ESTIMATING REDSHIFT FOR A MIXTURE OF TYPES OF GALAXY

Merrilee Hurn (Bath)

Peter Green (Bristol)

Fahimah Al-Awadhi (Kuwait)

Estimation/calibration using two types of data

- Background to the redshift problem
  - The Sloan Digital Sky Survey data
  - Goals
- Bayesian mixture model
  - Fitting to (almost all) the galaxies
  - Adding additional galaxies
- MCMC problems
- Results

Sloan Digital Sky Survey aims to generate 3D maps of more than a quarter of the sky...
How do you estimate how far away a galaxy is?

Redshift

Galaxies relative motion due to both peculiar velocity and spacetime expansion ("expanding rubber sheet universe").

Effect on wavelength of light is

$$\lambda_{\text{observed}} = (1 + z)\ \lambda_{\text{emitted}}$$

where (dimensionless) z is redshift. For distant galaxies, $(1 + z)$ interpreted as factor by which universe expanded while the photon travelled.

How is this useful?

$$\text{Distance} \overset{\text{Hubble's law}}{\propto} \text{redshift}$$

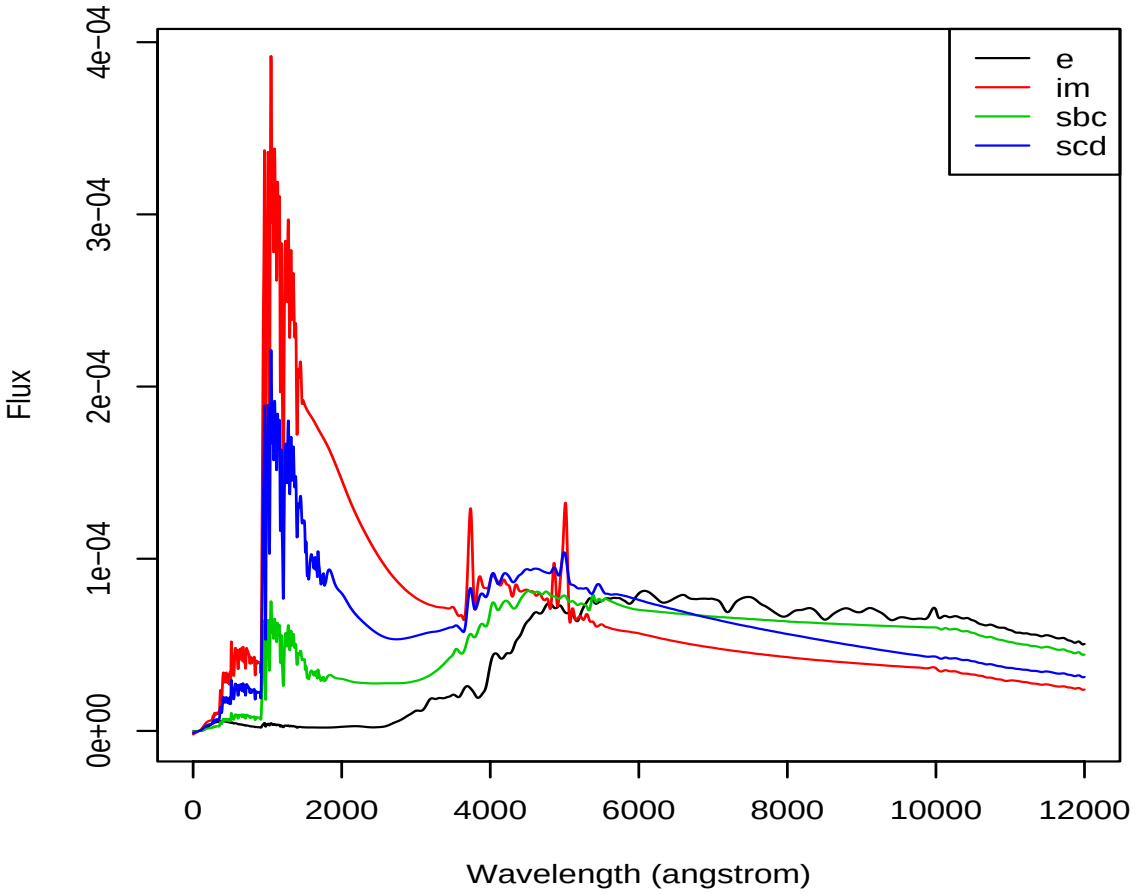
("scatter" due to peculiar velocity).

How to estimate the redshift?

