Reionization constraints using principal component analysis

Sourav Mitra,[1]★ T. Roy Choudhury[1]★ and Andrea Ferrara[2]★

[1]*Harish-Chandra Research Institute, Chhatnag Road, Jhusi, Allahabad 211019, India*
[2]*Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy*

ABSTRACT

Using a semi-analytical model developed by Choudhury & Ferrara we study the observational constraints on reionization via a principal component analysis (PCA). Assuming that reionization at $z > 6$ is primarily driven by stellar sources, we decompose the unknown function $N_{ion}(z)$, representing the number of photons in the intergalactic medium per baryon in collapsed objects, into its principal components and constrain the latter using the photoionization rate, Γ_{PI}, obtained from Lyα forest Gunn–Peterson optical depth, the 7 yr *Wilkinson Microwave Anisotropy Probe* (*WMAP*7) electron scattering optical depth τ_{el} and the redshift distribution of Lyman-limit systems dN_{LL}/dz at $z \sim 3.5$. The main findings of our analysis are as follows. (i) It is sufficient to model $N_{ion}(z)$ over the redshift range $2 < z < 14$ using five parameters to extract the maximum information contained within the data. (ii) All quantities related to reionization can be severely constrained for $z < 6$ because of a large number of data points whereas constraints at $z > 6$ are relatively loose. (iii) The weak constraints on $N_{ion}(z)$ at $z > 6$ do not allow to disentangle different feedback models with present data. There is a clear indication that $N_{ion}(z)$ must increase at $z > 6$, thus ruling out reionization by a single stellar population with non-evolving initial mass function, and/or star-forming efficiency, and/or photon escape fraction. The data allow for non-monotonic $N_{ion}(z)$ which may contain sharp features around $z \sim 7$. (iv) The PCA implies that reionization must be 99 per cent completed between $5.8 < z < 10.3$ (95 per cent confidence level) and is expected to be 50 per cent complete at $z \approx 9.5$–12. With future data sets, like those obtained by *Planck*, the $z > 6$ constraints will be significantly improved.

Key words: intergalactic medium – cosmology: theory – dark ages, reionization, first stars – large-scale structure of Universe.

1 INTRODUCTION

The importance of studying hydrogen reionization at high redshifts lies in the fact that it is tightly coupled to properties of first luminous sources and subsequent galaxy formation (for reviews, see Barkana & Loeb 2001; Loeb & Barkana 2001; Choudhury & Ferrara 2006a; Choudhury 2009). In recent years, studies in reionization have been boosted by (i) the availability of a wide range of data sets and (ii) the expectation that the volume of data would increase rapidly over the next few years (for reviews, see Fan, Carilli & Keating 2006a; Furlanetto, Oh & Briggs 2006). Given such a large amount of data, it is important to develop theoretical and statistical methods so that maximum information can be extracted.

Theoretically, reionization is modelled either semi-analytically or by numerical simulations. Unfortunately, the physical processes relevant to reionization are so complex that neither of the two approaches can capture the overall picture entirely. The simulations are indispensable for understanding detailed spatial distribution of ionized regions and topology of reionization. However, if one is interested in the evolution of globally averaged quantities, then semi-analytical models prove to be very useful in providing insights. The main reason for this is that these models can probe a wide range of parameter space which can be quite large depending on our ignorance of the different processes.

At present, our understanding of reionization is that it is primarily driven by ultraviolet (UV) radiation from stellar sources forming within galaxies. The major uncertainty in modelling reionization is to model the star formation history and transfer of radiation from the galaxies to the intergalactic medium (IGM) which is usually parametrized through N_{ion}, the number of photons entering the IGM per baryon in collapsed objects. This parameter, in principle, has a dependence on z which can arise from evolution of star-forming efficiency, fraction of photons escaping from the host halo and chemical and radiative feedback processes. Note that this

★E-mail: smitra@hri.res.in (SM); tirth@hri.res.in (TRC); andrea.ferrara@sns.it (AF)

parameter remains uncertain even in numerical simulations, hence the semi-analytical models can become handy in studying a wide range of parameter values and the corresponding agreement with data sets. In analytical studies, $N_{ion}(z)$ is either taken to be a piece-wise constant function (Wyithe & Loeb 2003; Choudhury & Ferrara 2005) or parametrized using some known functions (Chiu, Fan & Ostriker 2003; Pritchard, Loeb & Wyithe 2010) or modelled using a physically motivated prescription (Choudhury & Ferrara 2006b). In particular, a model involving metal-free and normal stars with some prescription for radiative and chemical feedback can match a wide range of observations (Choudhury & Ferrara 2006b; Gallerani, Choudhury & Ferrara 2006) and possibly make prediction regarding search for reionization sources by future experiments (Choudhury & Ferrara 2007).

However, the fact remains that many of the physical processes involved in modelling N_{ion} are still uncertain. Given this, it is worthwhile doing a detailed probe of the parameter space and determines the range of reionization histories that are allowed by the data. In other words, rather than working out the uncertain physics, one can ask the question as to what are the forms of $N_{ion}(z)$ implied by the data itself. It is expected that in near future, with more data sets becoming available, the allowed range in the forms of $N_{ion}(z)$ would be severely constrained, thus telling us exactly how reionization occurred. Now, it is obvious that the constraints on $N_{ion}(z)$ will not be same for all redshifts, points where there are more and better data available, the constraint would be more tight. Similarly, since we deal with a heterogeneous set of data, it is expected that the constraints would depend on the nature of data used. It is thus important to know which aspects of reionization history can be constrained by what kind of data sets. A method which is ideally suited to tackle this problem is to use the principal component analysis (PCA); this is a technique to compute the most meaningful basis to re-express the unknown parameter set and the hope is that this new basis will reveal hidden detailed statistical structure.

In this work, we make a preliminary attempt to constrain $N_{ion}(z)$ using PCA and hence estimate the uncertainties in the reionization history. The main objective of the work would be to find out the widest possible range in reionization histories allowed by the different data sets.

Throughout the paper, we assume a flat Universe with cosmological parameters given by the 7 yr *Wilkinson Microwave Anisotropy Probe* (*WMAP*7) best-fitting values: $\Omega_m = 0.27$, $\Omega_\Lambda = 1 - \Omega_m$, $\Omega_b h^2 = 0.023$ and $h = 0.71$. The parameters defining the linear dark matter power spectrum we use are $\sigma_8 = 0.8$, $n_s = 0.96$ and $dn_s/d\ln k = 0$ (Larson et al. 2011).

2 SEMI-ANALYTICAL MODEL OF REIONIZATION

2.1 Features of the model

The semi-analytical model used in this work is based on Choudhury & Ferrara (2005) and Choudhury & Ferrara (2006b). Let us first summarize the main features of the model along with the modifications made in this work.

(i) The model accounts for IGM inhomogeneities by adopting a lognormal distribution according to the method outlined in Miralda-Escudé, Haehnelt & Rees (2000); reionization is said to be complete once all the low-density regions (say, with overdensities $\Delta < \Delta_{crit} \sim 60$) are ionized. The mean free path of photons is thus determined essentially by the distribution of high-density regions:

$$\lambda_{mfp}(z) = \frac{\lambda_0}{[1 - F_V(z)]^{2/3}},\qquad(1)$$

where F_V is the volume fraction of ionized regions and λ_0 is a normalization parameter. In our earlier works, the value of this parameter was fixed by comparing with low-redshift observations while in this work, we treat it as a free parameter. We follow the ionization and thermal histories of neutral, H II and He III regions simultaneously and self-consistently, treating the IGM as a multiphase medium.

