A Bayesian Hierarchical Model for Photometric Redshifts

Merrilee Hurn*, Peter J. Green†, and Fahimah Al-Awadhi ‡

March 6, 2007

Abstract

The Sloan Digital Sky Survey (SDSS) is an extremely large astronomical survey conducted with the intention of mapping more than a quarter of the sky (`http://www.sdss.org/`). Among the data it is generating are spectroscopic and photometric measurements, both allowing estimation of the redshift of galaxies. The former is precise but expensive to gather, the latter is far cheaper but correspondingly gives far less accurate estimates. A recent paper by Csabai *et al.* (2003) describes various calibration techniques aiming to predict spectroscopic redshift from photometric measurements. In this paper, we investigate what a structured Bayesian approach to the problem can add. In particular, we are interested in providing uncertainty bounds associated with the underlying redshifts and the classifications of the galaxies. We find that a quite generic statistical modelling approach, using for the most part standard model ingredients, can compete with much more specific custom-made and highly-tuned techniques already available in the astronomical literature.

Keywords: Bayesian Modelling, Calibration, Hierarchical Modelling, Markov Chain Monte Carlo, Photometric Redshifts

1 Introduction

The Sloan Digital Sky Sky (SDSS) is a huge international collaboration. It describes itself as the most ambitious astronomical survey ever undertaken (`http://www.sdss.org/`), and eventually it will provide optical images covering more than a quarter of the sky. It will also provide a 3-dimensional map of roughly a million galaxies and quasars. The key to producing a 3-dimensional map from 2-dimensional

*Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK; Email: M.A.Hurn@bath.ac.uk; Tel.: +44 (0)1225 386001

†Department of Mathematics, University of Bristol, Bristol, BS8 1TW, UK; Email: P.J.Green@bristol.ac.uk; Tel.: +44 (0)117 928 7967

‡Department of Statistics and OR, Kuwait University, Kuwait, 13060; Email: falawadi@kuc01.kuniv.edu.kw; Tel.: +965 4837332

images of galaxies lies in measuring their *redshift*. The redshift measures how much light emitted from the galaxy has been shifted in wavelength by the speed at which the galaxy is moving relative to Earth (a more familiar concept would be that of the Doppler effect changing the apparent pitch of an ambulance's siren as it moves towards and then away from an observer). Once the redshift is known, the distance of the galaxy from Earth can be inferred using standard cosmological theory. The SDSS uses a spectograph to measure redshift; this technique measures the light the galaxy emits at different wavelengths in the range 3800Å(blue) to 9200Å(near infrared). By searching the spectra for certain characteristic features, the redshift can be measured. Such measurements are known as spectroscopic redshifts; they are believed to be very accurate but are expensive to collect. As a result there has been much interest in calibrating cheaper photometric data to predict redshift. Where spectroscopic data are accurate measurements of a whole section of the emission spectrum, photometric data are gathered by binning the emissions received in a small number of wavelength windows; in the case of the SDSS data 5 filter bins are used. As the bins are quite wide, reasonable signal-to-noise ratios can be achieved with short exposure times, so this form of data is far cheaper to collect.

A recent paper by Csabai *et al.* (2003) provides background to the use of photometric redshifts, and reviews the main existing approaches to deriving photometric redshifts. These approaches can be categorised into empirical or template based. Of the former, one approach is nearest neighbour matching where galaxies are assigned the spectroscopic redshift value of the galaxy whose photometric data most closely resemble its own (and efficient search techniques are required for this as the number of galaxies involved is huge). The second main empirical method is multiple regression, regressing the spectroscopic redshift data on second or third order polynomial functions of the photometric data. The template matching approaches use theoretical "spectral energy distributions" (SEDs) which have been derived to show the shape of the emissions at rest for various types of galaxies. For example, Figure 1 shows the templates proposed by Coleman, Wu and Weedman (1980) for galaxies of the four types Elliptical, Irregular, Barred Spirals and Spirals (E, Im, Sbc, Scd). Photometric redshift and galaxy type are then estimated by a least squares fit of the data to the observations predicted by these spectra. The paper by Csabai *et al.* (2003) also proposes some iterative techniques for improving the template matching algorithm.

What are we trying to achieve by a Bayesian approach to the problem? One of the drawbacks of the existing approaches seems to be a lack of both interpretability (in terms of the polynomial regression and nearest neighbour methods) and uncertainty estimates. A hierarchical Bayesian model will allow us to capture more of the structure of the application, and to generate estimates of redshift which include uncertainty over galaxy type, brightness, measurement error. A second key motivation is to examine the extent to which a quite generic statistical modelling approach, using for the most part standard model ingredients,

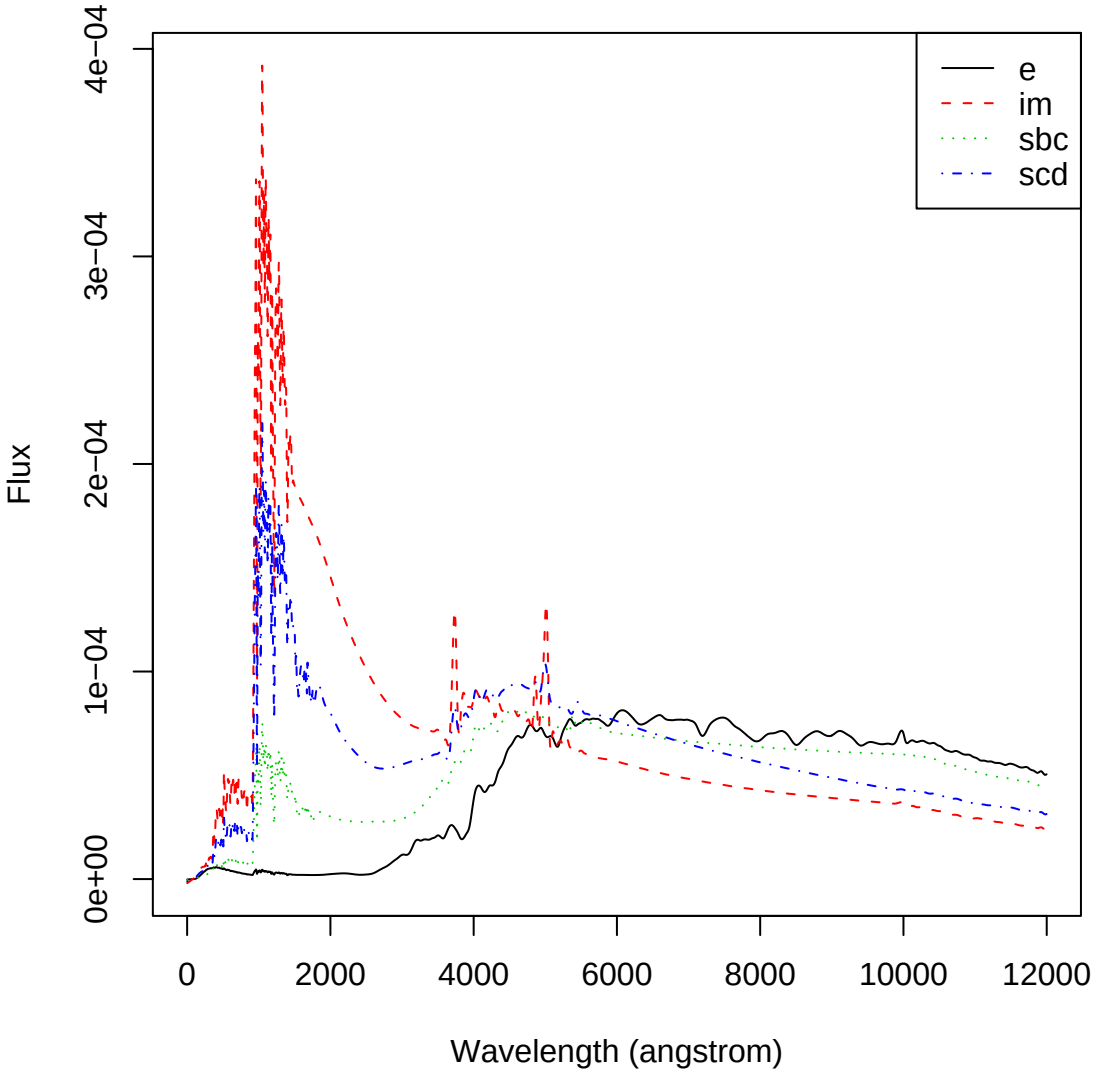

Figure 1: The four CWW galaxy spectra energy distributions (SED)

can compete with much more specific custom-made and highly-tuned techniques already available in the astronomical literature.

In the following section, we set up the structured statistical model we will assume, with particular emphasis on prior specification, and describe the MCMC computational techniques we use to fit the model and produce the required predictions. We will begin analysis of the model's performance in Section 3, with a small scale calibration problem, selecting a small subset of the data to formulate and evaluate the model and MCMC algorithm. The priors used at this stage are fairly vague although the form of them, in particular that for the brightness of the galaxies as a function of redshift, is driven by the application. With the model and algorithm tuned, in Section 4 we then evaluate the model in the way it might be used in practice, that is fitting using a large number of galaxies to provide very precise priors for the model parameters when estimating redshift for additional galaxies with only photometric data.

