ARTICLE IN PRESS

YNIMG-04451; No. of pages: 11; 4C: 4, 6

ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage xx (2007) xxx–xxx

Robust Bayesian general linear models

W.D. Penny,* J. Kilner, and F. Blankenburg

Wellcome Department of Imaging Neuroscience, University College London, 12 Queen Square, London WC1N 3BG, UK

Received 22 September 2006; revised 20 November 2006; accepted 25 January 2007

We describe a Bayesian learning algorithm for Robust General Linear Models (RGLMs). The noise is modeled as a Mixture of Gaussians rather than the usual single Gaussian. This allows different data points to be associated with different noise levels and effectively provides a robust estimation of regression coefficients. A variational inference framework is used to prevent overfitting and provides a model order selection criterion for noise model order. This allows the RGLM to default to the usual GLM when robustness is not required. The method is compared to other robust regression methods and applied to synthetic data and fMRI.

Keywords: Bayesian; fMRI; Artefact; Mixture model; Robust estimation

Introduction

Neuroimaging data contain a host of artifacts arising from 'physiological noise', e.g. subject respiration, heartbeat or movement of the head, eye, tongue or mouth or 'non-physiological noise', e.g. EEG electrodes with poor electrical contact, spikes in fMRI or extracranial magnetic sources in MEG. The presence of these artifacts can severely compromise the sensitivity with which we can detect the neuronal sources we are interested in. The optimal processing of artifactual data is therefore an important issue in neuroimaging analysis and a number of processing methods have been proposed. One approach is visual inspection and removal of trials deemed to contain artefacts. In the analysis of Event-Related Potentials (ERPs), however, this can lead to up to a third of the trials being removed. Because the statistical inferences that follow are based on fewer data points, this results in a loss of sensitivity.

In fMRI, signal processing methods exist for the removal of k-space spikes (Zhang et al., 2001; Greve et al., 2006), and Exploratory Data Analysis (EDA) methods have been proposed for removal of outliers in the context of mass-univariate modeling (Luo and Nichols, 2003). Alternatively, Independent Component Analysis (ICA) can be used to isolate 'noise sources' and remove them from the data (Jung et al., 1999). This is, however, a non-automatic process and will typically require user intervention to disambiguate the discovered components. In fMRI, autoregressive (AR) modeling can be used to downweight the impact of periodic respiratory or cardiac noise sources (Penny et al., 2003). More recently, a number of approaches based on robust regression have been applied to imaging data (Wager et al., 2005; Diedrichsen and Shadmehr, 2005). These approaches relax the assumption underlying ordinary regression that the errors be normally (Wager et al., 2005) or identically (Diedrichsen and Shadmehr, 2005) distributed. In Wager et al. (2005), for example, a Bisquare or Huber weighting scheme corresponds to the assumption of identical non-Gaussian errors. The method was applied to group-level fMRI analysis and was found to lead to more sensitive inferences.

Interestingly, Wager et al. (2005) tested a number of standard robust estimation methods by generating data from a known mixture process as this was thought to capture the essence of signal embedded in artefactual data. In this paper we take this idea one step further and develop an optimal robust estimation procedure for the case of mixture errors.

Specifically, we propose a Robust General Linear Model (RGLM) framework in which the noise is modeled with a Mixture of Gaussians. This allows different data points to be associated with different noise levels and provides a robust estimation of regression coefficients via a weighted least squares approach. Data points associated with high noise levels are downweighted in the parameter estimation step. Moreover, a Bayesian estimation framework (Attias, 2000) is used to prevent model overfitting and provides a model order selection criterion for noise model order. This allows selection of the usual GLM, i.e. a noise mixture with a single component, when an outlier model is not appropriate. This work is based on a similar algorithm for robust estimation of autoregressive processes (Roberts and Penny, 2002).

Theory

We define the General Linear Model (GLM) in the usual way

$$\boldsymbol{y} = \mathbf{X}\,\boldsymbol{w} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is an $N\times 1$ vector of data points, $\mathbf{X}$ is an $N\times p$ design matrix, $\boldsymbol{w}$ is a $p\times 1$ vector of regression coefficients and $\boldsymbol{e}$ is an

* Corresponding author. Fax: +44 020 7813 1420.
E-mail address: w.penny@fil.ion.ucl.ac.uk (W.D. Penny).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.01.058

$N \times 1$ vector of errors. We can also write this relationship for the nth data point

$$\boldsymbol{y}_n = x_n \, \boldsymbol{w} + \boldsymbol{e}_n \quad (2)$$

where $\boldsymbol{y}_n$ is the nth data point, x_n is the nth row of $\mathbf{X}$ and $\boldsymbol{e}_n$ is the nth error.

In the standard GLM, the noise $\boldsymbol{e}_n$ is modeled as a Gaussian. This implies that the regression coefficients can be set by minimizing a least squares cost function. Least squares, however, is known to be sensitive to outliers. Therefore, if our data are even marginally contaminated by artifacts the resulting regression coefficient estimates will be seriously degraded. See Bishop (Bishop, 1995, page 209) and Press et al. (Press et al., 1992, page 700) for a general discussion of this issue and a number of proposed solutions.

In this paper, we define a Robust GLM (RGLM) as one in which the noise is modeled as a Mixture of Gaussians (MoGs) having m-components. This includes the standard $m=1$ case, i.e. the single Gaussian that is assumed for the usual GLM. The overall generative model is shown in Fig. 1. For data containing outliers m will be equal to 2, comprising a low noise variance, signal-bearing component ($s=1$) and a high noise variance, outlier component ($s=2$). Component s has mixing coefficient $\boldsymbol{\pi}_s$, mean 0 and precision (inverse variance) $\boldsymbol{\beta}_s$. We can write the parameters collectively as the vectors $\boldsymbol{w}$, $\boldsymbol{\pi}=[\pi_1, \pi_2, \ldots, \pi_m]$, and $\boldsymbol{\beta}=[\beta_1, \beta_2, \ldots, \beta_m]$. We concatenate all the parameters into the overall vector $\boldsymbol{\theta}=\{\pi, \beta, w\}$.

Using the above equations, and the generative model shown in Fig. 1, we can generate data from a mixture process. As previously mentioned, Wager et al. (Wager et al., 2005) followed this procedure to *generate* data in a simulation study which, they imply, caricatures the essence of signal embedded in artefactual data. They then compared the performance of standard robust estimation methods on this data. In this paper we take the mixture model more literally and ask the question, assuming that data were generated from such a mixture process what are the optimal parameter *estimation* and statistical *inference* procedures.

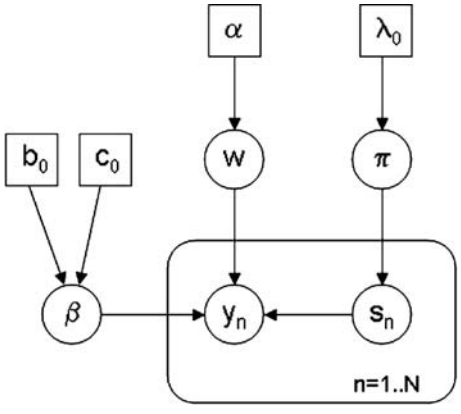

Fig. 1. Generative model with circles denoting random variables and squares denoting constants. An $[N \times p]$ design matrix $\mathbf{X}$ multiplies a $[p \times 1]$ vector of regression coefficients $\boldsymbol{w}$ to produce model prediction $\mathbf{X}\boldsymbol{w}$. Samples $\boldsymbol{y}_n$ are then formed by adding noise, $\boldsymbol{e}_n$, from a mixture distribution. Sample n is drawn from mixture component s if label $s_n=s$. These labels are drawn with class probabilities $\boldsymbol{\pi}$, an $[m \times 1]$ vector. Each noise sample has zero mean and a precision given by the appropriate entry in $\boldsymbol{\beta}$, an $[m \times 1]$ vector. The hyperparameters α, λ_0, b_0 and c_0 are set so as to produce vague priors (see Appendix B).