(ii) The model assumes that reionization is driven by stellar sources. The stellar sources can further be divided into two classes, namely, (i) metal-free (i.e. Population III) stars having a Salpeter initial mass function (IMF) in the mass range 1–100 $M_\odot$: they dominate the photoionization rate at high redshifts; (ii) Population II stars with subsolar metallicities also having a Salpeter IMF in the mass range 1–100 $M_\odot$.

(iii) Reionization by UV sources is accompanied by photoheating of the gas, which can result in a suppression of star formation in low-mass haloes. We compute such (radiative) feedback self-consistently from the evolution of the thermal properties of the IGM.

(iv) Furthermore, the *chemical feedback* including Population III→Population II transition is implemented using merger-tree-based genetic approach (Schneider et al. 2006). Under this approach, it is assumed that if a given star-forming halo has a progenitor which formed Population III stars, then the halo under consideration is enriched and cannot form Population III stars. In this work, we introduce an analytical formula for the transition from Population III to Population II phase using the conditional probability of Press–Schechter mass function (Lacey & Cole 1993). The probability that a halo of mass M at z never had a progenitor in the mass range $[M_{min}(z), M + M_{res}]$ is given by

$$f_{III}(M, z) = \frac{2}{\pi} \tan^{-1} \left[\frac{\sigma(M + M_{res}) - \sigma(M)}{\sigma(M_{min}(z)) - \sigma(M + M_{res})} \right],\qquad(2)$$

where M_{min} is the minimum mass of haloes which are able to form stars and M_{res} represents the minimum increase in mass (either by accretion or by merger) of an object so that it may be identified as a new halo. The fraction of collapsed haloes which are able to form Population II and Population III stars at redshift z are given by the following relations:

$$
\begin{aligned}
f_{coll,II}(z) &= \frac{1}{\bar{\rho}_m} \int_{M_{min}(z)}^{\infty} dM\, [1 - f_{III}(M, z)] M \frac{\partial n(M, z)}{\partial M}, \\
f_{coll,III}(z) &= \frac{1}{\bar{\rho}_m} \int_{M_{min}(z)}^{\infty} dM\, f_{III}(M, z) M \frac{\partial n(M, z)}{\partial M},
\end{aligned}\qquad(3)
$$

with $f_{coll,II}(z) + f_{coll,III}(z) = f_{coll}(z)$. The quantity $\bar{\rho}_m$ is the comoving density of dark matter and $\partial n/\partial M$ is number density of collapsed objects per unit comoving volume per unit mass range (Press & Schechter 1974).

(v) Given the collapsed fraction, this model calculates the production rate of ionizing photons in the IGM as

$$\dot{n}_{ph}(z) = n_b \left[N_{ion,II} \frac{df_{coll,II}}{dt} + N_{ion,III} \frac{df_{coll,III}}{dt} \right],\qquad(4)$$

where n_b is the total baryonic number density in the IGM and $N_{ion,II}(N_{ion,III})$ is the number of photons from Population II (Population III) stars entering the IGM per baryon in collapsed objects. The parameter N_{ion} can actually be written as a combination of various

other parameters:

$$N_{\mathrm{ion}} \equiv \epsilon_* f_{\mathrm{esc}} m_{\mathrm{p}} \int_{\nu_{H_{\mathrm{I}}}}^{\infty} \mathrm{d}\nu \left[\frac{\mathrm{d}N_\nu}{\mathrm{d}M_*} \right] \equiv \epsilon m_{\mathrm{p}} \int_{\nu_{H_{\mathrm{I}}}}^{\infty} \mathrm{d}\nu \left[\frac{\mathrm{d}N_\nu}{\mathrm{d}M_*} \right], \qquad (5)$$

where ϵ_* denotes the star-forming efficiency (fraction of baryons within collapsed haloes going into stars), f_{esc} is the fraction of photons escaping into the IGM, $[\mathrm{d}N_\nu/\mathrm{d}M_*]$ gives the number of photons emitted per frequency range per unit mass of stars (which depends on the stellar IMF and the corresponding stellar spectrum) and $\epsilon \equiv \epsilon_* f_{\mathrm{esc}}$. For Population II stars with subsolar metallicities having a Salpeter IMF in the mass range 1–100 $M_\odot$, we get $N_{\mathrm{ion,II}} \approx 3200 \, \epsilon_{\mathrm{II}}$, while for Population III stars having a Salpeter IMF in the mass range 1–100 $M_\odot$, we get $N_{\mathrm{ion,III}} \approx 35000 \, \epsilon_{\mathrm{III}}$.

In this section, we take $\epsilon_{\mathrm{II}}, \epsilon_{\mathrm{III}}$ (or, equivalently $N_{\mathrm{ion,II}}, N_{\mathrm{ion,III}}$) to be independent of z and M, which implies that the star-forming efficiencies and the escape fractions do not depend on the mass of the star-forming halo and also do not evolve. However, note that the effective N_{ion} (which is the appropriately weighted average of $N_{\mathrm{ion,II}}$ and $N_{\mathrm{ion,III}}$) evolves with z:

$$N_{\mathrm{ion}}(z) = \frac{N_{\mathrm{ion,II}}(\mathrm{d}f_{\mathrm{coll,II}}/\mathrm{d}t) + N_{\mathrm{ion,III}}(\mathrm{d}f_{\mathrm{coll,III}}/\mathrm{d}t)}{(\mathrm{d}f_{\mathrm{coll,II}}/\mathrm{d}t) + (\mathrm{d}f_{\mathrm{coll,III}}/\mathrm{d}t)}. \qquad (6)$$

At high redshifts, we expect $\mathrm{d}f_{\mathrm{coll,II}}/\mathrm{d}t \to 0$, hence $N_{\mathrm{ion}}(z) \to N_{\mathrm{ion,III}}$, and similarly at low redshifts where chemical enrichment is widespread, we have $N_{\mathrm{ion}}(z) \to N_{\mathrm{ion,II}}$.

(vi) We also include the contribution of quasars based on their observed luminosity function at $z < 6$ (Hopkins, Richards & Hernquist 2007); we assume that they have negligible effects on IGM at higher redshifts. They are significant sources of photons at $z \lesssim 4$ and are particularly relevant for studying helium reionization.

(vii) The free parameters for this analysis would be $\epsilon_{\mathrm{II}}, \epsilon_{\mathrm{III}}$ (or, equivalently $N_{\mathrm{ion,II}}, N_{\mathrm{ion,III}}$) and λ_0, the normalization which determines the mean free path of photons.

(viii) Usually, the model is constrained by comparing with a variety of observational data, namely, (1) redshift evolution of Lyman-limit absorption systems (LLS), (2) IGM Lyα and Lyβ optical depths, (3) electron scattering optical depth, (4) temperature of the mean intergalactic gas and (5) cosmic star formation history. However, most of the constraints on the model come from a subset of the above data sets. In this work, we would like to carry out a detailed likelihood analysis of the parameters. Hence to keep the analysis simple, the likelihood analysis is done using only three particular data sets which are discussed as follows.