Starting point is what we would see in the "at rest" case (ie $z = 0$)

Templates for flux emissions of 4 galaxy types:

elliptical, irregular, barred spiral, spiral



3800Å(blue) $\rightarrow$ 9200Å(near infrared)

(Templates mixtures of theory and observation)

Templates stretched for $z > 0$: z=0.5, emission at λ_{emit}= 4000Å observed at λ_{obs} = 6000Å

Sloan Digital Sky Survey collects two types of data:

Spectroscopic data Correlates peaks in observed spectrum against known lines

eg hydrogen peak at 3970Å observed at 6000Å implies redshift ≈ 0.5

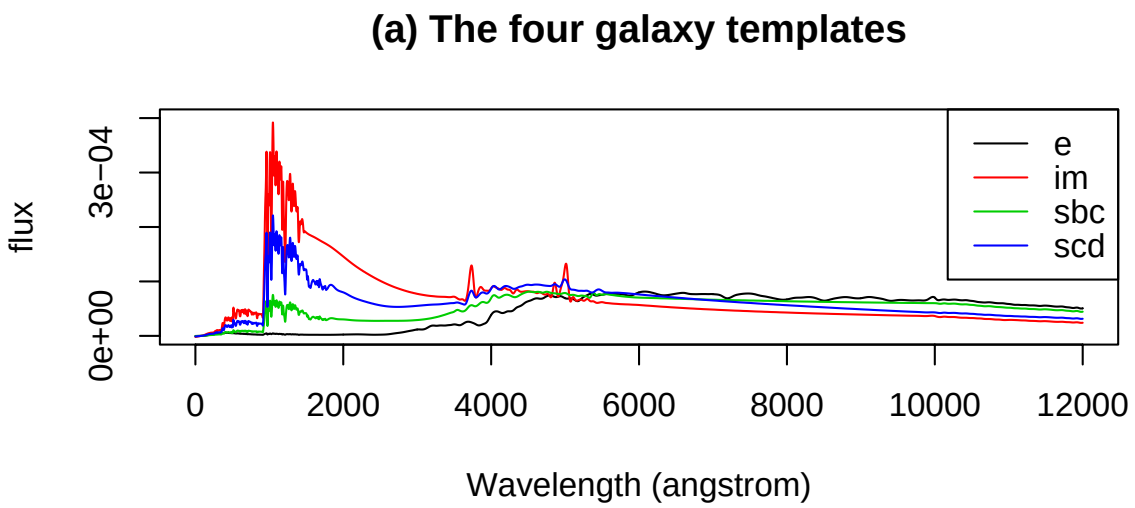
expensive, slow, accurate

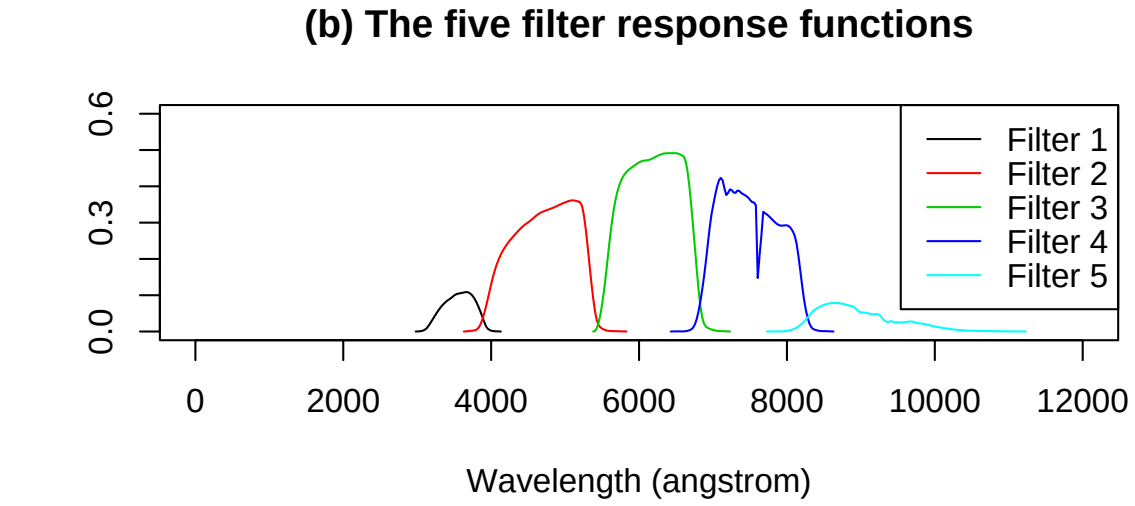
Photometric data Measures over five observation "windows" $\rightarrow$ convolution of observed flux with known filter response function

(comparatively) cheap, quick, but how to extract information about redshift?

Can we use a training set with both types of data to predict redshift for a galaxy with only photometric data?

Photometric data:



