NIH Public Access
Author Manuscript
Neuroimage. Author manuscript; available in PMC 2010 August 1.

Published in final edited form as:
Neuroimage. 2009 August 1; 47(1): 231–261. doi:10.1016/j.neuroimage.2009.02.035.

Enhanced False Discovery Rate using Gaussian Mixture Models for thresholding fMRI statistical maps

Gautam Pendse[1], **David Borsook**[2], and **Lino Becerra**[3]

[1]Gautam V. Pendse is with the Pain and Analgesia Imaging and Neuroscience (P.A.I.N) Group, McLean Hospital, Belmont, MA, U.S.A. (E-mail: gpendse@mclean.harvard.edu, Ph: (617) 855-3181, Fax: (617) 855-3772)

[2]David Borsook is with the Pain and Analgesia Imaging and Neuroscience (P.A.I.N) Group, McLean Hospital, Belmont, MA, U.S.A. and also with the Department of Psychiatry, Harvard Medical School, Cambridge, MA, U.S.A. (E-mail: dborsook@mclean.harvard.edu)

[3]Lino Becerra is with the Pain and Analgesia Imaging and Neuroscience (P.A.I.N) Group, McLean Hospital, Belmont, MA, U.S.A and also with the Department of Psychiatry, Harvard Medical School, Cambridge, MA, U.S.A. (E-mail: lbecerra@mclean.harvard.edu)

Abstract

A typical fMRI data analysis proceeds via the generalized linear model (GLM) with Gaussian noise using a model based on the experimental paradigm. This analysis ultimately results in the production of *z*-statistic images corresponding to the contrasts of interest. Thresholding such *z*-statistic images at uncorrected thresholds suitable for testing activation at a single voxel results in the problem of multiple comparisons. A number of methods which account for the problem of multiple comparisons have been proposed including Gaussian random field theory, mixture modeling and false discovery rate (FDR). The focus of this paper is on the development of a generalized version of FDR (GFDR) in an empirical Bayesian framework, specially adapted for fMRI thresholding, that is more robust to modeling violations as compared to traditional FDR. We show theoretically as well as by simulation that for real fMRI data various factors lead to a mixture of Gaussians (MOG) density for the "null" distribution. Artificial data was used to systematically study the bias of FDR and GFDR under varying intensity of modeling violations, signal to noise ratios and activation fractions for a range of *q*-values. GFDR was able to handle modeling violations and produce good results when FDR failed. Real fMRI data was also used to confirm GFDR capabilities. Our results indicate that it is very important to account for the form and fraction of the "null" hypothesis adaptively from the data in order to obtain valid inference.

Keywords

Functional magnetic resonance imaging (fMRI); False discovery rate (FDR); Adaptive null hypothesis; Empirical Bayes; Gaussian mixture modeling (GMM); Expectation maximization (EM)

Correspondence to: Gautam Pendse.

Conflict of interest: The authors have no conflicts of interest to declare.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

2 Introduction

Most fMRI analyses base their results on thresholded statistical maps that take into account the problem of multiple comparisons. The correction for multiple comparisons is a long standing one that tries to balance the number of false positives as well as the number of false negatives. Several approaches, some very rigorous, some heuristics in nature, have been used over the years.

In the early years of fMRI, Bonferroni type corrections [Simes, 1986] were common in part to impose high rigurosity in the results [Kwong et al., 1992; Weisskoff et al., 1993], yet, as the field grew it was evident that strong correction for false positives came at the cost of a high false negative rate [Worsley et al., 1996]. New approaches were implemented including Gaussian random field theory [Worsley et al., 1996], mixture modeling [Everitt and Bullmore, 1999], [Woolrich et al., 2005], permutation testing [Nichols, 2002] and false discovery rate (FDR) analysis [Genovese et al., 2002]. A comparative review of many techniques is presented in [Nichols and Hayasaka, 2003].

FDR has gained popularity due to its simplicity in implementation and reasonable results obtained with it. Conventional application of FDR technique to fMRI makes several simplifying assumptions including the form and fraction of "null" distribution in the data (section 3.1.1 - 3.1.2). It is not unusual to find in fMRI experiments, situations in which, such assumptions are not satisfied and the use of FDR is compromised. In this article we develop a generalized form of FDR (GFDR) for fMRI in an empirical Bayesian framework inspired by the work of Efron et al. [2001] that adaptively estimates the form and fraction of "null" from the data (section 3.2). In particular, we show that under the Gaussian noise model for massively univariate GLM analysis (using the same design matrix) of fMRI data, various factors such as modeling violations, signal inhomogeneities, variance in vascular flow and/or BOLD response (onset, strength, duration, extent), presence of coherent resting state network (RSN) type activity typically lead to a mixture of Gaussians (MOG) density for the "null" distribution (section 3.3 and Appendix B). Failure to account for this empirical "null" could result in misleading conclusions. Significant performance improvements are observed using GFDR as compared to FDR in experiments on both simulated and real data sets (section 3.4.1).

This paper is laid out as follows: (1) First, we cast FDR in a mixture modeling framework to examine its limitations as well as present prior work and motivation for the development of GFDR (sections 3.1.1-3.1.3). (2) Next, we present the algorithm to perform inference using GFDR (section 3.2). (3) In section 3.3.1 we examine the identifiability and separability of the mixture model used in GFDR as well as test the MOG hypothesis of GFDR for the "null" distribution in a simulation study with data generation in the presence of multiple confounds at locally varying SNRs. (4) In section 3.4.1, we present a second simulation study that compares GFDR and FDR. (5) In section 3.4.2, we describe application of GFDR to a real fMRI data-set.

3 Materials and methods

3.1 Motivation, limitations and prior work

In this section we present some background information that is necessary for understanding the developments in later sections. We present FDR theory in a mixture modeling framework, followed by an examination of its limitations. We review previously suggested improvements to FDR and present our novel contributions.

3.1.1 FDR in a mixture modeling framework—FDR theory assumes existence of a fixed mechanism to generate *p*-values corresponding to deviations from the null-hypothesis. Here

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

we present a derivation of FDR based on a mixture modeling framework without making this assumption. This will help us in understanding its assumptions as well as presenting its generalization in the next section.

Consider a Z statistic image with z-stat values z_i, $i = 1, 2, \ldots, n$ which are realizations of random variables Z_i, $i = 1, 2, \ldots, n$, where Z_i are independent, identically distributed random variables with probability distribution function $p(z)$.

Assume the image is thresholded at α and all voxels with $z_i > \alpha$ are declared "active". Suppose each z_i is associated with a latent variable c_i which indicates its membership in one of the three classes defined by the set $S = \{1, 2, 3\}$. $c_i = 1$ means membership in the "deactivation" class, $c_i = 2$ means membership in the "null" class and finally $c_i = 3$ indicates membership in the "activation class". Suppose each class is associated with a prior probability:

$$\pi_k = p(c_i = k) \tag{1}$$

where $i = 1, 2, \ldots, n$ and $k = 1, 2, 3$. With these assumptions, the distribution $p(z)$ can be decomposed as follows:

$$p(z) = \sum_{k=1}^{3} \pi_k \, p\left(z \middle| c = k\right) \tag{2}$$

Let $v_{ca}(\alpha)$ be the number of voxels considered active at threshold α. Then

$$v_{ca}(\alpha) = \sum_{i=1}^{n} I(Z_i > \alpha) \tag{3}$$

where $I(A)$ is the indicator function defined as:

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases} \tag{4}$$

If $v_{fa}(\alpha)$ be the number of voxels falsely declared as active at threshold α, then

$$v_{fa}(\alpha) = \sum_{i=1}^{n} \sum_{k=1}^{2} I(Z_i > \alpha) \, I(c_i = k) \tag{5}$$

FDR was first introduced as a general method of thresholding by Benjamini and Hochberg [Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001]. FDR is a function of the chosen threshold α and is defined as:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\mathrm{FDR}(\alpha) = \mathrm{E}\left(\frac{v_{fa}(\alpha)}{v_{ca}(\alpha)}\right) \tag{6}$$

where the expectation is taken with respect to Z_i.

Many equivalent but slightly different variations of the FDR procedure have been proposed over the years. For example, Storey [2003, 2002] proposes the positive false discovery rate (pFDR) which is defined as the expectation of the fraction of rejected nulls given that a non-zero number of rejections occur.

$$\mathrm{pFDR}(\alpha) = \mathrm{E}\left(\frac{v_{fa}(\alpha)}{v_{ca}(\alpha)} | v_{ca}(\alpha) > 0\right) \tag{7}$$

It is shown in Storey [2003] under some general conditions and under the assumption of independent and identically distributed tests that:

$$\mathrm{pFDR}(\alpha) = \frac{\mathrm{E}[v_{fa}(\alpha)]}{\mathrm{E}[v_{ca}(\alpha)]} \tag{8}$$

They show that pFDR comes quite close to the above form even under dependence when the number of tests is large. Efron et al. [2001] view FDR in an empirical Bayesian framework and define the false discovery rate in terms of cumulative distribution functions (CDF) as:

$$\mathrm{Fdr}(\alpha) = \frac{p_0(1 - F_0(\alpha))}{1 - \overline{F}(\alpha)} \tag{9}$$

where p_0, F_0 and $\overline{F}$ are the fraction of null, the CDF for the null hypothesis and the empirical overall CDF respectively. They show that $Fdr(\alpha)$ is biased upward ("conservative bias" theorem in their paper) as an estimator of conventional FDR as defined by Benjamini and Hochberg [1995]. Mathematically

$$\mathrm{FDR}(\alpha) \leq \mathrm{E}[\mathrm{Fdr}(\alpha)] \tag{10}$$

In addition, Efron et al. [2001] also note that the independence of test statistics plays no essential role in the empirical Bayes framework. In this paper, we will view FDR from this empirical Bayesian viewpoint. An alternative interpretation of $Fdr(\alpha)$ is that it is a non-parametric estimate of the probability $P(Z_i$ is "not active" $|Z_i > \alpha)$ using the "empirical" CDF of the observed data (Efron [2003]). Thus, $\mathrm{Fdr}(\alpha)$ can be written as:

$$\mathrm{Fdr}(\alpha) = \frac{\mathrm{E}[v_{fa}(\alpha)/n]}{1 - \overline{F}(\alpha)} \tag{11}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Define

$$n(\alpha) = \sum_{i=1}^{n} I(z_i > \alpha) \tag{12}$$

The denominator of (11) can be approximated using the standard empirical estimate:

$$1 - \overline{F}(\alpha) = \frac{n(\alpha)}{n} \tag{13}$$

and the numerator is given by:

$$\mathrm{E}[v_{fa}(\alpha)/n] = \sum_{k=1}^{2} \pi_k \mathrm{P}(Z_i > \alpha | c_i = k) \tag{14}$$

Combining (11), (14) and (13)

$$\mathrm{Fdr}(\alpha) = \frac{n \sum_{k=1}^{2} \pi_k \mathrm{P}(Z_i > \alpha | c_i = k)}{n(\alpha)} \tag{15}$$

It has been shown by Efron et al. [2001] ("equivalence theorem" in their paper) that for a given q, choosing smallest possible α (i.e., largest rejection region possible) such that $Fdr(\alpha) \leq q$ also implies $FDR(\alpha) \leq q$. Thus, the goal of controlling FDR in an empirical Bayesian framework is to bound the above by a user-defined fraction q:

$$\mathrm{Fdr}(\alpha) \leq q \tag{16}$$

Conventional FDR [Genovese et al., 2002; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001] controlling procedure makes the following assumptions:

- **Assumption 1:** $\pi_2 \gg \pi_1$ i.e., the fraction of the null distribution is very large as compared to the deactivation (and activation) distribution. Thus, one can use the approximation $\pi_2 \approx 1$ and hence force $\pi_1 = 0$.
- **Assumption 2:** The "null" distribution has a fixed parametric form. As applies to GLM based fMRI analysis this assumption translates to: The "null" distribution is the standard Gaussian distribution with mean 0 and variance 1, $N(0, 1)$, i.e., $\mathbf{P}(Z_i|c_i = 2) = N(0, 1)$.

With these assumptions, if $p(\alpha)$ is the p-value associated with the z-threshold α for a standard Gaussian $N(0, 1)$ distribution, then $Fdr(\alpha)$ is bounded by q if:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\frac{np(\alpha)}{n(\alpha)} \leq q \tag{17}$$

In conventional FDR, one usually finds the largest value of $p(\alpha)$ satisfying the above equation and thresholds the image at α. If the p-values are sorted in an increasing order, $p(1), p(2), \ldots, p(n)$ and if the kth p-value $p(k)$ is taken as a threshold then the above equation can be equivalently written as $np(k)/k \leq q$ or $p(k) \leq qk/n$. This is illustrated graphically in Figure 1 for the p-domain and in Figure 2 for the z-domain. From now on, we will use the acronym FDR to mean "conventional" FDR with the above simplifying assumptions and GFDR to mean the empirical Bayesian view in equation (15).

3.1.2 Limitations of conventional FDR—Based on the exposition of conventional FDR in section 3.1.1, it is immediately apparent that FDR makes some strong assumptions.

Assumption 1 is invalid if:

- There is significant amount of activation and/or deactivation i.e., when π_1 and π_3 are not "small" relative to π_2.

Assumption 2 is invalid if:

- There is un-modeled signal not accounted for by the design matrix. See Figure 15 for a list of factors that contribute to modeling violations.
- There is a local variability of the Haemodynamic Response Function (HRF).

Since different spatial locations in the brain have different functions, it is very likely that timeseries at different voxels are a mixture of different basis signals, some of which may not be stimulus driven. This invalidates modeling signals in the entire brain via a single design matrix. Since most conventional fMRI analyses do not postulate a different design matrix at different spatial locations, assumption (2) is very likely violated. Using a constrained basis set [Woolrich et al., 2004] in a single design matrix provides more flexibility in fitting the fMRI response that is stimulus driven. Performing an F-test on the fitted coefficient vector results in a loss of directionality information meaning that we detect both the activation and deactivation jointly. Woolrich et al. [2004] propose to use pseudo z-statistics to recover directionality but they point out that the resulting null distribution is not $N(0, 1)$ because of the constrained HRF priors used in the Bayesian inference.

A combined effect of one or more of these violations is that the "null" distribution becomes an MOG density instead of $N(0, 1)$. (see Appendix B for mathematical proof). Efron [2006] show that even when the individual null voxels follow the theoretical null distribution $N(0, 1)$ (perfect modeling at all voxels), the presence of correlations between the voxels can make the ensemble null behave as $N(0, \sigma^2)$. where σ is far from one. In view of these important considerations, conventional FDR procedure must be used carefully as a method of controlling false discoveries.

This was the motivation for developing a generalization of FDR that makes none of the assumptions made by FDR but instead attempts to estimate all quantities adaptively from the data that are needed to get a robust estimate of FDR given in equation (15).

3.1.3 Prior work and novel contributions—The problem of estimating the fraction of truly null hypothesis π_2 has been tackled before using a number of approaches. For example, Allison et al. [2002] uses a mixture model comprising of a uniform distribution (to model the null distribution) and a beta distribution (to model non-null distributions) to fit the distribution

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

of *p*-values using maximum likelihood (ML). The fraction of truly null hypothesis is then estimated as the ML solution. Another estimator was proposed by Storey [2003] who used the fact that *p*-values above a specified threshold λ are mostly draws from a uniform distribution corresponding to the null. They propose the estimator:

$$\widehat{\pi}_2(\lambda) = \frac{\text{Number of } p-\text{values} > \lambda}{n(1-\lambda)} \tag{18}$$

This estimate depends on the tuning parameter λ. An average estimate of $\hat{\pi_2}$ is obtained by averaging over different values of λ or via a bootstrap. Benjamini et al. [2006] propose a two stage procedure for estimating $\hat{\pi_2}$. These and other approaches make the following assumptions

1. There exists a reliable mechanism of estimating the *p*-values based on the raw data. This usually means that there are some strong assumptions about the null hypothesis $\mathbf{H_0}$ [for example $\mathbf{H_0} = N(0, 1)$]. As we show later these assumptions are not always valid in real fMRI data sets.
2. There are two types of data points, "null" and "non-null". In fMRI data analysis, it is more common to have "activation", "deactivation" and "null" data points.

The novel contributions of this article are the following:

1. Enable estimation of $\mathbf{H_0}$ adaptively from the data as a mixture of Gaussians (MOG) density.
2. Enable estimation of the fraction of null and deactivation distributions $\{\pi_k, k = 1, 2\}$.

The problem of estimating $\mathbf{H_0}$ is at the heart of empirical Bayesian approaches. For example, Efron [2004] discuss a drug mutation study that was analyzed using logistic regression and where the *z*-values were computed using maximum likelihood estimates for the logistic coefficients and large sample estimates of their standard error. Here they propose to estimate the null distribution as a Gaussian distribution fitted to the peak of the histogram of *z*-values. This approach first requires the identification of a maximum δ_0 of the histogram via poisson regression. Next, it involves the estimation of standard deviation σ_0 of the "null" by fitting a quadratic model to log $f(z)$ for z within 1.5 units of maximum δ_0. The "null" fraction π_0 is chosen to be the fraction of *z*-values falling in the interval $(\delta_0 - \alpha\sigma_0, \delta_0 + \alpha\sigma_0)$, where α is a user specified constant (e.g., $\alpha = 1.64$). The main disadvantage of this approach is the need to specify the "width" around the maximum for estimation of σ_0 as well as the need to specify α for the estimation of π_0.

How should one choose $\mathbf{H_0}$ to enable its estimation from the data? The answer depends on the type of analysis being carried out to generate the *z*-stat images as well as the effect of modeling violations (see 15 for examples of modeling violations) on the analysis process i.e., estimation of $\mathbf{H_0}$ from the data should be application specific. In fMRI, a massively univariate analysis of a large number of voxels using the same design matrix X via a GLM (with Gaussian noise) is typical. It is easy to see how, in the presence of confounding signals, this analysis results in altered densities of the "null", "activation" and "deactivation" distributions (see Appendix B). A simple mathematical calculation reveals that that a mixture of Gaussians (MOG) density is a natural model for the data histogram as well as for the "activation", "null" and "deactivation" distributions (see Appendix B) given this analysis scheme. We were also able to numerically confirm this phenomenon via a large simulation study (section 3.3) for a range of "activation" fractions and Signal to Noise Ratios (SNRs) of the multiple unmodeled confounds using a realistic locally variable mixing process. Our approach in estimating $\mathbf{H_0}$ involves the MOG

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

hypothesis mentioned above. One of the key strengths of this approach is that it does not require any parameter setting by the user.