(1) We use estimates for the photoionization rates Γ_{PI} obtained using Lyα forest Gunn–Peterson optical depth observations and a large set of hydrodynamical simulations (Bolton & Haehnelt 2007). The error bars in these data points take into account the uncertainties in the thermal state of the IGM in addition to the observational errors in the Lyα optical depth. The data points have a mild dependence on the cosmological parameters which has been taken into account in this work. We also find that although the error bars on Γ_{PI} are highly asymmetric, those on $\log(\Gamma_{\mathrm{PI}})$ are relatively symmetric; hence we use values of $\log(\Gamma_{\mathrm{PI}})$ and the corresponding errors in our likelihood analysis. The photoionization rate can be obtained in our model from $\dot{n}_{\mathrm{ph}}(z)$ using the relation

$$\Gamma_{\mathrm{PI}}(z) = (1+z)^3 \int_{\nu_{H_{\mathrm{I}}}}^{\infty} \mathrm{d}\nu \, \lambda_{\mathrm{mfp}}(z;\nu)\dot{n}_{\mathrm{ph}}(z;\nu)\sigma_H(\nu), \qquad (7)$$

where ν the frequency of radiation, $\nu_{H_{\mathrm{I}}}$ is the threshold frequency for photoionization of hydrogen and $\sigma_H(\nu)$ is the photoionization cross-section of hydrogen.

(2) The second set of observations we have used corresponds to the *WMAP*7 data on electron scattering optical depth τ_{el} (Larson

et al. 2011). The reported value of this quantity depends on the background cosmological model used. In this work, we restrict ourselves to the flat cold dark matter (CDM) universe with a cosmological constant and use the corresponding constraints on τ_{el}. Furthermore, the τ_{el} constraint is treated as a single data point which should be thought as a simplification because cosmic microwave background (CMB) polarization observations are, in principle, sensitive to the shape of the reionization history (Burigana et al. 2008). However, we have checked and found that the range of reionization histories considered in this paper would hardly make any difference to the currently observed large angular scale polarization anisotropies other than the value of τ_{el}. The quantity τ_{el} can be obtained from our model given the global reionization history, in particular the comoving density of free electrons $n_{\mathrm{e}}(z)$:

$$\tau_{\mathrm{el}}(z) = \sigma_{\mathrm{T}} c \int_0^{z[t]} \mathrm{d}t \, n_{\mathrm{e}} (1+z)^3, \qquad (8)$$

where σ_{T} is the Thomson scattering cross-section.

(3) Finally, we use the redshift distribution of LLS $\mathrm{d}N_{\mathrm{LL}}/\mathrm{d}z$ at $z \sim 3.5$ (Prochaska, O'Meara & Worseck 2010).[1] The data points are obtained using a large sample of quasi-stellar object (QSO) spectra which results in extremely small statistical errors. However, there are various systematic effects arising from effects like the incidence of proximate LLS and uncertainties in the continuum. Usually, these effects contribute to about 10–20 per cent uncertainty in the data points. The quantity $\mathrm{d}N_{\mathrm{LL}}/\mathrm{d}z$ can be calculated in our model from the mean free path:

$$\frac{\mathrm{d}N_{\mathrm{LL}}}{\mathrm{d}z} = \frac{c}{\sqrt{\pi}\,\lambda_{\mathrm{mfp}}(z)H(z)(1+z)}. \qquad (9)$$

Note that inclusion of the Lyman-limit systems in the analysis is crucial for constraining the parameter λ_0.

The likelihood function used in our calculations is given by

$$L \propto \exp(-\mathcal{L}), \qquad (10)$$

where $\mathcal{L}$ is the negative of the log-likelihood. It is estimated using the relation

$$\mathcal{L} = \frac{1}{2} \sum_{\alpha=1}^{n_{\mathrm{obs}}} \left[\frac{\mathcal{G}_\alpha^{\mathrm{obs}} - \mathcal{G}_\alpha^{\mathrm{th}}}{\sigma_\alpha} \right]^2, \qquad (11)$$

where $\mathcal{G}_\alpha$ represents the set of n_{obs} observational data points described above, i.e. $\mathcal{G}_\alpha = \{\log(\Gamma_{\mathrm{PI}}), \, \tau_{\mathrm{el}}, \, \mathrm{d}N_{\mathrm{LL}}/\mathrm{d}z\}$ and σ_α are the corresponding observational error bars. We constrain the free parameters by maximizing the likelihood function. We impose a prior such that reionization should be complete by $z = 5.8$, otherwise it will not match that Lyα and Lyβ forest transmitted flux data.

2.2 Reionization constraints

The results of our likelihood analysis using the reionization model described above are summarized in Table 1. The evolution of various quantities for models which are allowed within 95 per cent confidence limit is shown in Fig. 1.

---

[1] We did not include the more recent measurements of $\mathrm{d}N_{\mathrm{LL}}/\mathrm{d}z$ by Songaila & Cowie (2010) because the values are systematically larger than the ones quoted in Prochaska et al. (2010) at $z \sim 3.5$; inclusion of both the data sets would lead to a bad fit for the model. The Songaila & Cowie (2010) set has a data point at $z \sim 6$ which is not present in other data sets, however, the present error bar on that particular point is relatively large and hence excluding it does not affect our constraints significantly.

Table 1. The marginalized posterior probabilities with 95 per cent confidence limits errors of all free parameters (top three parameters) and derived parameters (from the fourth parameter down) for the reionization model with Population II and Population III stars.

Parameters	Mean value	95 per cent confidence limits
ϵ_{II}	0.003	[0.001, 0.005]
ϵ_{III}	0.020	[0.000, 0.043]
λ_0	5.310	[2.317, 9.474]
$z(Q_{HII} = 0.5)$	9.661	[7.894, 11.590]
$z(Q_{HII} = 0.99)$	6.762	[5.800, 7.819]

The top left-hand panel of the figure shows the evolution of the effective N_{ion} as given by equation (6). One can see that the quantity attains a constant value ≈ 10 at $z < 6$ which is a consequence of the fact that the photon emissivity at those epochs are purely determined by Population II stars. However at higher redshifts, the value of N_{ion} increases with z because of the presence of Population III stars. It is clear that the data cannot be fitted with Population II stars with constant $N_{ion,II}$ alone, one requires a rise in N_{ion} at higher redshifts. For the kind of chemical feedback employed in the model, the rise is rather smooth and gradual.

The mean values of parameters quoted in Table 1 are similar to the best-fitting model described in Choudhury & Ferrara (2006b) and hence the corresponding reionization history is similar to those described in the same paper. This can be readily verified from Fig. 1 where we see that reionization starts around $z \approx 15$ driven by

Population III stars, and it is 90 per cent complete by $z \approx 7.5$. After a rapid initial phase, the growth of the volume filled by ionized regions slows down at $z \lesssim 10$ due to the combined action of chemical and radiative feedback, making reionization a considerably extended process completing only at $z \approx 6$. We refer the reader to our earlier papers for a discussion of this model. Our likelihood analysis shows that reionization is 50 (99) per cent complete between redshifts $z = 7.9–11.6$ (5.8–7.8) at 95 per cent confidence level. Hence, under the assumptions made in the model, we find that completion of reionization cannot occur earlier than $z \approx 8$, essentially ruling out models of very early reionization. The reason for this is that the number of photons in the IGM at $z = 6$ is very low as implied by the Lyα forest data. In order to take the data point into account, the models typically cannot have too high an emissivity at $z \sim 6$. On the other hand, the constraints on τ_{el} imply that reionization must be initiated early enough. Thus the IGM has to go through a gradual reionization phase. As we discussed above, the gradual reionization is maintained by a combined action of radiative and chemical feedback effects.