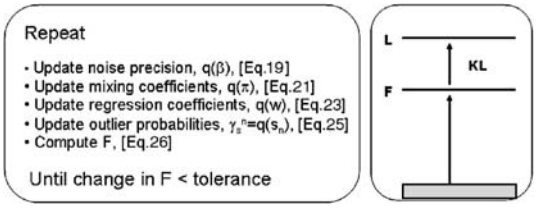

Fig. 2. Pseudo-code for Bayesian fitting of the RGLM model. The parameters of the model are estimated by updating the approximate posteriors, $q()$, until the negative free energy, F, is maximized to within a certain tolerance (left panel). At this point, because the log evidence, $L=\log p(Y)$, is fixed, the approximate posteriors will best approximate the true posteriors in the sense of KL divergence (right panel), as described in Appendix E.

This question is framed within the context of Bayesian inference. The optimal estimation and inference procedures are described in the appendices to this paper. These comprise descriptions of the model likelihood (Appendix A), priors over model parameters (Appendix B), and approximate inference procedures based on variational Bayes (Appendices C and D). We also derive an approximation to the model evidence, $p(y|m)$ (Appendix E). This allows for Bayesian model comparison and will be used to select how many mixture components to use in the RGLM.

Our approximation to the model evidence, as described in Eq. (26) in Appendix E, comprises two terms. The first term, the average likelihood, can be thought of as the accuracy of the model. The second term, composed of Kullback–Liebler (KL) divergences, describes the complexity of the model. Thus, good models have to both fit the data well and be as simple as possible. If two models fit the data equally well, then the simpler one (e.g. one with fewer parameters) will be preferred. This trade-off is also embodied in frequentist model comparisons, where extra degrees of freedom must result in better model fits to get higher statistic values (Kleinbaum et al., 1988).

Fig. 2 shows pseudo-code for Bayesian fitting of the RGLM. This updates the approximate posterior distributions over model parameters (see Appendix D) so as to maximize a lower bound on

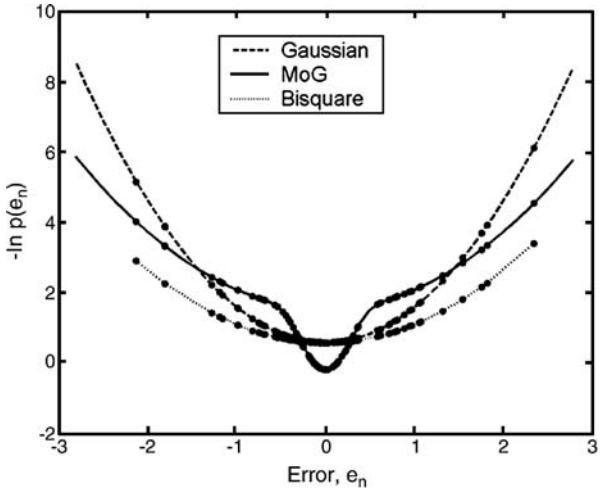

Fig. 3. Cost functions for different error models.

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058

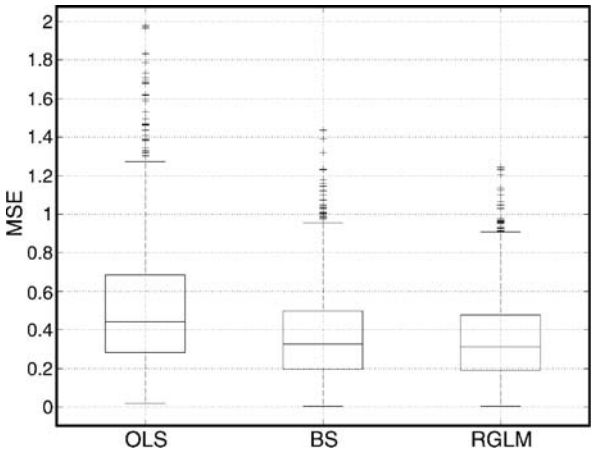

Fig. 4. Box and whisker plots showing mean squared error (MSE) in estimating the first regression coefficient. The boxes have lines at the lower, median and upper quartile values. The whiskers extend out to the most extreme value within a distance of one and a half times the interquartile range from the box. Data points outside the whiskers are drawn as crosses.

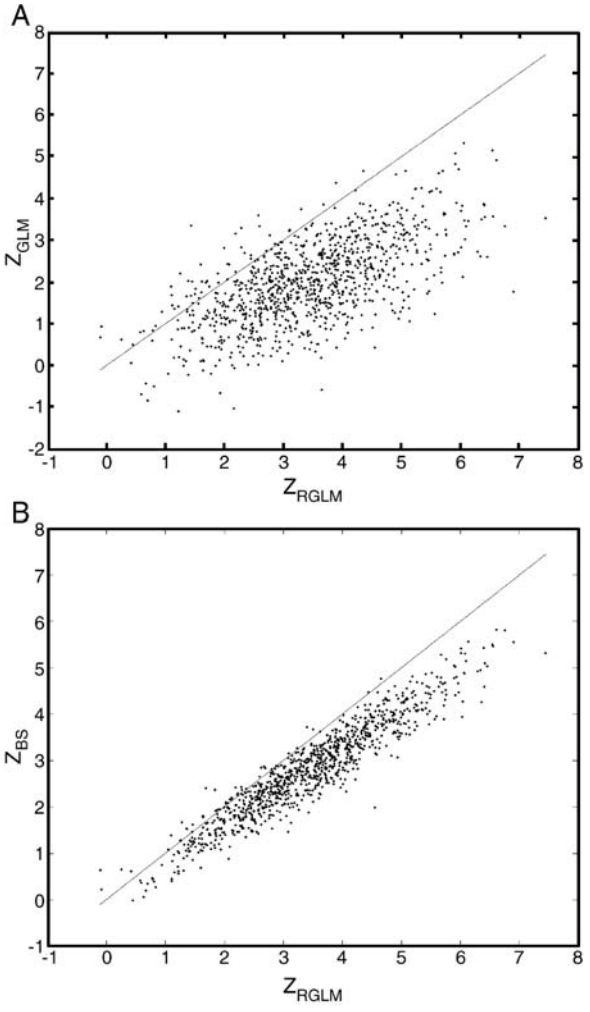

Fig. 5. Sensitivity on data containing outliers: *Z*-statistics for RGLM versus (A) GLM and (B) Bisquare show RGLM to be the most sensitive method.