3.2 GFDR Algorithm

The GFDR algorithm has several processing steps, each of which will be explained in detail in the following discussion. For notational convenience, let $\varphi(z; \theta)$ be the probability density function of a Gaussian distribution with mean μ and variance σ^2. Here, $\theta = \{\mu, \sigma\}$ is the parameter set associated with this distribution.

$$\varphi(z;\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}}, \text{ where } \theta = \{\mu, \sigma\} \tag{19}$$

We assume that the histogram of observed z-values can be well modeled as a mixture of Gaussian (MOG) distributions-even in the presence of modeling violations. In addition, we also assume that within the full MOG fit, the null distribution will be modeled as a mixture of 1 or more Gaussian distributions, ditto for "activation" and "deactivation" distributions. The assumption that the MOG density remains valid even in the presence of modeling violations is a key to the GFDR framework. For theoretical motivations of these assumptions see Appendix B. For numerical experiments confirming these assumptions see section 3.3. In our experience, we find that the "null" distribution is typically modeled by 1 or 2 Gaussians, while the "activation" and "deactivation" distributions are typically modeled by a single Gaussian. Thus for each type of category ("null", "activation" and "deactivation") we have:

$$\pi_k \mathrm{P}(z_i | y_i = k) = \sum_{j=1}^{J_k} \gamma_{kj} \varphi(z_i; \theta_{kj}) \tag{20}$$

where $k = 1, 2, 3$ and $i = 1, 2, \ldots, n$. The term on the left of equation (20) represents the contribution of the kth class to the full MOG fit in equation (2). We have still not specified exactly how the individual Gaussian components in the full MOG fit will be labelled as parts of "activation", "null" or "deactivation". We will address this issue in section 3.3.1. For the moment, assume that we have correctly labelled the individual subcomponents of MOG as "activation", "null" or "deactivation". From equations (20) and (2), the unknown probability density $p(z)$ can be approximated as:

$$p(z_i) = \sum_{k=1}^{3} \sum_{j=1}^{J_k} \gamma_{kj} \varphi(z_i; \theta_{kj}) \tag{21}$$

where $i = 1, 2, \ldots, n$. Let $\Omega = \{J_k, k = 1, 2, 3\}$ be the set containing the unknown number of Gaussians in each class and let $\Theta = \{\gamma_{kj}, \theta_{kj}, k = 1, 2, 3, j = 1, 2, \ldots, J_k\}$ be the set of unknown parameters. The log-likelihood of observing the data $D = \{z_i, i = 1, 2, \ldots, n\}$ is

$$\ell(D; \Omega, \Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1,2,3} \sum_{j=1}^{J_k} \gamma_{kj} \varphi(z_i; \theta_{kj}) \right) \tag{22}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

The first step in GFDR algorithm involves estimation of the associated parameters {Ω, Θ} using a joint maximization of model evidence and data likelihood. The goal of this step is **not** to identify J_k individually but to just identify the total number of Gaussians $(\sum_{k=1}^{3} J_k)$ in the overall MOG fit. Essentially, the procedure involves:

1. Postulate an overall MOG model with only 1 Gaussian.
2. Estimate the model parameters $\hat{\Theta}$ using Expectation Maximization (EM).
3. Measure the model evidence using Bayes Information Critetion (BIC).
4. Add another Gaussian and go to step 2.

The procedure ends when a model with lower BIC than the previous model is found (see Appendix).

Maximization in (22) can be accomplished via the Expectation-Maximization (EM) algorithm [Dempster et al., 1977] with automatic initialization using the *k*-means algorithm. We used the Bayes Information Criterion (BIC) because it is asymptotically consistent [Hastie et al., 2001]. fMRI data-sets typically have on the order of 10^5 voxels which is quite large to justify the choice of BIC over others. See [Lanterman, 2001] for a review on model order estimation.

With the final estimates of parameters at hand, one can proceed to develop equations for controlling the FDR. It will be ultimately important to jointly identify the "null" and "deactivation" distributions ($J_1 + J_2$). We will return to the issue of calculating $J_1 + J_2$ in the section 3.3.1. Suppose for now that we know $J_1 + J_2$. Then the subcomponent corresponding to "deactivation" from the full MOG fit is given by:

$$\pi_1 \mathrm{P}\left(z \middle| y=1\right) = \sum_{j=1}^{J_1} \widehat{\gamma}_{1j} \varphi(z; \widehat{\theta}_{1j}) \tag{23}$$

Similarly, the subcomponent for "null" distribution from the full MOG fit is:

$$\pi_2 \mathrm{P}\,(z|y=2) = \sum_{j=1}^{J_2} \widehat{\gamma}_{2j} \varphi(z; \widehat{\theta}_{2j}) \tag{24}$$

From (15), (23) and (24):

$$\mathrm{Fdr}(\alpha) = \frac{n\{\sum_{j=1}^{J_1} \widehat{\gamma}_{1j} \int_{\alpha}^{\infty} \varphi(z; \widehat{\theta}_{1j})\, dz + \sum_{j=1}^{J_2} \widehat{\gamma}_{2j} \int_{\alpha}^{\infty} \varphi(z; \widehat{\theta}_{2j})\, dz\}}{n(\alpha)} = \frac{nF(\alpha)}{n(\alpha)} \tag{25}$$

In terms of the *p*-value of a standard Gaussian distribution $N(0, 1)$, this can be rewritten as:

$$\mathrm{Fdr}(\alpha) = \frac{n\{\sum_{j=1}^{J_1} \widehat{\gamma}_{1j}\, p(\frac{\alpha - \widehat{\mu}_{1j}}{\widehat{\sigma}_{1j}}) + \sum_{j=1}^{J_2} \widehat{\gamma}_{2j}\, p(\frac{\alpha - \widehat{\mu}_{2j}}{\widehat{\sigma}_{2j}})\}}{n(\alpha)} \tag{26}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Controlling FDR at q means restricting:

$$\frac{n\{\sum_{j=1}^{J_1}\widehat{\gamma}_{1j}\,p(\frac{\alpha-\widehat{\mu}_{1j}}{\widehat{\sigma}_{1j}})+\sum_{j=1}^{J_2}\widehat{\gamma}_{2j}\,p(\frac{\alpha-\widehat{\mu}_{2j}}{\widehat{\sigma}_{2j}})\}}{n(\alpha)}\leq q \tag{27}$$

Define the GFDR threshold as:

$$\alpha_{\mathrm{GFDR}}=\inf_{\alpha}\left\{\alpha>0:F(\alpha)=\sum_{j=1}^{J_1}\widehat{\gamma}_{1j}\,p\left(\frac{\alpha-\widehat{\mu}_{1j}}{\widehat{\sigma}_{1j}}\right)+\sum_{j=1}^{J_2}\widehat{\gamma}_{2j}\,p\left(\frac{\alpha-\widehat{\mu}_{2j}}{\widehat{\sigma}_{2j}}\right)\leq\frac{qn(\alpha)}{n}\right\} \tag{28}$$

Solution of equation (28) can be found graphically by plotting α vs. $nF(\alpha) - qn(\alpha)$ and finding the smallest value of α for which the curve crosses the α-axis. If there is more than one such α, then it means that the desired false discovery rate is attained at multiple α values and in this case one would choose the smallest α that satisfies (28). An illustration of this procedure is shown in Figure 3. Note that (28) depends only on $J_1 + J_2$ since the summation can be written as $\sum_{j=1}^{J_1+J_2}$.

The GFDR algorithm has been coded as a toolbox in MATLAB (www.mathworks.com). The average running time for the algorithm on standard fMRI images is < 5 minutes. The standard FDR procedure on the other hand runs in only a few seconds since it does not have to estimate a model for the empirical "null".

3.3 Numerical validation of the MOG hypothesis under modeling violations

The goal of this section is to test the validity of the MOG hypothesis in the presence of multiple confounds under a locally variable random mixing process. We assume data generation from R classes such that class k occurs with prior probability π_k. Let X_k be the design matrix generating class k at minimum SNR $\delta^k_{\min}$ and maximum SNR $\delta^k_{\max}$. Suppose there are M confounds in the data such that confound s occurs with prior probability π_{Ws}. Let W_s be the design matrix generating confound s at minimum SNR $\rho^s_{\min}$ and maximum SNR $\rho^s_{\max}$. Suppose observed data at an example voxel is generated using true data from class c and confound data from class c_W.

$$y=X_c\,\delta^c+W_{c_W}\,\rho^{c_W}+\varepsilon \tag{29}$$

where $\varepsilon \sim N(0, I_p)$. Here c and c_W are random variables such that $P(c = k) = \pi_k$ and $P(c_W = s) = \pi_{W_S}$. We are also given an analyzing matrix Z and contrast c_Z. We generate n_p points of size R^p. Observed data at point i is generated as follows:

1. Randomly select class 1, class 2, …, class R as per prior probabilities $\pi_1, \pi_2, \pi_3, \ldots, \pi_R$.
2. If class k is selected, generate data $X_k\delta^k$ using design matrix X_k at SNR δ^k selected randomly from the uniform distribution $U(\delta^k_{\min}, \delta^k_{\max})$.
3. Randomly select a confound from $W_1, \ldots, W_M$ as per probabilities $\pi_{W_1}, \ldots, \pi_{W_M}$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4. If confound class s is selected, generate confound $W_s\rho^s$ using W_s at SNR ρ^s selected randomly from the uniform distribution $U(\rho^s_{\min}, \rho^s_{\max})$.
5. Generate noise ε from $N(0, I_p)$ of length p and add it true data and confound to get observed data $y = X_k \delta^k + W_s \rho^s + \varepsilon$.
6. Analyze y using the given design matrix Z and contrast c_Z and calculate a z-statistic.

Here we try to simulate data generated from a mixture of $R = 3$ classes in the presence of $M = 3$ confounds at various SNRs and examine if an MOG density adequately captures the classes. The data generating design matrix X and confounds W_s had $p = 140$ timepoints (shown in Figure 4). Confound W_1 was generated from a uniform distribution $U(0, 1)$, confound W_2 was generated from a gamma distribution $G\alpha(1, 1)$ and confound W_3 is a real fMRI signal extracted from a resting state network (RSN). Each confound W_s was demeaned and normalized to have the same energy as X. The dot products of these confounds with X are as follows: $W_1^T X = -3.63$, $W_2^T X = -0.32$ and $W_3^T X = 2.10$. The relative class fractions of these confounds were fixed as follows: $\pi_{W_1} = 0.2$, $\pi_{W_2} = 0.4$, $\pi_{W_3} = 0.4$. Each simulation consists of $n_p = 2000$ vectors generated as describe above for each combination of $\rho^s_{\max}$ and π_k. The class fraction for "activation" and "deactivation" distributions were taken to be equal in the data generation process i.e., $\pi_1 = \pi_3$. z-statistics were generated by analyzing the generated data pointwise using the design matrix $Z = X$ and contrast $c_Z = [1]$. Each simulation is repeated $n_s = 100$ times to estimate the statistics of quantities of interest. The parameters for the simulation 3.3 are shown in Tables 1 and 2.

We first examine the accuracy of determining the number of classes using BIC as we propose in this paper. Results are shown in Figures 5 - 8. Next, we examine the true positive rates for Bayes aposteriori (maximum posterior probability based) classification using the estimated MOG fit for each class. Results are shown in Figure 9 - 11. Finally we show distribution fits for some cases when BIC identifies a 4 component fit to the data. We show that in these cases the "null" is actually well-described by a 2-component MOG density. Using this 2-component MOG density for "null" we get an average true positive rate for all classes > 96%. (see Figures 17 - 14).

3.3.1 Identification of class distributions (Estimation of $J_1 + J_2$)—First we review some typical cases that arise when using MOG density modeling.

1. Figures 16(a) - 16(d) show some typical MOG fits obtained under various confound corruption conditions using data from section 3.3. Figure 16(a) shows 3-component MOG fit with a centered "null" component (primary corruption using W_2). Figure 16 (b) shows a 3-component MOG fit with a right shifted "null" component (primary corruption using W_3). Figure 16(c) shows a 3-component MOG fit with a left shifted "null" component (primary corruption using W_1). Figure 16(d) shows a "split" null component where the "null" itself is described by a 2-component MOG fit.
2. Figure 17(a) shows a 2-component MOG fit where the "activation" and "deactivation" are jointly captured by a single Gaussian distribution. In such cases, it is reasonable to force BIC to identify at least 3 classes. With this constraint, we get a 3-component best fit for the same data set. This is shown in 17(b).

The "null" density is modeled in most cases as a 1 or 2 component MOG density. This can be identified easily in most cases by looking for 1 or 2 "large volume" components near $z = 0$. The "activation" and "deactivation" components are usually much smaller in size and are found to be centered near the "tails" (positive "tail" for "activation" and negative "tail" for "deactivation"). Note that these comments are intended as "heuristic" rules that should work in many but not all cases.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

We also describe here another strategy that is also "heuristic" in nature but more quantitative. Calculation in (28) requires the identification of the activation and non-activation distributions. Suppose C is the set of classes identified by EM. For each voxel v_i, the posterior probability of membership in each class c is defined as:

$$P(v_i \in c|z_i)=\frac{\pi_c P(z_i|v_i \in c)}{\sum_{l\in C}\pi_l P(z_i|v_i \in l)} \tag{30}$$

where π_l are the estimated class fractions. Suppose t_i be the raw demeaned time courses associated with voxel v_i. We define a probability weighted time course for class c as follows:

$$u_c=\sum_{i=1}^{n} P(v_i \in c|z_i)\, \mathcal{I}\,(P(v_i \in c|z_i)>0.5)\, t_i \tag{31}$$

If e is the explanatory variable used in generating the current z-stat map, we calculate the correlation coefficient:

$$\rho_c=\frac{(u_c-\bar{u}_c)^T (e-\bar{e})}{\|u_c-\bar{u}_c\|\|e-\bar{e}\|} \tag{32}$$

where $\bar{u}_c$ and $\bar{e}$ are the means of u_c and e respectively.

We assess the distribution of the correlation coefficient by bootstrapping pairs [Davison and Hinkley, 1997] from the demeaned vectors u_c and e and calculate a confidence interval $\{L, U\}$ for the true value of ρ_c. We suggest classifying a class c as activation if $L > 0$, i.e. whenever there is a statistically positive correlation. Similarly, a class c is classified as deactivation if $U < 0$, i.e. whenever there is a statistically significant negative correlation. For the purposes of calculating the GFDR threshold, J_1 and J_2 can be chosen as any values that satisfy $J_1 + J_2 = J_T - J_3$ where J_T is the total number of classes found using BIC. Figures 30, 31, 32 and 33 (shown in the appendix) demonstrate the application of bootstrapping to a real audio visual fMRI dataset. For the visual stimulus, class 2 was identified as "activation" (bootstrap CI approx. [0.64, 0.83]) and class 1 was identified as "null" (bootstrap CI approx. [-0.23, 0.2]). For the auditory stimulus, we identified classes 3 and 4 as "activation" (bootstrap CIs approx. [0.58, 0.83], [0.77, 0.93] respectively), class 2 as "null" (bootstrap CI approx. [-0.25, 0.25]) and class 1 as "deactivation" (bootstrap CI approx. [-0.7, -0.45]).

For the simulation study, we chose $J_1 = 0$, $J_2 = 1$ and $J_3 = 0$ or $J_3 = 1$ depending on whether we found a mixture model with 1 component or 2 components.

3.3.2 Precautions to be taken when using GFDR—In this section, we briefly want to review the precautions that should be taken when using GFDR.

1. **Validity of the "null" model:** Modeling of empirical "null" via a MOG density relies on the assumption that the stationary noise in fMRI data is Gaussian. If this assumption is invalid then the MOG hypothesis may not be valid. We want to emphasize that an attempt should be made to derive the correct form of parametric "null" based on the analysis under consideration. For example, if some analysis proceeds via logistic regression, then one should attempt to derive how modeling violations affect the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

distribution of the "null" hypothesis. If such a derivation is possible, then the correct parametric form of "null" under these violations should be used for empirical estimation. One cannot expect one form of "null" to be valid in all analyses.

2. **Gross misspecification of design matrix:** Under the section titled "Identification of class distributions" we suggested a bootstrapping approach based on the correlation coefficients of the EV w.r.t a probability weighted time course. This strategy might fail if the magnitude of modeling violations is so large as to disable our ability to detect positive or negative correlations with the EV. These would typically be the scenarios where the model is grossly misspecified. In these cases, one should consider modifying the design matrix appropriately.

3. **Being conservative:** The number of voxels in fMRI data n should be large enough for the estimated model $\hat{\Omega}$ to be close to the true model using BIC. This means that the data should contain voxels from all three categories, "activation", "deactivation" and "null". In fact, it was seen in numerical experiments described in section 3.3 that when the fraction of "activation" and "deactivation" is ≤ 0.02 and when the modeling violations are "large" ($\rho^{s}_{\max} \geq 4.5$), BIC fails to identify the correct number of classes. However, this does not necessarily mean that GFDR will perform poorly. As shown in 3.4.1 when the activation fraction was small 0.01, BIC often identified only 1 distribution. In this case we took this distribution to be the "null" distribution and in these cases GFDR produced results very similar to FDR (see Figure 22). GFDR inference is conservative so long as the "null" is chosen in a conservative fashion i.e., any "class" that cannot be positively identified as "activation" should be included in the MOG for "null" hypothesis.

3.4 Performance Evaluation

3.4.1 Simulation Study—The goal of this simulation study was two fold:

1. Assess the bias of GFDR compared to FDR under varying values of q, degrees of modeling violations and activation fraction.

2. Assess the robustness of GFDR based inference under varying SNRs.

To this end, we generated 4-D artificial data according to the General Linear Model (GLM) as follows:

$$y = x\beta + w\eta + \varepsilon \tag{33}$$

where $y \in \mathbf{R}^p$, $x \in \mathbf{R}^p$, $w \in \mathbf{R}^p$ and $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_p)$. In the above equation, y is the observed artificial data, x is a randomly chosen design vector with elements $U(0, 1)$, w is a randomly chosen unmodeled signal vector with elements $U(0, 1)$ and η measures the strength of unmodeled signal. Data was generated on a $64 \times 41 \times 64$ image consisting of nearly 36854 non-zero voxels. The length of simulated timecourses was $p = 74$. For each voxel (i, j, k), we chose $\beta = 1$ for active voxels and $\beta = 0$ for inactive voxels such that the overall fraction of active voxels was f (user specified). Since $\beta = 1$ for "activation" class, SNR becomes $\frac{\beta}{\sigma} = \frac{1}{\sigma}$. We chose $\sigma = 1$ to get an approximate SNR of 1. The artificial data was analyzed using only the design vector x, **ignoring** the unmodeled signal $w\eta$.

$$y = x\beta + \varepsilon \tag{34}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Artificial datasets were generated by varying the intensity of modeling violations η in the set $\{0, \pm 0.05, \pm 0.1\}$ and the fraction of true activation f in the set $\{0.1, 0.2\}$. The resulting z-statistic images were thresholded using both FDR and GFDR for multiple values of q. For GFDR calculation, the number of Gaussian distributions in the mixture model selected by BIC was either 1 or 2. In the case when only one distribution was fit we took $J_1 = 0$, $J_2 = 1$ and $J_3 = 0$. In the case when two distributions were fit we took the large component near the origin as the "non-activated" distribution and took $J_1 = 0$, $J_2 = 1$ and $J_3 = 1$. Since the true activation was known in each dataset, we were able to calculate the true false discovery rate attained at thresholds chosen by FDR and GFDR. Ideally, this value should be close to q on the average.

We also investigated the bias and true positive rate (TPR) of GFDR and FDR under different SNRs and various combinations of f and η. We varied SNR ($1/\sigma$) in the set {0.9, 0.95, 1, 1.05, 1.1, 1.15, 1.2}, the true activation fraction f in the set {0.01, 0.05, 0.1, 0.15} and the intensity of modeling violation η in the set {0, ±0.05, ±, 0.1}. For each combination of σ, f and η, we generated 100 artificial datasets and for each dataset we calculated the FDR and GFDR based thresholds for a range of q values. For each q value, we define the mean absolute bias for each technique (GFDR and FDR) as:

$$B(\sigma, f, \eta, q) = \frac{1}{m}\sum_{i=1}^{m} |\widehat{q}(\sigma, f, \eta, q, i) - q)| \tag{35}$$

Here $\hat{q}(\sigma, f, \eta, q, i)$ is the actual value of FDR attained at the threshold calculated by one technique (GFDR or FDR) for given values of σ, f, η and q at iteration i and m is the number of simulations. Note the definition of $B(\sigma, f, \eta, q)$ has small values for a tight control of the true FDR near the chosen q value and large values otherwise. Since for each artificial dataset, we know which voxels are truly active, we can also calculate a mean TPR (T) for each technique as a function of f, η, σ and q as follows:

$$T(\sigma, f, \eta, q) = \frac{1}{m}\sum_{i=1}^{m} \frac{D_i}{A_i} \tag{36}$$

where D_i is the number of voxels declared active by one technique (GFDR or FDR) for given values of σ, f, η and q at iteration i and A_i is the number of truly active voxels at iteration i. As before m is the number of simulations. Ideally we would like a high TPR T at a low bias B.