Interestingly, we find that a couple of data points for dN_{LL}/dz lie above the 2σ limits of our analysis. In models where these points agree with the data, the photon mean free path λ_{mfp}, and hence the photoionization rate Γ_{PI}, are relatively smaller. These lead to larger Gunn–Peterson optical depths which then violate the Lyα forest constraints. This discrepancy can arise either (i) because of some unaccounted systematics present in the data or (ii) from the simplifying assumptions made in our models for calculating λ_{mfp}. The actual reason needs to be investigated further.
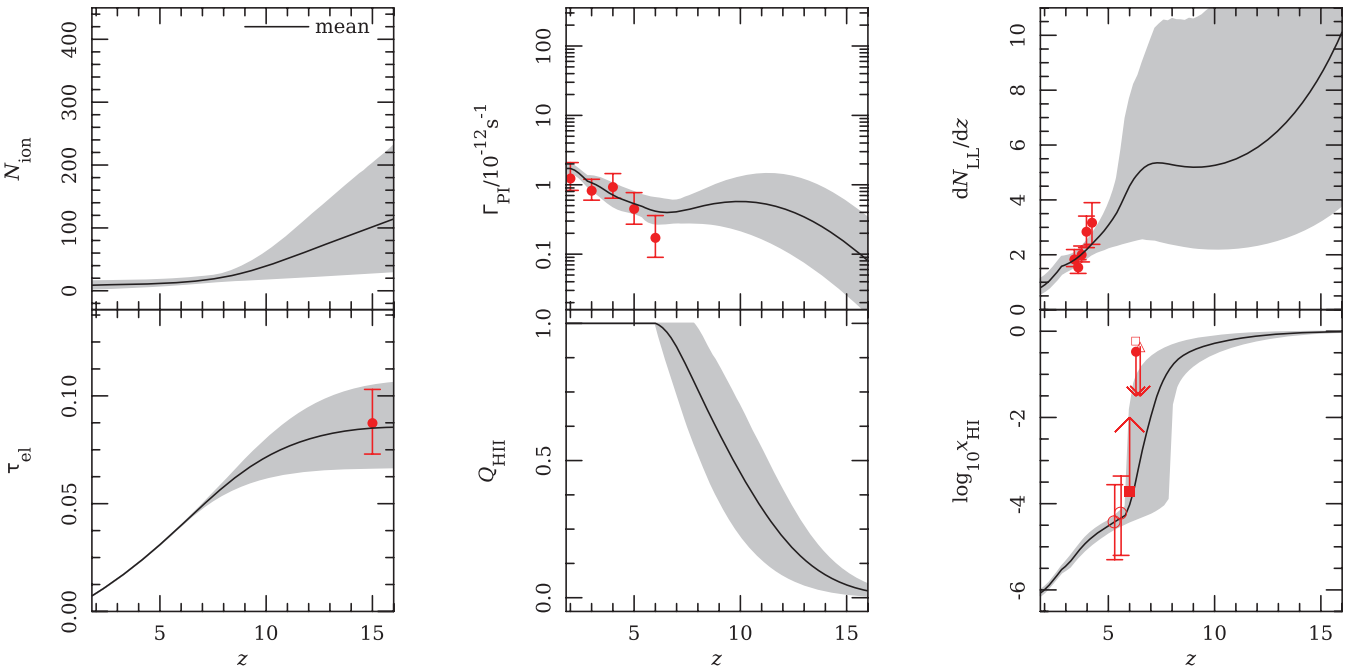
Figure 1. The marginalized posterior distribution of various quantities related to reionization history for a model with chemical feedback (Choudhury & Ferrara 2006b). The solid lines correspond to the model described by mean values of the parameters while the shaded regions correspond to 2σ limits. The points with error bars denote the observational data points. Top-left: the evolution of the effective $N_{ion}(z)$; top-middle: the hydrogen photoionization rate $\Gamma_{PI}(z)$ along with the constraints from Bolton & Haehnelt (2007); top-right: the LLS distribution dN_{LL}/dz with data points from Prochaska et al. (2010); bottom-left: the electron scattering optical depth τ_{el} with the *WMAP7* constraint (Larson et al. 2011); bottom-middle: the volume filling factor of H II regions $Q_{HII}(z)$; bottom-right: the global neutral hydrogen fraction $x_{HI}(z)$ with observational limits from QSO absorption lines (Fan et al. 2006a; filled square), Lyα emitter luminosity function (Kashikawa et al. 2006; open triangle) and GRB spectrum analysis (Totani et al. 2006; open square). Also shown are the constraints using dark gap statistics on QSO spectra (Gallerani et al. 2008a; open circles) and GRB spectra (Gallerani et al. 2008b; filled circle).

3 PRINCIPAL COMPONENT ANALYSIS

3.1 Motivation

It is most likely that the star-forming efficiencies and escape fractions and hence N_{ion} are functions of halo mass and redshift; however, since the dependencies are not well understood, they were taken to be constant for each considered stellar population in the previous section. The question one can ask is that how would the constraints on reionization histories of the previous section change when the evolution of N_{ion} is taken into account. Ideally one would like to do a rigorous likelihood analysis with N_{ion} varying with z and see the possible ranges of reionization histories consistent with available data. One possible approach could be to parametrize $N_{ion}(z)$ using some (known) function and constrain the parameters of the function (Pritchard et al. 2010). However, it is possible that the reionization constraints thus obtained could depend on the nature of the function chosen. In addition, it is not clear as to how many parameters should be used to parametrize the function.

An alternative approach is to assume $N_{ion}(z)$ to be completely arbitrary and decompose it into principal components. These principal components essentially filter out components of the model which are most sensitive to the data. Obviously, these components are the ones which can be constrained most accurately, while the others cannot be done so. This PCA, thus, should give an idea as to which aspects of N_{ion} can be constrained with available data. This implies that one should get a clear idea about the optimum number of parameters required to model N_{ion} to fit the data most accurately.

In order to carry out such analysis, we modify the model described in the previous section in following respects.

(i) We take N_{ion} to be a function of z. Unlike in the previous section, we do not explicitly assume the presence of two population of stars but rather we include only one stellar population; any change in the characteristics of these stars over time would be accounted for in the evolution of N_{ion}.

(ii) Clearly, the chemical feedback prescription has to abandoned in this model, as there are no two different populations of stars anymore. The chemical feedback is rather taken into account indirectly by the evolution of N_{ion}. However, we retain radiative feedback in the model given its weak dependence on the specific stellar population properties.