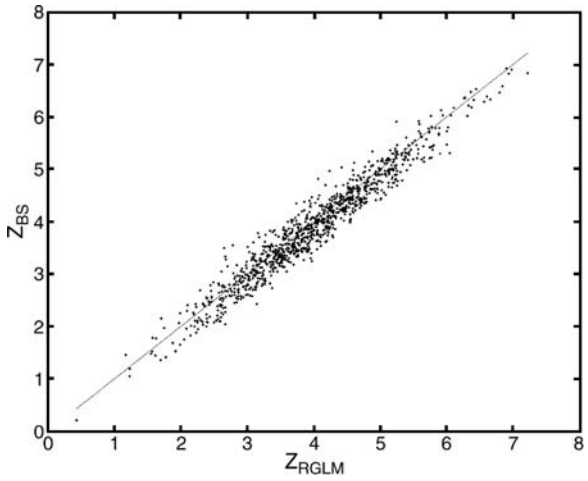

Fig. 6. Sensitivity on normal data: *Z*-statistics for RGLM versus Bisquare. A significantly ($p<1e-6$) higher proportion of *Z* scores are higher for RGLM than Bisquare.

the model evidence, *F*. As the update for the regression coefficients (Eq. (22)) is computationally more intensive than the updates for the other parameters (as it involves a matrix inversion) we perform this step only once every W_t iterations. The value of W_t only affects the computer time taken during estimation and in our experiments we used $W_t=5$. We evaluate *F* every W_t iterations (after the regression updates) and terminate optimization if the proportionate increase from one evaluation to the next is less than a tolerance value of 0.01%.

Summary of method

Before presenting applications of the method, we briefly describe the approach for readers unfamiliar with the above Bayesian terminology.

Occasionally, fMRI time series are corrupted with outliers, as can be seen in Fig. 12. Fitting a GLM to this data results in inflated estimates of the error variance. This, in turn, leads to smaller *Z* scores and a loss of sensitivity.

By fitting GLMs with mixture error processes, outliers can be soft-assigned to an outlier class. This results in an estimate for the error covariance *V* that allows for weighted least squares estimation

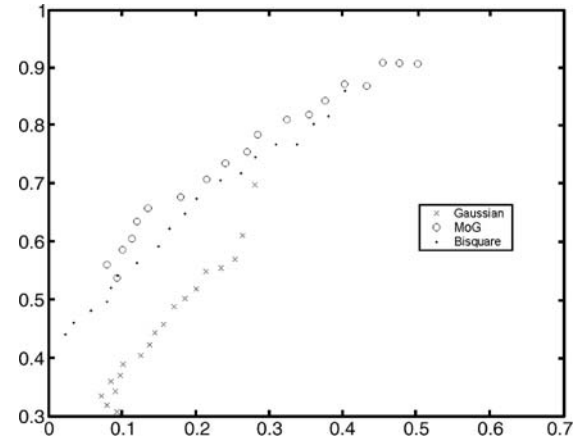

Fig. 7. Receiver Operating Characteristic (ROC) curves.

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058

of the regression coefficients $\hat{w}_{WLS}=(X^TV^{-1}X^T)^{-1}\ X^TV^{-1}y$ in which outlier samples are downweighted. In principle, this can improve the sensitivity at both the first (subject) and second (group) level.

In algorithmic terms, the RGLM approach is summarized as follows. Fit GLMs with mixture errors according to the pseudo-code in Fig. 2 for models with $m=1$ and $m=2$ error components. These are referred to as Mix-1 and Mix-2 GLMs. The one with the highest evidence is then referred to as the RGLM and is used for subsequent inference. This allows RGLM to default to the standard GLM for data without outliers.

Results

Exemplar data

This section compares the standard GLM, robust regression using a Bisquare cost function and the RGLM using synthetic data. The aim of this section is to demonstrate the potential of the RGLM approach.

Data were generated from a GLM with a design matrix **X** comprising two regressors (i) a boxcar of period 10 samples and (ii) a constant column. The regression coefficients were set to be

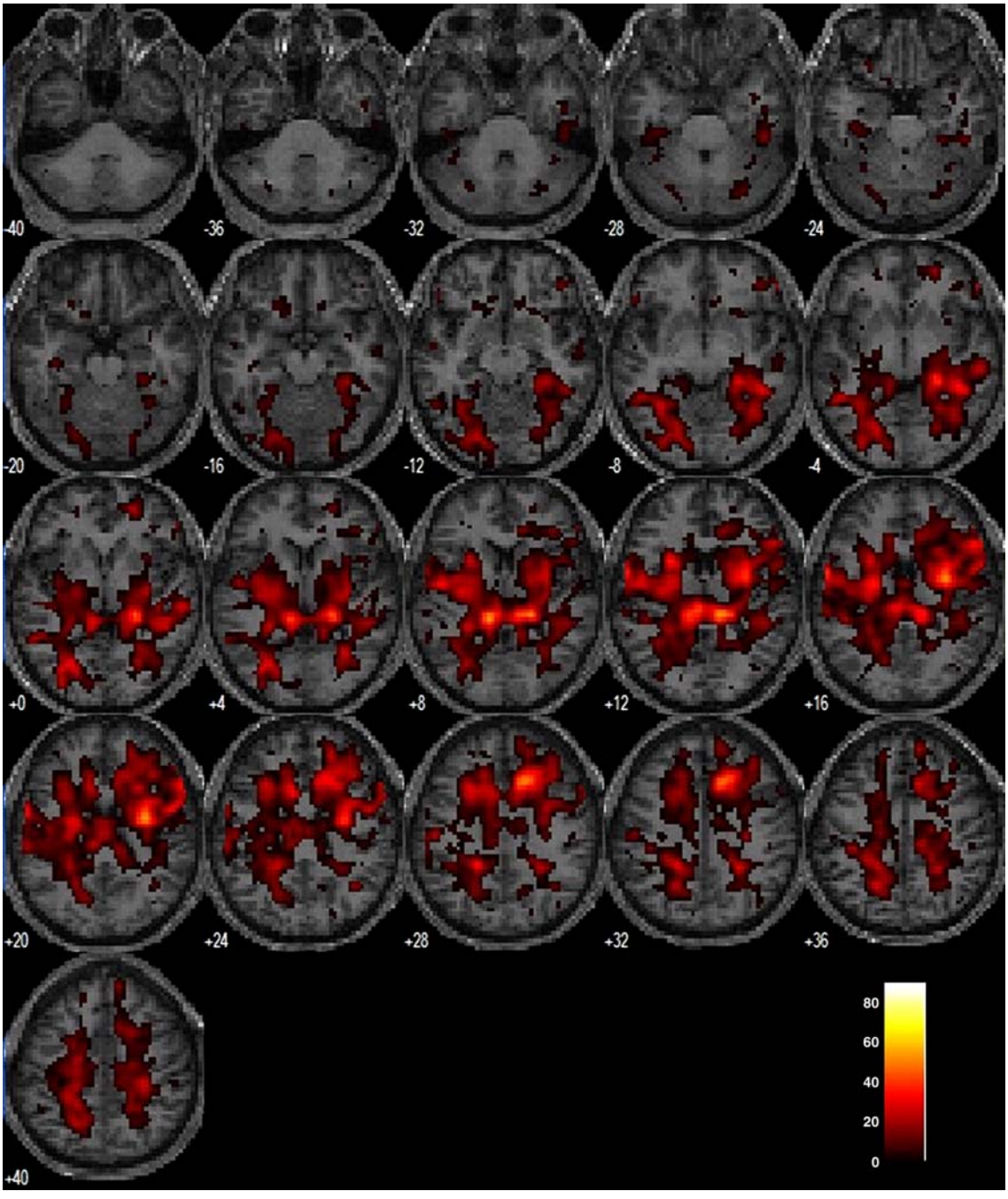

Fig. 8. Log Bayes Factor, BF_{21}, showing voxels where Mix-2 GLM is preferred to Mix-1 GLM.