3.4.2 fMRI Data—To demonstrate how GFDR performs on a real dataset, we used the "FSL Evaluation and Example Data Suite" (FEEDS) from FMRIB Image Analysis Group, Oxford University. The URL for this data suite is:
http://www.fmrib.ox.ac.uk/fsl/feeds/doc/index.html

One of the datasets in the example suite contains an audio visual experiment with two explanatory variables, the visual stimulus (30s off, 30s on) and an auditory stimulus (45s off, 45s on). Analysis was carried out using FEAT (FMRI Expert Analysis Tool) Version 5.4, part of FSL (FMRIB's Software Library). www.fmrib.ox.ac.uk/fsl

The following pre-statistics processing was applied; motion correction using MCFLIRT [Jenkinson 2002]; non-brain removal using BET [Smith 2002]; spatial smoothing using a Gaussian kernel of FWHM 5mm; mean-based intensity normalisation of all volumes by the same factor; highpass temporal filtering (Gaussian-weighted LSF straight line fitting, with

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

sigma=50.0s). Time-series statistical analysis was carried out using FILM with local autocorrelation correction [Woolrich 2001]. To study the effect on results produced by GFDR and FDR for real fMRI data under modeling violations, we introduced an unmodeled signal (see Figure 27) with maximum amplitude of 1% signal change into the fMRI data and performed a second statistical analysis without any autocorrelation correction.

4 Results

4.1 Simulation Study

1. Figure 18 depicts a comparison between FDR and GFDR in which no unmodeled data has been included at SNR = 1, only relative fraction of activation is changed. The true expected bound lies on the line with slope 1. FDR curves approached well the true bound for small values of q and deviated from it the larger the q value and the larger the fraction of activation. GFDR, on the other hand, maintained good control staying very close to the true bound for large values of q for both small and large activation.

2. In Figure 19, unmodeled data was included at SNR = 1. FDR departed significantly from the true bound by becoming overly conservative or liberal depending on the unmodeled data (positive or negative). The departure was significant regardless of the value of q, becoming larger with larger q values. GFDR, however, maintained good control and resulted in q values very close to the true ones even for large values of q.

3. Figures 20 and 21 show the combined effect of small and large activation in combination with positive and negative unmodeled signals at SNR = 1. In both cases, FDR was unable to control false discoveries satisfactorily. GFDR was able to achieve good control even under this joint violation.

4. Figure 22 shows the mean absolute bias (B) attained by GFDR and FDR under varying SNR's for the fraction of activation $f = 0.01$ at $q = 0.1$. It can be seen that GFDR and FDR produce identical results when there are no modeling violations $\eta = 0$. The mean absolute bias of FDR increases dramatically with SNR for all values of $\eta \neq 0$ (going as high as $B = 0.3$) whereas GFDR is able to maintain almost the same mean absolute bias ($B < 0.03$) as in the case of $\eta = 0$. As expected, the bias B decreases with increasing SNR for GFDR. For FDR we do not see a similar trend, the bias B even increasing in some cases with increasing SNR.

   Figure 23 shows the mean true positive rate (T) attained by GFDR and FDR under varying SNR's for the fraction of activation $f = 0.01$. The mean true positive rate T produced by FDR varies a lot for the same SNR under varying η. In comparison, GFDR produces a much smaller variation in T for the same SNR under varying η. The smallest mean TPR at SNR 1 produced by FDR was close to 0.9 while that produced by GFDR was close to 0.95. Thus GFDR was able to attain a small mean absolute bias and a high mean true positive rate under all SNR's and η for $f = 0.01$.

5. Figure 24 shows the mean absolute bias of GFDR and FDR under varying SNR's for no modeling violations $\eta = 0$ and various fractions of activation f and $q = 0.1$. It should be noted that both GFDR and FDR produce a relative low mean absolute bias ($B < 0.03$). Even so, it can be seen that as f increases, the mean absolute bias B of FDR increases whereas GFDR produces a much smaller B for all but one case (SNR = 0.9, $f = 0.01$, $B = 0.03$). Figure 25 shows the mean true positive rate T for $\eta = 0$ and varying f for both techniques. Both FDR and GFDR produce almost identical mean true positive rate ($T > 0.95$) at all f and SNR values.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4.2 fMRI Data

1. Figure 28 shows the results produced by GFDR and FDR on FSL Feeds Data for the visual and auditory stimulus before and after the introduction of 1% unmodeled signal. We used a *q*-value of 0.1 for this comparison. 28(a) shows the voxels declared active by GFDR and FDR for the visual stimulus. As can be seen there is good agreement between the two techniques. 28(b) shows the voxels declared active by GFDR and FDR for the visual stimulus under 1% unmodeled signal. GFDR produces results almost identical to 28(a) whereas FDR declares many more voxels as active.

   Figure 28(c) shows the voxels declared active by GFDR and FDR for the auditory stimulus. Again, there is good agreement between the two techniques. 28(d) shows the voxels declared active by GFDR and FDR under modeling violations. Again GFDR produces almost identical results to 28(d). In this case, FDR also produces almost identical results to 28(c). This can be understood by looking at the GMM fits for the *z*-stat distributions with and without modeling violations.

2. Figures 29(a) and 29(b) show the distribution of *z*-stats for visual stimulus and the corresponding GMM fits. It can be seen that introduction of modeling violations has shifted the distribution away from 0. This shift causes FDR to declare many more voxels as active as compared to the case without modeling violations. GFDR on the other hand is able to produce identical results even under these modeling violations.

   Figures 29(c) and 29(d) show the distribution of *z*-statistic for the auditory stimulus. There is almost no difference between the two distributions and hence there is also no difference in the GMM fits and the inference produced by GFDR and FDR as compared to the case without modeling violations.

5 Discussion and Conclusions

In this paper, we investigated whether the performance of FDR can be improved in the presence of modeling violations. Our results indicate that accounting for the empirical "null" and its fraction in fMRI via a MOG density using the GFDR algorithm leads to much better performance compared to traditional FDR that uses a fixed "null" of $N(0, 1)$. We have shown how the assumptions made by conventional FDR (section 3.1.2) are violated in a real fMRI data analysis. Presence of modeling violations can occur due to a number of reasons (see Figure 15) including (1) using the same design matrix to perform a GLM analysis of all brain voxels (2) physics related effects or (3) biology related effects. When modeling violations are present the assumption of a fixed null $N(0, 1)$ is not valid. Hence any method that performs inference under this assumption may produce incorrect results. It is easy to show mathematically (Appendix B) as well as via numerical simulation (section 3.3) that the density of "null" is well described by an MOG density. We find that the MOG density describes all three classes quite accurately (see Figure 5 - 11). This is the key to the improved performance of GFDR (Figures 18 - 25 and 28 - 29).

GFDR is a procedure for controlling false discoveries in an empirical Bayesian framework [Efron et al., 2001] specially adapted to fMRI thresholding. It is a generalization of previously suggested corrections to FDR [Allison et al., 2002; Storey, 2003] where we not only estimate the fraction of the "null" distribution but also the "null" distribution itself adaptively from the data. The problem of estimating $\mathbf{H_0}$ is at the heart of empirical Bayesian approaches based on the work of Efron et al. [2001]. Ours is not the first approach to suggest accounting for an empirical "null". For example, Efron [2004] discuss a drug mutation study that was analyzed using logistic regression and where the *z*-values were computed using maximum likelihood estimates for the logistic coefficients and large sample estimates of their standard error. Here they propose to estimate the "null" distribution as a Gaussian distribution fitted to the peak of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the histogram of *z*-values. This approach assumes that the "null" can be well described by a single Gaussian distribution. Morever, this approach requires specification of additional parameters by the user (see section 3.1.3 for details) for the determination of the variance and fraction of the "null" distribution. As we have shown here the assumption of a single Gaussian distribution for "null" is not always true (see Figures 12(a) - 13(d)), especially for fMRI data. In addition, the form of parametric "null" prescribed in GFDR is derived theoretically from the underlying GLM model in the presence of modeling violations for fMRI. This is also verified via a large simulation study as shown in section 3.3. GFDR also does not require any parameter setting by the user.

GFDR fits the histogram of *z*-values using a MOG model with the number of components in the MOG estimated using the BIC criterion. While the BIC criterion often estimates a correct model, it might underestimate the number of classes in the data when the "activation" fraction is small. In these cases, GFDR becomes slightly conservative. We also found in our simulations situations where BIC identifies a 2 component MOG fit to the data, with one component modeling the "null" and the other component jointly capturing both the "activation" and "deactivation" (see Figure 17(a) - 17(b)). In these cases, it is necessary to force separate identification of "activation" and "deactivation" by fitting at least a 3 component Gaussian to the data via BIC. In Figure 16(a) - 16(d) we show example MOG fits to the data where identification of "null" and "activation" distributions is relatively easy. However, this is not always the case. In situations where it is questionable as to whether certain components should be considered "activation" or "null", one should be conservative and consider them to make up the "null" class. We described a correlation based heuristic approach to classify components as members of "activation" or "null" class automatically. This approach is also a heuristic and might fail in cases of large modeling violations. Again, we recommend being conservative in the "definition" of "null" making it as large as possible without any ambiguities.

The artificial data results indicate that conventional FDR produces good results when model violations are not strong, however, once activation fraction becomes relatively large or unmodeled data is not accounted for, severe deviations from true *q* values are obtained. Correcting approaches include revising and modifying EV's. In some instances this is possible, for example by adding motion parameteres as covariates of no interest, in others, the source is not known and hence a proper EV is not possible to be generated. One can attempt to compute sources using a technique such as Independent Component Analysis (ICA) [Beckmann and Smith, 2004]. In our experience, these techniques produce a very large number (typically 30 - 100) of independent components, making it difficult to pick out a meaningful source component. The exact origin of these source signals is difficult to quantify but could be related to physiological noise, motion artifacts, scanner noise as well as local variability of the Haemodynamic Response Function (HRF) and drift. These un-modeled signals could have a significant impact on standard methods of inference.

For instance, cluster-based methods such as Gaussian random field (GRF) theory approaches (Worsley et al. [1996]) assume the *z*-stat image as a realization of a smooth $N(0, 1)$ Gaussian random field. Inference proceeds by first selecting a cluster forming threshold (e.g., $z > 2.3$) followed by testing for the size of the resulting blobs. As noted before, the assumption of $N(0, 1)$ smooth Gaussian random field is a strong one may not be valid in the presence of modeling violations which is typical given the nature of fMRI analysis. Other approaches such as permutation tests (Nichols [2002]) do not make any assumptions about the "null" distribution but instead make strong assumptions about the "exchangeability" of the data. This is again a strong assumption that may not be valid in the presence of **unknown** modeling violations. Exchangeability is also violated in the case of correlated data (e.g., temporal correlation in timeseries). GFDR inference, on the other hand has the desirable properties of enabling estimation of "null" from the data as well as providing a sharp control of the true FDR at a user

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

specified level ("equivalence theorem" of Efron et al. [2001]) as long as the empirical "null" is estimated correctly. In cases when the estimated "null" is larger than the true "null", GFDR inference becomes conservative.

The MOG fit from GFDR can also be used to perform alternative hypothesis testing via the identified "activation" distributions. Posterior probabilities of membership in the joint MOG "activation" class can be computed followed by a user selected "cutoff" (e.g., $p > 0.5$) to identify activation. A potential drawback of this approach is that when the "activation" class fraction is small, the "activation" class might not be identified as a separate component in the MOG fit resulting in overly conservative inference.

To summarize, we postulated and validated the MOG density "null" model for real fMRI data. The GFDR algorithm achieves enhanced control of FDR by adaptively accounting for the form and fraction of the "null" from the data. In conclusion, GFDR is a useful technique capable of handling model violations and producing robust results. Its use could significantly improve practical fMRI studies involving massively univariate GLM analyses in which perfect voxelwise modeling is not possible.

Acknowledgments

This work was supported in part by an unrestricted Grant from Merck and Co. and in part by National Institutes of Health (NIH) under Grant NS042721.

Appendix

Appendix: A fMRI Modeling using the GLM

Consider an FMRI dataset with vectors of p dimensional signals y_i obtained over the set of voxels $i = 1, 2, \ldots, n$ spanning the entire brain. Suppose that the experimental paradigm and associated covariates of interest are modeled via a p by q design matrix X. Each column of X is either an explanatory variable (EV) or a covariate of no interest used to model known non-EV related signal in the brain. Assuming Gaussian noise, one can hypothesize the generation of data at voxels $i = 1, 2, \ldots, n$ via a General Linear Model (GLM) [Friston et al., 1995] as follows:

$$y_i = X\beta_i + \varepsilon_i \text{ where } y_i \in R^p, X \in R^{p\times q}, \beta_i \in R^q \tag{A-1}$$

The random term in the above model ε_i is assumed to be have a Gaussian distribution:

$$\varepsilon_i \sim N(0, \sigma_i^2 V_i) \tag{A-2}$$

Each component of β_i measures the strength of the corresponding column of X in the measured response y_i. Assuming V_i is estimated or modified using some strategy like prewhitening or coloring for each voxel i, equation (A-1) can be solved to yield:

$$\widehat{\beta_i} = (X^T V_i^{-1} X)^{-1} X^T V_i^{-1} y_i \tag{A-3}$$

The variance σ_i^2 is estimated unbiasedly as follows:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\widehat{\sigma}_i^2=(y_i - X\widehat{\beta}_i)^T V_i^{-1}(y_i - X\widehat{\beta}_i)/(p-q) \tag{A-4}$$

One is normally interested in testing if a particular contrast of these regression parameters β_i are zero. Let $c \in \mathbf{R}^q$ be a contrast vector. Suppose the hypothesis test of interest is:

$$\mathrm{H}_0: c^T \beta_i = 0, i=1,2,\ldots,n \tag{A-5}$$

One can perform, for example, the following *t*-test at each voxel *i* to assess the hypothesis H_0:

$$\text{Under } \mathrm{H}_0: \frac{c^T \widehat{\beta}_i}{\widehat{\sigma}_i \sqrt{c^T (X^T V_i^{-1} X)^{-1} c}} \sim t_{p-q} \tag{A-6}$$

Since this test is carried out at each voxel *i*, one has an image of the *T* test statistic. These *T* values can be converted into equivalent *Z* values corresponding to the standard Gaussian distribution by using the following *T* to *Z* transformation function:

$$g(T_i)=z_i \text{ if } t_i, z_i \text{ satisfy } \mathrm{P}(T_{p-q}>t_i)=\mathrm{P}(N(0,1)>z_i), i=1,2,\ldots,n \tag{A-7}$$

In the resulting *Z* statistic image, the null hypothesis $\mathbf{H}_0$ can be rejected and voxel *i* can be declared as active if:

$$z_i \geq z_h \tag{A-8}$$

Similarly, voxel *i* can be declared as deactive if:

$$z_i \leq z_l \tag{A-9}$$

Here, z_h and z_l are the upper and lower *Z* thresholds. The difficulty with choosing thresholds z_h and z_l to detect activation and deactivation arises because of the problem of multiple comparisons. Since the same tests (A-8) and (A-9) are carried out at all *n* voxels, and *n* is typically very large, there is a significant amount of false “activation” or “deactivation” purely by chance (Type I error). Many schemes exist to correct for multiple comparisons, such as Bonferroni correction [Simes, 1986;Shaffer, 1995], Gaussian Random Field (GRF) theory [Worsley et al., 1996] and False Discovery Rate (FDR) [Benjamini and Yekutieli, 2001]. See Nichols and Hayasaka [2003] for a comparative review.

Appendix: B Effect of Modeling Violations

Please refer to Appendix A for an introduction to basic GLM based fMRI modeling. This section discusses the effect of modeling violations on the statistical maps produced by a GLM

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

with Gaussian noise. It should be noted that model (A-1) assumes that we use the same design matrix for each voxel i and that the noise is stationary at all voxels. The signal at any voxel i can be thought of as a combination of stimulus related signal ($X\beta_i$), structured noise ($w\rho$) and stationary noise ε_i. The origins of structured noise for FMRI are not fully understood. To better understand the effect of structured noise on FMRI analysis, consider that the true model describing data y_i is given by:

$$y_i = X\beta_i + w\rho + \varepsilon_i \tag{A-10}$$

where y_i, $w \in \mathbf{R}^p$, $X \in \mathbf{R}^{p\times q}$, $\beta_i \in \mathbf{R}^q$, $\rho \in \mathbf{R}$ and ε_i is distributed according to (A-2). In the equations below, it is assumed that the expectations, variances and distributions are conditional given w and ρ.

If (A-1) is the assumed model the $\hat{\beta}_i$ has a Gaussian distribution with mean:

$$\mathrm{E}(\widehat{\beta}_i) = \beta_i + (X^T V_i^{-1} X)^{-1} X^T V_i^{-1} w\rho \tag{A-11}$$

The estimate is no longer unbiased. Define b_i to be the bias of β_i. Thus

$$b_i = (X^T V_i^{-1} X)^{-1} X^T V_i^{-1} w\rho \tag{A-12}$$

The variance of $\hat{\beta}_i$ remains unchanged at:

$$\mathrm{Var}(\widehat{\beta}) = (X^T V_i^{-1} X)^{-1} \sigma_i^2 \tag{A-13}$$

Also $\dfrac{(p-q)\widehat{\sigma}_i^2}{\sigma_i^2}$ has a non-central χ^2 distribution with $p - q$ degrees of freedom:

$$\frac{(p-q)\widehat{\sigma}_i^2}{\sigma_i^2} \sim \chi^2(p-q, \mu_i) \tag{A-14}$$

and with non-centrality parameter

$$\mu_i = \frac{\rho^2}{\sigma_i^2}\{w^T V_i^{-1} w - w^T V_i^{-1} X (X^T V_i^{-1} X)^{-1} X^T V_i^{-1} w\} \tag{A-15}$$

The estimate $\widehat{\sigma}_i^2$ is also biased:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\mathrm{E}\left[\frac{\widehat{\sigma}_i^2}{\sigma_i^2}\right]=1+\frac{\mu_i}{p-q} \tag{A-16}$$

However, if $p-q$ is large then the estimate is approximately unbiased. One can also show that $\hat{\beta}_i$ and $\hat{\sigma}_i$ are independent random variables. With these facts in place, it is easy to see that:

$$\text{Under } \mathrm{H}_0: \frac{c^T\widehat{\beta}_i}{\widehat{\sigma}_i\sqrt{c^T\left(X^T V_i^{-1} X\right)^{-1}c}} \sim \frac{\mathrm{N}\left(\frac{c^T b_i}{\sigma_i\sqrt{c^T\left(X^T V_i^{-1} X\right)^{-1}c}},1\right)}{\sqrt{\frac{\chi^2(p-q,\mu_i)}{p-q}}} \tag{A-17}$$

If the degrees of freedom $p-q$ is sufficiently large then from Central Limit Theorem, one can show that:

$$\frac{\chi^2(p-q,\mu_i)}{p-q} \sim \mathrm{N}\left(\frac{\mu_i}{p-q}+1, \frac{4\mu_i}{(p-q)^2}+\frac{2}{p-q})\right) \tag{A-18}$$

For large $p-q$ this distribution is sharply peaked at its mean value. Thus when $p \gg q$, the following holds true:

$$\text{Under } \mathrm{H}_0: \frac{c^T\widehat{\beta}_i}{\widehat{\sigma}_i\sqrt{c^T\left(X^T V_i^{-1} X\right)^{-1}c}} \sim \mathrm{N}\left(\frac{c^T b_i f_i}{\sigma_i\sqrt{c^T\left(X^T V_i^{-1} X\right)^{-1}c}}, f_i^2\right) \tag{A-19}$$

where

$$f_i=\sqrt{\frac{p-q}{\mu_i+p-q}} \tag{A-20}$$

The mean of this null distribution is $m_i=\dfrac{c^T b_i f_i}{\sigma\sqrt{c^T\left(X^T V_i^{-1} X\right)^{-1}c}}$ and variance $s_i=f_i^2$. Equation (A-19) proves that incorrect modeling assumptions or unmodeled signals result in a shifted and scaled Gaussian null distribution. Locally variable effects due to signal inhomogeneities, variance in vascular flow and/or BOLD response (onset, strength, duration, extent), background effects due to coherent activity similar to resting state networks (Luca et al. [2005]) etc. in general result in a mixture of Gaussians (MOG) density for the “null” distribution. Similarly, the distribution of “activation” and “deactivation” can also be argued to be a scaled and shifted Gaussian densities. The histogram of *z*-values can thus be modeled

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

as a MOG density with 1 or more subcomponents representing “activation”, “deactivation” and “null” distributions. See Figure 15 for a summary of these ideas.