In recent years there has been a wide use of this method in cosmological data analysis. The first set of works was mostly related to CMB data where e.g. Efstathiou & Bond (1999) and Efstathiou (2002) used PCA of CMB anisotropy measurements to investigate degeneracies among cosmological parameters. Kadota et al. (2005) applied PCA to study how accurately CMB observables can constrain inflaton potential in a model-independent manner. Leach (2006) used PCA techniques for measuring departures from scale invariance in the primordial power spectrum of density perturbations using CMB C_l data. Mortonson & Hu (2008) developed a model-independent method to study the effects of reionization on the large-scale E-mode polarization for any reionization history with the help of PCA followed by the earlier work by Hu & Holder (2003). In the context of weak lensing surveys, Munshi & Kilbinger (2006) studied the degeneracies between cosmological parameters and measurement errors from cosmic shear surveys using PCA. The PCA has also been employed as an effective tool in the context of Type Ia supernova observations to constrain the equation of state of dark energy (Huterer & Starkman 2003; Huterer & Cooray 2005; Crittenden, Pogosian & Zhao 2009; Clarkson & Zunckel 2010).

3.2 Basic theory of PCA

Consider a set of n_{obs} observational data points labelled by $\mathcal{G}_\alpha$, $\alpha = 1, 2, \ldots, n_{obs}$. Recall that $\mathcal{G}_\alpha$ can represent combinations of different data sets, e.g. in our case $\mathcal{G}_\alpha = \{\log(\Gamma_{PI}), \ \tau_{el}, \ dN_{LL}/dz\}$.

Now, let us assume that our model contains an unknown function $N_{ion}(z)$, which we wish to constrain through observations. We can divide our entire redshift interval $[z_{min}, z_{max}]$ into (equal) bins of width Δz and represent $N_{ion}(z)$ by a set of n_{bin} discrete free parameters:

$$N_{ion}(z_i) \equiv N_i; \ i = 1, 2, \ldots, n_{bin}, \tag{12}$$

where

$$z_i = z_{min} + (i - 1)\Delta z, \tag{13}$$

and the bin width is given by

$$\Delta z = \frac{z_{max} - z_{min}}{n_{bin} - 1}. \tag{14}$$

In other words, we have modelled reionization using the value of N_{ion} in each redshift bin. We can also include other free parameters apart from $N_{ion}(z_i)$ in the analysis, like the normalization of the mean free path λ_0, cosmological parameters etc. However, for the moment let us assume that these parameters are fixed (known from other observations) and concentrate on $N_{ion}(z_i)$ only. We will address the inclusion of other parameters later in this section.

The next step is to assume a fiducial model for $N_{ion}(z_i)$, which we denote by $N_{ion}^{fid}(z_i)$. The fiducial model should be chosen such that it is close to the 'true' model. The departure from the fiducial model is denoted by

$$\delta N_{ion}(z_i) = N_{ion}(z_i) - N_{ion}^{fid}(z_i) \equiv \delta N_i. \tag{15}$$

We can then construct the $n_{bin} \times n_{bin}$ Fisher matrix:

$$F_{ij} = \sum_{\alpha=1}^{n_{obs}} \frac{1}{\sigma_\alpha^2} \frac{\partial \mathcal{G}_\alpha^{th}}{\partial N_i} \frac{\partial \mathcal{G}_\alpha^{th}}{\partial N_j}, \tag{16}$$

where $\mathcal{G}_\alpha^{th}$ is theoretical value of $\mathcal{G}_\alpha$ modelled using the N_i and σ_α is the observational error on $\mathcal{G}_\alpha$. The derivatives in the above relation are evaluated at the fiducial model $N_i = N_i^{fid}$.[2]

Once the Fisher matrix is constructed, we can determine its eigenvalues and corresponding eigenvectors. The principal value decomposition is then given by the eigenvalue equation

$$\sum_{j=1}^{n_{bin}} F_{ij} S_{jk} = \lambda_k S_{ik}, \tag{17}$$

where λ_k are the eigenvalues and the eigenfunctions corresponding to λ_k are the kth column of the matrix S_{ik}, these are the principal components of N_i. They can be thought of a function of z, i.e. $S_{ik} = S_k(z_i)$.

The eigenvalues λ_k are usually ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{nbin}$, i.e. λ_1 corresponds to the largest eigenvalue while λ_{nbin} the smallest. The eigenfunctions are both orthonormal and complete and hence we can expand any function of z as linear combinations of them. In particular we can expand the departure from the fiducial model as

$$\delta N_i = \sum_{k=1}^{n_{bin}} m_k S_k(z_i); \ \ m_k = \sum_{i=1}^{n_{bin}} \delta N_{ion}(z_i) S_k(z_i), \tag{18}$$

---

[2] It is worthwhile to mention that any analysis based on the Fisher matrix F_{ij}, in principle, depends on the fiducial model chosen. The PCA, which essentially involves diagonalizing F_{ij}, is thus dependent on the choice of N_i^{fid} too. In this sense, the PCA is not completely model independent.

where m_k are the expansion coefficients with $m_k = 0$ for the fiducial model. We can now describe our model by the coefficients m_k rather than the original parameters δN_i. The advantage is that, unlike N_i, the coefficients m_k are uncorrelated with variances given by the inverse eigenvalue:

$$\langle m_i \, m_j \rangle = \frac{1}{\lambda_i} \delta_{ij}. \tag{19}$$

The accuracy with which we can determine δN_{ion} at a particular z_i is determined by the Cramer–Rao bound:

$$\langle \delta N_{ion}^2(z_i) \rangle \geq \sum_{k=1}^{n_{bin}} \frac{S_k^2(z_i)}{\lambda_k}. \tag{20}$$

So, the largest eigenvalues correspond to minimum variance. The eigenvalues which are smaller would essentially increase the uncertainty in determining $\delta N_{ion}(z_i)$. Hence, most of the information relevant for the observed data points $\mathcal{G}_\alpha$ is contained in the first few modes with the largest eigenvalues. One may then attempt to reconstruct the function $\delta N_{ion}(z_i)$ using only the first $M \leq n_{bin}$ modes:

$$\delta N_i^{(M)} = \sum_{k=1}^{M} m_k S_k(z_i). \tag{21}$$

However, in neglecting the last $n_{bin} - M$ terms, one introduces a bias in determining $\delta N_{ion}(z_i)$. One has to then use a carefully chosen M to perform the analysis; the choice usually depends on the particular problem in hand. We shall discuss our choice of M in the next section.

In realistic situations, there will be other free parameters (apart from m_k or δN_i) in the model; these could be e.g. the normalization of the mean free path λ_0, cosmological parameters etc. Let there be n_{ext} number of extra parameters other than m_k; this means that we are now dealing with a total of $n_{tot} = n_{bin} + n_{ext}$ parameters. In this case, we can still form the Fisher matrix of $n_{tot} \times n_{tot}$ dimensions which can be written as

$$\mathcal{F} = \begin{pmatrix} \mathbf{F} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{F}' \end{pmatrix}, \tag{22}$$

where $\mathbf{F}$ is the $n_{bin} \times n_{bin}$-dimensional Fisher matrix for the δN_i, $\mathbf{F}'$ is the $n_{ext} \times n_{ext}$-dimensional Fisher matrix for the other parameters and $\mathbf{B}$ is a $n_{bin} \times n_{ext}$-dimensional matrix containing the cross-terms. One can then invert the above $\mathcal{F}$ to obtain the corresponding Hessian matrix $\mathcal{T} = \mathcal{F}^{-1}$. Following that, one simply retains the sub-block $\mathbf{T}$ corresponding to δN_i whose principal components will be 'orthogonalized' to the effect of the other parameters. The resulting 'degraded' sub-block will be (Press et al. 1992)

$$\tilde{\mathbf{F}} = \mathbf{T}^{-1} = \mathbf{F} - \mathbf{B}\mathbf{F}'^{-1}\mathbf{B}^T. \tag{23}$$

In this work we keep the cosmological parameters fixed; however, we still need to use the above formalism to marginalize over λ_0. In that case, obviously $n_{ext} = 1$.