2 The calibration problem

2.1 Available data

The data we are using in this study come from the Early Data Release of the Sloan Digital Sky Survey. The larger of the two sets, which we denote MAIN, contains 15477 galaxies, the smaller, denoted LRG, contains a further 7861 galaxies. This smaller set is composed of galaxies which have been selected as having the characteristics of elliptical galaxies with high redshifts (the set's name, Luminous Red Galaxies, indicates that these are galaxies with a red appearance). The MAIN sample has no such selection process. For each galaxy i in the data sets we have both the spectroscopic measurement of redshift, y_i, and the five photometric measurements, $x_{i1}, \ldots, x_{i5}$. Figure 2 shows a single set of photometric data, as well as 2000 MAIN and LRG samples.

2.2 Modelling

Photometric data are recorded on the magnitude scale, with magnitude measured relative to a designated bright source. The relationship between magnitude and the flux emitted by a galaxy is

$$\begin{aligned} M &= -2.5 \log_{10}(F/F_0) \\ &= -2.5 \log_{10} F + C \end{aligned} \tag{1}$$

where F_0 is the flux of the "zero-point" source, and C is a constant (Smith (1995)). The plots in Figure 1 represent the shape of the emitted flux; the actual flux will be a multiple of that template depending on the brightness of the galaxy. Denote the spectral energy distribution function for the l^{th} galaxy type by $\phi_l(\lambda)$

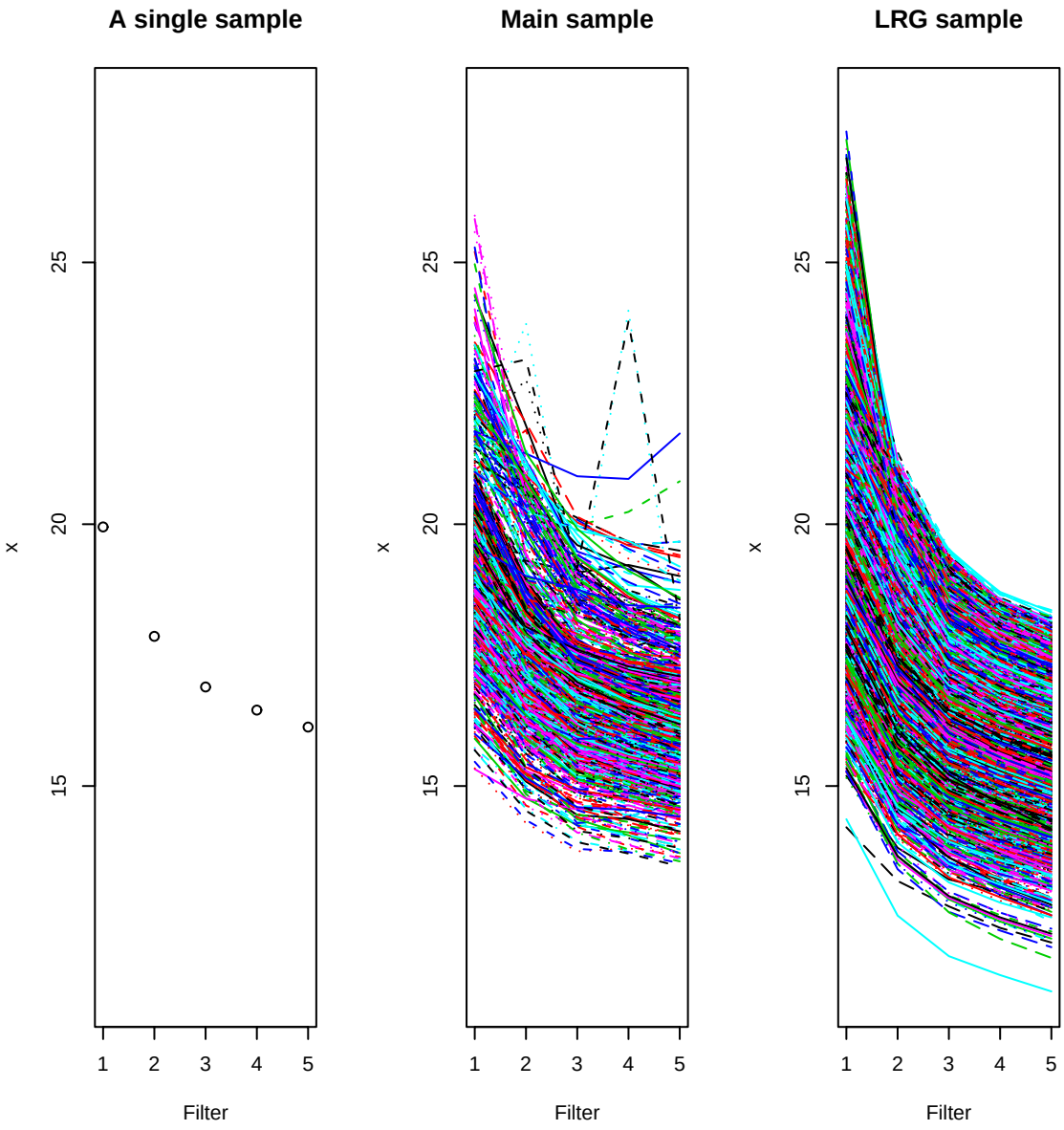

Figure 2: Plots of the photometric data x_{ij}.

where λ is the wavelength. The filter response functions $h_j(\lambda)$ for the five filters are known (see Figure 3). Then a straightforward physical argument using the definition of redshift z as the relative stretching of the wavelength scale due to motion of the source, convolution with the response function, and transformation using (1), implies that the unscaled response on the magnitude scale when a spectra is shifted by redshift z is given by

$$\psi_{lj}(z) \quad = \quad -2.5\log_{10}\int_{\lambda} h_j(\lambda)\phi\left(\frac{\lambda}{1+z}\right)d\lambda, \quad l=1,\ldots,4,\ j=1,\ldots,5 \tag{2}$$

These filtered redshifted templates are the key ingredient in our model for the observed photometric data. The ith astronomical object observed is regarded as a weighted linear combination of the elementary galaxy types l, using weights w_{il}, although in practice we generally impose the restriction that each observed object is of a single pure type so that for each i, $w_{il}=1$ for a single l. We also need to allow for differences in sensitivity in the different filters, and different underlying brightnesses among the observed objects. These effects are additive on the magnitude (logarithmic) scale, so the observational model proposed for the photometric measurements is

$$x_{ij} \quad = \quad a_j + b_i + \sum_{l=1}^{4} w_{il}\psi_{lj}(z_i) + \epsilon_{ij} \tag{3}$$

$$\text{with } \epsilon_{ij} \quad \sim \quad N(0,\sigma_j^2),\ \ i=1,\ldots,n,\ j=1,\ldots,5. \tag{4}$$

Here z_i is the true redshift of the i^{th} galaxy, b_i is a measure of galaxy brightness, and the $\{a_j\}$ allow for different filter sensitivities (constrained for identifiability to $\sum_{j=1}^{5} a_j = 0$). The error terms are mutually independent.

We also propose an observational model for how the spectroscopic redshift measurements arise; we simply assume independent symmetric Normal measurement errors

$$y_i|z_i \quad \sim \quad N(z_i,\sigma_y^2). \tag{5}$$

2.3 Prior specifications

The prior specification for the problem is a combination of quite standard choices assuming independence, and some more specific models that we find are needed to capture some of the structure required for the calibration. We suppose

$$\begin{aligned}\pi(\{z_i\},\{w_{il}\},\{a_j\},\{b_i\},\{\sigma_j^2\},\sigma_y^2|\{y_i\},\{x_{ij}\}) \quad \propto \quad & \pi(\{y_i\}|\{z_i\},\sigma_y^2)\times \\ & \pi(\{x_{ij}\}|\{a_j\},\{b_i\},\{z_i\},\{w_{il}\},\{\sigma_j^2\})\times \\ & \pi(\{w_{il}\})\times\pi(\{a_j\})\times\prod_{j=1}^{5}\pi(\sigma_j^2)\times \\ & \pi(\sigma_y^2)\times\pi(\{b_i\}|\{z_i\})\times\pi(\{z_i\}).\end{aligned} \tag{6}$$