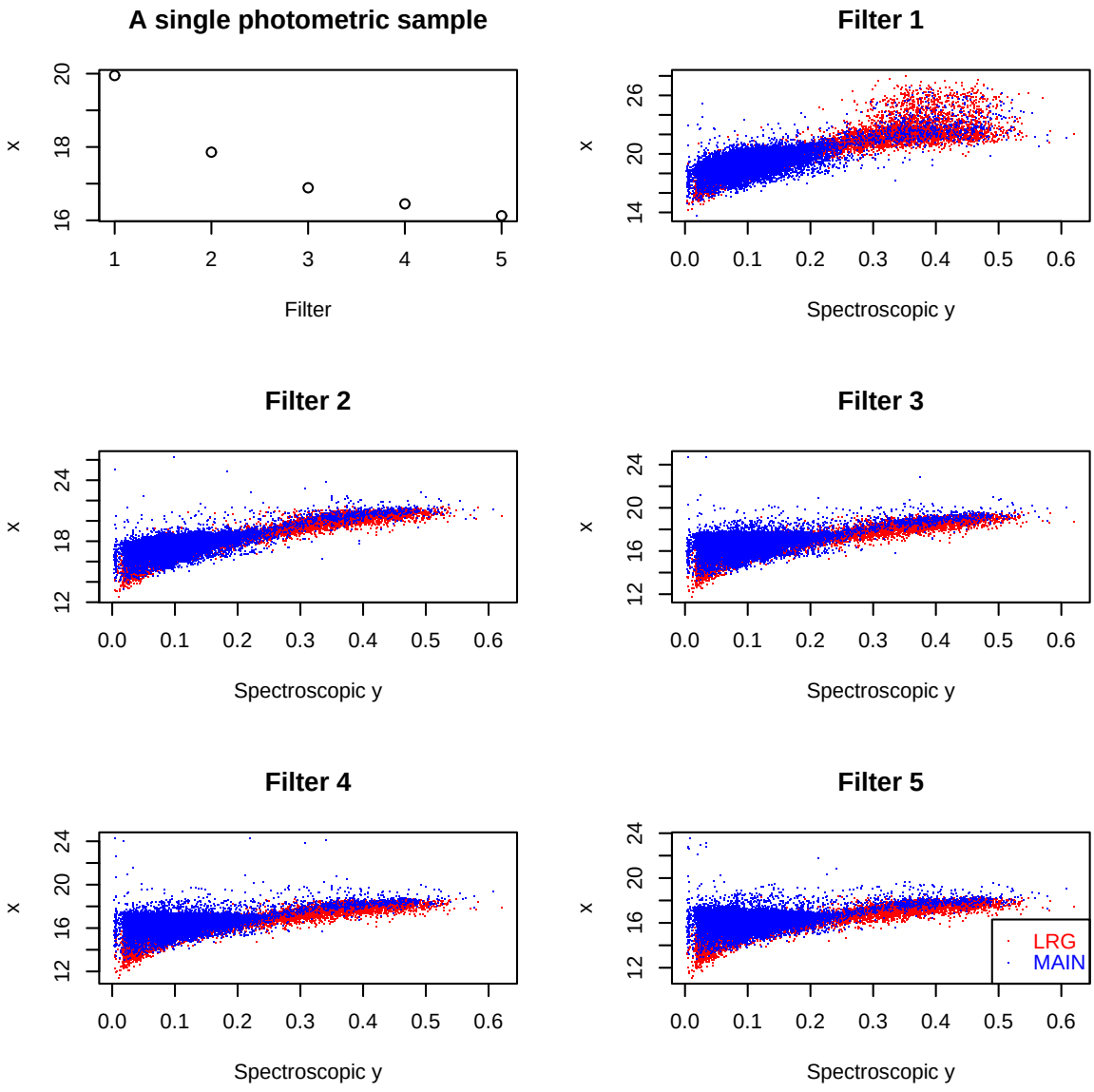


Convolution of j^{th} filter response $h_j(\ )$ with l^{th} galaxy template $\phi_l(\ )$ at redshift z

$$\psi_{lj}(z) = -2.5 \log_{10} \int_\lambda h_j(\lambda) \phi_l \left(\frac{\lambda}{1+z} \right) d\lambda$$

$$(\text{magnitude} = -2.5 \log_{10}(\text{Flux}/F_0))$$

The data available from the SDSS (23338 galaxies all with spectroscopic y and photometric $x_1, \ldots, x_5$)



MAIN = No particular selection

LRG = Luminous Red Galaxy

Existing methods

Empirical

- Nearest neighbour matching
- Polynomial regression of spectroscopic data on photometric data

Template based

- Least squares fit of photometric data to galaxy template

Statistical goals

- Uncertainty estimates for redshift and galaxy type
- (Use fairly standard components)

Bayesian mixture model

$$
\begin{aligned}
y_i|\ldots &\sim N(z_i, \sigma_y^2) \ \textit{spectroscopic} \\
x_{ij}|\ldots &= a_j + b_i + \sum_{l=1}^{4} w_{il}\psi_{lj}(z_i) + \epsilon_{ij} \ \textit{photometric}
\end{aligned}
$$

$$
\begin{aligned}
a_j &: \text{filter sensitivity, } \sum_{j=1}^{5} a_j = 0 \\
b_i &: \text{galaxy brightness} \\
\{w_{il}\} &: \text{zero-one allocation to galaxy type} \\
\psi_{lj}(z_i) &: \text{filtered spectrum at redshift } z_i \\
\epsilon_{ij} &\sim N(0, \sigma_j^2)
\end{aligned}
$$