4 RESULTS

The detailed results of our PCA are presented in this section.

4.1 Fiducial model

The first task is to make an assumption for the fiducial model N_{ion}^{fid} (z). The model should match the Γ_{PI} and dN_{LL}/dz data points at $z < 6$ and also produce a τ_{el} in the acceptable range. Unfortunately, the simplest model with N_{ion} being constant does not have these requirements (recall models with only Population II stars were disfavoured

in the previous section). We have found earlier that the effective N_{ion} should be higher at early epochs dominated by Population III stars and should approach a lower value at $z \sim 6$ determined by Population II stars. In this work we take N_{ion}^{fid} to be the model given by mean values of the free parameters in Section 2.2.

The choice of this N_{ion}^{fid} may seem somewhat arbitrary as there could be many other forms of N_{ion} which may match the data equally well. We have chosen this to be our fiducial model because of the following reasons: (i) it is obtained from a physically motivated model of star formation which includes both metal-free and normal stars; (ii) it is characterized by a higher N_{ion} at higher redshifts and hence produces a good match with different observations considered in this work and (iii) the transition from higher to lower values is smooth (i.e. there is no abrupt transition or sharp features). The final conclusions of this work (to be presented later in the section) would hold true for any fiducial model having these three properties (though the actual functional form might be different). The match with the data for our fiducial model is similar to fig. 2 of Choudhury (2009).

We have run the reionization models over a redshift range $[z_{min} : z_{max}] = [0:30]$, with a bin width of $\Delta z = 0.2$. This gives $n_{bin} = 151$. We have checked and found that our main conclusions are unchanged if we vary the bin width between 0.1 and 0.5.

The Fisher matrix F_{ij} defined in equation (16) is evaluated at the fiducial model and is shown as a shaded plot in the z–z plane in Fig. 2. First, the components of the matrix vanish for $z < 2$ because there are no data points considered at these redshifts. The plot shows different characteristics for F_{ij} at redshift intervals $2 < z < 6$ and $z > 6$. For $z < 6$, the values of F_{ij} are considerably higher because it is determined by the sensitivity of Γ_{PI} and dN_{LL}/dz on N_{ion} and it turns out that Γ_{PI} is extremely sensitive to changes in N_{ion}. One can see a band-like structure in the information matrix which essentially corresponds to the presence of data points. The regions where data points are sparse (or non-existent, like between $z = 2$ and 3), the value of F_{ij} is relatively smaller, implying that one cannot constrain N_{ion} from the data in those redshift bins. On the other hand, the information at $z > 6$ is determined by the sensitivity of τ_{el} on N_{ion}. One can see that $F_{ij} \to 0$ at the highest redshifts considered; this is expected because the collapsed fraction of haloes is negligible at those redshifts and hence there exist no free electrons to contribute to τ_{el}. The precise redshift range at which F_{ij} become negligible depends on the (measured) value of τ_{el}. For the *WMAP*7 measurements, we find that F_{ij} is negligible for $z > 14$; if e.g. the measured value of τ_{el} were higher, F_{ij} would be non-negligible
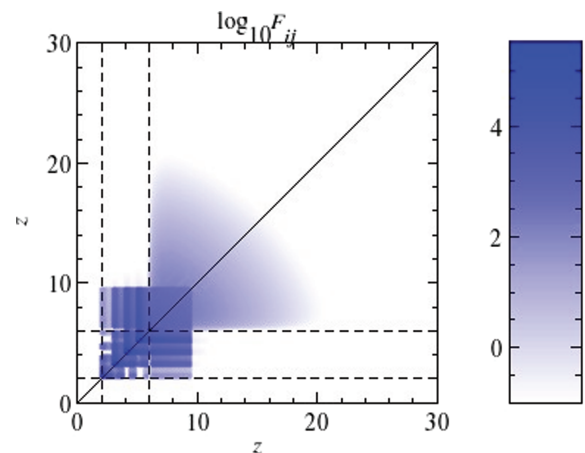

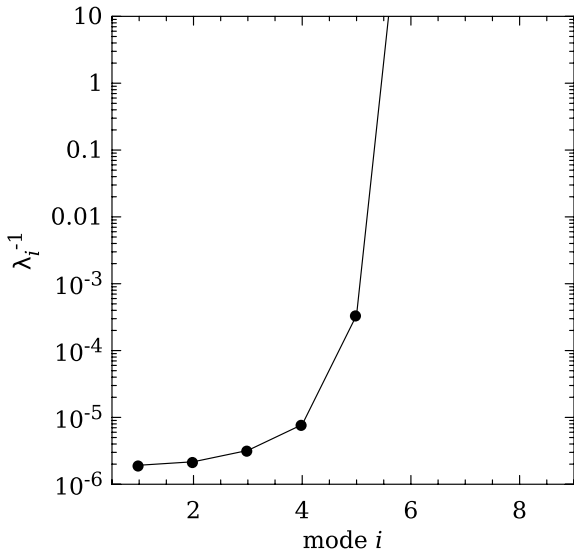**Figure 2.** The Fisher matrix F_{ij} in the z–z plane.

Figure 3. The inverse of the eigenvalues which essentially measures the variance on the corresponding coefficient m_i. For modes larger than five, the eigenvalues are extremely small.

Figure 4. The first five eigenmodes of the Fisher matrix, i.e. $S_k(z)$; $k = 1, \ldots, 5$.

till relatively higher redshifts. We can thus conclude that it is not possible to constrain any parameters related to star formation at redshifts $z > 14$ using the data sets we have considered in this work.

Once we diagonalize the matrix F_{ij}, we obtain its eigenvalues and the corresponding eigenmodes. The inverse of the eigenvalues, which are essentially the variances of the corresponding modes, is plotted in Fig. 3. Since the eigenvalues λ_i are sorted in ascending order, the variances are larger for higher modes. For modes $i > 5$, the eigenvalues are almost zero and the variances are extremely large. This implies that the errors on N_{ion} would increase dramatically if we include modes $i > 5$.