Appendix: C Identifiability of the Mixture Model

The parameter set Θ in the mixture model is estimated via the maximization of $\ell(D; \Omega, \Theta)$, the likelihood of observing the data D for a given model Ω. If Ω_1 and Ω_2 are two nested models with the number of unknown parameters Θ_2 in Ω_2 more than Θ_1 those in Ω_1 then it is easy to show using the estimated parameters that:

$$\ell(D;\Omega_2, \widehat{\Theta}_2) \geq \ell(D;\Omega_1, \widehat{\Theta}_1) \tag{A-21}$$

i.e, the maximized likelihood of observing the data in the larger model is at least as much as that in the smaller model. Since our model identification depends on maximizing the likelihood, how can we be sure that our estimate is close to the true model Ω_0?. This is where Bayesian model selection comes in. Model evidence for a particular model Ω is the logarithm of the probability of jointly observing the model Ω and the data D. Ideally, one would want to choose a model with the largest model evidence:

$$\widehat{\Omega} = \arg\ \max_{\Omega} \log P(\Omega, D) \tag{A-22}$$

Of the several approximations to the model evidence [Lanterman, 2001], we chose the Bayes Information Criterion (BIC) which has the following definition:

$$\log P(\Omega, D) = \ell(D;\Omega, \widehat{\Theta}) - \frac{\left|\widehat{\Theta}\right|}{2} \log n \tag{A-23}$$

where $\hat{\Theta}$ is the maximizer of $\ell(D; \Omega, \hat{\Theta})$ and n is the size of data D. The BIC criterion has the following property:

$$\widehat{\Omega} \rightarrow \Omega_0 \text{ as } n \rightarrow \infty \tag{A-24}$$

- It was shown in the section titled “Effect of Modeling Violations” that the null distribution can be represented as a mixture of scaled and shifted Gaussian distributions. Thus the true model Ω_0 is well defined.
- It is well known that optimization algorithms converge to different local solutions depending on the initialization. The maximization that we use is based on the Expectation Maximization (EM) algorithm [Dempster et al., 1977] which guarantees convergence to a local solution. We initialize EM using the *k*-means algorithm [MacQueen, 1967]. In addition, we use the apriori knowledge about the true model, namely that activations have the highest *z*-stat values and deactivations have the lowest *z*-stat values. The initial centers for *k*-means are chosen uniformly distributed between (z_{min}, z_{max}), so that the center with highest *z*-value corresponds to activation and that with the lowest *z*-value corresponds to deactivation.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

fMRI data sets have on the order of $n \sim 10^5$ voxels. Combining this with the fact that Ω_0 is well defined, maximization of likelihood is well initialized and convergence to true model guaranteed by BIC (A-24), we can be confident that the estimated model is $\hat{\Omega}$ is identified correctly. Figure 26 shows the application of this logic to a real audio-visual fMRI dataset (described later). The optimal number of mixtures was identified as 4 at which the BIC attained a maximum.

Appendix: D Description of EM/BIC procedure used in GFDR

1. Initialize, $\Omega^0 \leftarrow \{J_k, k=1,2,3: \sum_{k=1}^{3} J_k=1\}, m \leftarrow 0$.
2. At step m, estimate the full parameter set Θ^m by maximizing the likelihood of observing the data D:

$$\widehat{\Theta}^m=\text{argmax}_{\Theta}\ell(D;\Omega^m,\Theta) \tag{A-25}$$

3. Estimate model evidence log $\mathbf{P}(\Omega^m, D)$, i.e., the joint probability of observing the model Ω^m and the data D using the Bayes Information Criterion (BIC) [Lanterman, 2001]:

$$\log P(\Omega^m, D)=\ell(D;\Omega^m,\widehat{\Theta}^m) - \frac{|\widehat{\Theta}^m|}{2}\log n \tag{A-26}$$

where $|\hat{\Theta}^m|$ is the number of unknown parameters in the model Ω^m:

$$|\widehat{\Theta}^m|=\text{Card}(\widehat{\Theta}^m) \tag{A-27}$$

4. For $m \geq 1$, if log $\mathbf{P}(\Omega^m, D)$ > log $\mathbf{P}(\Omega^{m-1}, D)$, update

$$\Omega^{m+1} \leftarrow \{J_k, k=1,2,3: \sum_{k=1}^{3} J_k \leftarrow \sum_{k=1}^{3} J_k+1\} \tag{A-28}$$

$$m \leftarrow m+1 \tag{A-29}$$

and go to step 2. Otherwise estimate:

$$\{\widehat{\Omega}, \widehat{\Theta}\}=\{\widehat{\Omega}^m, \widehat{\Theta}^m\} \tag{A-30}$$

Appendix: E Bootstrap figures for fMRI data set

References

Allison DB, Gadbury GL, Heo M, Fernandez JR, Lee CK, Prolla TA, Weindruch R. A mixture model approach for the analysis of microarray gene expression data. Computational Statistics and Data Analysis 2002;39:1–20.

Beckmann CF, Smith SM. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans on Medical Imaging 2004;23(2):137–152.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society 1995;57(1):289–300.

Benjamini Y, Yekutieli D. The control of the False Discovery Rate in multiple testing under dependency. The Annals of Statistics 2001;29(4):1165–1188.

Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 2006;93(3):491–507.

Davison, AC.; Hinkley, DV. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press; 1997. Bootstrap Methods and Their Application.

Dempster AP, Larid NM, Bubin DB. Maximum Likelihood From Incomplete Data via EM Algorithm. Journal of Royal Statistical Society, Series B 1977;39:1–38.

Efron B. Robbins, Empirical Bayes and Microarrays. The Annals of Statistics 2003;31(2):366–378.

Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association 2004;99:96–104.

Efron B. Correlation and Large-Scale Simultaneous Significance Testing. Journal of the American Statistical Association 2006;102(477):93–103.

Efron B, Storey JD, Tibshirani R. Microarrays, Empirical Bayes Methods, and False Discovery Rates. Stanford Technical Report. 2001

Everitt B, Bullmore E. Mixture model mapping of brain activation in functional magnetic resonance images. Human Brain Mapping 1999;7:1–14. [PubMed: 9882086]

Friston K, Holmes A, Worsley K, Poline JB, Frith C, Frackowiak R. Statistical parametric maps in functional imaging: A general linear approach. Human Brain Mapping 1995;2:189–210.

Genovese CR, Lazar NA, Nichols T. Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. NeuroImage 2002;15:870–878. [PubMed: 11906227]

Hastie, T.; Tibshirani, R.; Friedman, JH. Springer Series in Statistics. Springer; 2001. The elements of Statistical Learning.

Kwong KK, Beliveau JW, Chesler DA, Goldberg IE, Weisskoff RM, Poncelet BP, Kennedy DN, Hoppel BE, Cohen MS, Turner R. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. Proceedings of the National Academy of Sciences 1992;89:5675–5679.

Lanterman A, Schwarz Wallace. Rissanen: Intertwining Themes in Theories of Model Order Estimation. International Statistical Review August;2001 69:185–212.

De Luca M, Beckmann CF, De Stefano N, Matthews PM, Smith SM. fMRI resting state networks define distinct modes of long-distance interactions in the human brain. NeuroImage November;2005 29:1359–1367. [PubMed: 16260155]

MacQueen, JB. Some methods for classification and analysis of multivariate observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability; Berkeley: University of California Press; 1967. p. 281-297.

Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research 2003;12:419–446. [PubMed: 14599004]

Nichols TE. Nonparametric Permutation Tests for Functional Neuroimaging: A Primer with Examples. Human Brain Mapping 2002;15:1–25. [PubMed: 11747097]

Shaffer J. Multiple hypothesis testing: A review. Annual Review of Psychology 1995;46:561–584.

Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986;73:751–754.

Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol 2002;64:479–498.

Storey JD. The Positive False Discovery Rate: A Bayesian Interpretation and the q-value. The Annals of Statistics 2003;31(6):2013–2035.

Weisskoff RM, Baker J, Belliveau J, Davis TL, Kwong KK, Cohen MS, Rosen BR. Power spectrum analysis of functionally-weighted MR data: what's in the noise? Proceedings of the Society of Magnetic Resonance in Medicine 1993;1(7)

Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal Autocorrelation in Univariate Linear Modeling of FMRI data. NeuroImage December;2001 14(6):1370–1386. [PubMed: 11707093]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Woolrich MW, Behrens TEJ, Smith SM. Constrained linear basis sets for HRF modeling using Variational Bayes. NeuroImage 2004;21:1748–1761. [PubMed: 15050595]

Woolrich MW, Behrens TEJ, Beckmann CF, Smith SM. Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data. IEEE Trans on Medical Imaging January;2005 24:1–11.

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. A unified statistical approach for determining significant voxels in images of cerebral activation. Human Brain Mapping 1996;4:58–73.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

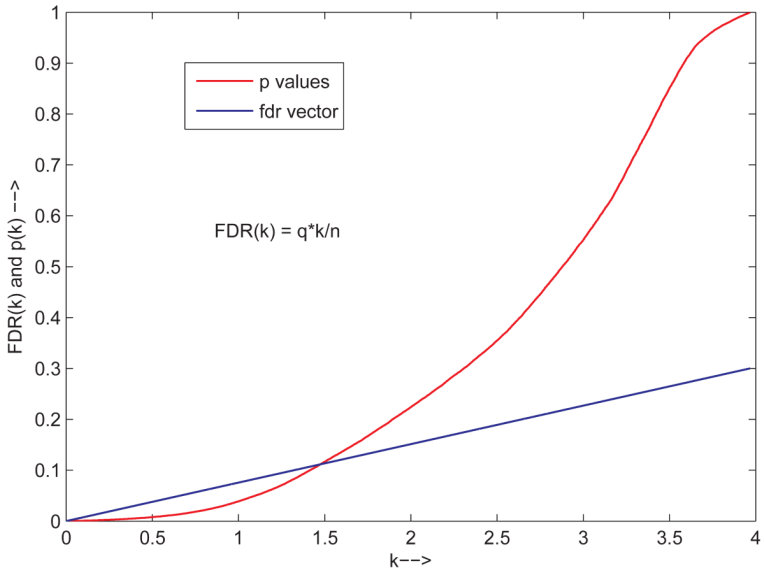

Figure 1.

Graphical depiction of FDR controlling procedure in the *p*-domain. The figure shows a plot of *k* versus the sorted *p*-values $p(k)$ and the quantity $\mathrm{FDR}(k) = \frac{qk}{n}$. The maximum *p*-value at which $p(k)$ crosses $FDR(k)$ becomes the threshold. The scale on the *k*-axis in the above figure is $\times 10^4$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

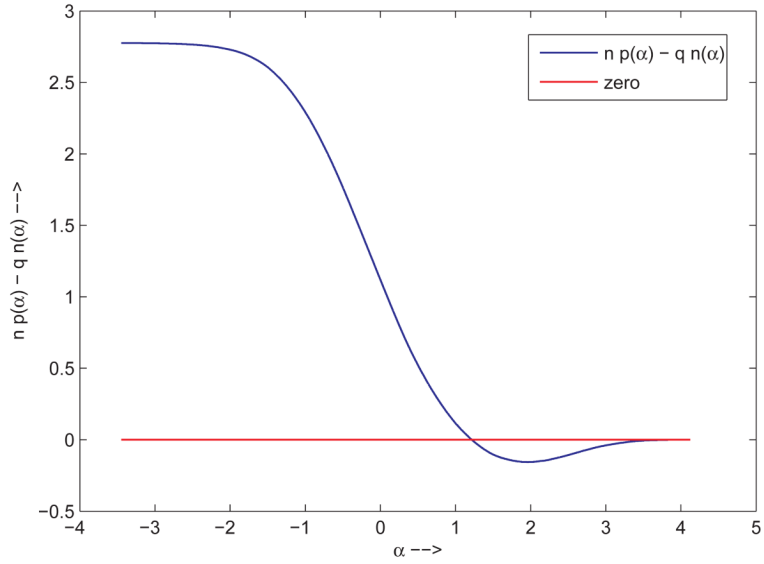

Figure 2.

Graphical depiction of FDR controlling procedure in the *z*-domain. This is equivalent to Figure 1 except that it is in the *z*-domain. The figure shows a plot of α versus $np(\alpha) - qn(\alpha)$. The threshold is determined as the smallest α for which $np(\alpha) - qn(\alpha) \leq 0$

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

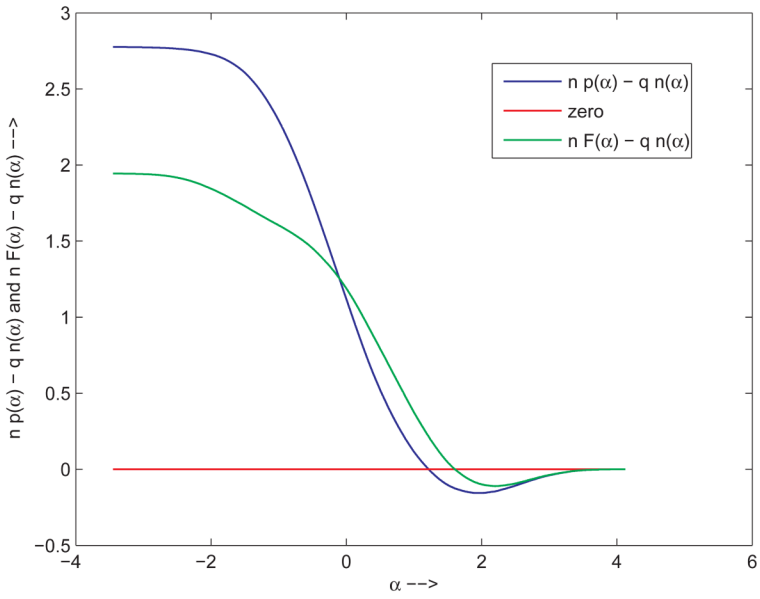

Figure 3.
Graphical comparison of GFDR and FDR in the *z*-domain. The figure shows a plot of α versus $np(\alpha) - qn(\alpha)$ for FDR (blue) and $nF(\alpha) - qn(\alpha)$ for GFDR (green). The thresholds for both methods are determined as the smallest values of α such that the corresponding curves cross the zero line (red) from positive to negative values.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

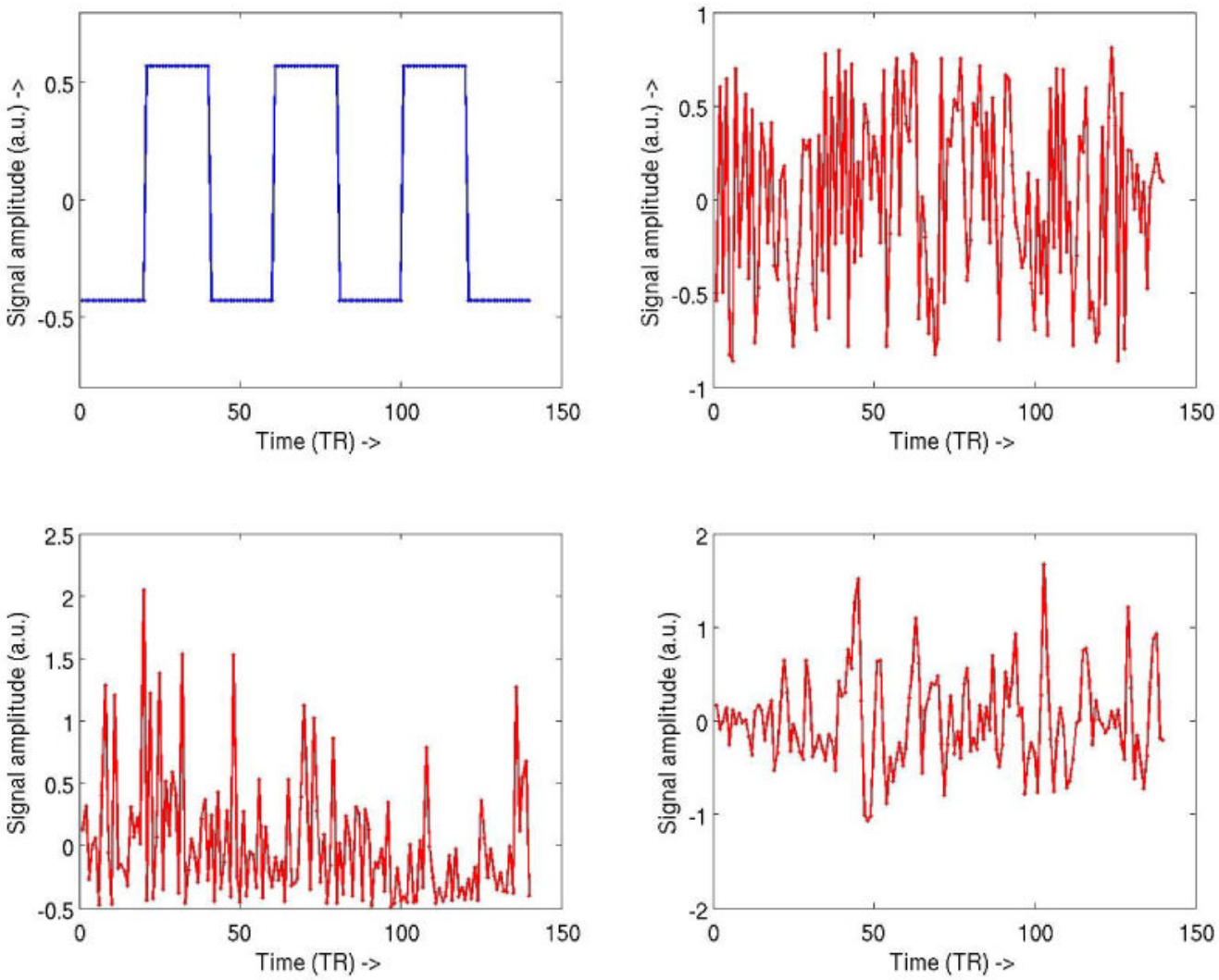

Figure 4.