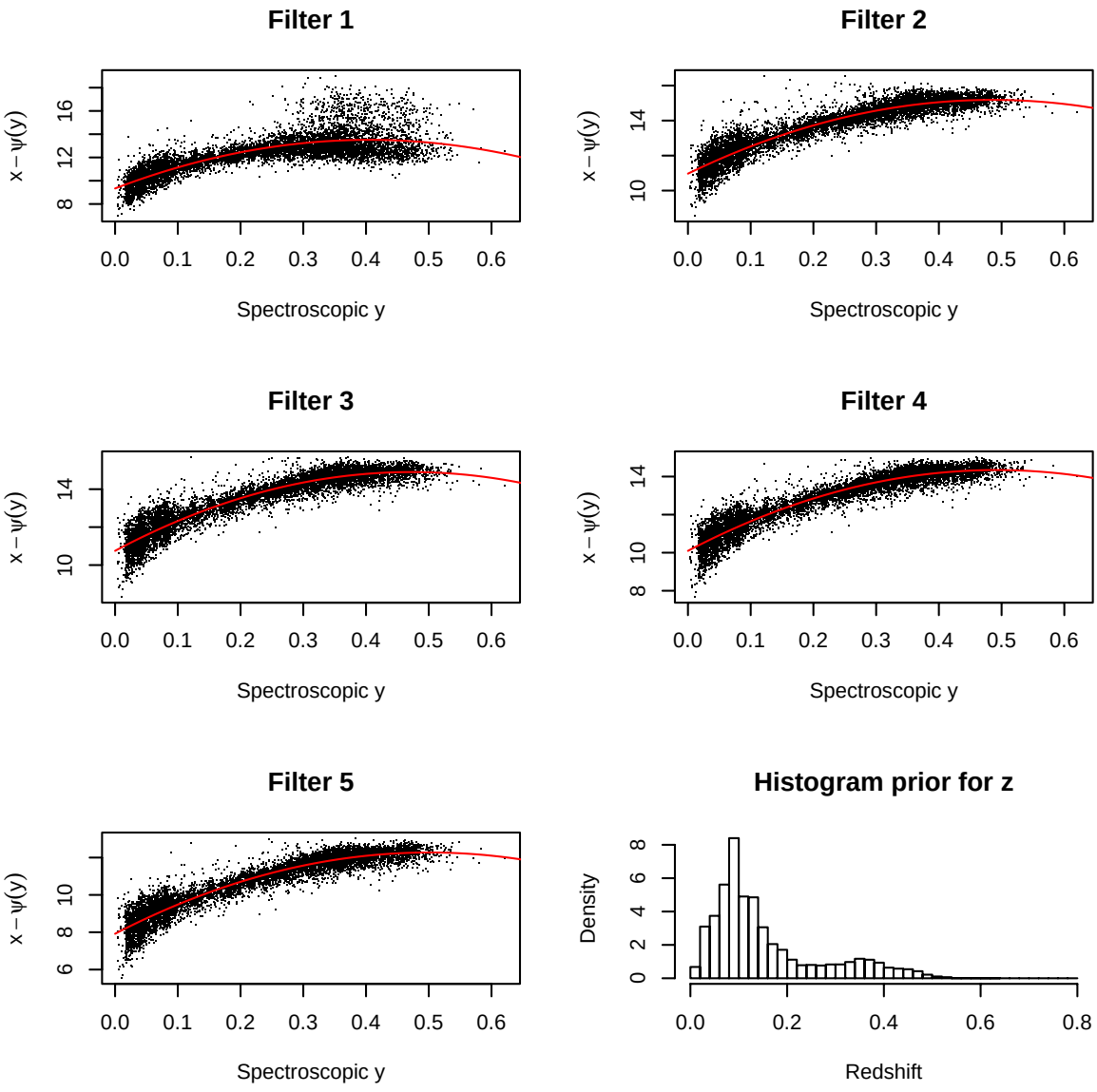
$$
\begin{aligned}
&\pi(\{z_i\}, \{w_{il}\}, \{a_j\}, \{b_i\}, \{\sigma_j^2\}, \sigma_y^2 | \{y_i\}, \{x_{ij}\}) \\
&\quad \propto f(\{y_i\}|\{z_i\}, \sigma_y^2) \\
&\quad \times f(\{x_{ij}\}|\{a_j\}, \{b_i\}, \{z_i\}, \{w_{il}\}, \{\sigma_j^2\}) \\
&\quad \times p(\{w_{il}\}) \times p(\{a_j\}) \times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

Mainly standard conjugate priors used...

Exceptions... $p(\{b_i\}|\{z_i\}) \times p(\{z_i\})$

$$\text{LRG}: \quad x_{ij} - \psi_{1j}(z_i) \quad = \quad a_j + b_i + \epsilon_{ij}$$

$$\text{plot} \quad x_{ij} - \psi_{1j}(y_i) \quad vs \quad y_{ij}$$



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

[standard conjugate priors for α, β, γ, σ_b^2]

Strategy

1. Randomly select 1000 galaxies from 23338
2. Fit model to remaining 22338 galaxies using photometric AND spectroscopic data
3. Update priors (using posterior marginals of parameters)
4. Fit model using ONLY photometric data to excluded 1000 galaxies using these informative priors