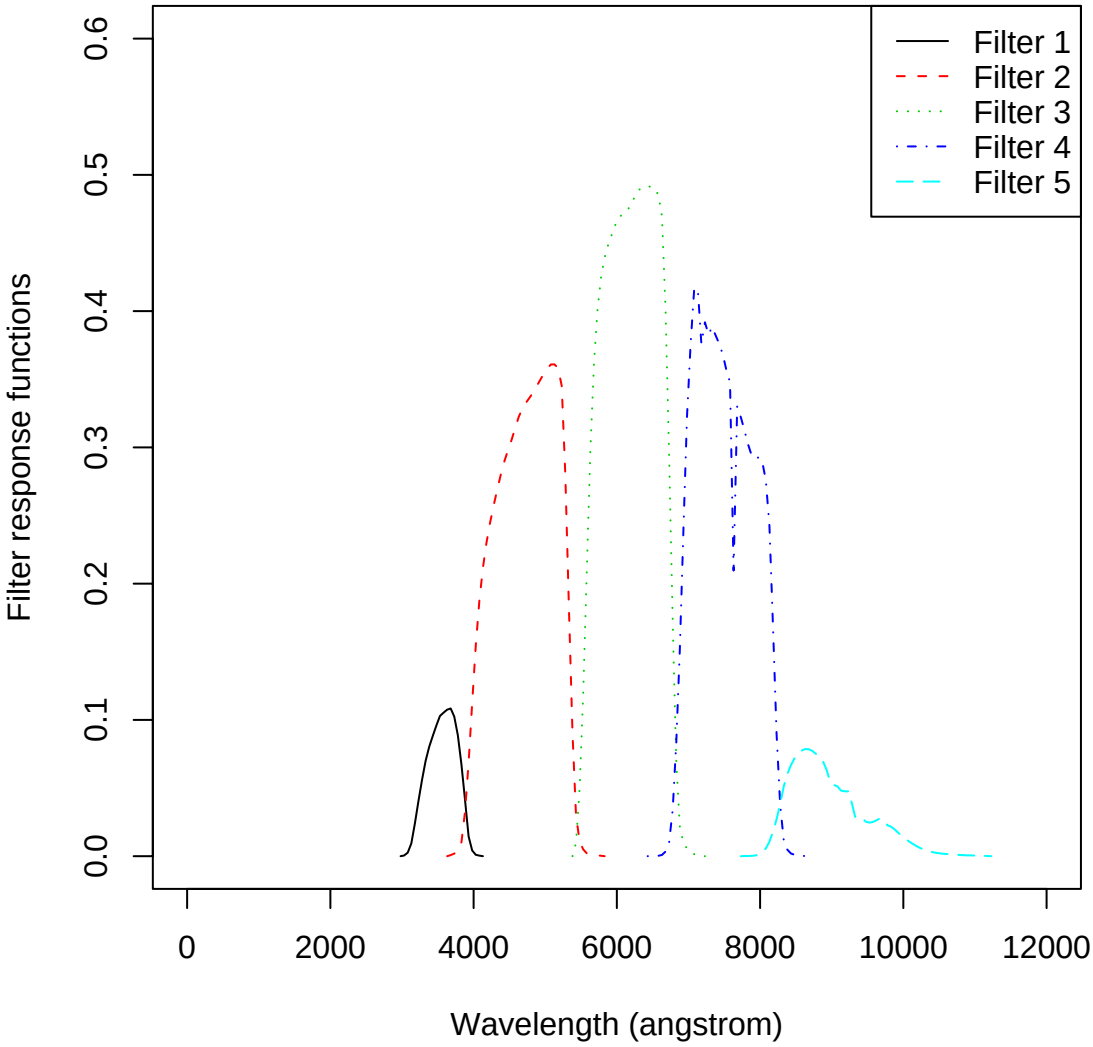

Figure 3: The five filter response functions

The prior for the galaxy type allocation $\pi(\{w_{il}\})$ is given by

$$P(w_{il} = 1) = p_l,\ l = 1, \ldots, 4,\ i = 1, \ldots, n \tag{7}$$

where $\sum_{l=1}^{4} p_l = 1$, with the allocation probabilities $\{p_l\}$ given a Dirichlet$(1, 1, 1, 1)$ prior. The filter sensitivity parameters $\{a_j\}$ are given an improper uniform prior (subject to the identifiability constraint). The inverse variance parameters are given proper Gamma priors, $\Gamma(5, 5/1000)$ for σ_y^2 which is expected to be very small, and $\Gamma(1, 5/100)$ for the σ_j^2. Using improper priors here for the variance terms can lead to problems with variance terms tending to zero, in particular σ_y^2, with obvious consequences for any MCMC mixing (Spiegelhalter, Best, Gilks, and Inskip (1996)).

One of the difficulties revealed in earlier experiments using a set of independent standard priors for all the parameters, was a lack of pooling of information between galaxies with and without recorded y_i. One possible improvement to the model would be to modify the priors on the brightnesses $\{b_i\}$ to depend on the redshift $\{z_i\}$. (A possible alternative would be instead to make the type allocation distribution $\{w_{il}\}$ depend on brightness, but since we have limited information about the galaxy types, other than that the LRG data set has been selected to consist generally of type 1, i.e. elliptical, galaxies, we do not pursue this line.) For each of the five filters, Figure 4 shows a scatter plot for the LRG data of $x_{ij} - \psi_{1j}(y_i)$ (the recorded data minus the filtered type 1 spectra at the observed spectroscopic redshift) against y_i (the spectroscopic redshift). If the likelihood model for x_{ij} is reasonable, on the y-axis we have an indication of $a_j + b_i + \epsilon_{ij}$, while on the x-axis assuming that the spectroscopic redshift is quite accurate we have an indication of z_i. There is a clear relationship, as might be expected by noting that a higher redshift may indicate that a galaxy is further away and thus less bright. Fitting a regression for each of the five filters gives the following five quadratic regression curves:

[Filter 1] $9.335535 + 20.626274 y_i - 25.464682 y_i^2$

[Filter 2] $10.96678 + 17.36172 y_i - 17.87796 y_i^2$

[Filter 3] $10.75092 + 17.58528 y_i - 18.62522 y_i^2$

[Filter 4] $10.09591 + 17.26753 y_i - 17.55181 y_i^2$

[Filter 5] $7.918389 + 17.379245 y_i - 17.346116 y_i^2$

Although there are clearly some discrepancies, in particular for filter 1, we suggest the prior

$$b_i | z_i \sim N(\alpha + \beta z_i + \gamma z_i^2, \sigma_b^2)$$

with α, β, γ all having improper priors, and the inverse of σ_b^2 having an $\Gamma(1, 5/100)$ prior in common with the $\{\sigma_j^2\}$.

For the purposes of these calibration experiments, the prior $\pi(\{z_i\})$ is taken to be the normalised histogram of the combined LRG and MAIN data sets, using 40 equal-sized bins between 0 and 0.8 (see Figure 5).

2.4 Tuned MCMC moves

Inference from the posterior distribution

$$\begin{aligned}\pi(\{z_i\},\{w_{il}\},\{a_j\},\{b_i\},\{\sigma_j^2\},\sigma_y^2,\{p_l\},\alpha,\beta,\gamma,\sigma_b^2 \quad | \quad & \{y_i\},\{x_{ij}\}) \propto \\ & \pi(\{y_i\}|\{z_i\},\sigma_y^2) \times \\ & \pi(\{x_{ij}\}|\{a_j\},\{b_i\},\{z_i\},\{w_{il}\},\{\sigma_j^2\}) \times \\ & \pi(\{w_{il}\}) \times \pi(\{a_j\}) \times \prod_{j=1}^{5} \pi(\sigma_j^2) \times \\ & \pi(\sigma_y^2) \times \pi(\{b_i\}|\{z_i\}) \times \pi(\{z_i\}) \times \\ & \pi(\{p_l\}) \times \pi(\alpha) \times \pi(\beta) \times \pi(\gamma) \times \pi(\sigma_b^2) \quad (8)\end{aligned}$$

will require Markov chain Monte Carlo (MCMC) methods (Gilks, Richardson and Spiegelhalter (1996)). Many of the components can be updated using standard single-component Gibbs sampler moves (α, β, γ, and all of the variance terms: σ_b^2, σ_y^2, $\{\sigma_j^2\}$). The constrained parameters $\{p_l\}$ and $\{a_j\}$ can also be updated using Gibbs moves from their joint conditional distributions; the former is another Dirichlet distribution, the latter uses the result that if the multivariate normal $a \sim N(\mu, V)$ subject to $1'a = 0$ (i.e. the sum of a's = 0), then the resulting conditioned distribution can be written as $N(R\mu, RVR')$, and samples can be generated by drawing $z \sim N(\mu, V)$ and premultiplying by R, where $R = I - V1(1'V1)^{-1}1'$.

The standard approach for updating the remaining parameters $\{b_i\}$, $\{z_i\}$ and $\{w_{il}\}$ would be single site proposals, Gibbs for the brightnesses and symmetric Metropolis proposals for the redshifts and galaxy types. However, as we will demonstrate in the next section, this can lead to mixing problems. Consider Figure 6 which shows the smoothed filter spectra as a function of z for the five filters, ie $\psi_{lj}(z), j = 1,\ldots,5, j = 1,\ldots,4$. Changing z while holding all other variables fixed means a simultaneous shift along the x-axis for the five graphs. Altering the label $\{w_{il}\}$ while holding all other variables fixed means a horizontal jump between curves. In the former case, an acceptable acceptance rate can be maintained by choosing the spread of the Metropolis proposal to be sufficiently small. However for the label changing, no such tuning mechanism exists, potentially leading to the chain becoming trapped with the wrong label. If this happens, the values of the other parameters including z will evolve to match the data, and results far from the spectroscopic redshift may result. Ideally the chain should visit all reasonable interpretations of the data as the goal so as to generate reliable interval estimates of z.