The first five eigenmodes which have the lowest variances are shown in Fig. 4. Clearly, all these modes tend to vanish at $z > 14$, which is because of F_{ij} being negligible at these redshifts. Also, modes are identically zero at $z < 2$ because we have not used any data points at these redshifts. The first four modes essentially trace the sensitivity of Γ_{PI} and dN_{LL}/dz at $z < 6$ on the value of N_{ion}. One can see a number of spikes and troughs in these modes whose positions correspond to the presence of data points and amplitudes correspond to the error bars on these data points (smaller the error, larger the amplitude). The shape of the fifth mode is vary much different from the previous four. This mode essentially contains the behaviour of N_{ion} at $z > 6$ and hence it characterizes the sensitivity of τ_{el} on N_{ion}. Since τ_{el} is obtained by integrating the reionization history over the whole redshift range, the sensitivity covers a wide range of redshifts (which is unlike the sensitivity of Γ_{PI}). The sensitivity is maximum around $z \approx 8$, which is determined by the nature of the fiducial model. The sensitivity falls at $z > 8$ because there is a reduction in the number of sources and free electrons. Interestingly the sensitivity falls at $z < 8$ too which is due to the fact that reionization is mostly complete at these redshifts $x_e \to 1$ and hence changing N_{ion} does not change the value of τ_{el} significantly.

The modes with smaller eigenvalues have large variances and hence introduce huge uncertainties in the determination of N_{ion}. The modes are characterized by sharp features at different redshifts and they do not contain any significant information about the overall reionization history.
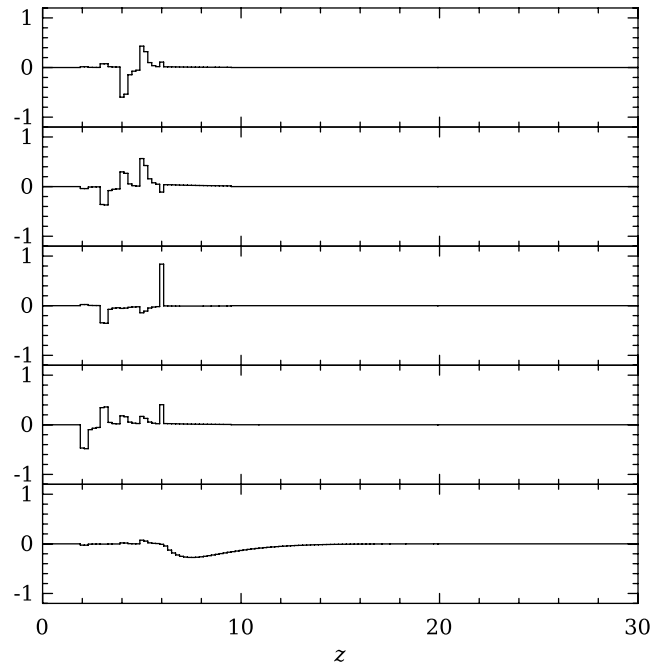
4.2 Choice of the number of modes

The next step in our analysis is to decide on how many modes M to use. In the case where $M = n_{bin}$, all the eigenmodes are included in the analysis and no information is thrown away. However, this would mean that modes with very small eigenvalues (and hence large uncertainties) are included and thus the errors in recovered quantities would be large. Reducing M is accompanied by a reduction in the error, but an increased chance of getting the recovered quantities wrong (which is known as bias).

It is thus natural to ask what could be the optimum value of M for calculations. The most straightforward way, which is used often, is to determine it by trial and error, i.e. more and more terms are added till one gets some kind of convergence in the recovered quantities (Mortonson & Hu 2008). Let us first work out the simplistic trial-and-error approach to fix M and as we shall see that this would be helpful in understanding recovery of various parameters using PCA. We have already discussed that inclusion of modes > 5 implies drastic rise in the errors. Hence, it seems that $M \leq 5$ would be a good choice. The question is whether throwing away such a large number of modes ($n_{bin} - M$) would introduce large biases in the recovered quantities.

In order to examine these issues in more detail, let us assume that the underlying 'true' form of N_{ion} is very different from the fiducial model we have chosen and then try to estimate the errors we make in recovering this underlying model using only the first few modes. In order to put our method to test, it is then natural to assume an underlying model which is noticeably different from the fiducial one and study its recovery using only the first few modes. Recall that the fiducial model represents a smoothly varying N_{ion}, so we assume the underlying input model to be one having an abrupt transition e.g. a step function:

$$N_{ion}^{inp}(z) = 10 \quad \text{for } z < 7$$
$$= 40 \quad \text{for } z \geq 7. \tag{24}$$

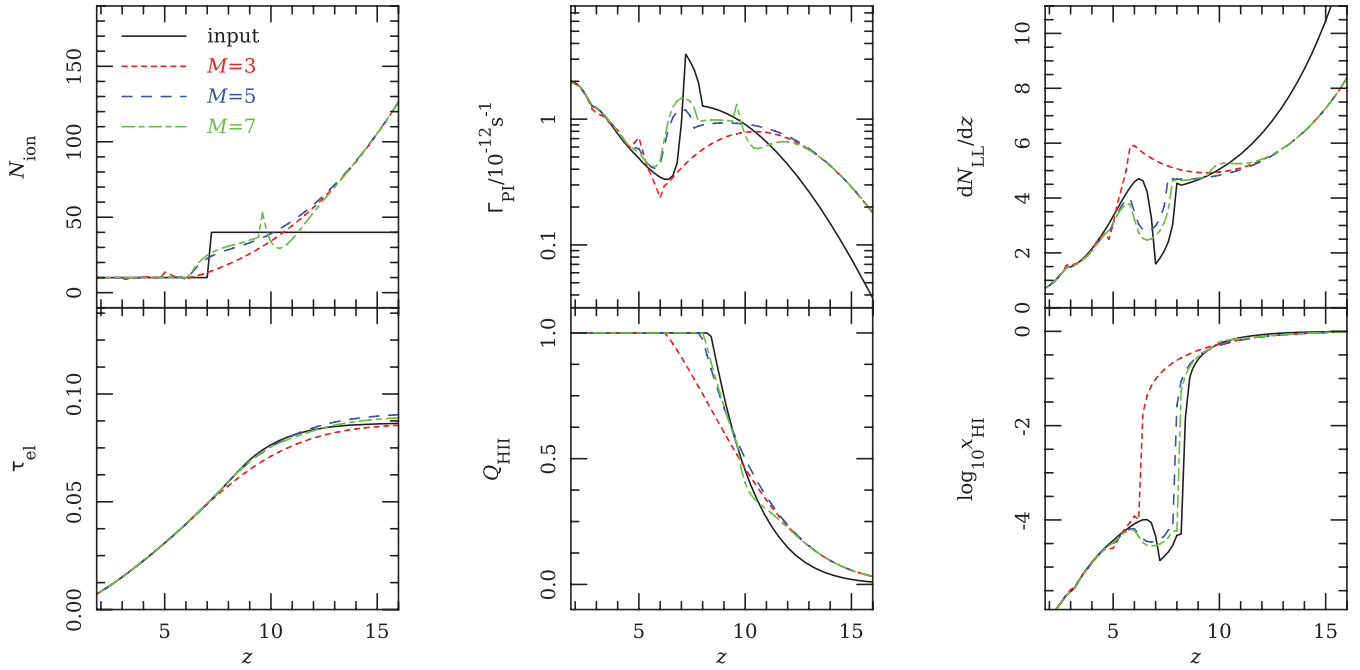**Figure 5.** Recovery of various quantities related to reionization when the input underlying model of $N_{ion}(z)$ is assumed to be a step function (shown by solid lines). The extent of recovery is shown when the first three (short-dashed lines), five (long-dashed lines) and seven (short-long-dashed lines) PCA modes are included in the analysis.