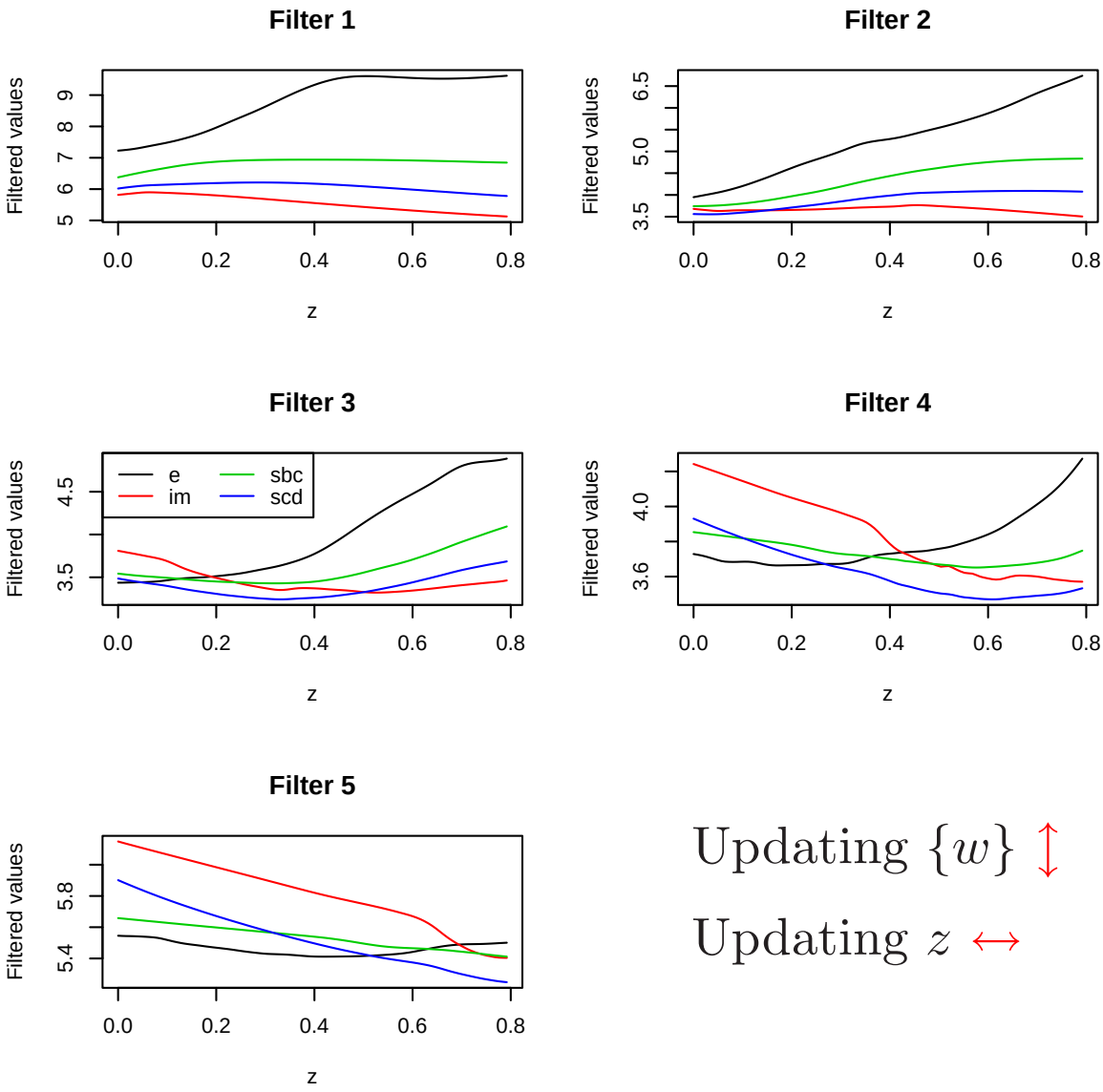
MCMC

Mainly standard single-site Metropolis or Gibbs steps

Poor mixing over galaxy labels

→ poor mixing for galaxy redshifts

$$X_{ij}|\ldots \sim N(a_j + b_i + \sum_{l=1}^{4} w_{il}\psi_{lj}(z_i), \sigma_j^2)$$