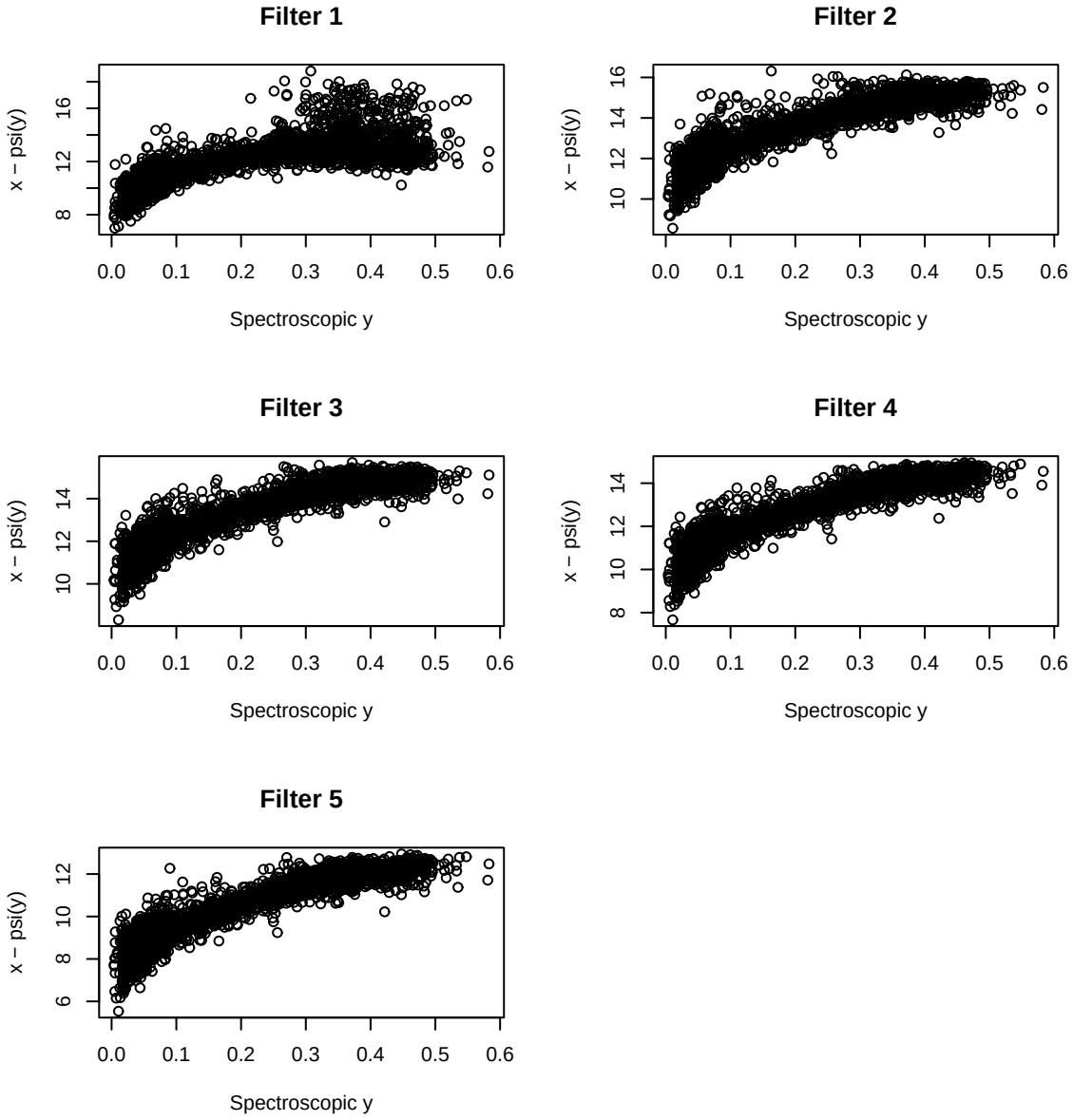

Figure 4: Plots for the five filters of $x_{ij} - \psi_{1j}(y_i)$ against y_i for the LRG data set

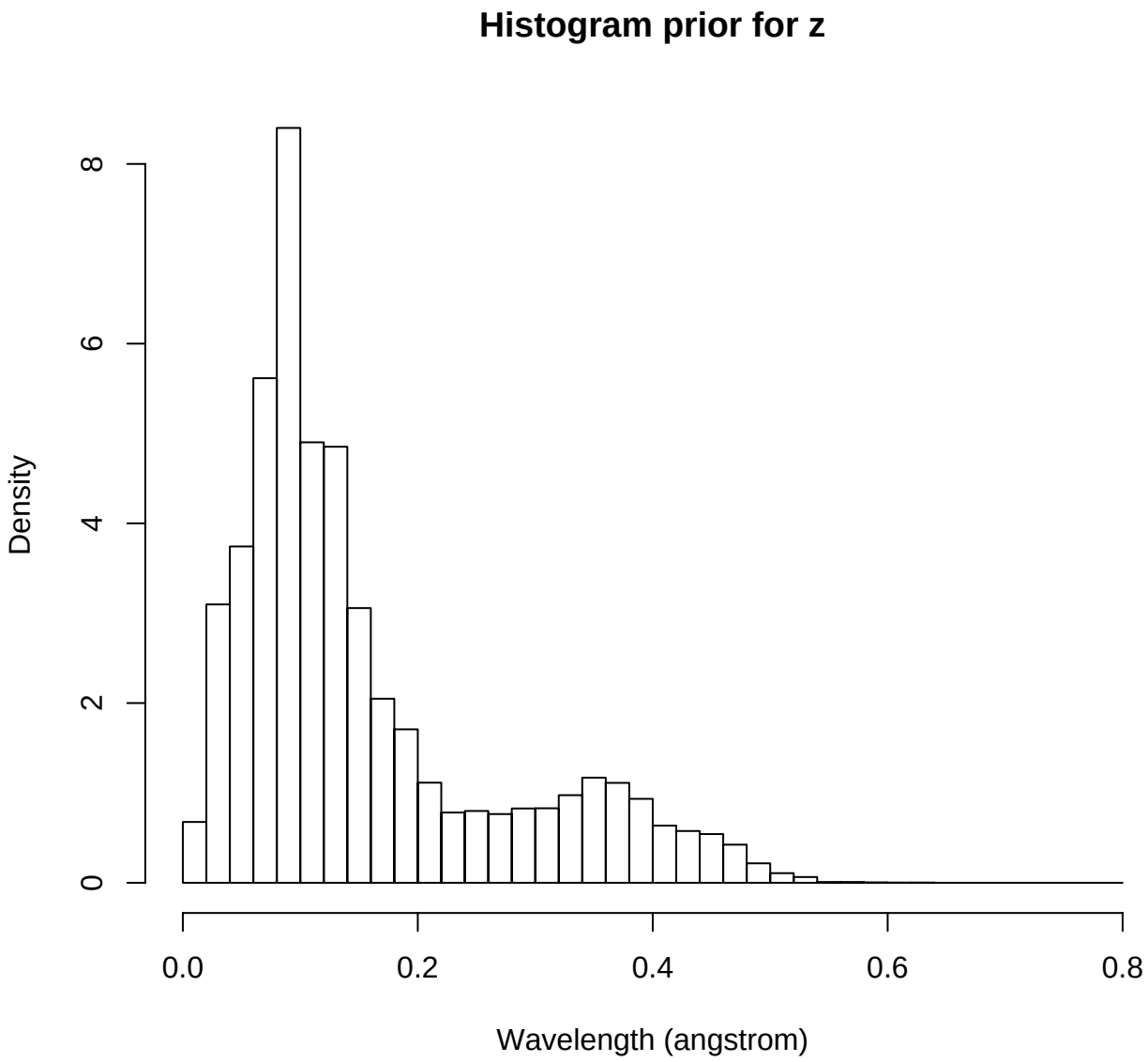

Figure 5: The histogram of the combined LRG and MAIN y_i values

The obvious choice of joint update would be that of z_i and $\{w_{il}\}$ together. Unfortunately this is hard to tune; roughly speaking we would be trying to move along the x-axis and between curves in Figure 6 in such a way as to maintain the level of the convolved filter values. Instead we construct proposals for joint updates involving either the labelling and the brightness b_i, or the redshift and the brightness, in the following way:

$$p(z_i, b_i \to z_i', b_i') = p(z_i \to z_i') \times p(b_i \to b_i'|z_i') \quad (9)$$

$$p(\{w_{il}\}, b_i \to \{w_{il}'\}, b_i') = p(\{w_{il}\} \to \{w_{il}'\}) \times p(b_i \to b_i'|\{w_{il}'\}) \quad (10)$$

ie propose a change in the label or the redshift, and then conditional on this new value, propose a new b_i. If we use a symmetric Metropolis proposal for either z_i or $\{w_{il}\}$, we follow this with a proposal for b_i which is effectively a draw from its full conditional (conditioned on either z_i or $\{w_{il}\}$, and all the other parameters):

$$b_i'|z_i' \ldots \sim N\left(\frac{\sum_{j=1}^{J} \frac{x_{ij} - a_j - \sum_l w_{lj}\psi(z_i')}{\sigma_j^2} + \frac{\alpha + \beta z_i' + \gamma z_i'^2}{\sigma_b^2}}{\sum_{j=1}^{J} 1/\sigma_j^2 + 1/\sigma_b^2}, \frac{1}{\sum_{j=1}^{J} 1/\sigma_j^2 + 1/\sigma_b^2}\right) \quad (11)$$

$$b_i'|\{w_{il}'\} \ldots \sim N\left(\frac{\sum_{j=1}^{J} \frac{x_{ij} - a_j - \sum_l w_{lj}'\psi(z_i)}{\sigma_j^2} + \frac{\alpha + \beta z_i + \gamma z_i^2}{\sigma_b^2}}{\sum_{j=1}^{J} 1/\sigma_j^2 + 1/\sigma_b^2}, \frac{1}{\sum_{j=1}^{J} 1/\sigma_j^2 + 1/\sigma_b^2}\right) \quad (12)$$