Top left: “activation”, “deactivation” and “null” data was generated using a block design EV with 140 timepoints (*X*). Top right and bottom row: Confounds generated from a uniform distribution *U* (0, 1) (top right), a Gamma distribution *Gα*(1, 1) (bottom left) and signal from a resting state network in real fMRI data (bottom right). All confounds were demeaned and normalized to have the same energy as the data generating EV *X*. For each combination of π_k and $\rho^s_{\max}$ 100 simulated data-sets were generated. Each simulated data-set consisted of 2000 vectors simulated as described in 3.3. Each data-set was then analyzed using only the block design EV *X* (i.e., ignoring the point-wise variable confounds) to calculate *z*-statistics at the 2000 points corresponding to the parameter estimate of the only column in *X*. For each data-set a Gaussian mixture model was fit to the full data. When GMM identified $R = 3$ classes via BIC we also calculated the true positive rates of each simulated class using Bayes aposteriori classification.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

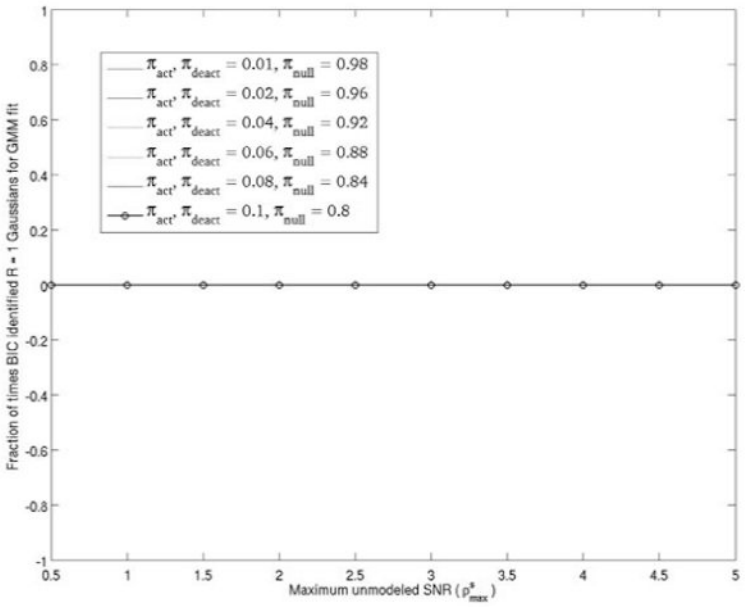

Figure 5.
Figure showing the fraction of times that GMM identified $R = 1$ classes using the BIC criterion over 100 simulated datasets of 2000 vectors each as per section 3.3. BIC did not select $R = 1$ in any of the simulations.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

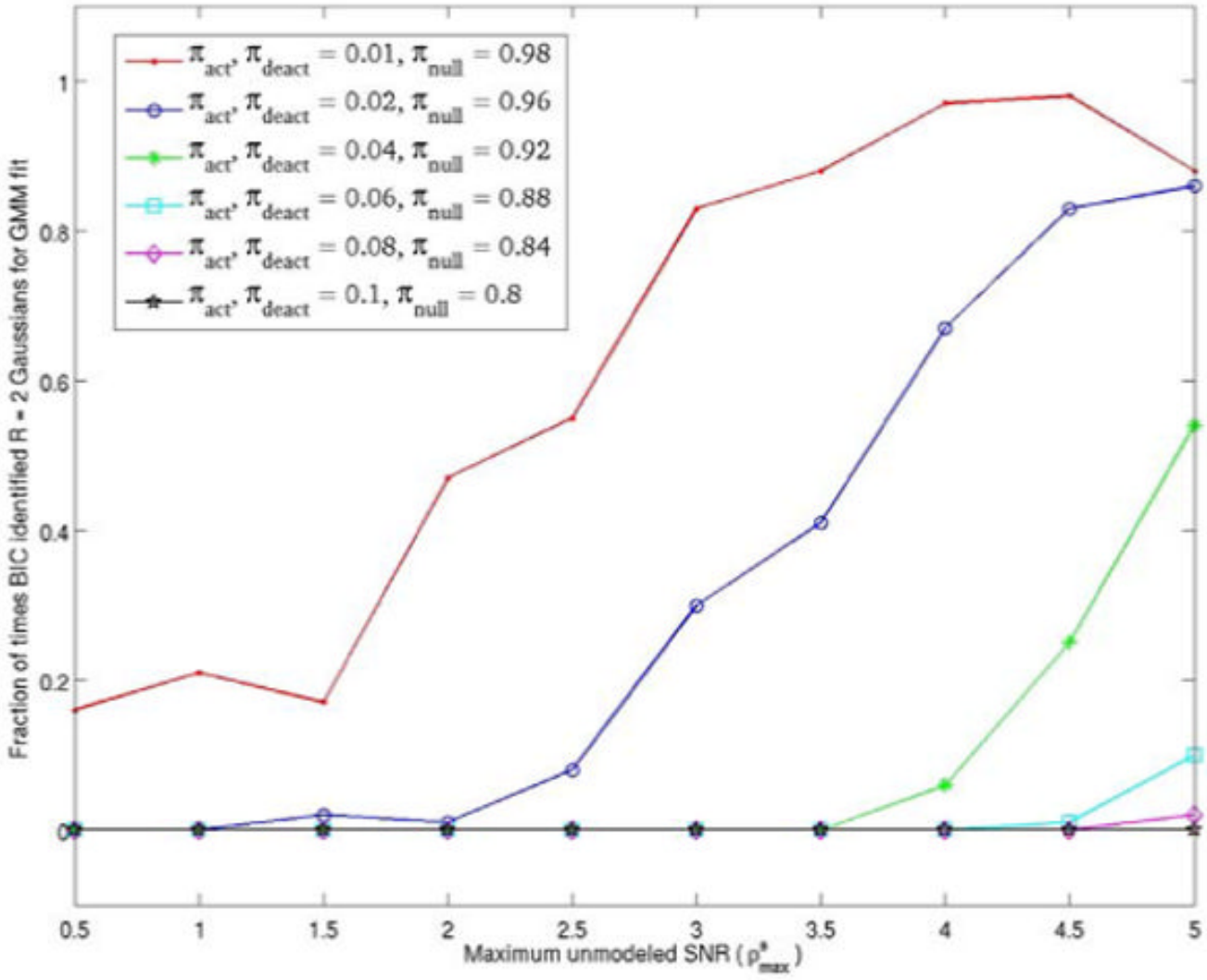

Figure 6.

Figure showing the fraction of times that GMM identified $R = 2$ classes using the BIC criterion over 100 simulated datasets of 2000 vectors each as per section 3.3. It is seen that an increase in maximum SNR of unmodeled signal ($\rho^s_{\max}$) results in more frequent misidentification. It is also seen the BIC misidentification increases when the fraction of "activation" and "deactivation" decreases.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

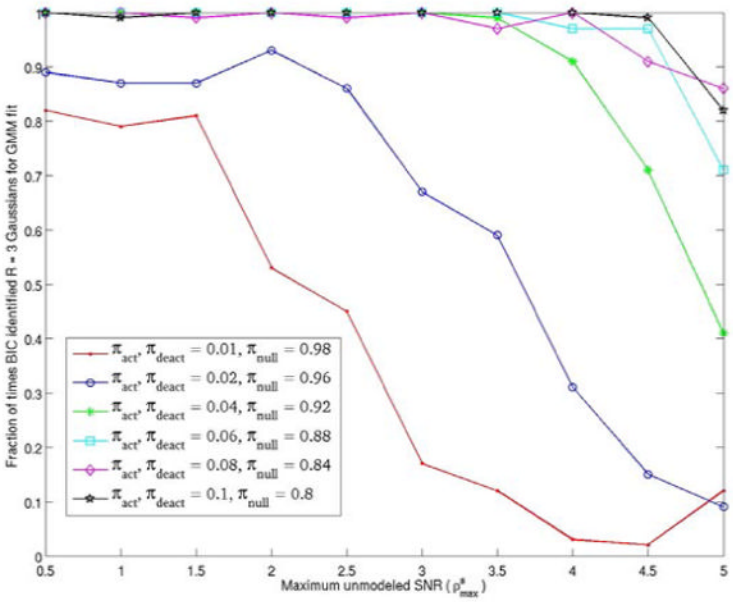

Figure 7.

Figure showing the fraction of times that GMM correctly identified $R = 3$ classes using the BIC criterion over 100 simulated datasets of 2000 vectors each as per section 3.3. It is seen that when the fraction of "activation" and "deactivation" $\pi_{act}, \pi_{deact} \geq 0.02$ and maximum SNR of unmodeled signal $\rho^s_{\max} \leq 2.5$, BIC achieves a correct identification accuracy of $\geq 85\%$. Also when $\pi_{act}, \pi_{deact} \geq 0.04$ and $\rho^s_{\max} \leq 4$, BIC achieves a correct identification accuracy of $\geq 90\%$.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

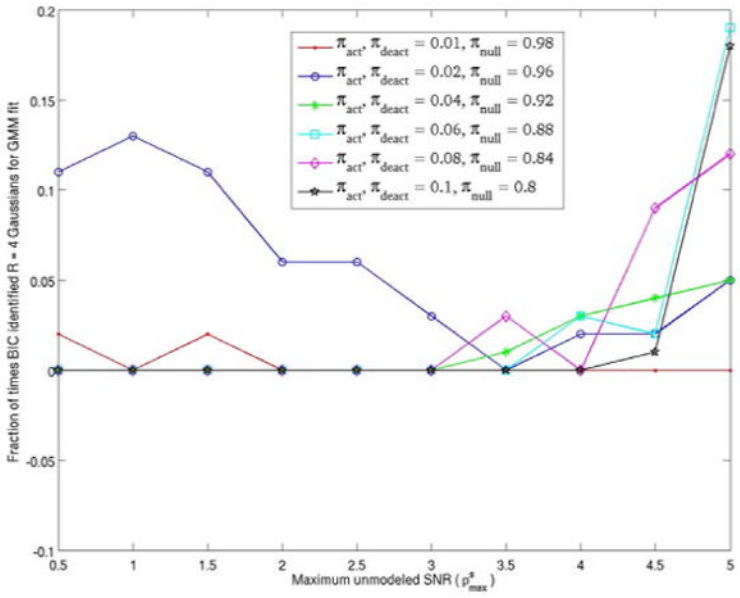

Figure 8.

Figure showing the fraction of times that GMM identified $R = 4$ classes using the BIC criterion over 100 simulated datasets of 2000 vectors each as per section 3.3. It is seen that when the fraction of "activation" and "deactivation" $\pi_{act}, \pi_{deact} \geq 0.04$, BIC identifies 4-classes with increasing probability as maximum SNR of unmodeled signal $\rho^s_{\max}$ increases reaching a value of around 0.2 for $\rho^s_{\max}=5$. BIC also misidentifies the number of classes when $\pi_{act}, \pi_{deact} \leq 0.02$ reaching a maximum probability of around 0.12 for $\rho^s_{\max}=1$.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

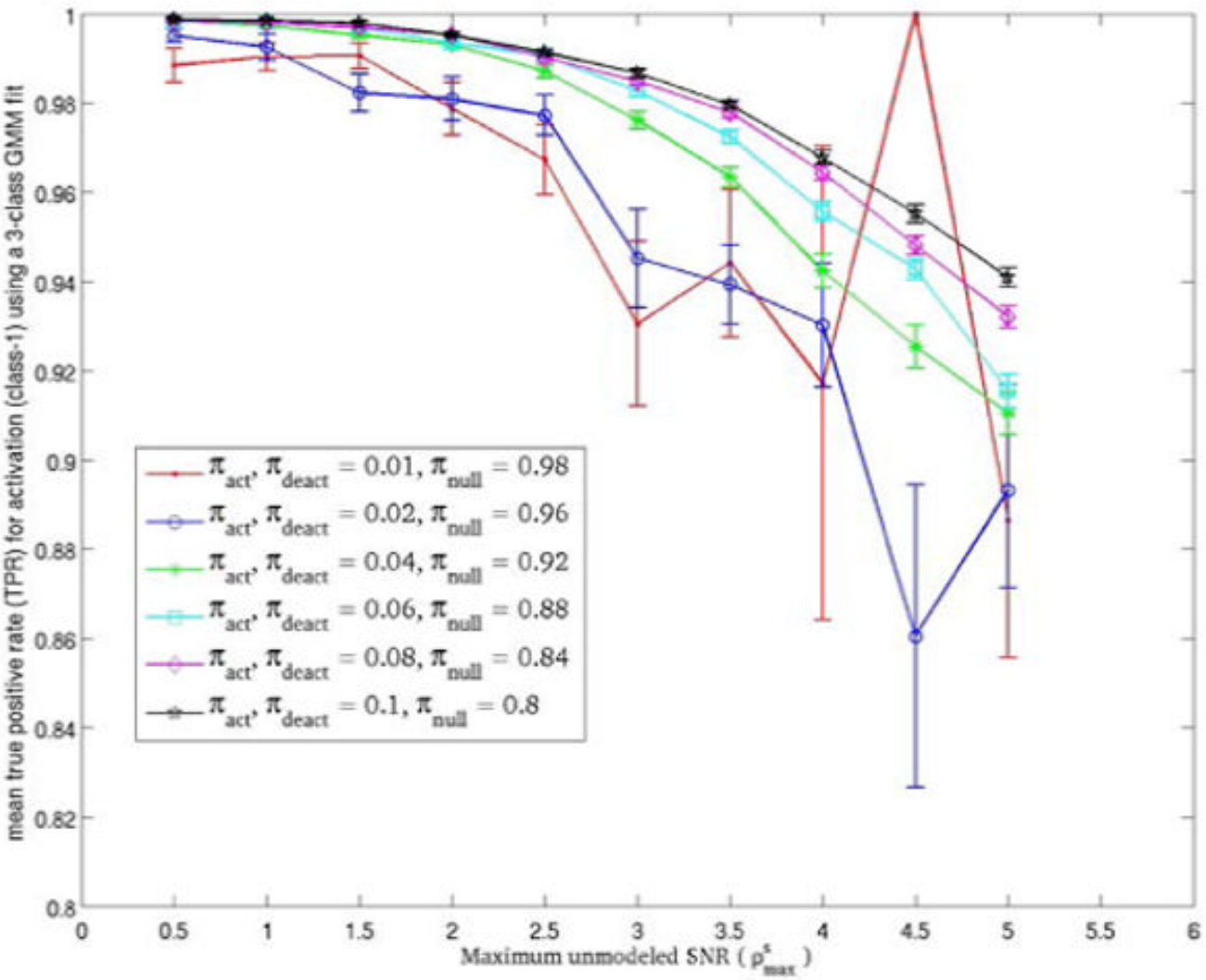

Figure 9.

Figure showing the true positive rate (maximum posterior probability based) for the "activation" class when BIC correctly identified $R = 3$ as a function of maximum SNR of unmodeled signal $\rho^s_{\max}$ and the "activation" fraction. It is seen that for $\pi_{act} = \pi_{deact} \geq 0.04$ a TPR rate of ≥ 90% is obtained for all $\rho^s_{\max}$. For $\pi_{act} = \pi_{deact} \geq 0.02$ a TPR rate of ≥ 92% is obtained for $\rho^s_{\max} \leq 3.5$. The values for $\pi_{act} = \pi_{deact} \leq 0.02$ and $\rho^s_{\max} \geq 4$ are not reliable because of the small fraction of correct identification by BIC in these cases.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

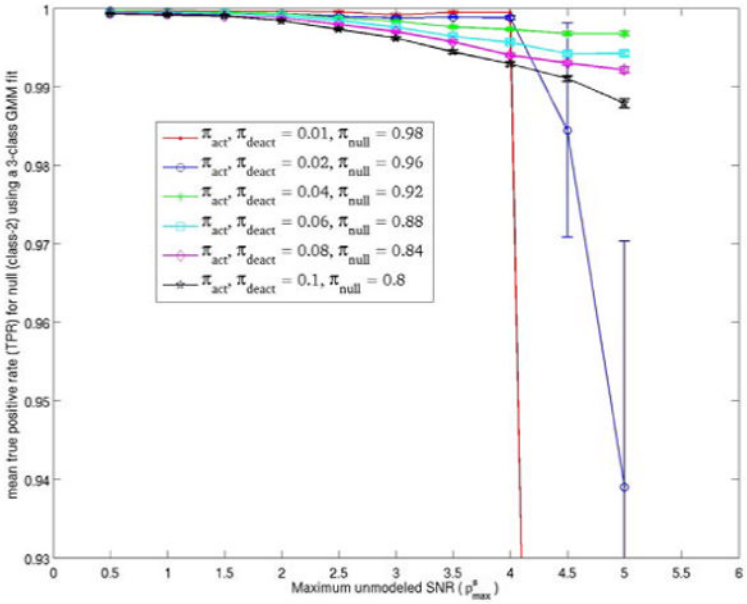

Figure 10.

Figure showing the true positive rate (maximum posterior probability based) for the "null" class when BIC correctly identified $R = 3$ as a function of maximum SNR of unmodeled signal $\rho^s_{\max}$ and the "null" fraction. It is seen that for $\pi_{act} = \pi_{deact} \geq 0.04$ a TPR rate of $\geq 98\%$ is obtained for all $\rho^s_{\max}$. For $\pi_{act} = \pi_{deact} \geq 0.02$ a TPR rate of $\geq 99\%$ is obtained for $\rho^s_{\max} \leq 3.5$. The values for $\pi_{act} = \pi_{deact} \leq 0.02$ and $\rho^s_{\max} \geq 4$ are not reliable because of the small fraction of correct identification by BIC in these cases.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

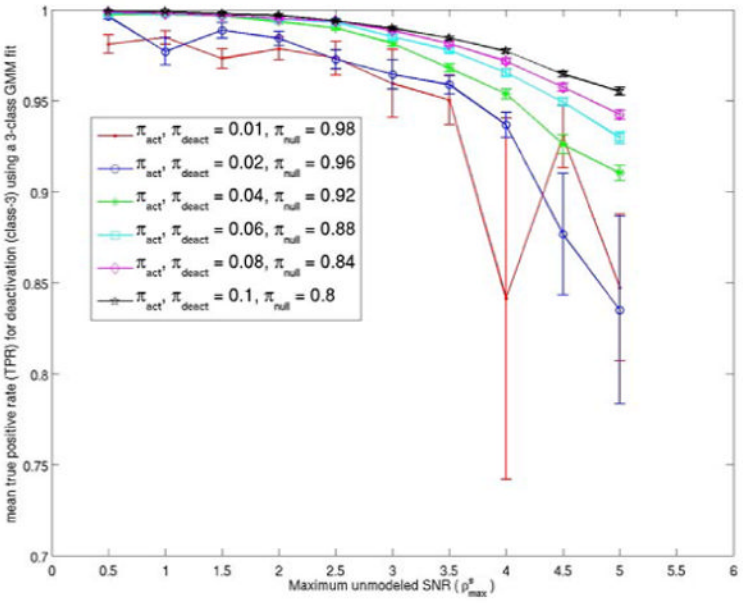

Figure 11.