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058

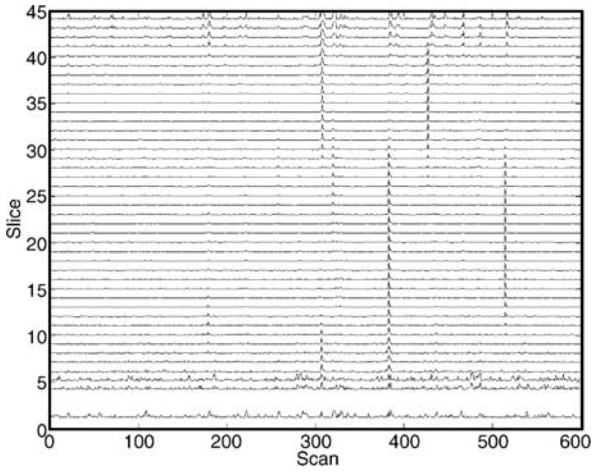

Fig. 9. Outlier probability, averaged within slice, as a function of slice number and scan, $<\gamma_2>$. This shows clear spatial variability. For slices 15 to 20, scans 383 and 514 are regarded as outliers. Moving towards the bottom of the brain, scans 179 and 307 appear problematic but scan 514 less so. For slices 30 to 40, scans 308 and 427 are the outliers. At the top and bottom of the brain the outlier pattern is more complex but does show local spatial homogeneity.

$\boldsymbol{w}=[1,\ 1]^T$ and the errors were drawn from a two-component mixture process.

So that the simulations are realistic, the parameters of the mixture process were set identical to those observed in a preliminary analysis of fMRI time series from a 'face fMRI data set', see Penny and Kilner (2006) for details. The mixing proportions were $\boldsymbol{\pi}=[0.73,\ 0.27]^T$, and standard deviations were $\sqrt{1/\boldsymbol{\beta}} = [2.4, 8.4]^T$.

We generated $T=1000$ data sets, each containing $N=351$ samples, and fitted GLMs, GLMs with Bisquare cost functions, and the RGLM. For all 1000 generated data sets, the mixture GLM with 2 components had higher model evidence than the single-component GLMs. RGLM therefore used the two-component GLM in all cases.

Fig. 3 shows the implied cost function for each approach. The Gaussian cost function, used in the GLM, pays the highest cost for large errors. This means that outliers have a big influence on signal estimation. Bisquare plays an increasingly smaller cost for larger errors. The RGLM approach, which employs an MoG-2 error model, has two operating regimes, each defined by a separate Gaussian. The narrower Gaussian allows RGLM to pay a higher-cost for larger error signal-bearing samples than does GLM or Bisquare. This is the mechanism by which RGLM can provide higher sensitivity as signal samples have a greater influence on regression coefficient estimation.

Fig. 4 shows boxplots of the squared error in the estimate of the first regression coefficient. Overall, RGLM is the most accurate method with, on average 115% smaller error than the GLM and 15% smaller error than the Bisquare approach. The sensitivity of the approaches can be assessed by computing Z-statistics, the ratio of effect size to effect standard deviation. Fig. 5 shows that larger Z-statistics are obtained for the RGLM model which therefore offers greater sensitivity.

We also compared the sensitivity of RGLM and Bisquare approaches on data with purely Gaussian errors. To this end, we repeated the above simulations but the noise on each sample was drawn from a single zero-mean Gaussian with variance 2.4. For all 1000 generated data sets, the mixture GLM with one component had higher model evidence than the two-component GLM. RGLM therefore defaulted to the standard GLM in all cases. Fig. 6 compares RGLM and Bisquare Z scores. On average the RGLM Z scores are 3% higher. For 757 out of 1000 of these data sets, the RGLM scores were higher. This is a significantly higher proportion ($p<1e{-}6$) than is to be expected if the two methods are equally sensitive.

We finish this section by addressing the sensitivity/specificity trade-off. The slopes of the data points in Fig. 5 suggest that the RGLM may lack specificity. To address this issue we performed the following simulation. We generated $T=10{,}000$ data sets in which the first regression coefficient was drawn from a uniform distribution between 0 and 1. Other than this, the parameters were the same as in the previous simulation. A threshold, t_w, was then defined such that values below this corresponded to a null hypothesis. True positives were then deemed to occur for $w_t \geq t_w$, $\hat{w}_t \geq t_w$ and false positives for $w_t < t_w$, $\hat{w}_t \geq t_w$. Receiver Operating Characteristic (ROC) curves were then formed by varying t_w and plotting true positive rate (sensitivity) versus false positive rate (1-specificity). Fig. 7 compares ROC curves for each method, showing that RGLM has higher sensitivity over a broad range of specificities.

Mismatch Negativity fMRI

This data set was acquired simultaneously with EEG, during a Mismatch Negativity study using a roving paradigm. Stimuli comprised a sequence of tones, with inter-stimulus interval 650 ms, whose frequency was changed after between 2 and 32 repetitions. The frequencies followed a random walk comprising 15 different tones in the range 200 to 2000 Hz. The tones were therefore either 'oddballs', the first of a new frequency, or 'standards'. In this paper we model only the oddball responses.

Images were acquired from a 3 T Allegra system (Siemens, Erlangen, Germany) which produced T2*-weighted transverse Echo-Planar Images (EPIs) with BOLD contrast. Whole brain EPIs consisting of 34 transverse slices were acquired every 2.21 s resulting in a total of $N=606$ scans. The first 6 scans were

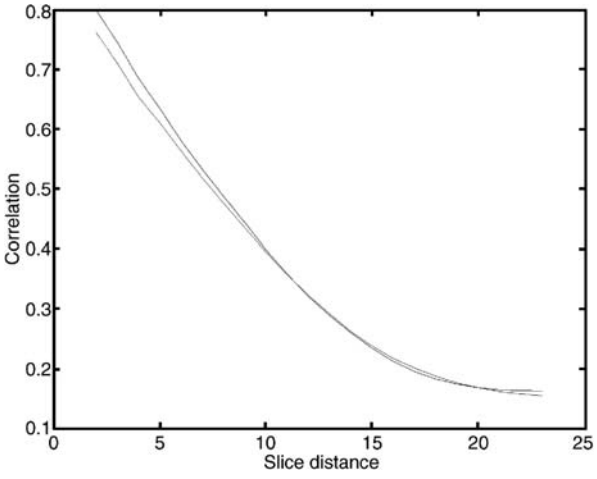

Fig. 10. Correlation in outlier probability as a function of number of slices apart, for all voxels (higher correlation at small distance) and gray matter voxels (lower correlation at small distance). The slice separation distance is 3 mm.

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058

discarded prior to subsequent analysis. All functional images were realigned to the first functional image using a six-parameter rigid-body transformation. This transformation embodies 'movement parameters', three translations and three rotations, which are later used in the design matrix. Images were then spatially normalized to a standard EPI template using a nonlinear warping method (Ashburner and Friston, 2003). Each time series was then high-pass filtered using a set of discrete cosine basis functions with a filter cut-off of 128 s.

The data were then analyzed with robust and standard GLMs using a design matrix comprising eight columns. The first column models oddball responses and was formed by convolving a canonical HRF with delta functions located at times when oddballs were presented. The following 6 columns contain the movement parameters, and the final column contains a vector of 1's to model the average response at each voxel.