This particular construction is used because in the likelihood term, equation (3), the galaxy specific part of the mean of x_{ij} is given by $b_i + \sum_{l=1}^{4} w_{il}\psi_{lj}(z_i)$. The intention is that drawing b_i in this way compensates for changes in the mean of x_{ij} due to the proposed update to z_i or $\{w_{il}\}$ respectively. In terms of the acceptance ratio, it is still true that the initial symmetric proposal part cancels, but the ratio of the b_i proposals is required, eg

$$\frac{p(z_i, b_i \to z_i', b_i')}{p(z_i', b_i' \to z_i, b_i)} = \frac{p(b_i \to b_i'|z_i')}{p(b_i' \to b_i|z_i)} \quad (13)$$

There is some cancellation with the ratio of the posteriors. A separate single-component b_i using a Gibbs sampler is also used as the acceptance ratio of the combined moves may still not be high.

3 Small sample calibration results

In order to assess the general feasibility of our method, we consider first a small sample problem. We select 300 galaxies at random from the full set of 23338 galaxies, drawing from the MAIN and LRG sets in proportion to their size. For the first two–thirds of the galaxies, the spectroscopic redshift y_i will be retained in the model. For the remaining third, it will be used only in assessing how well the model fits. No use is made in fitting of the knowledge of which data set the galaxies are from, but again this information is used in assessing performance (since we would expect most LRG galaxies to be assigned to the elliptical category).

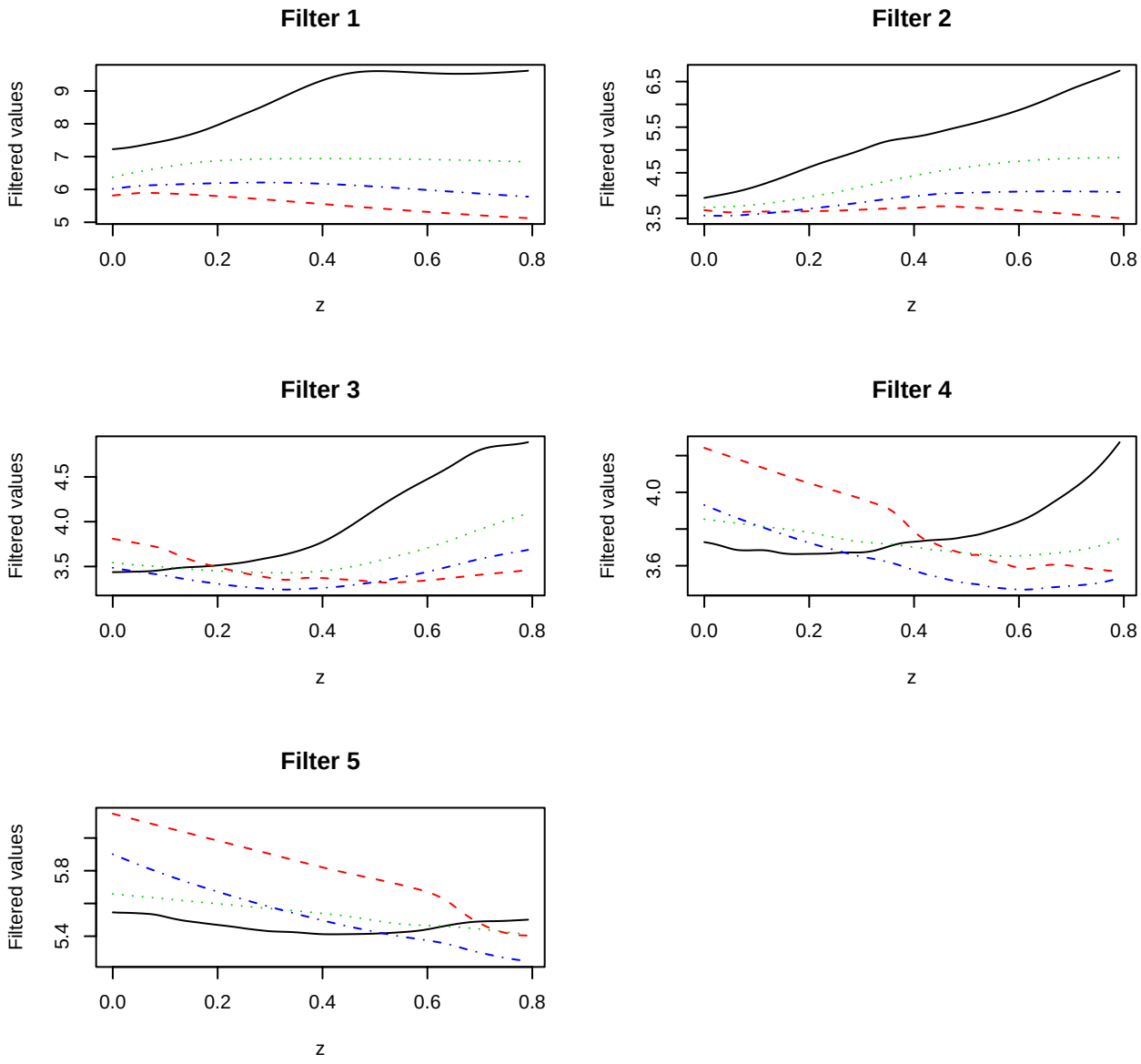

Figure 6: The convolved values observed at the five filters as a function of the redshift z

We work with the same measure of success in estimating z_i for the final 100 galaxies used by Csabai *et al.* (2003), the "root mean square error" (RMS)

$$RMS = \sqrt{\sum_{i=201}^{300} (y_i - \hat{z}_i)^2/100} \tag{14}$$

that is, the root mean square distance between the fitted redshift and the observed spectroscopic redshift.

Figure 7 shows the RMS results over 1000 randomly drawn samples of 300 galaxies. For each sample, 25000 MCMC iterations were run using either single-site updates (Standard) or the joint b_i moves described in the previous section (Tuned). The first 10% of each run is discarded as burn-in. Clearly there is much variation between samples, but generally, in the cases where the two algorithms do not agree, the Tuned MCMC moves lead to lower values of the RMS. The average RMS for the Standard case is 0.0455, and for the Tuned case it is 0.0426 (corresponding medians 0.0442 and 0.0409). Since the model is the same in both cases, this is a mixing issue. So, what is going on? Figure 8 shows traces, thinned by a factor of 10, of the 25000 iterations for one particular galaxy using the two sampling approaches. For this particular galaxy, the model supports two potential label-redshift interpretations of the data. The Standard sampler is unable to move between these two interpretations, while the Tuned move types overcome the problem.

The average RMSs found here are comparable to the results reported in Csabai *et al.* (2003) for template matching methods (although our method also has the benefit of delivering associated uncertainty estimates). These results are from comparatively small sample sizes, and we would expect to improve on them by using more than 150 known y_i galaxies to fit 100 unknown y_i ones. In particular, there is significant variability in the fitted values for the regression of b_i on z_i, α, β and γ, with some of the worse RMS results displaying quite distinct values of these parameters. In the next section, we discuss the application of these ideas to large samples of galaxies.

4 Adding additional galaxies

The SDSS has already generated both spectroscopic and photometric redshift estimates for many thousands of galaxies. The overall goal of this work is to use these existing data to estimate z_i for additional galaxies without spectroscopic data. We propose the following strategy: The entire data set will be split into a training set, of size 22338, and a test set of size 1000 galaxies (chosen at random from the LRG and MAIN subsets in proportion to the size of these sets). The model as described in previous sections will be fitted to the training set including spectroscopic measurements for all 22338 galaxies. Histograms of some of the resulting marginal posterior distributions are shown in Figure 9. The posterior distributions of the model parameters $\{a_j\}, \{\sigma_j^2\}, \{p_l\}, \alpha, \beta, \gamma, \sigma_b^2$ are approximated by parametric models which are then to be used as highly-