Figure showing the true positive rate (maximum posterior probability based) for the "deactivation" class when BIC correctly identified $R = 3$ as a function of maximum SNR of unmodeled signal $\rho^s_{\max}$ and the "deactivation" fraction. It is seen that for $\pi_{act} = \pi_{deact} \geq 0.04$ a TPR rate of $\geq 90\%$ is obtained for all $\rho^s_{\max}$. For $\pi_{act} = \pi_{deact} \geq 0.02$ a TPR rate of $\geq 92\%$ is obtained for $\rho^s_{\max} \leq 3.5$. The values for $\pi_{act} = \pi_{deact} \leq 0.02$ and $\rho^s_{\max} \geq 4$ are not reliable because of the small fraction of correct identification by BIC in these cases.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

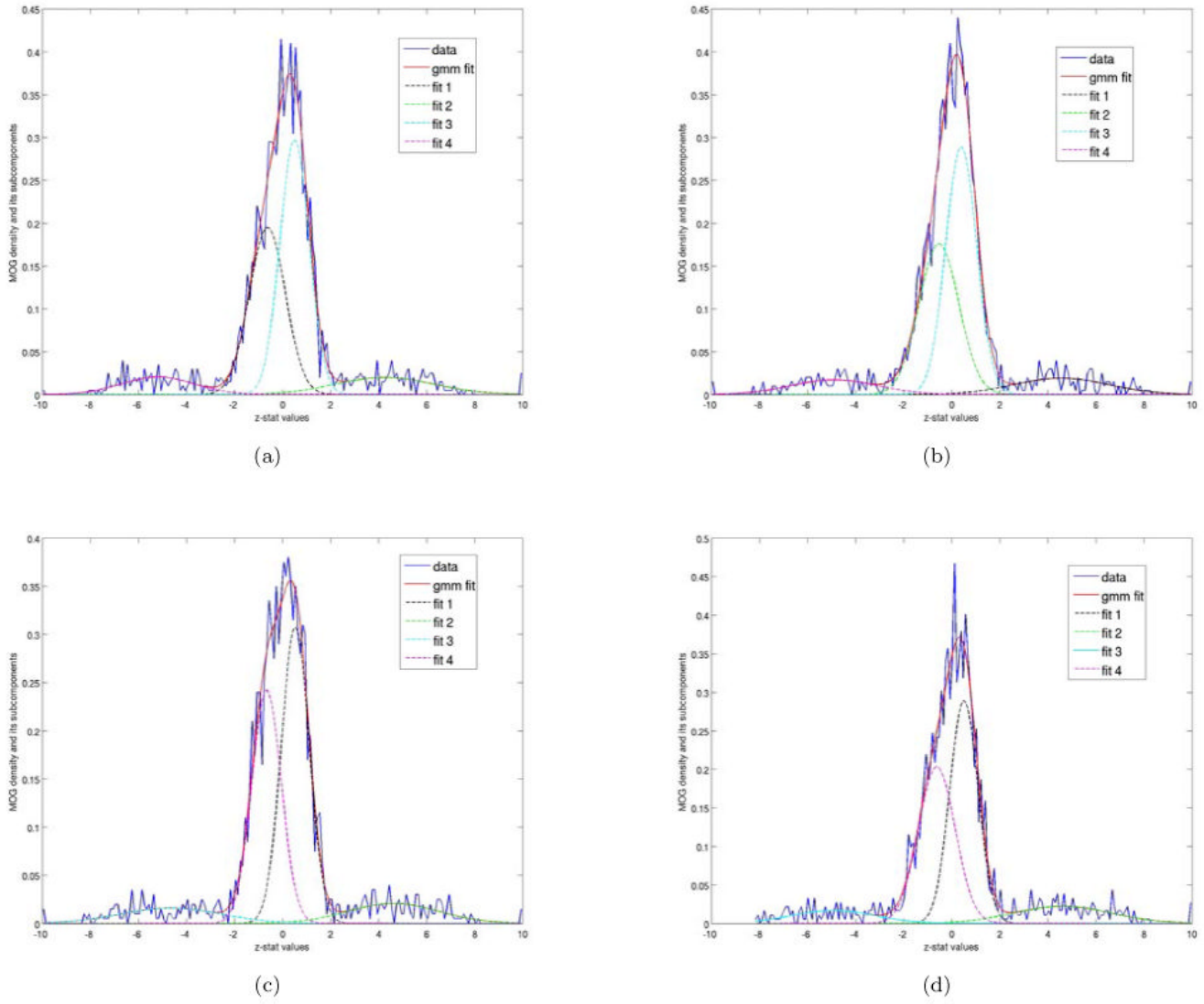

Figure 12.

Figures 12(a) - 12(d) show the MOG density fit to the histogram of *z*-values for data from simulation study described in section 3.3. Each figure shows an example of a "split null" for multiple simulation runs with $\pi_{act} = \pi_{deact} = 0.08$, $\pi_{null} = 0.84$ and $\rho^{s}_{\max}$=4.5. For example, in Figure 12(a), subcomponents 2 and 4 are taken to be "activation" and "deactivation" respectively while the "null" distribution is taken to be well-described by an MOG density with subcomponents labelled 1 and 3. Similar comments apply to other Figures 12(b) - 12(d). When the "null" is assumed to be an MOG density as described above then we obtain an average true positive rate (TPR) for all classes > 96% using Bayes aposteriori classification.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

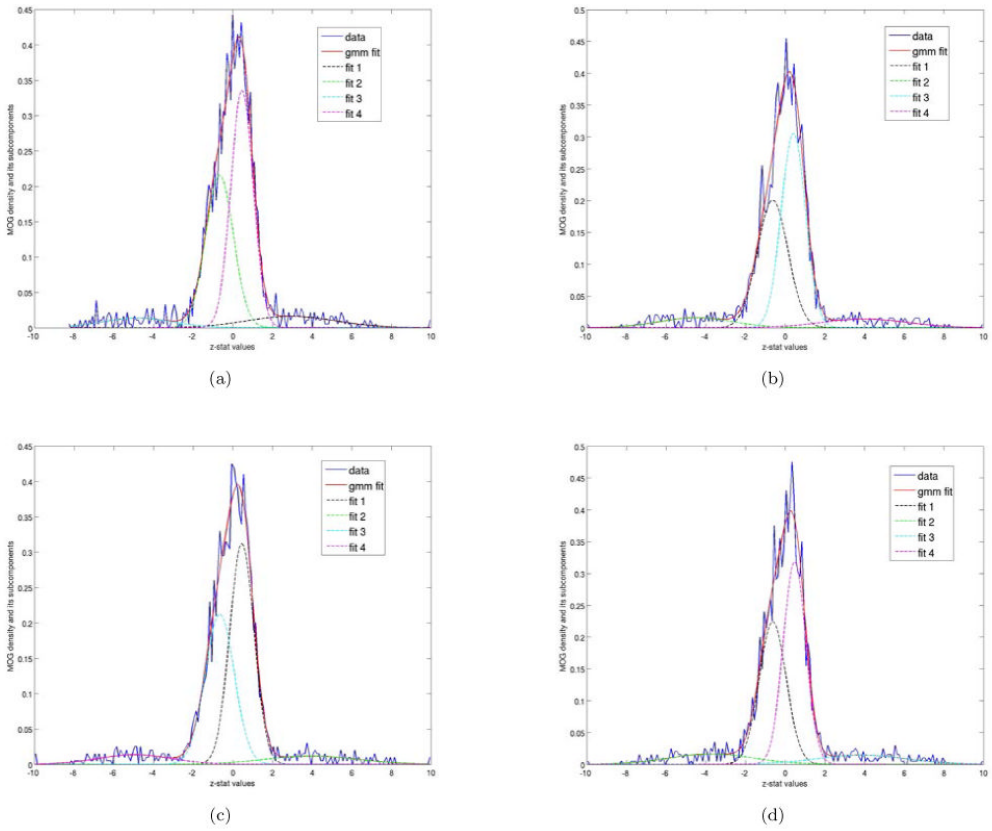

Figure 13.

Figures 13(a) - 13(d) show the MOG density fit to the histogram of *z*-values for data from simulation study described in section 3.3. Each figure shows an example of a "split null" for multiple simulation runs with $\pi_{act} = \pi_{deact} = 0.06$, $\pi_{null} = 0.88$ and $\rho^{s}_{\max}=5$. For example, in Figure 13(a), subcomponents 1 and 3 are taken to be "activation" and "deactivation" respectively while the "null" distribution is taken to be well-described by an MOG density with subcomponents labelled 2 and 4. Similar comments apply to other Figures 13(b) - 13(d). When the "null" is assumed to be an MOG density as described above then we obtain an average true positive rate (TPR) for all classes > 95% using Bayes aposteriori classification.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

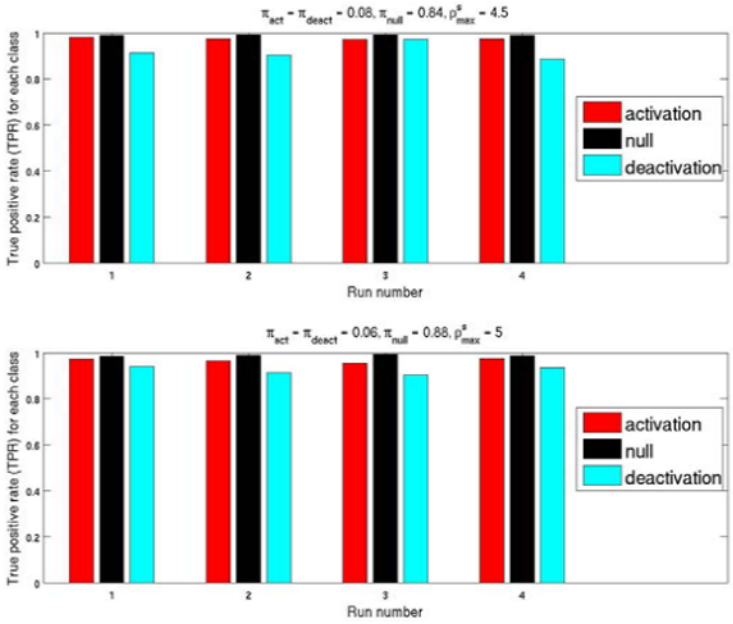

Figure 14.

Figure showing the true positive rates (TPR) (maximum posterior probability based) attained for examples shown in Figures 17-13 when using a 2 component MOG density for the "null" distribution. Results for 12(a) - 12(d) are shown from left to right in the top figure when $\pi_{act} = \pi_{deact} = 0.08$, $\pi_{null} = 0.84$ and $\rho^s_{\max}=4.5$ while results for 13(a) - 13(d) are shown from left to right in the bottom figure when $\pi_{act} = \pi_{deact} = 0.06$, $\pi_{null} = 0.88$ and $\rho^s_{\max}=5$ It was found that in both cases a high TPR was obtained for all classes. For the top figure, the average TPR over all classes and runs was > 96% while for the bottom figure it was > 95%.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

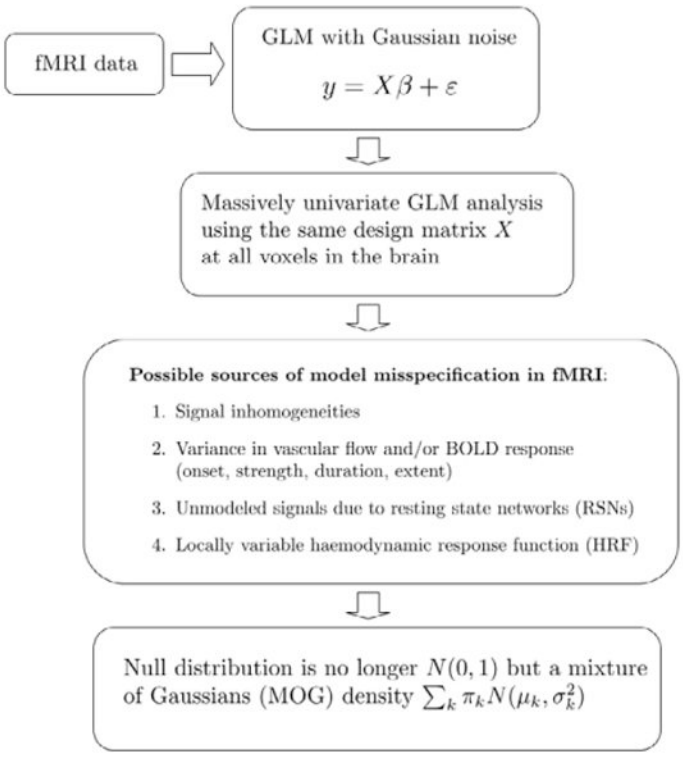

Figure 15.
Figure explaining that the empirical "null" distribution for GLM (with Gaussian noise) based analysis of fMRI data is a MOG density.

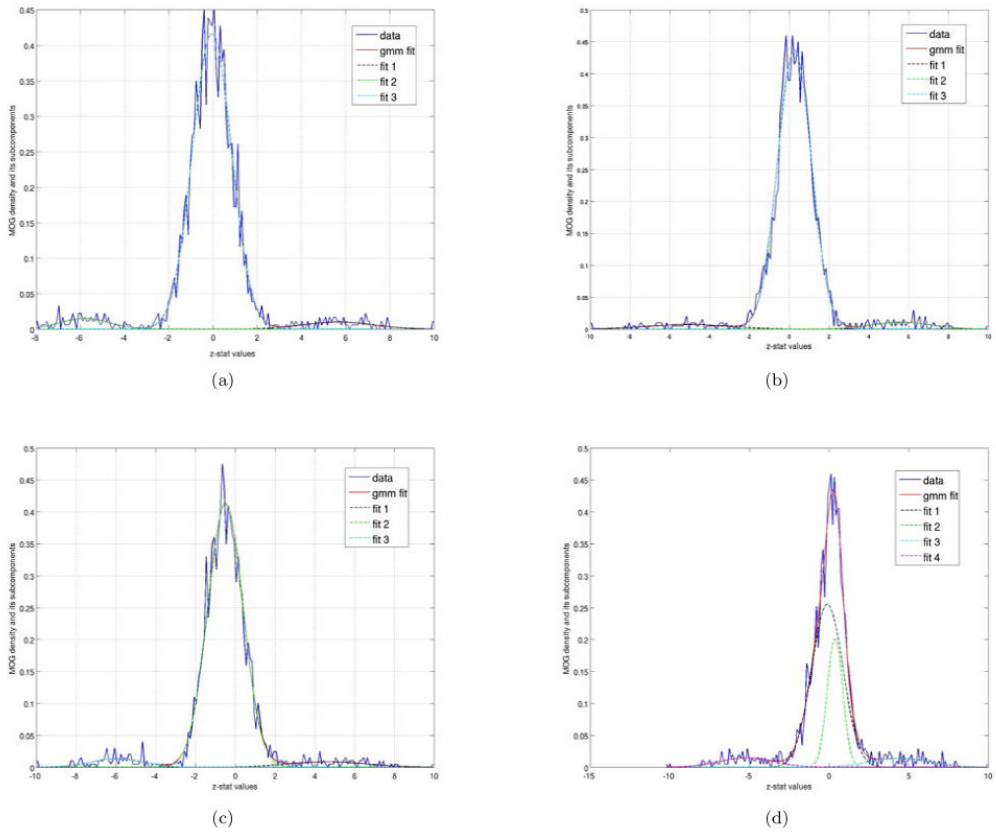

Figure 16.

Figures illustrating commonly observed MOG fits. Figures 16(a) - 16(b) show data from simulation scheme of section 3.3. 16(a) shows the case of a "centered" 3 component MOG fit for $\pi_{null} = 0.92$, $\pi_{act} = \pi_{deact} = 0.04$, $\rho^{s}_{\max}=3$ and primary corruption via confound 2 ($\pi_{W1} = 0.1$, $\pi_{W2} = 0.8$, $\pi_{W3} = 0.1$). 16(b) shows the case of a "right shifted" 3 component MOG fit for $\pi_{null} = 0.92$, $\pi_{act} = \pi_{deact} = 0.04$, $\rho^{s}_{\max}=3$ and primary corruption via confound 3 ($\pi_{W1} = 0$, $\pi_{W2} = 0.2$, $\pi_{W3} = 0.8$). 16(c) shows the case of a "left shifted" 3 component MOG fit for $\pi_{null} = 0.92$, $\pi_{act} = \pi_{deact} = 0.04$, $\rho^{s}_{\max}=3$ and primary corruption via confound 1 ($\pi_{W1} = 0.8$, $\pi_{W2} = 0.2$, $\pi_{W3} = 0$). 16(d) shows the case of a "split null" MOG fit where the "null" distribution is well-modeled by a 2 component MOG distribution for $\pi_{null} = 0.88$, $\pi_{act} = \pi_{deact} = 0.06$ and $\rho^{s}_{\max}=5$.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

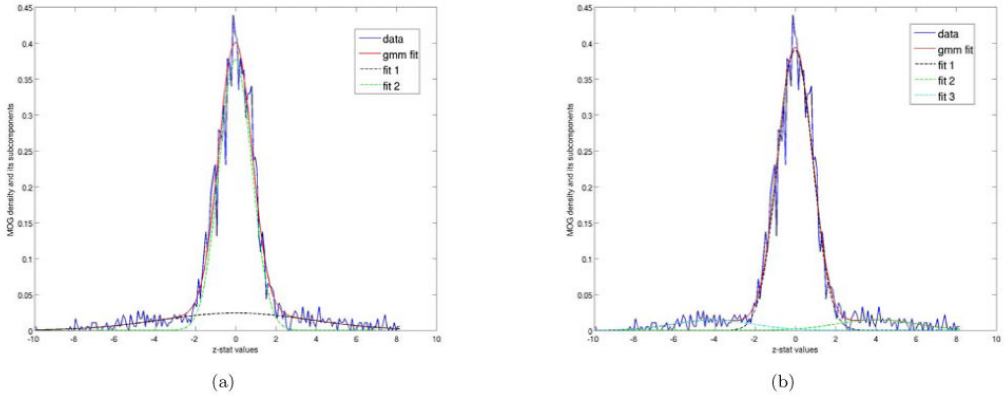

Figure 17.

Figure 17(a) shows the case when BIC identifies 2 classes for $\pi_{null} = 0.88$, $\pi_{act} = \pi_{deact} = 0.06$ and $\rho^{s}_{\max}=5$. Here the "activation" and "deactivation" have been jointly identified by 1 Gaussian subcomponent. In these cases, we find it reasonable to force BIC to fit at least 3 classes. Refitting starting from 3 classes we find BIC determines the optimal number of classes to be 3. The MOG fit using this 3 component fit is shown in 17(b). Now the "activation" and "deactivation" have been identified correctly as separate classes.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

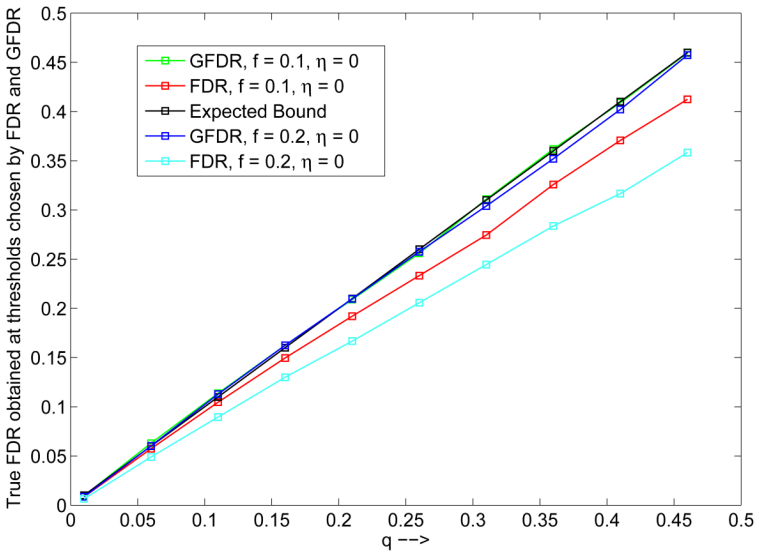

Figure 18.