Fig. 8 plots the Log Bayes factor, BF_{21} (see Appendix E), overlaid on slices of the subjects structural image. Positive Bayes factors indicate that the Mix-2 GLM is favored over the Mix-1 GLM. Over the whole volume, Mix-2 is favored at 17,349 of the 68,448 voxels. That is 25.4% of voxels. Of these, 8938 are in gray matter.

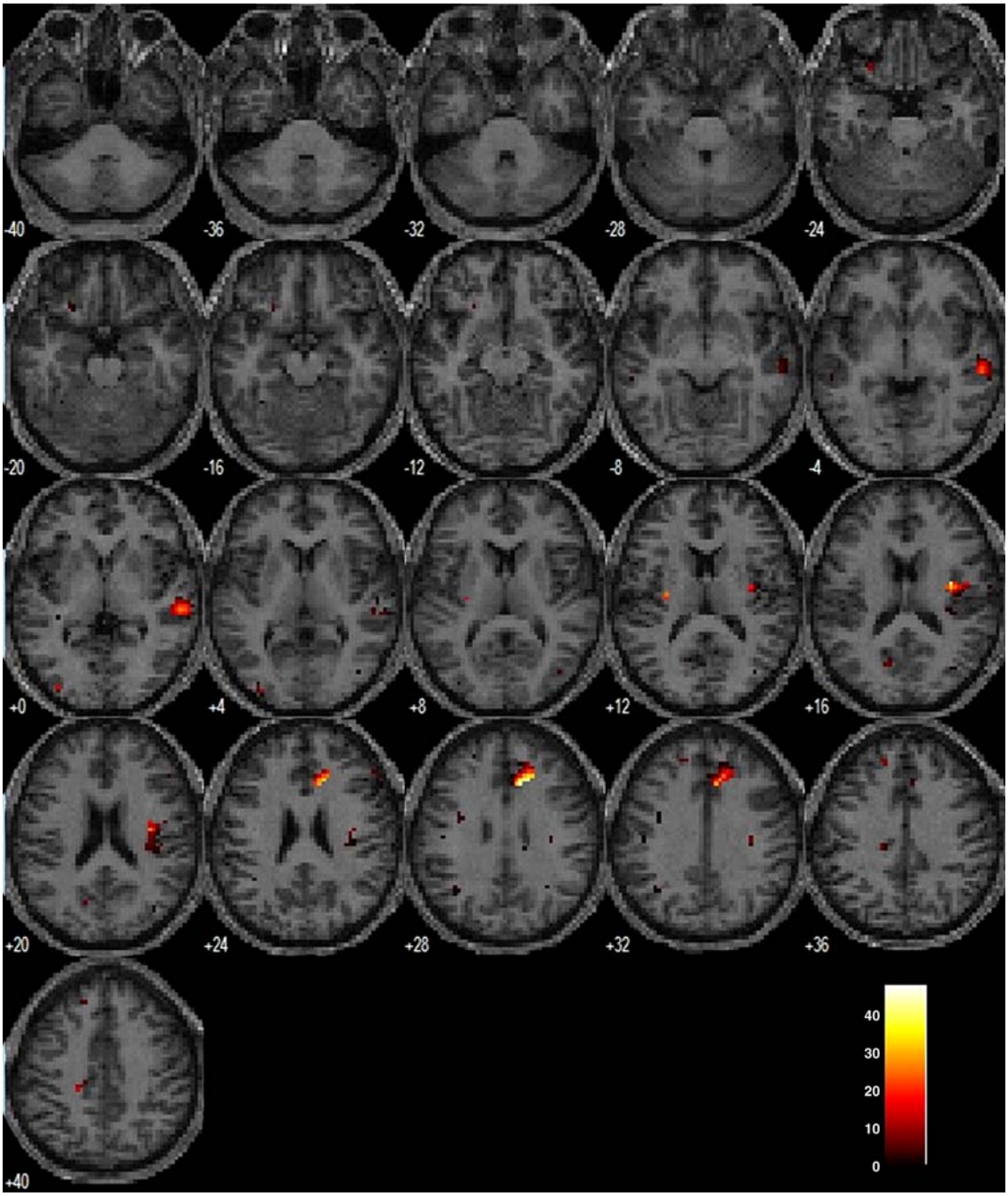

Fig. 11. Log Bayes Factor, BF_{21}, showing where Mix-2 GLM is preferred to Mix-1 GLM (as in Fig. 8) but restricted to voxels showing an oddball response ($p<0.05$, uncorrected).

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058

To characterize the spatial variability of the outliers we averaged the outlier probabilities, γ_2^n (see Appendix D), within slices. This averaging was restricted to gray matter voxels. Fig. 9 plots the average outlier probability as a function of slice and scan number. In different parts of the brain different scans are regarded as outliers. This spatial dependence is quantified in Fig. 10 which plots the correlation in average outlier probability as a function of distance between slices. This provides strong evidence of spatial heterogeneity.

Fig. 11 shows a map of the log Bayes Factor, BF_{21}, where Mix-2 is preferred to Mix-1, but restricted to voxels showing an oddball response ($p<0.05$, uncorrected). Over the whole volume this comprises 294 voxels.

Fig. 12 shows a time series at a voxel in auditory cortex ($x=60$, $y=-27$, $z=0$ mm) showing an oddball response, along with the outlier probability time series, γ_2^n, from the Mix-2 model. For the Mix-1 GLM the oddball effect size is 0.45, with standard deviation 0.15 giving rise to a Z score of 3.00. For the (favored) Mix-2 GLM the estimated effect size is 0.54, standard deviation 0.13, giving rise to a Z score of 4.08.

This improved sensitivity is evident in almost all of the 294 voxels. Fig. 13 compares the Z scores for RGLM and GLM methods. On average, the RGLM Z scores are 50% larger. This improved sensitivity is also shared by the Bisquare robust estimation method. The corresponding Z scores are shown in Fig. 14. On average the RGLM Z scores are 2% larger.

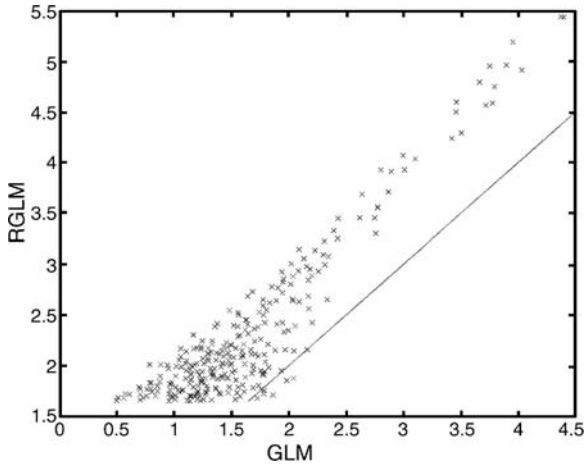

Fig. 13. Z scores for RGLM versus Z scores for GLM at voxels showing an oddball response ($Z_{RGLM}>1.65$, $p<0.05$) and where RGLM is the favored model. On average the RGLM Z scores are 50% larger.

