow Metab. 16, 7–22.

Lindsay, B.G., 1983. The ge.ometry of mixture likelihoods: a general theory. Ann. Stat. 11 (1), 86–94.

Magder, L.S., Zeger, S.L., 1996. A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. J. Am. Stat. Assoc. 91 (435), 1141–1151.

Meng, X.L., Rubin, D.B., 1993. Maximum likelihood via the ECM algorithm: a general framework. Biometrika 80, 267–278.

Mériaux, S., Roche, A., Dehaene-Lambertz, G., Poline, J.-B., 2006a. When do mixed-effect models fail to improve detection sensitivity in fMRI group activation maps? NeuroImage (HBM'06), Florence, Italy.

Mériaux, S., Roche, A., Dehaene-Lambertz, G., Thirion, B., Poline, J.-B., 2006b. Combined permutation test and mixed-effect model for group average analysis in fMRI. Hum. Brain Mapp. 27 (5), 402–410.

Mériaux, S., Roche, A., Thirion, B., Dehaene-Lambertz, G., 2006c. Robust statistics for nonparametric group analysis in fMRI. Proc. 3th Proc. IEEE ISBI, pp. 936–939. Arlington, VA.

Neumann, J., Lohmann, G., 2003. Bayesian second-level analysis of functional magnetic resonance images. NeuroImage 20 (2), 1346–1355.

Nichols, T., Hayasaka, S., 2003. Controlling the family-wise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res. 12 (5), 419–446.

Nichols, T., Holmes, A., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

Owen, A.B., 2001. Empirical Likelihood. CRC Press, New York.

Poline, J.-B., Strother, S.C., Dehaene-Lambertz, G., Egan, G.F., Lancaster, J.L., 2006. Motivation and synthesis of the FIAC experiment: reproducibility of fMRI results across expert analyses. Hum. Brain Mapp. 27 (5), 351–359.

Rorden, C., Bonilha, L., Nichols, T.E.E., 2007. Rank-order versus mean based statistics for neuroimaging. NeuroImage 35 (4), 1531–1537.

Rousseeuw, P., Leroy, A., 1987. Robust Regression and Outlier Detection. Wiley.

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.-B., 2006. Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. Hum. Brain Mapp. 27 (8), 678–693.

Wager, T., Keller, M., Lacey, S., Jonides, J., 2005. Increased sensitivity in neuroimaging analyses using robust regression. NeuroImage 26 (1), 99–113.

Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., Smith, S., 2004. Multi-level linear modelling for fMRI group analysis using Bayesian inference. NeuroImage 21 (4), 1732–1747.

Worsley, K., 1994. Local maxima and the expected Euler characteristic of excursion sets of χ^2, f, and t fields. Adv. Appl. Probab. 26, 13–42.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., Evans, A., 2002. A general statistical analysis for fMRI data. NeuroImage 15 (1), 1–15.