Updating $\{w\}$ $\updownarrow$

Updating z $\leftrightarrow$

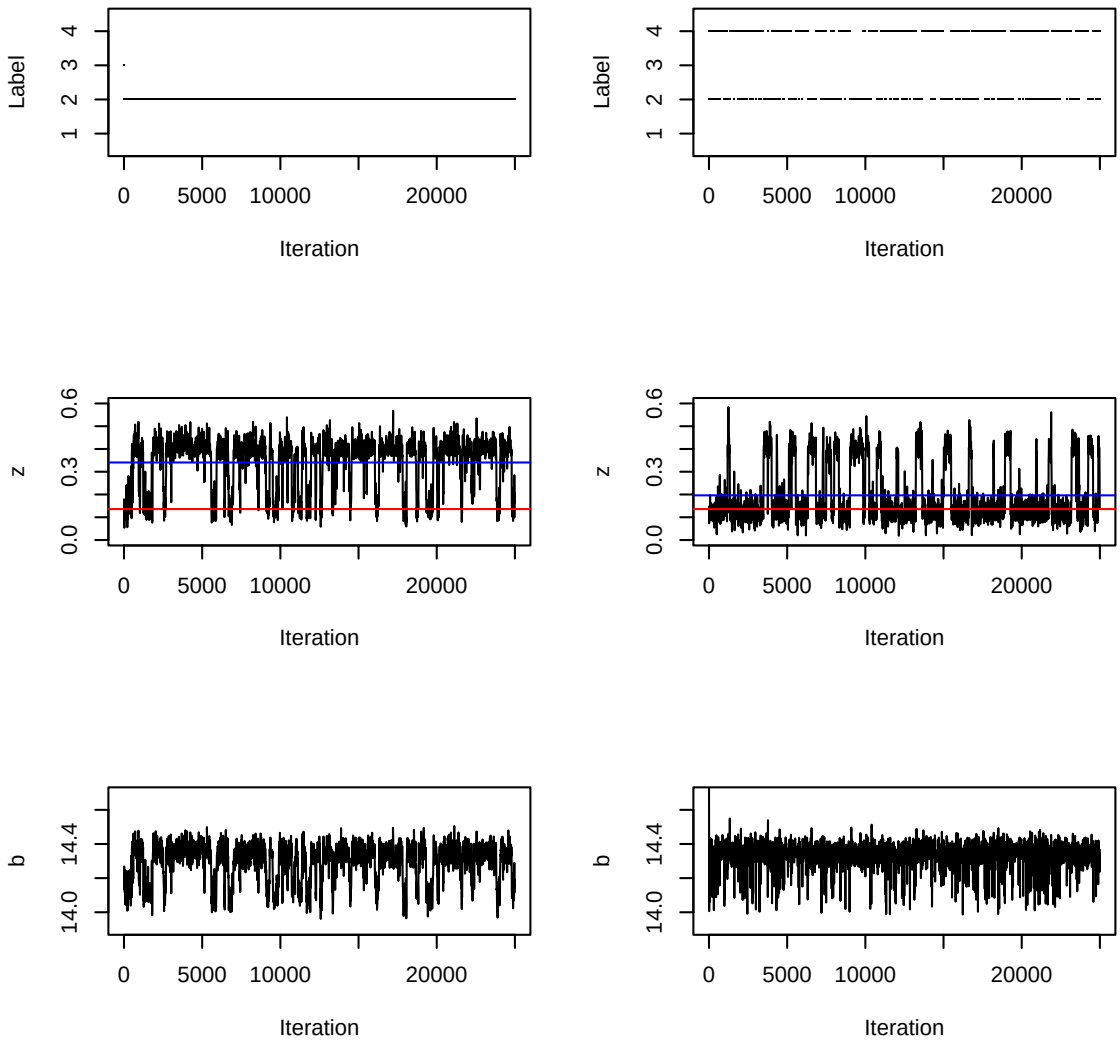
$$p(\{w\}, b \to \{w'\}, b') = p(\{w\} \to \{w'\})p(b \to b'|\{w'\})$$

$$p(\{w\} \to \{w'\}) : \text{Metropolis}$$

$$p(b \to b'|\{w'\}) : \text{“Gibbs”}$$

LEFT: Standard MCMC move types

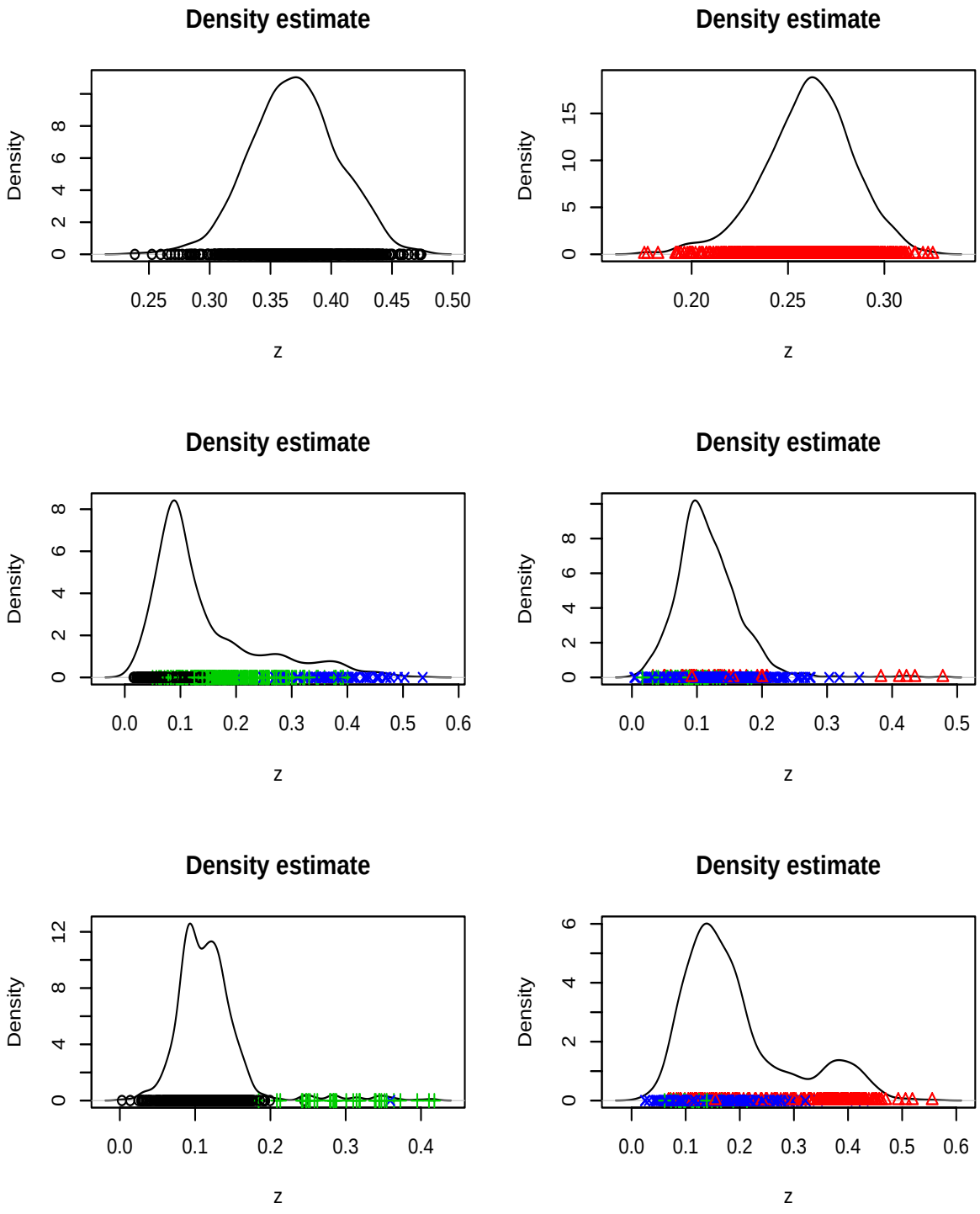
RIGHT: Moves using b to "mop up"