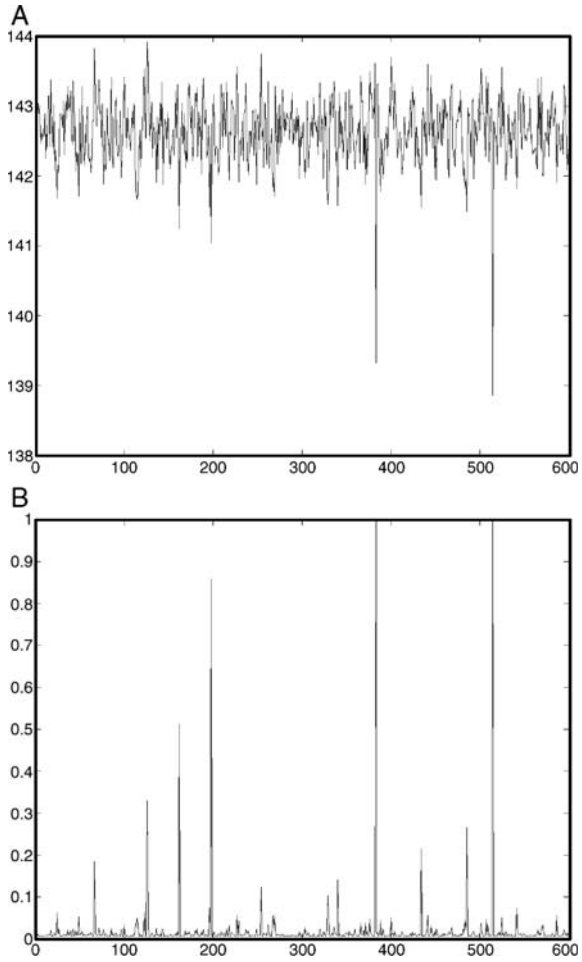

Fig. 12. Top: Time series at voxel in auditory cortex ($x=60$, $y=-27$, $z=0$ mm) showing oddball response. Bottom: Outlier probability time series γ^n from RGLM model.

Discussion

We have described a Bayesian learning algorithm for Robust General Linear Models (RGLMs), based on Roberts and Penny (2002), in which the noise is modeled as a Mixture of Gaussians. This allows different data points to be associated with different noise levels and effectively provides a robust estimation of regression coefficients.

A Bayesian inference framework is used to prevent overfitting and provides a model selection criterion for noise model order, e.g. to select noise mixtures with one, two or more components. This allows the RGLM to automatically default to the usual GLM when robustness is not required.

Our simulations, based on statistical characteristics of artefactual fMRI time series, suggest that the RGLM approach

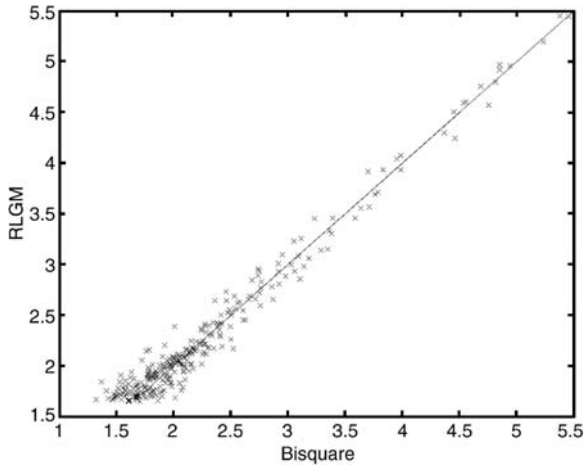

Fig. 14. Z scores for RGLM versus Z scores for Bisquare robust estimation at voxels showing an oddball response ($Z_{RGLM}>1.65$, $p<0.05$) and where RGLM is the favored model. On average the RGLM Z scores are 2% larger.

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058

can be more sensitive to underlying signal than the Bisquare robust estimation procedure (Wager et al., 2005). This is the case if the weighting scheme implied by the mixture error process is a better description of fMRI errors than is the Bisquare process. This includes the case where errors are purely Gaussian as RGLM defaults to a non-robust approach in the absence of outliers.

The advantage of the method over the Restricted Maximum Likelihood (ReML) approach described in Diedrichsen and Shadmehr (2005) is that the outlier profile is allowed to vary across space. That is, scans which are 'corrupted' in one part of the brain are not assumed to be corrupted in another. The fMRI data presented in this paper provide evidence that this is indeed a good assumption.

The statistical model employed in RGLM renders it particularly robust to impulsive noise. The model could be extended in a number of ways, primarily to account for temporally correlated noise sources. First, if the heteroscedasticity were autoregressive then a Generalized Autoregressive Conditional Heteroscedastic (GARCH) process would be appropriate. Alternatively, one could replace the mixture process with a Markov process.

A second area for further work is to embed RGLM into previous algorithms developed for fMRI. Incorporation of a spatial prior, for example, should improve ROC performance yet further (Penny et al., 2005). Standard robust estimation procedures, such as Bisquare, cannot be readily improved in such a manner.

A third area is to apply the method to group data, as in Wager et al. (2005). Before this is possible, however, the approach would need to be modified such that inferences would be based on t rather than Gaussian distributions. The use of Gaussian posteriors is a good approximation for first-level fMRI data, where the number of data points equals the number of scans, a typically large number. But this is a poor approximation at the second (group) level, where the number of data points is given by the number of subjects and is typically small. This modification to RGLM can be made by removing the factorization (in the prior and posterior) between regression coefficients and noise precisions (Box and Tiao, 1977).

We have applied RGLMs to fMRI data and concluded that our data contain artefacts but have made no claims about the origin of these artefacts. This 'phenomenological' approach to modeling is in the spirit of modeling fMRI errors using autoregressive processes. Neither of these approaches claim to model artefacts in any causal or mechanistic way. The approach is rather general in nature and we hope it can find application beyond fMRI.

Acknowledgments

The authors are funded by the Wellcome Trust and we are grateful to Mohamed Seghier and Jorn Diedrichsen for helpful comments.

Appendix A. Likelihood

The likelihood of a data point is given by the mixture model

$$p(\boldsymbol{y}_n|\boldsymbol{\theta}) = \sum_{s=1}^{m} p(\boldsymbol{y}_n|, s_n = s, \boldsymbol{\beta}_s, \boldsymbol{\mu}_s, \boldsymbol{w}) p(s_n = s|\boldsymbol{\pi}) \quad (3)$$

where s_n is a variable indicating which component is selected for which data point. A-priori these are chosen probabilistically according to

$$p(s_n = s|\boldsymbol{\pi}) = \boldsymbol{\pi}_s \quad (4)$$

Each component is a Gaussian with

$$p(\boldsymbol{y}_n|s_{n=s}, \boldsymbol{\beta}_s, \boldsymbol{\mu}_s, \boldsymbol{w}) = (2\boldsymbol{\pi})^{-1/2} \boldsymbol{\beta}_s^{1/2} \exp\left(\frac{-\boldsymbol{\beta}_s}{2}(\boldsymbol{y}_n - x_n \boldsymbol{w})^2\right) \quad (5)$$

The joint likelihood of a data point and indicator variable is

$$p(\boldsymbol{y}_n, s_n|\boldsymbol{\theta}) = p(s_n|\boldsymbol{\pi}) p(\boldsymbol{y}_n|s_n, \boldsymbol{\beta}_s, \boldsymbol{w}) \quad (6)$$

which, given that the data are Independent and Identically Distributed (IID), gives

$$p(Y, S|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(s_n|\boldsymbol{\pi}) p(\boldsymbol{y}_n|s_n, \boldsymbol{\beta}_s, \boldsymbol{w}) \quad (7)$$

over the whole date set, where $Y=[y_1, y_2, \ldots, y_N]^T$ and $S=[s_1, s_2, \ldots, s_N]^T$.