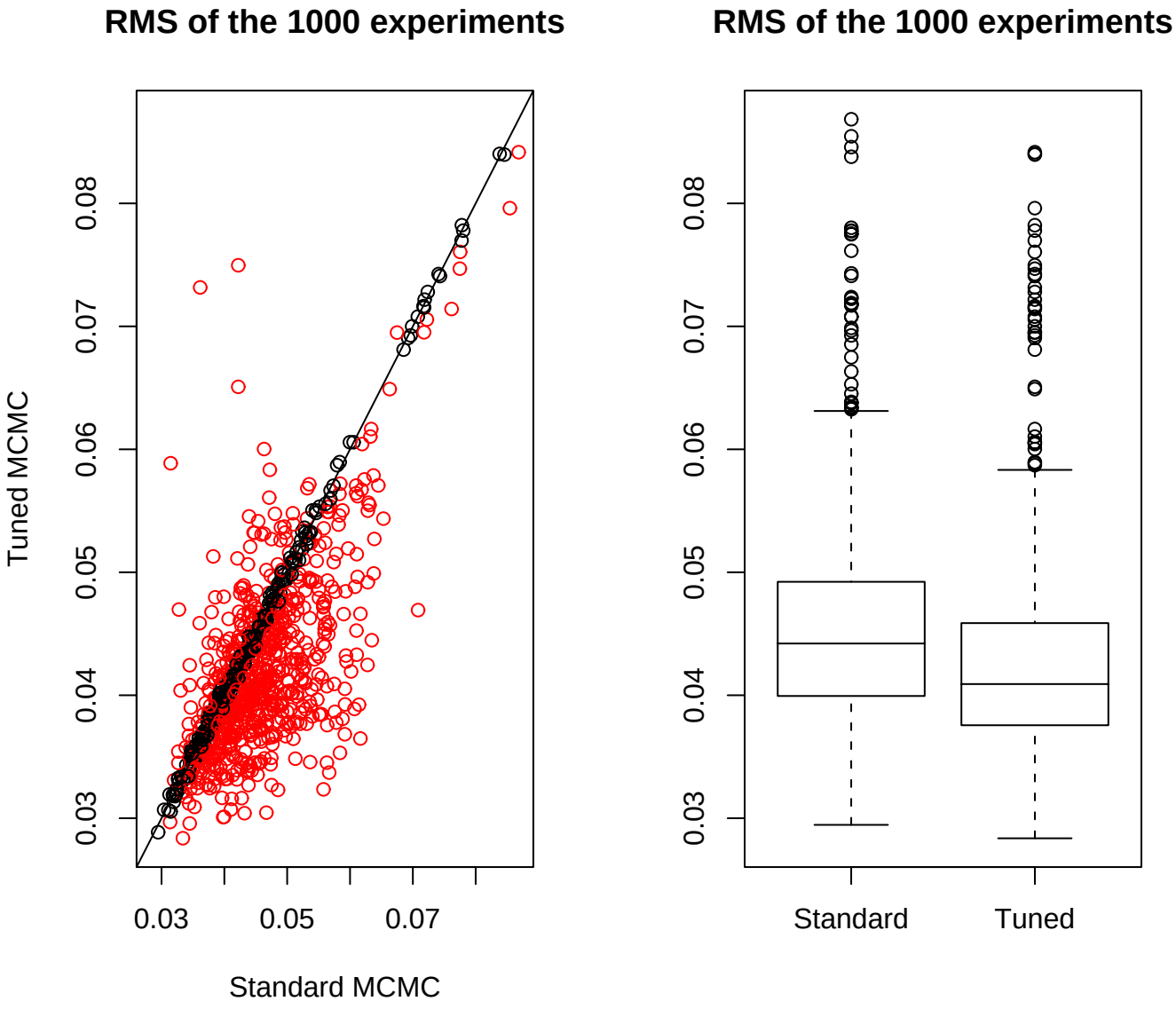

Figure 7: The RMS results for the small calibration problem. In the left image, discrepancies of more than 0.001 are marked in red.

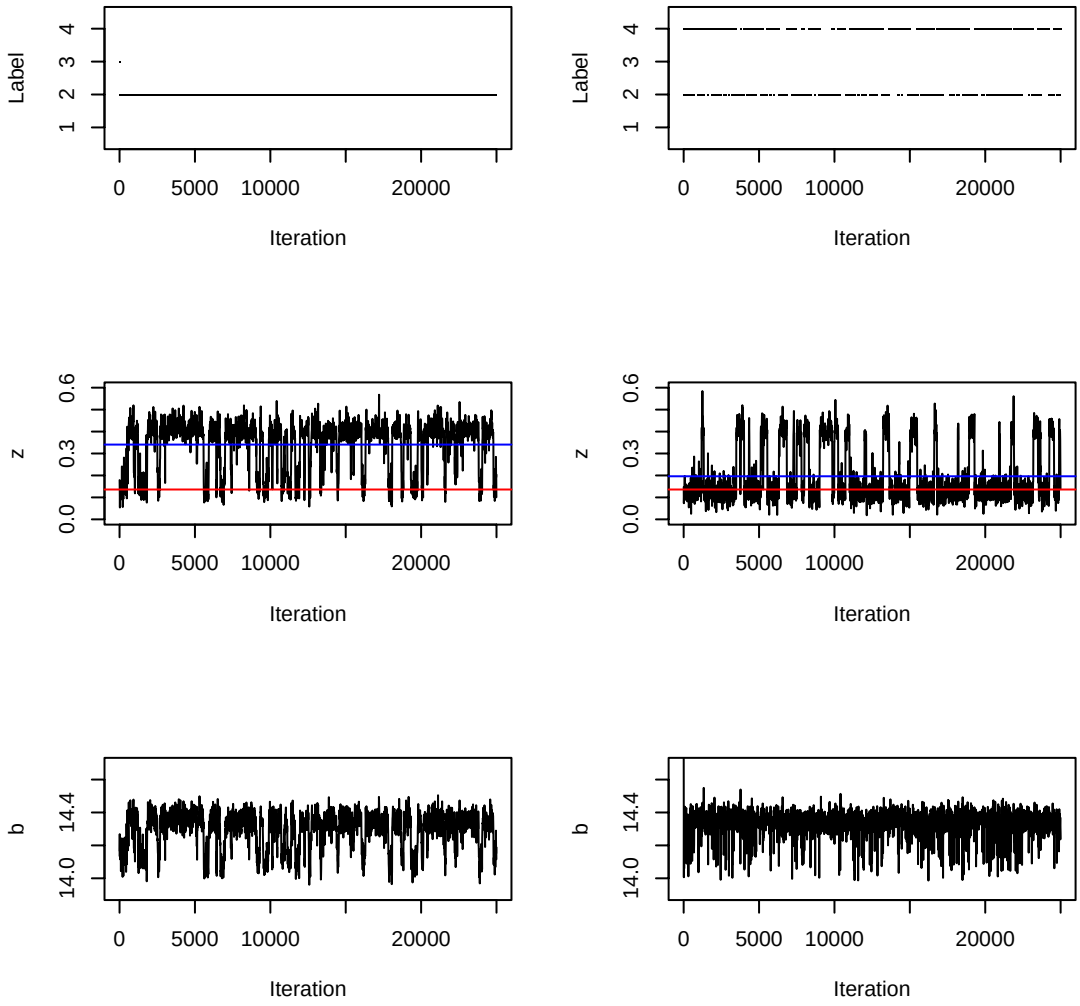

Figure 8: Thinned output traces for the Standard (left) and Tuned (right) MCMC for one galaxy of the 1000 small data sets. Traces are for $\{w_{il}\}$ (top), z_i (middle) and b_i (bottom). For the z_i plots, the ergodic average is marked in blue, and the measured spectroscopic redshift y_i in red

informative priors in estimating the redshift for the test set of 1000 galaxies; the following distributions are fitted

$$\begin{aligned}
\sigma_j^2 &\sim \Gamma(\lambda_j, t_j),\ j = 1, \ldots, 5 && (15)\\
\{p_l\} &\sim \text{Dirichlet}(\delta_1, \delta_2, \delta_3, \delta_4) && (16)\\
\alpha &\sim N(\mu_\alpha, \sigma_\alpha^2) && (17)\\
\beta &\sim N(\mu_\beta, \sigma_\beta^2) && (18)\\
\gamma &\sim N(\mu_\gamma, \sigma_\gamma^2) && (19)\\
\sigma_b^2 &\sim \Gamma(\lambda_b, t_b) && (20)\\
a_j &\sim N(\mu_{a_j}, \sigma_{a_j}^2),\ \ j = 1, \ldots, 5 && (21)
\end{aligned}$$

with all model parameters estimated from the posterior marginals of the training set, and where in the case of the $\{a_j\}$, the constraint $\sum a_j = 0$ will be imposed on top of these distributions.

Figure 10 illustrates a typical fitting of the test sample using 10000 iterations, discarding the first 1000 as burn-in. The RMS of this particular fit is 0.0397. Figure 10 also gives approximate 95% interval estimates of the redshift using $\pm$ 1.96 times the estimated standard deviation. Of the 1000 intervals, 44 do not cover the spectroscopic redshift; however, as spectroscopic redshift is not exactly equal to true redshift, this is only a rough indication of whether the nominal coverage is accurate. Also perhaps of some use are density estimates of the posterior distributions of the z_i, Figure 11, which indicate the degree of uncertainty over z related to the uncertainty of galaxy classification. The skewness of the examples seen suggests that credible interval estimates computed from the quantiles of the sampled z_i might be more appropriate than the approach used here, although considerably more expensive in computer storage.