$\eta = 0$, no unmodeled signal (see text) at SNR = 1. FDR and GFDR perform similarly for small q values ($q \leq 0.1$). For larger values of q, FDR becomes more conservative while GFDR maintains tight control.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

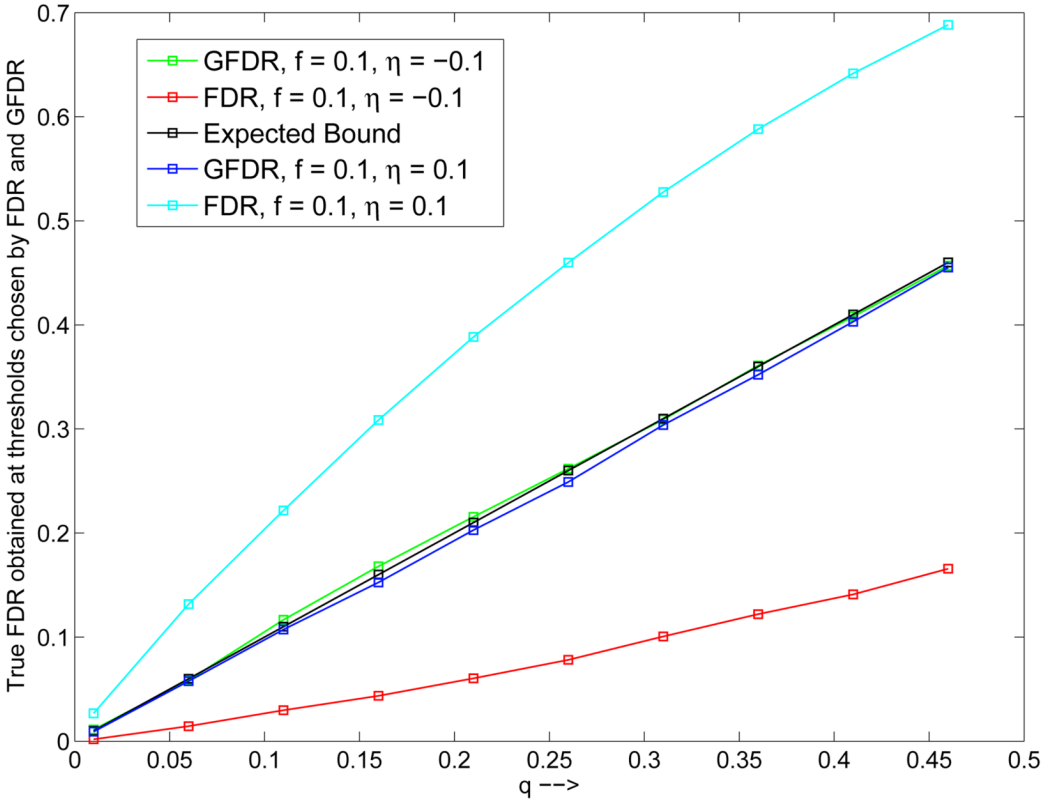

Figure 19.

$|\eta| = 0.1$, $f = 0.1$ (see text). In the presence of unmodeled effects, FDR is significantly affected, becoming overly liberal for $\eta = 0.1$ and overly conservative for $\eta = -0.1$. This effect becomes more pronounced at higher q-values. On the other hand, GFDR maintains good control for both $\eta = 0.1$ and $\eta = -0.1$ even for large values of q.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

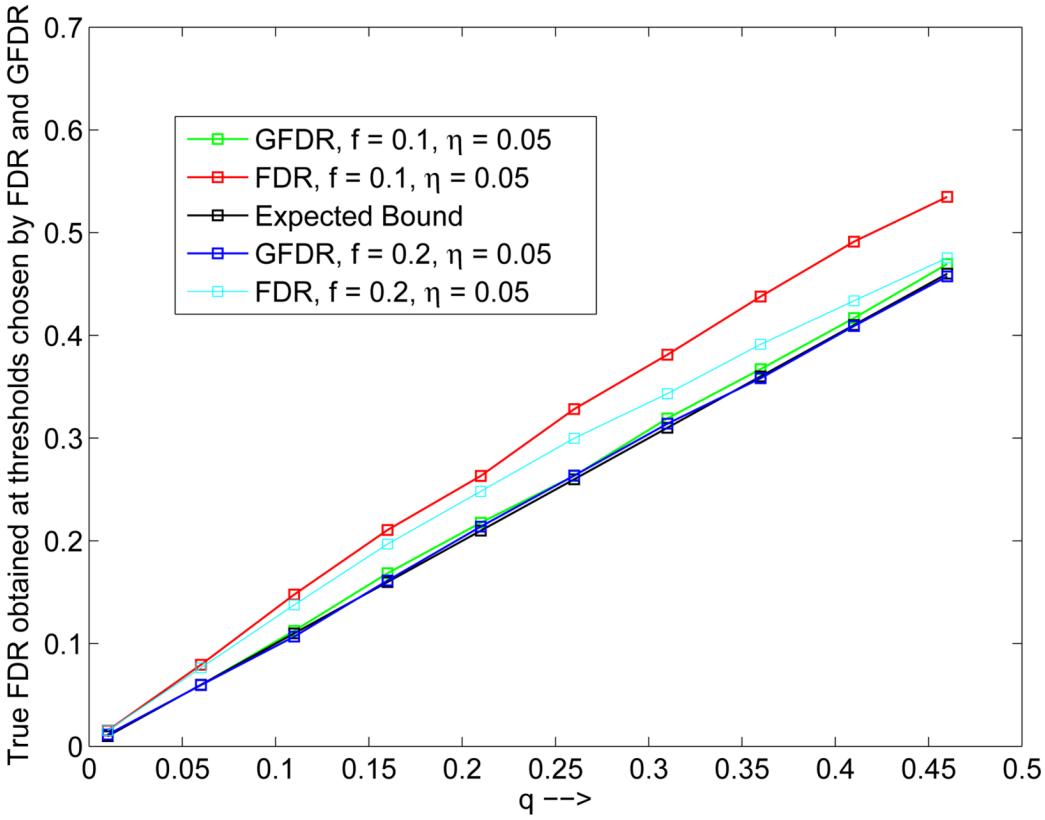

Figure 20.

$\eta = 0.05$, $f = 0.1, 0.2$ (see text). This figure shows the combined effect of positive unmodeled signals and strength of activation. For positive unmodeled effect ($\eta = 0.05$), FDR becomes overly liberal at all values of q for both small and large activation ($f = 0.1$ and $f = 0.2$). The degree of liberality is higher for smaller activation than larger activation. GFDR maintains a tight control for both small and large activation at all values of q.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

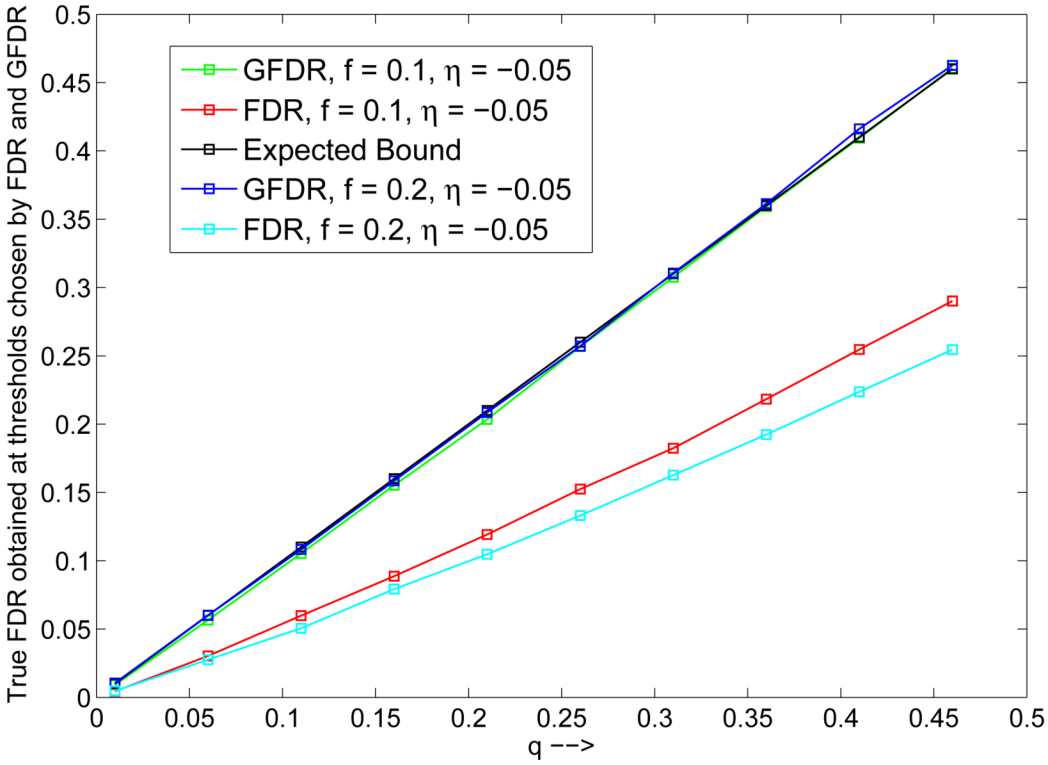

Figure 21.

$\eta = -0.05$, $f = 0.1, 0.2$ (see text). This figure shows the combined effect of negative unmodeled signals and strength of activation. For negative unmodeled effect ($\eta = -0.05$), FDR becomes overly conservative at all values of q for both small and large activation ($f = 0.1$ and f = 0.2). The degree of conservativeness is larger for larger activation. GFDR is able to maintain good control for both small and large activation at all values of q.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

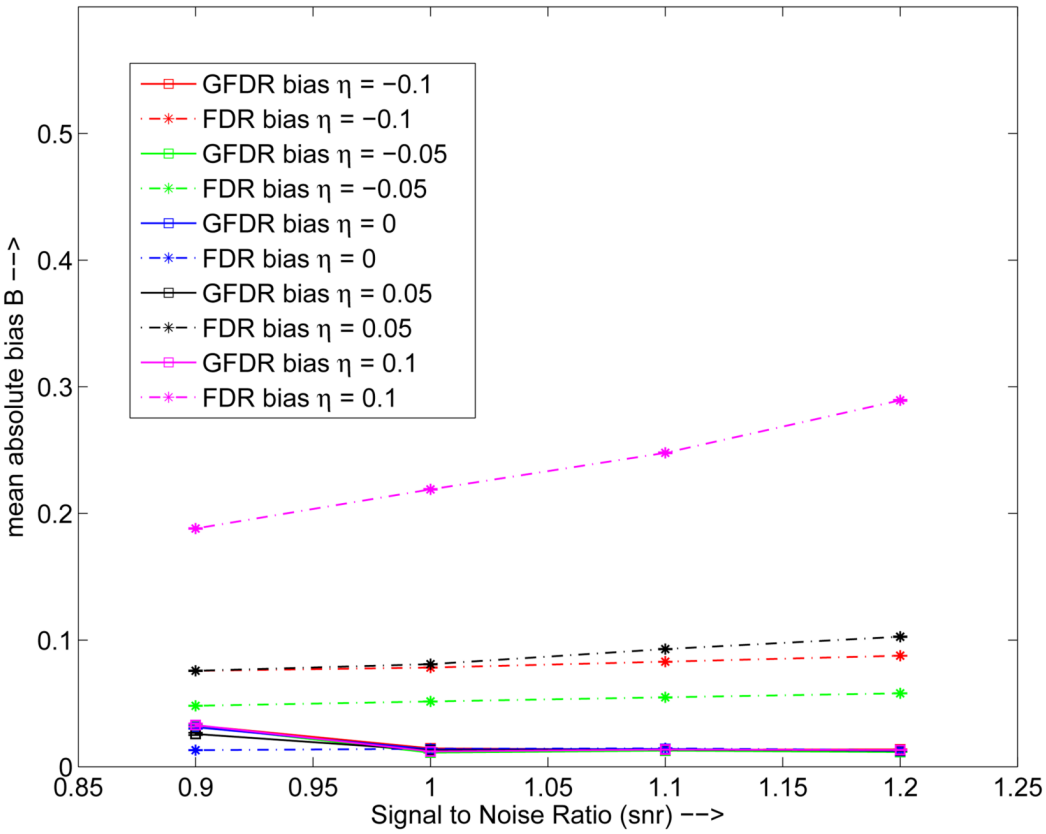

Figure 22.
Simulation Results for CASE I: $q = 0.1$, $f = 0.01$. Figure shows the mean absolute bias introduced by GFDR and FDR for various SNRs at various values of unmodeled signal intensity η

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

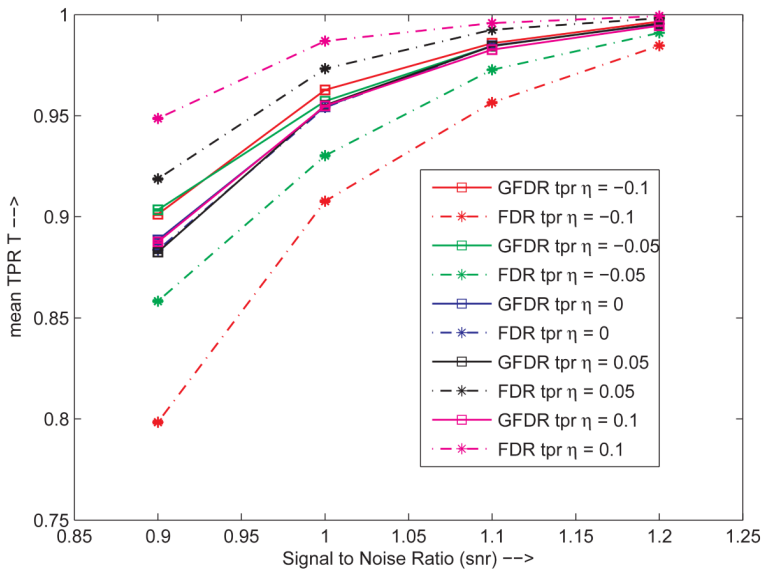

Figure 23.

Simulation Results for CASE I: $q = 0.1$, $f = 0.01$. Figure shows the mean true positive rate attained by GFDR and FDR for various SNRs at various values of unmodeled signal intensity η

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

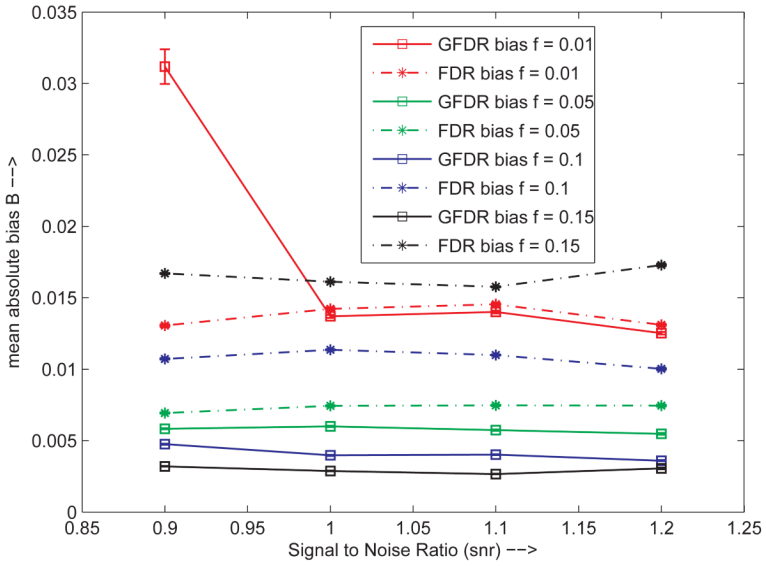

Figure 24.

Simulation Results for CASE II: $q = 0.1$, $\eta = 0$. Figure shows the mean absolute bias introduced by GFDR and FDR for various SNRs at various values of activation fraction f

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

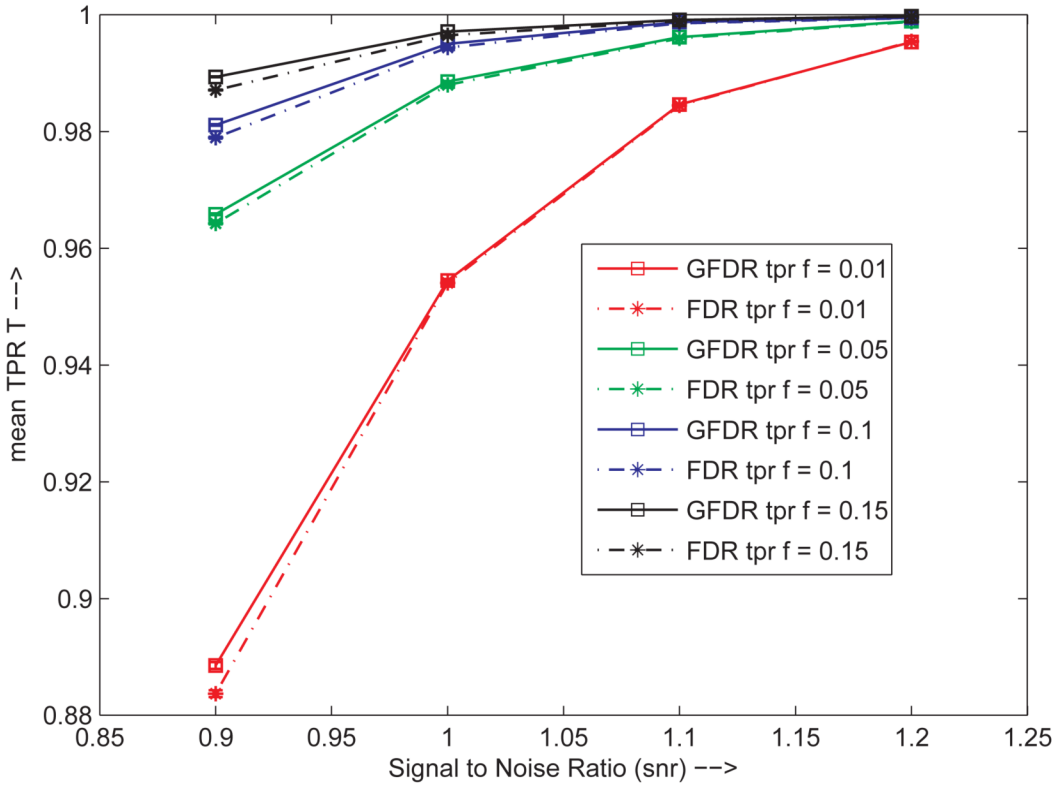

Figure 25.
Simulation Results for CASE II: $q = 0.1$, $\eta = 0$. Figure shows the mean true positive rate attained by GFDR and FDR for various SNRs at various values of activation fraction f

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

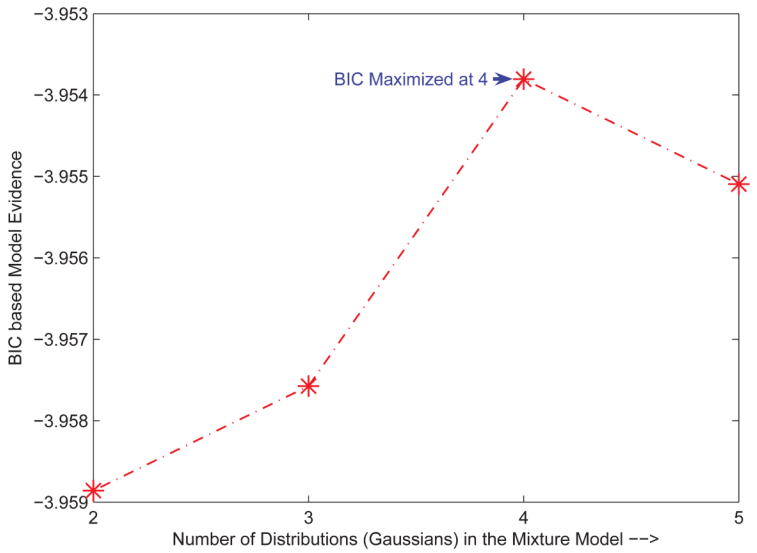

Figure 26.