Appendix B. Priors

We specify prior distributions over model parameters to incorporate appropriate domain specific knowledge if available (e.g. approximate proportion of outliers).

The prior on the model parameters is

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) p(\boldsymbol{w}|\alpha) \prod_s p(\boldsymbol{\beta}_s) \quad (8)$$

where the mixing prior is a symmetric Dirichlet

$$p(\boldsymbol{\pi}) = \frac{\Gamma(m\lambda_0)}{\Gamma(\lambda_0)^m} \prod_{s=1}^{m} \boldsymbol{\pi}_s^{\lambda_{0-1}} \quad (9)$$

and $\Gamma(x)$ is the Gamma function (Press et al., 1992). This means that we assign λ_0 pseudo-counts to each component. This is readily extended to allow asymmetric distributions if, for example, we had prior information that there were typically fewer samples in the outlier class.

The prior over the precisions is a Gamma

$$p(\boldsymbol{\beta}_s) = \Gamma(\boldsymbol{\beta}_s; b_0, c_0) \quad (10)$$

where the Gamma density is defined as

$$\Gamma(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right) \quad (11)$$

The prior over regression coefficients is a zero-mean Gaussian with an isotropic covariance having precision α

$$p(\boldsymbol{w}|\alpha) = \left(\frac{\alpha}{2\boldsymbol{\pi}}\right)^{p/2} \exp\left(-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right) \quad (12)$$

To obtain a practical algorithm we must choose parameters for the prior distributions. For the experiments in this paper, we used vague priors so that inferences are largely unaffected by information external to the current data, although in future work (see Discussion) we envisage the use of informative priors based,

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058

e.g. on spatial models. To this end, we set $b_0=10^3$, $c_0=10^{-3}$ for $p(\boldsymbol{\beta}s)$ (giving a prior mean of 1 and prior variance of 1000), $\lambda_0=5$ for $p(\boldsymbol{\pi})$ to give a uniform distribution and set $\alpha=0.001$ (i.e. a very low prior precision on the regression coefficients).

Appendix C. Variational Bayes

We estimate model parameters using Variational Bayesian (VB) learning. The aim of VB is to maximize the negative free energy

$$F(m)=\left\langle \log\frac{p(Y,S|\boldsymbol{\theta})}{q(S)}\right\rangle_{S,\boldsymbol{\theta}}-\mathrm{KL}(q(\boldsymbol{\theta})\,||p(\boldsymbol{\theta})) \tag{13}$$

where $q(S)$ and $q(\boldsymbol{\theta})$ are approximate posterior distributions over hidden states and parameters.[1]

The first term corresponds to an average likelihood L_{av}, where the expectation is taken wrt. $q(S)$ and $q(\boldsymbol{\theta})$, and the second term is the Kullback–Liebler (KL) divergence between the approximate posteriors $q(\boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$. This objective function can be maximized via a two-step process (Attias, 2000). In the first step, $q(S)$ is updated according to

$$q(S)\propto\exp[I(S)] \tag{14}$$

where

$$I(S)=\langle\log p(Y,S|\boldsymbol{\theta})\rangle_{\boldsymbol{\theta}} \tag{15}$$

and the expectation is taken wrt. $q(\boldsymbol{\theta})$. In the second step $q(\boldsymbol{\theta})$ is updated according to

$$q(\boldsymbol{\theta})\propto\exp[I(\boldsymbol{\theta})]p(\boldsymbol{\theta}) \tag{16}$$

where

$$I(\boldsymbol{\theta})=\log\langle\log p(Y,S|\boldsymbol{\theta})\rangle_S \tag{17}$$

and the expectation is taken wrt. $q(S)$. The negative free energy, $F(m)$, is also a lower bound on the model evidence and can be used for model selection (see below).

C.1. Factorization

To apply VB to RGLMs we approximate the posterior distribution over parameters with the factorized density

$$q(\boldsymbol{\theta})=q(\boldsymbol{\pi})q(\boldsymbol{\beta})q(\boldsymbol{w}) \tag{18}$$

and the posterior distribution over hidden variables by $q(S)$. We then set each distribution so as to maximize $F(m)$. We omit the derivations due to lack of space but the procedure is similar to that used in (Penny et al., 2003). The optimal posteriors factorize into the parametric forms described in the following appendices. Each section describes how to update the sufficient statistics for each component of the approximate posterior. These updates may depend on parameters derived from sufficient statistics of other components. These quantities will then be defined in the relevant subsection below.

[1] In this nomenclature 'hidden states' are differentiated from 'parameters' by having as many instantiations as there are data points.

Appendix D. Approximate posteriors

D.1. Updating noise precisions

For the precisions we have

$$q(\boldsymbol{\beta})=\prod_s q(\boldsymbol{\beta}_s)$$

$$q(\boldsymbol{\beta}_s)=\Gamma(\boldsymbol{\beta}_s;b_s,c_s)$$

$$\frac{1}{b_s}=\frac{N}{2}\,\tilde{\sigma}_s^2+\frac{1}{b_0}$$

$$c_s=\frac{\bar{\mathrm{N}}_s}{2}+c_0$$

$$\tilde{\sigma}_s^2(n)=(\boldsymbol{y}_n-\hat{\mathrm{y}}_n)^2+x_n^T\hat{\mathbf{C}}x_n$$

$$\tilde{\sigma}_s^2=\frac{1}{N}\sum_{n=1}^{N}\gamma_s^n\,\tilde{\sigma}_s^2(n)$$

$$\bar{\mathrm{N}}_s=\sum_{n=1}^{N}\gamma_s^n$$

$$\bar{\beta}_s=b_sc_s \tag{19}$$

where $\hat{y}_n=x_n\hat{w}$ is the RGLM prediction for the nth data point. The quantity $\bar{\mathrm{N}}_s$ is the number of data points attributed to component s and $\tilde{\sigma}_s^2$ is the expected variance of component s. The quantities b_0 and c_0 are parameters of the prior distribution and b_s and c_s of the posterior distribution. The quantity $\hat{\boldsymbol{C}}$ is the posterior covariance matrix over regression coefficients (see Appendix D.3) and γ_s^n is the posterior probability that sample n belongs to mixture component s (see Appendix D.4).

We can understand these equations by looking at the corresponding mean variance (the inverse of the mean precision) which is given by $1/(b_sc_s)$. If we ignore terms involving the prior this comes out to be

$$\frac{\sum_{n=1}^{N}\gamma_s^n\,\tilde{\sigma}_s^2(n)}{\sum_{n=1}^{N}\gamma_s^n} \tag{20}$$

which is the expected variance of that component re-weighted according to the number of examples that the component is responsible for.

D.2. Updating mixing coefficients

For the mixing coefficients we have a Dirichlet

$$q(\boldsymbol{\pi})=\Gamma\left(\sum_{s'=1}^{m}\lambda_{s'}\right)\prod_{s=1}^{m}\frac{\boldsymbol{\pi}_s^{\lambda_s-1}}{\Gamma(\lambda_s)}$$

$$\lambda_s=\bar{\mathrm{N}}_s+\lambda_0 \tag{21}$$

The mixing hyperparameters, $\bar{\lambda}_s$, are updated by adding the data counts, $\bar{\mathrm{N}}_s$, to the prior counts, λ_0.