5 Discussion

The results obtained using the Bayesian model suggested here produce comparable RMSs to those obtained in Csabai *et al.* (2003) where they use the entire SDSS (LRG and MAIN) samples as training data and additional sets from other sources as test data. Our results are slightly better than the out–of–sample template matching they report but slightly worse than the in–sample template matching and empirical methods such as polynomial regression or nearest neighbour approaches fitted using large numbers of galaxies. For the out-of-sample galaxies they report results of "0.035 at $r^* < 18$ and rising to 0.1 at $r^* < 21$" (r^* is the third of the photometric measurements, here referred to as x_{i3}, and is generally used as an indicator of magnitude). Figure 12 plots the cumulative discrepancy between y_i and the fitted redshift against r^* for the

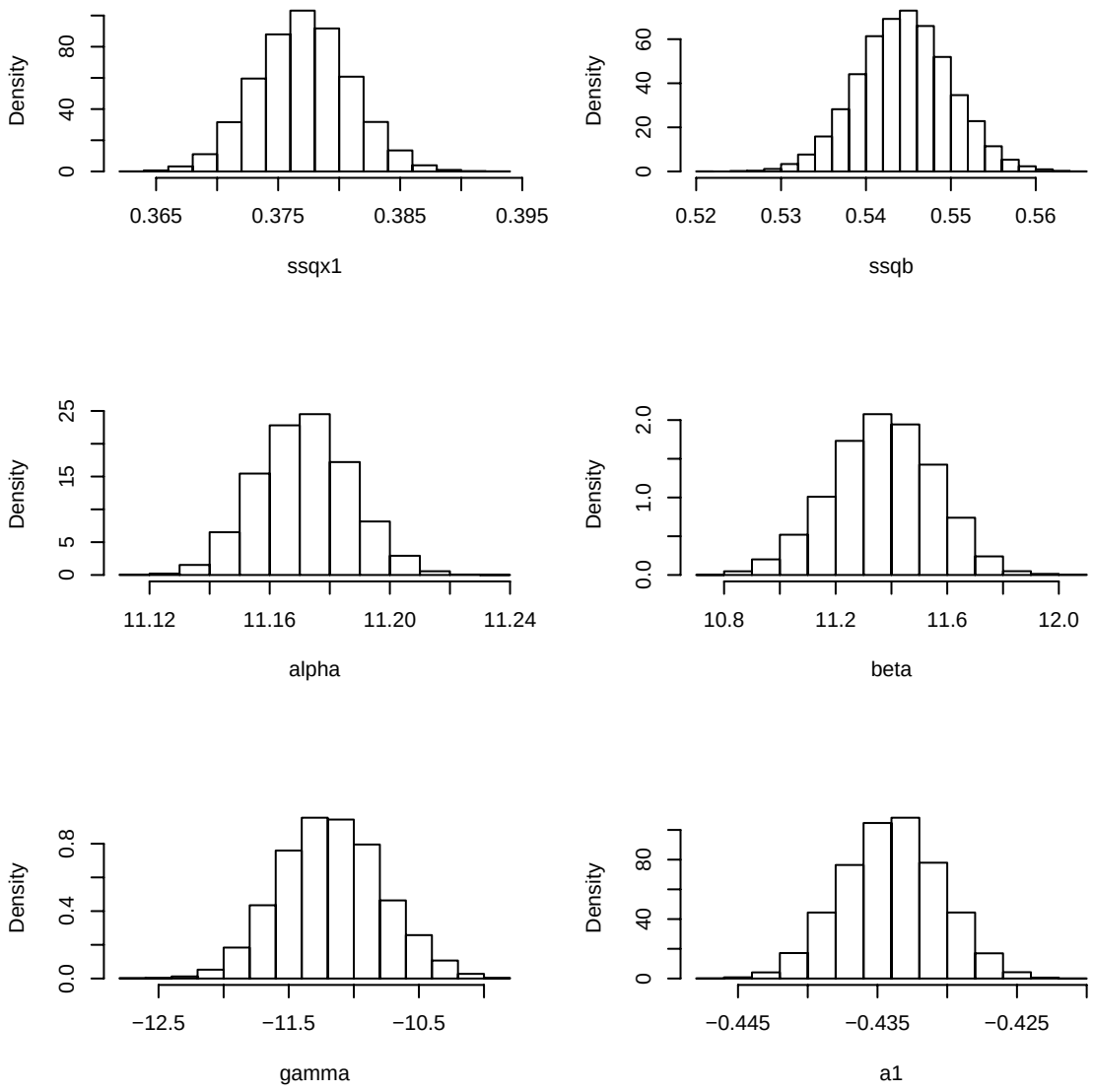

Figure 9: Histograms for some of the parameters from the MCMC run fitting the training set of galaxies

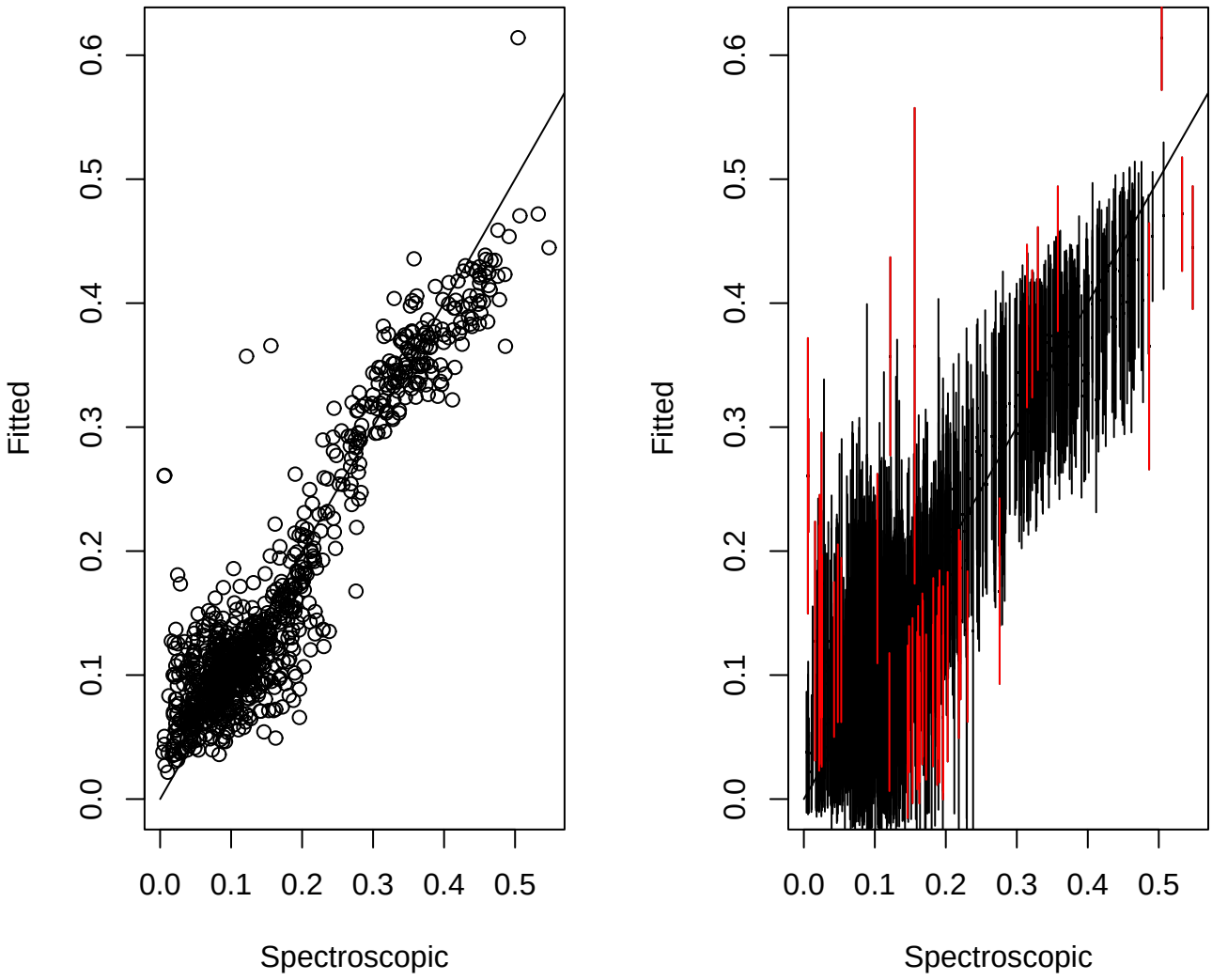

Figure 10: Left panel: fitted estimates of z for the 1000 galaxies in the test sample. Right panel: fitted estimates $\pm$ 1.96 estimated standard deviations, with intervals not containing the spectroscopic redshift marked in red