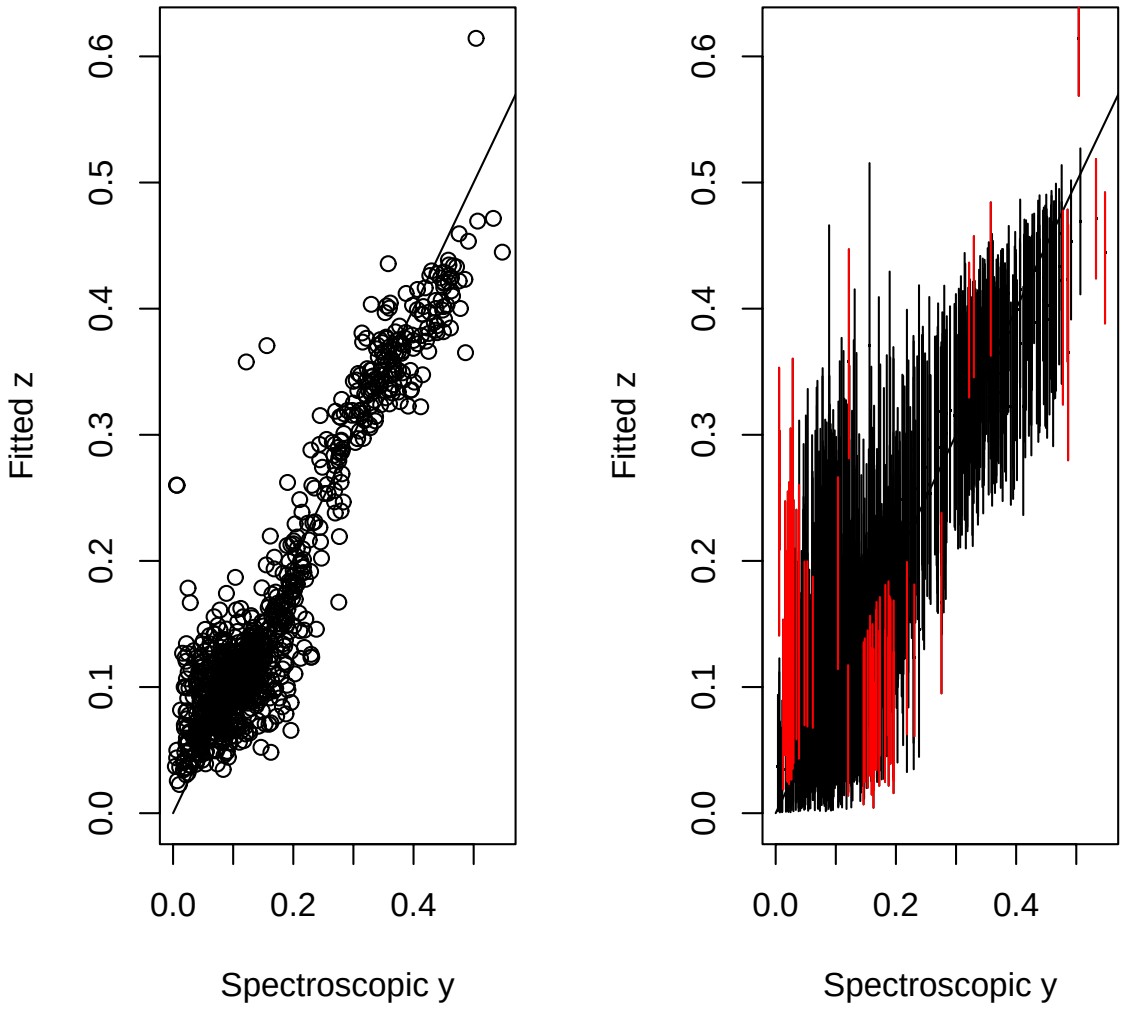
spectroscopic measurement y

ergodic average $\hat{z}_i$

The effect of the galaxy type uncertainty on the distribution of 6 galaxies' redshift