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058

D.3. Updating regression coefficients

For the regression coefficients we have

$$q(w) = N(w; \hat{\mathrm{w}}, \hat{\mathrm{C}})$$

$$V^{-1} = \sum_{s=1}^{m} \bar{\beta}_s \Gamma_s$$

$$\hat{\mathbf{C}} = \left(X^T V^{-1} X + \alpha\right)^{-1}$$

$$\hat{\mathrm{w}} = \hat{\mathrm{C}} X^T V^{-1} y \tag{22}$$

where V is our estimate of the error covariance matrix and $\Gamma_s = \mathrm{diag}([\gamma_s^1, \gamma_s^2, \ldots, \gamma_s^N])$.

Without a prior, $\alpha = 0$, the above equation reduces to weighted least squares, and with a single noise component, $m = 1$, this reduces to the least-squares estimate

$$\hat{\mathrm{w}}_{LS} = (X^T X)^{-1} X^T y \tag{23}$$

D.4. Updating indicators

For the indicator posteriors we have

$$q(S) = \prod_n q(s_n)$$

$$\gamma_s^n \equiv q(s_n) \tag{24}$$

The (approximate) posterior probability that component s is responsible for data point $\boldsymbol{y}_n$ is then updated using

$$\tilde{\gamma}_s^n = \tilde{\boldsymbol{\pi}}_s \beta_s^{1/2} \exp\left[-\frac{1}{2}\bar{\beta}_s \tilde{\sigma}_s^2(n)\right]$$

$$\gamma_s^n = \frac{\tilde{\gamma}_s^n}{\sum_{s'} \tilde{\gamma}_{s'}^n}$$

$$\tilde{\beta}_s = \exp(\Psi(c_s) + \log b_s)$$

$$\tilde{\boldsymbol{\pi}}_s = \exp\left(\Psi(\lambda_s) - \Psi\left(\sum_{s'} \lambda_{s'}\right)\right) \tag{25}$$

where $\Psi()$ is the digamma function (Press et al., 1992). If $s = 2$ corresponds to the outlier class, then γ_s^n is the probability that sample n is an outlier.

D.5. Initialization

The posterior distributions $q(\boldsymbol{\pi})$, $q(\boldsymbol{\beta}_s)$, and $q(w)$ have the parameters λ_s, b_s, c_s, $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{C}}$ which are initialized as follows. The posterior for the regression coefficients (with sufficient statistics $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{C}}$) is initialized using the Maximum Likelihood (ML) solution. The remaining posteriors are set as follows.

We first calculate the errors $e(n)$ from the ML model and then define a new variable $z(n) = |e(n) - \bar{\mathrm{e}}|$ which is the absolute deviation of each error from the mean error $\bar{\mathrm{e}}$. We then apply k-means clustering (Bishop, 1995) to $z(n)$ which results in mixing coefficients $\lambda_z(s)$ and means $m_z(s)$. We then set $\lambda_s = 100\lambda_z(s)$. The parameters b_s and c_s are then set so as to achieve means of $(1/m_z)(s)^2$ and variances of $\mathrm{Var}(1/m_z)$ (the mean and variance of a Gamma density are bc and b^2c respectively).

Appendix E. Free energy and model comparison

The negative free energy is computed as

$$F(m) = L_{\mathrm{av}} - \mathrm{KL}(\boldsymbol{w}) - \mathrm{KL}(\boldsymbol{\pi}) - \mathrm{KL}(\boldsymbol{\beta}) \tag{26}$$

where

$$L_{\mathrm{av}} = H(q(S)) + \sum_{s=1}^{m} \bar{\mathrm{N}}_s(\log \tilde{\boldsymbol{\pi}}_s + 0.5 \log \tilde{\beta}_s) - 0.5N \sum_{s=1}^{m} \bar{\beta}_s \tilde{\sigma}_s^2 - 0.5N \log 2\boldsymbol{\pi} \tag{27}$$

The entropy over the hidden variables is

$$H(q(S)) = -\sum_{n=1}^{N} \sum_{s=1}^{m} \gamma_s^n \log \gamma_s^n \tag{28}$$

The KL terms for Normal and Gamma densities can be computed from the equations given in Penny et al. (2003). The first term in $F(m)$, the average likelihood, can be thought of as the accuracy of the model. The second term, composed of Kullback–Liebler (KL) divergences, describes the complexity of the model. This is because KL increases as, e.g. the dimension of $\boldsymbol{\pi}$ increases. Thus, good models have to both fit the data well and be as simple as possible.

Models i and j can then be compared using the Bayes factor

$$\mathrm{BF}_{ij} = \frac{p(y|m=i)}{p(y|m=j)} \tag{29}$$

As the negative free energy can be used as a surrogate for the log evidence, i.e. $F(m) \approx \log p(y|m)$, we can write the log Bayes factor as

$$\log \mathrm{BF}_{ij} = F(i) - F(j) \tag{30}$$

This will be used to select how many mixture components to use in the RGLM.

References

Ashburner, J., Friston, K.J., 2003. Spatial normalization using basis functions. In: Frackowiak, R.S.J., Friston, K.J., Frith, C., Dolan, R., Friston, K.J., Price, C.J., Zeki, S., Ashburner, J., Penny, W.D. (Eds.), Human Brain Function, 2nd edition. Academic Press.

Attias, H., 2000. A variational Bayesian framework for graphical models. In: Leen, T., et al. (Eds.), NIPS, vol. 12. MIT Press, Cambridge, MA.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Box, G.E.P., Tiao, G.C., 1977. Bayesian Inference in Statistical Analysis. Addison Wesley.

Diedrichsen, J., Shadmehr, R., 2005. Detecting and adjusting for artifacts in fMRI time series data. NeuroImage 27, 624–634.

Greve, D.N., White, N.S., Gadde, S., 2006. Automatic spike detection for fMRI. NeuroImage. Proceedings of the 12th Annual meeting of the Organization of Human Brain Mapping, Florence, Italy, vol. 31, p. S1.

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058

Jung, T., Makeig, S., Humphries, C., Lee, T., McKeown, M.J., Iragui, V., Sejnowski, T.J., 1999. Removing electroencephalographic artifacts by blind source separation. Psychophysiology.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E., 1988. Applied Regression Analysis and Other Multivariable Methods. PWS-Kent, Boston.

Luo, Nichols, T., 2003. Diagnosis and exploration of massively univariate neuroimaging methods. NeuroImage 19, 1014–1032.

Penny, W.D., Kilner, J.M., 2006. Robust Bayesian general linear models. In: Corbetta, M., Nichols, T., Pietrini, P. (Eds.), 12th Annual Meeting Human Brain Mapping. NeuroImage, vol. 31, p. S1.

Penny, W.D., Kiebel, S.J., Friston, K.J., 2003. Variational Bayesian inference for fMRI time series. NeuroImage 19 (3), 727–741.

Penny, W.D., Trujillo-Bareto, N., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24 (2), 350–362.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.V.P., 1992. Numerical Recipes in C. Cambridge.

Roberts, S.J., Penny, W.D., 2002. Variational Bayes for generalised autoregressive models. IEEE Trans. Signal Process. 50 (9), 2245–2257.

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

Zhang, X., Van De Moortele, P.F., Pfeuffer, J., Hu, X., 2001. Elimination of k-space spikes in fMRI data. Magn. Reson. Imaging 19, 1037–1041.

Please cite this article as: Penny, W.D., et al., Robust Bayesian general linear models, NeuroImage (2007), doi:10.1016/j.neuroimage.2007.01.058