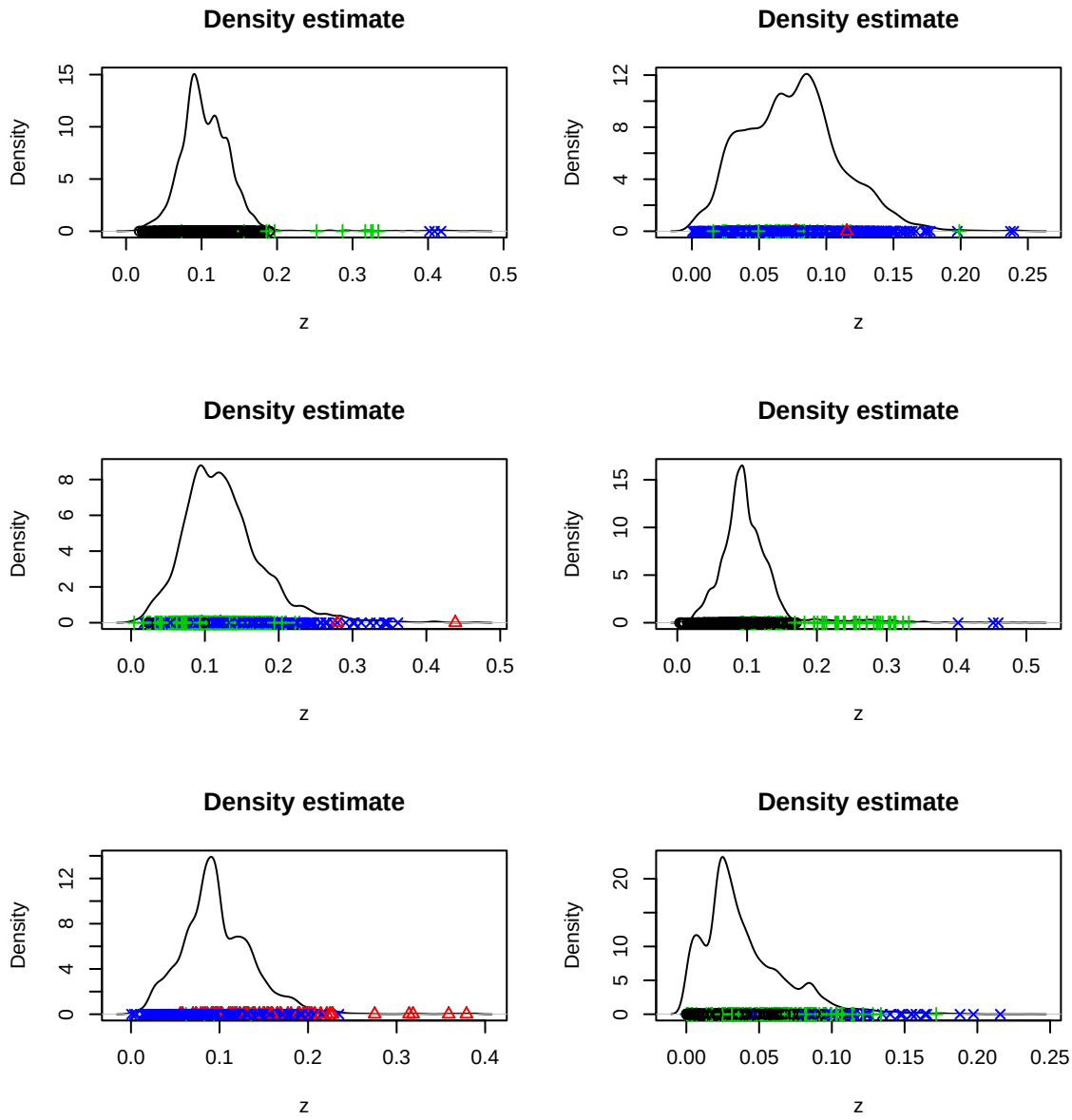

Figure 11: Density estimates for z_i for 6 of the test galaxies, overlaid by thinned MCMC realisations indicating the associated label at each iteration

1000 test galaxies. For comparison, $r^* = 18$ and a RMS of 0.035 are overlaid. For this set of galaxies, the majority have $r^* < 20$ and so a comparison for all $r^* < 21$ is not really appropriate.

The attraction of the approach proposed here is that interval estimates of redshift are generated, along with galaxy type probabilities. The training set can be expanded as more galaxies with both spectroscopic and photometric data become available, updating the prior for z and refitting the model parameters. Caveats are that the "out–of–sample" data used here are a subset of the combined SDSS sample. If data from other sources are added, it will be important to monitor whether the fitted model changes significantly. In terms of the existing data, Figure 10 suggests some residual structure to the fitting with more apparent underestimation in the mid-range of spectroscopic values. Given the reliance of this method on accuracy in the spectral energy templates, this may be due to inadequacies in these templates (hence the work on modifying these in Csabai *et al.* (2003)).

Acknowledgements

We are grateful to Bob Nichol and Andrew Connolly for providing both the data and background information about the application. F.A-A. was supported in a visit to Bath by the Bath Institute for Complex Systems.

References

[1] G.D. Coleman, C-C. Wu, and D.W. Weedman (1980), "Colors and Magnitudes Predicted for High Redshift Galaxies", *The Astrophysical Journal Supplement Series*, 43, 393–416.

[2] Istvan Csabai, Tamas Budavari, Andrew J. Connolly, Alexander S. Szalay, Zsuzsanna Gyry, Narciso Benitez, Jim Annis, Jon Brinkmann, Daniel Eisenstein, Masataka Fukugita, Jim Gunn, Stephen Kent, Robert Lupton, Robert C. Nichol, and Chris Stoughton (2003), "The Application of Photometric Redshifts to the SDSS Early Data Release", *The Astronomical Journal*, 125, 580–592.

[3] Gilks, W., Richardson, S. and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London.

[4] Smith, R.C. (1995), *Observational Astrophysics*. Cambridge University Press: Cambridge.

[5] Spiegelhalter, D. J., Best, N. G., Gilks, W. R. and Inskip, H. (1996), Hepatitis B: a case study in MCMC methods, Chapter 2 in *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London. (eds Gilks, W., Richardson, S. and Spiegelhalter, D.).

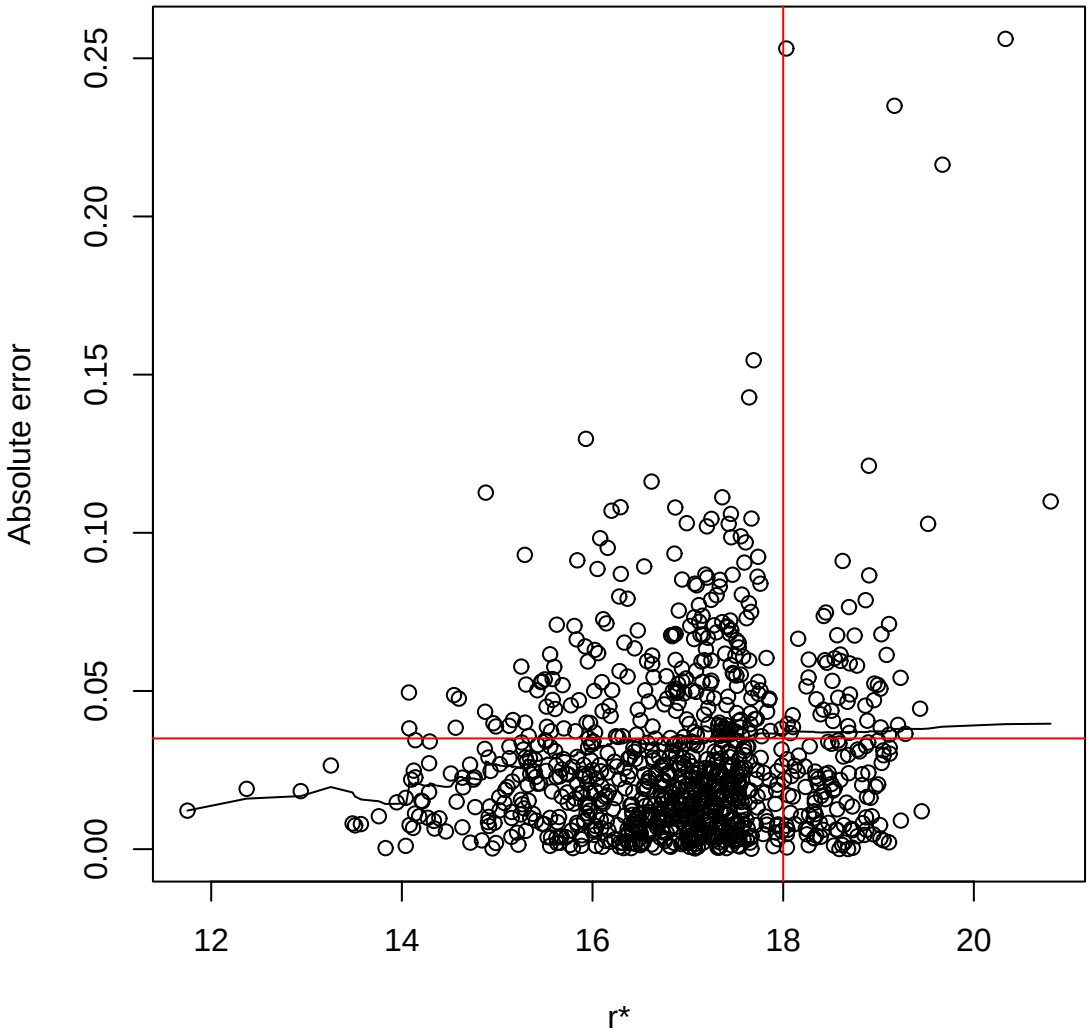

Figure 12: Absolute error between the fitted z_i and the spectroscopic redshift plotted against x_{i3}, referred to as r^*, an indicator of magnitude. The overlaid black curve is the cumulative rmse, and the red lines indicate the cumulative rmse error of 0.035 at $r^* = 18$