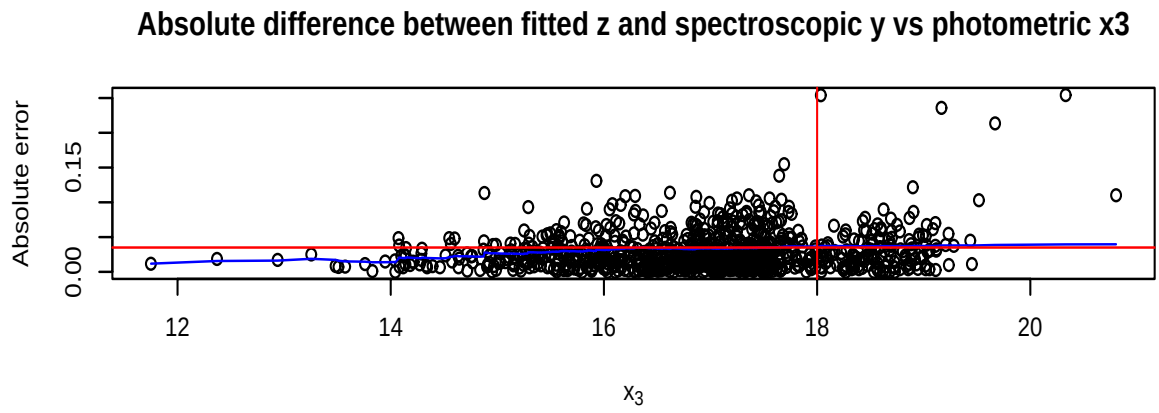
Point estimates and estimated 95% credible intervals plotted against spectroscopic redshift for the 1000 galaxies set

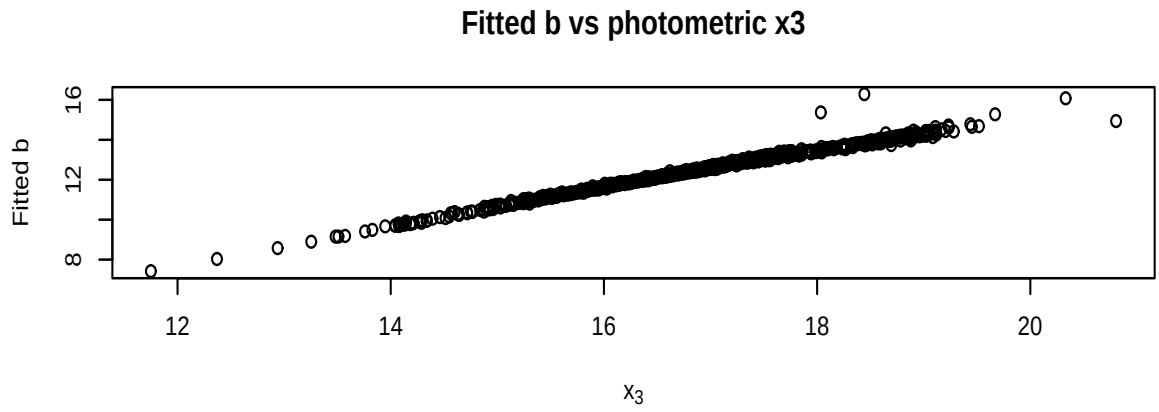


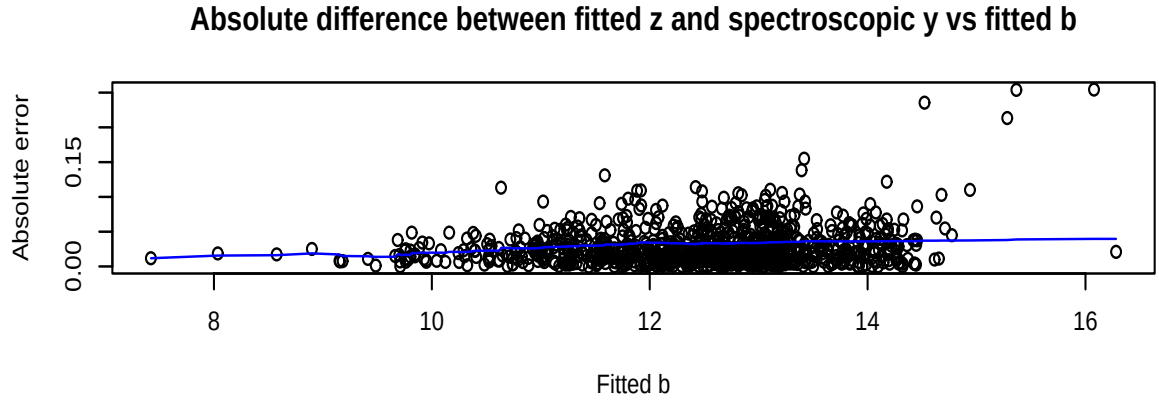
(58 of 1000 intervals for z do not contain y)

RMSE = 0.0396

Fainter galaxies have more redshift uncertainty







cumulative RMSE of 0.035 at $x_3 = 18$

cumulative RMSE of Bayesian model

– Pointwise RMSE about the same as polynomial regression model

\+ BUT adds interval estimates of redshift and galaxy type estimates

(could hopefully improve our results by getting more expert astronomical input into the model)