An example of Model Selection using Bayes Information Criterion (BIC) for the real fMRI dataset (zstat 2). Here, BIC attains a maximum for a mixture density with 4 Gaussians.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

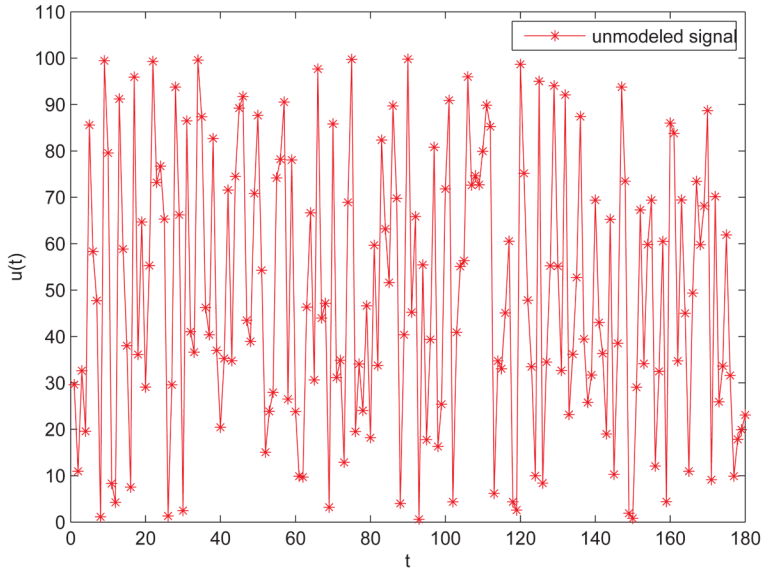

Figure 27.
Artificially introduced unmodeled random signal 1˜% signal change

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

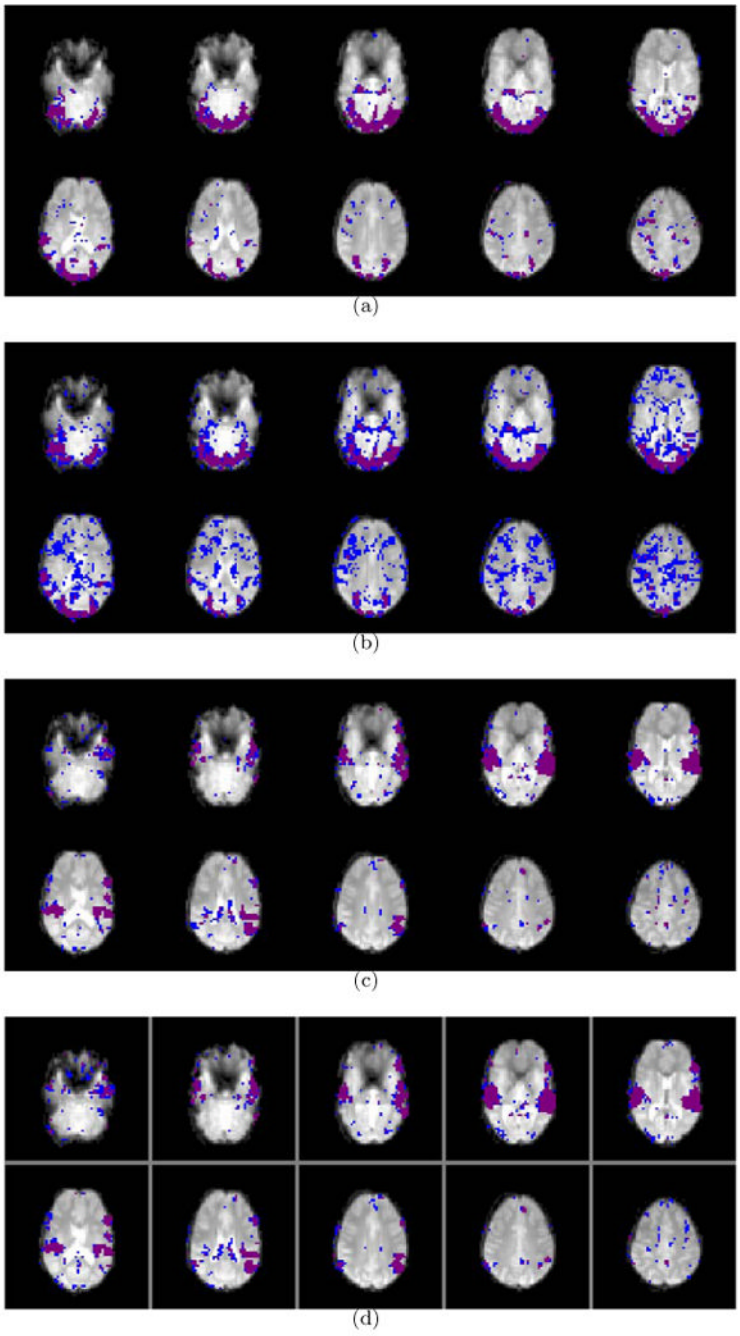

Figure 28.

Figures 28(a)-28(d) show the activation maps for visual and auditory stimuli thresholded using FDR and GFDR. The purple color shows the voxels declared "active" by GFDR and FDR whereas the blue color shows the voxels declared active by "FDR" but not declared active by GFDR. Figures 28(a)-28(b) show the thresholded activation map for visual stimulus using FDR and GFDR before and after the introduction of the artificial unmodeled signal shown in Figure 27. Figures 28(c)-28(d) show the thresholded activation map for the auditory stimulus using FDR and GFDR before and after the introduction of the artificial unmodeled signal shown in Figure 27.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

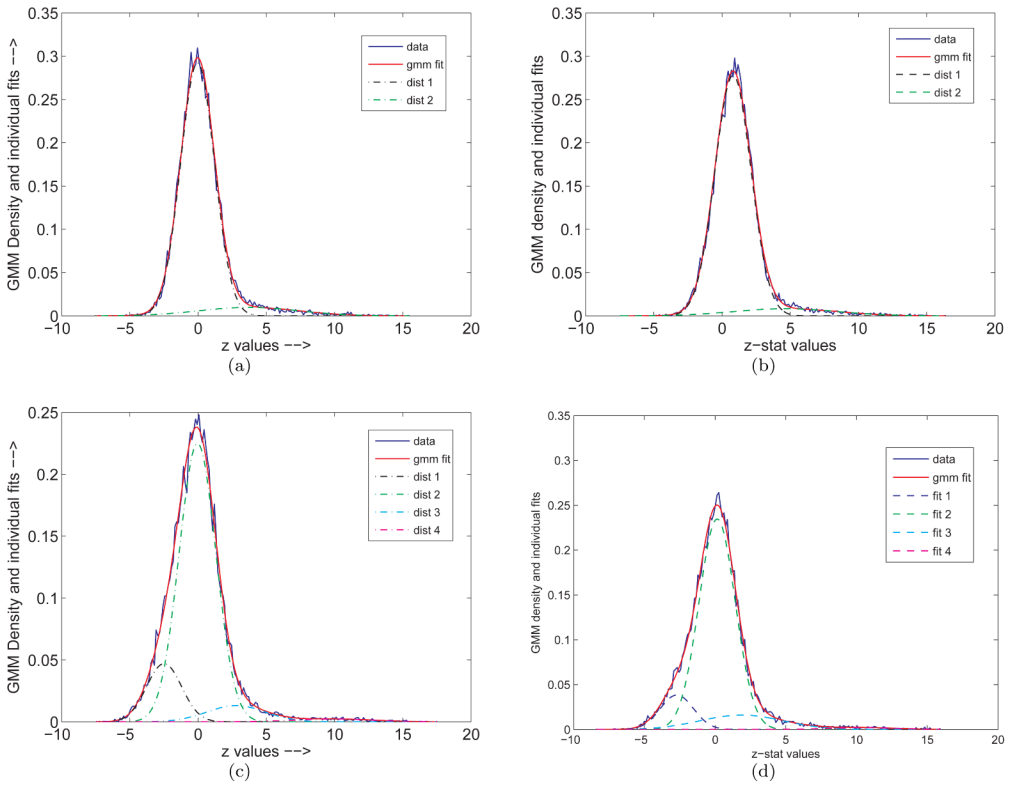

Figure 29.

Figures 29(a) - 29(d) show the GMM maximum likelihood (ML) fits to the distribution of *z*-values for visual and auditory stimulus. Figures 29(a) - 29(b) show the GMM ML fits for the visual stimulus without the unmodeled signal and with the unmodeled signal respectively. Figures 29(c) - 29(d) show the GMM ML fits for the auditory stimulus without the unmodeled signal and with the unmodeled signal respectively.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

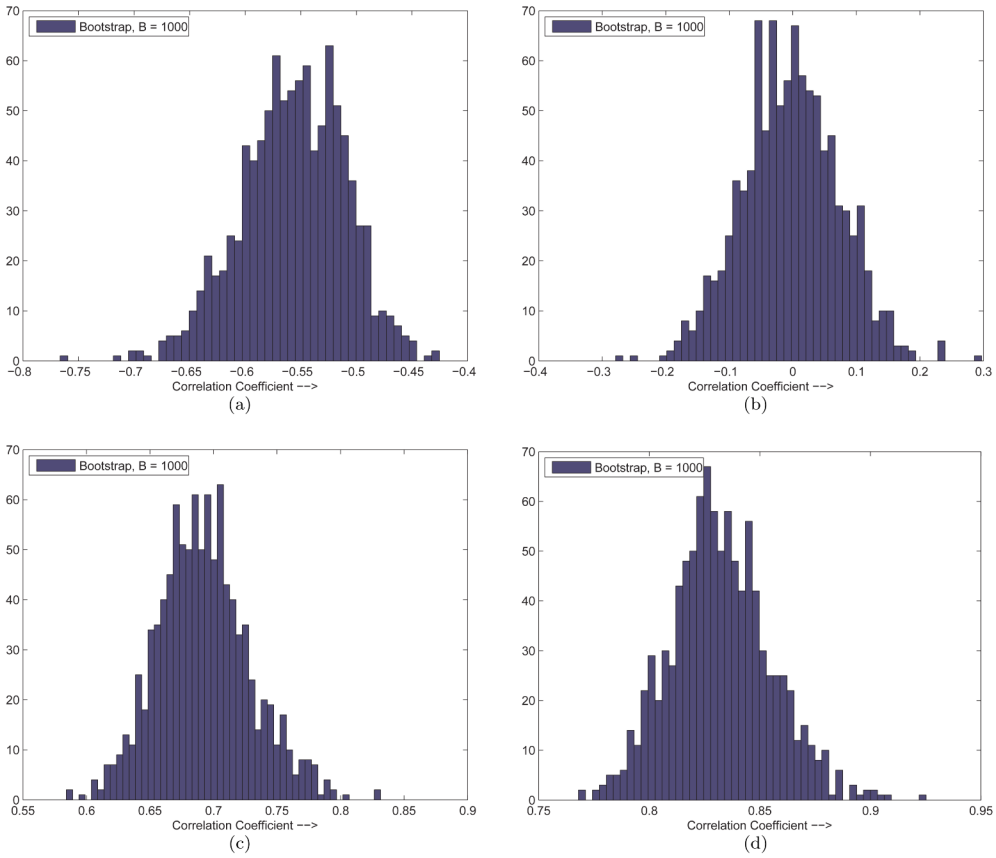

Figure 30.
Bootstrap simulations for the correlation coefficient between the EV for auditory stimulus and probability weighted time courses for each of the four maps without the unmodeled signal.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

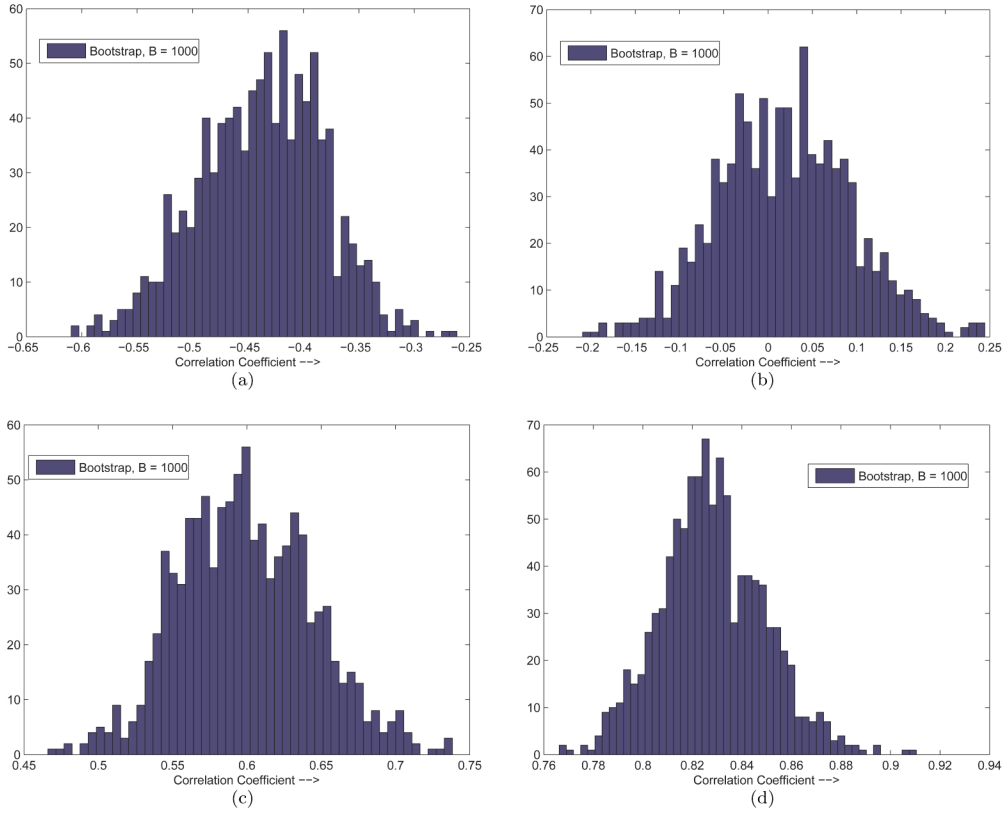

Figure 31.

Bootstrap simulations for the correlation coefficient between the EV for auditory stimulus and probability weighted time courses for each of the four maps with the unmodeled signal.

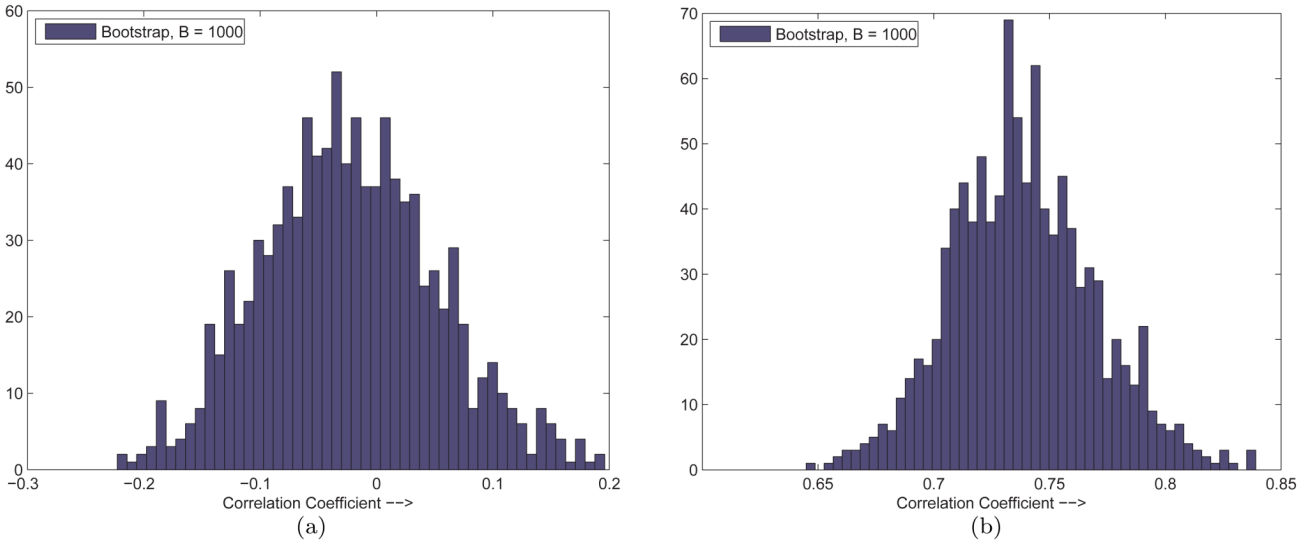

Figure 32.
Bootstrap simulations for the correlation coefficient between the EV for visual stimulus and probability weighted time courses for each of the two maps without the unmodeled signal.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

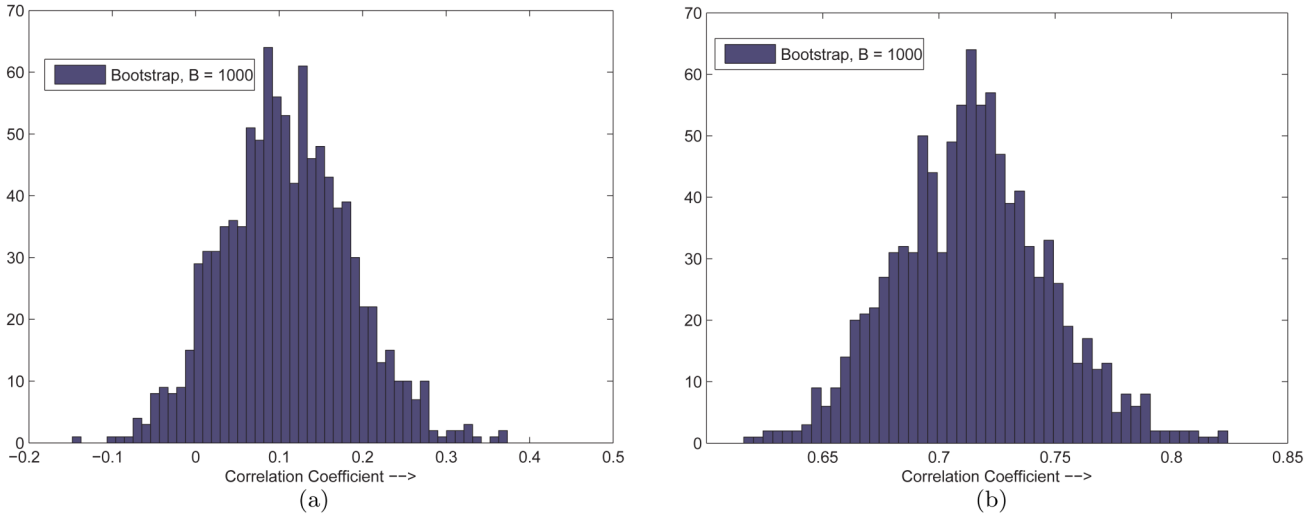

Figure 33.
Bootstrap simulations for the correlation coefficient between the EV for visual stimulus and probability weighted time courses for each of the two maps with the unmodeled signal.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Parameters selected for generating “activation”, “null” and “deactivation” classes.

	activation	**null**	**deactivation**
δ_{min}^{k}	1	0	-1.5
δ_{max}^{k}	1.5	0	-1
π_k	0.01 to 1	0.8 to 0.98	0.01 to 0.1
Data from	*X*	*X*	*X*

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Parameters selected for generating one of the three confounds.

	Confound 1	**Confound 2**	**Confound 3**
ρ^{s}_{min}	-0.5	-0.5	-0.5
ρ^{s}_{max}	0.5 to 5	0.5 to 5	0.5 to 5
π_{Ws}	0.2	0.4	0.4
Data from	W_1	W_2	W_3