The parameters in the model are adjusted so that it matches with observations of Γ_{PI} and dN_{LL}/dz at $z < 6$ and also gives the correct observed value of τ_{el}. The idea would be to check whether we are able to recover quantities of interest with reasonable accuracy with $M = 5$. The model chosen above is similar to the abrupt-transition model considered in Choudhury & Ferrara (2005).

The results of our analysis are shown in Fig. 5 and in Table 2. In the figure, we have plotted, as functions of redshifts, the six quantities relevant to reionization which we would like to recover, namely, N_{ion} (top left-hand panel), the photoionization rate Γ_{PI} (top middle panel), redshift distribution of Lyman-limit systems dN_{LL}/dz (top right-hand panel), electron scattering optical depth τ_{el} (bottom left-hand panel), the volume filling factor of ionized regions Q_{HII} (bottom middle panel) and the globally averaged neutral hydrogen fraction x_{HI} (bottom right-hand panel). Different curves represent the input step model (solid) and the recovered quantities for three values of $M = 3, 5$ and 7 (short-dashed, long-dashed and short-long-dashed, respectively). We have not shown results for intermediate values of M (i.e. $M = 4$ and 6) because the difference between successive plots is too small to be noticed. It is clear from the top left-hand panel that the recovered N_{ion} is excellent for $z < 6$ because the fiducial and input models agree at these redshifts, which is a manifestation of the fact that the value of N_{ion} is highly constrained by good quality data points at these redshifts. On the other hand, the

recovery is quite poor for $z > 6$. This is because the evolution at $z > 6$ is only weakly constrained by τ_{el}. In particular at $z > 14$, the modes are essentially zero and hence all models tend to the fiducial one implying that it is impossible to recover N_{ion} at $z > 14$ with the first few modes.

The top middle and top right-hand panels show the corresponding plots for the photoionization rate Γ_{PI} and the redshift distribution of Lyman-limit systems dN_{LL}/dz, respectively. The input model has a sharp feature around $z \approx 7$ in both the quantities arising mainly from the abrupt step in N_{ion}. The reionization is complete ($Q_{HII} = 1$) at $z \approx 8$ after which the photoionization rate rises sharply because of overlap of ionized regions and consequent rise in mean free path (which manifests itself as a sharp drop in the number of LLS). This rise in Γ_{PI} is suddenly halted at $z = 7$ where we see a sharp decline because of the corresponding step decline in N_{ion}. Following that, Γ_{PI} settles to a smaller value (corresponding to a smaller value of N_{ion}) and subsequently shows a gradual rise arising again from the rise in mean free path. Interestingly, this feature is completely missing in the recovered model for $M = 3$ (and also for $M = 4$, not shown in the figure). The feature shows up when M is increased to 5, though the exact nature of this feature is not identical to the input one. Increasing M to 7 introduces other sharp features at $z \sim 10$ which are not present in the input model. Of course, the recovery at $z > 14$ is poor as most of the eigenmodes hardly contain any information at these redshifts and the recovered models simply follow the fiducial model. Hence, the recovery of the photoionization rate and the LLS distribution is probably not satisfactory overall, however, we can recover it with reasonable accuracy for $z < 12$ by considering the first $M = 5$ modes.

The recovery of τ_{el} is shown in the bottom left-hand panel. It is clear that the recovery is good for all values of M. In most reionization studies, the quantities of main interest are the Q_{HII} and x_{HI}, which are plotted in the bottom middle and bottom right-hand panels, respectively. One can easily see from both the panels that

Table 2. The recovered quantities for a input model where N_{ion} is represented by a step function when only the first M PCA modes are included in the analysis.

Parameters	Input value	Recovered values		
		$M = 3$	$M = 5$	$M = 7$
$z(Q_{HII} = 0.5)$	9.817	9.722	10.004	9.728
$z(Q_{HII} = 0.99)$	8.337	6.311	7.875	8.037

that the agreement between the $M = 3$ case and the input model is quite poor (which is the case for $M = 4$ as well). In particular, reionization is complete at $z \approx 8.5$ for the input model, while it completes only at $z \approx 6$ for $M = 3$ case (see Table 2). However, the moment M is increased to 5, one has a remarkable match with the input model, e.g. the difference in $Q_{H\,II}$ is < 0.1 for $z < 12$ while the difference is <10 per cent for $z < 10$. Unfortunately, we cannot recover the sharp feature in $x_{H\,I}$ around $z = 7$ for the input model (which corresponds to a similar feature in Γ_{PI}, discussed above) for the $M = 5$ case, however, the overall agreement with the input model is still quite good. The agreement is further improved as we increase M (to 7 in the plot) but that comes at the cost of increasing errors. As far as recovering the basic reionization history (i.e. evolution of $Q_{H\,II}$ and $x_{H\,I}$) is concerned, $M = 5$ seems to be the optimum choice.

It is important to point out that the recovery of various quantities related to reionization is good (or excellent, in some cases) even when the recovered value of N_{ion} is incorrect. This may seems surprising as it is the value of N_{ion} that acts as a source for reionization. To understand this apparent paradox, note that the recovery of N_{ion} is poor mostly at $z > 12$. At these redshifts the collapsed fraction df_{coll}/dt is typically small, hence the source emissivity $N_{ion}df_{coll}/dt \rightarrow 0$ at these epochs. Hence even if we change the value of N_{ion}, the absolute change in the emissivity is negligible and hence the reionization process remains relatively unaffected. There is another way of looking at it: the extent of recovery of various quantities at $z > 6$ is determined by the behaviour of PCA modes at $z > 6$ which, in turn, is determined by the data set related to τ_{el}. Now τ_{el} is most sensitive to the ionized fraction $Q_{H\,II}(z)$ at $z > 6$. Hence, it is not surprising that $Q_{H\,II}(z)$ would be nicely recovered at these redshifts. Such arguments can be extended for other quantities too. This also brings out the fact that in order to recover $N_{ion}(z)$ (and thus star formation, escape fraction and chemical feedback) reliably, one requires data points at $z > 12$ related to quantities which are sensitive to N_{ion}, like say, hypothetically, a good constraint on Γ_{PI} at $z \sim 12$ can constrain N_{ion} at those redshifts.

To summarize our results on recovering the input step model, the recovery of all the quantities is excellent for $z < 6$. We find that recovery of N_{ion} at $z > 6$ is not satisfactory. The recovery of Γ_{PI} at $z < 12$ is quite reasonable by considering the first $M = 5$ modes. Fortunately, the recovery of $Q_{H\,II}$ and $x_{H\,I}$ turns out to be excellent for $M = 5$. Hence we can use the coefficients m_i of these five best-constrained eigenmodes as our model parameters instead of $N_{ion}(z_i)$ without significant loss of information.

We should mention that the above analysis depends on the choice of the input model which is taken to be the step function. In fact, the recovery is better if the input N_{ion} is a smoother function (provided it satisfies the observational constraints, of course). In particular, all models which are bracketed by the fiducial model and the step model would end up giving good agreements for Γ_{PI}, dN_{LL}/dz, τ_{el}, $Q_{H\,II}$ and $x_{H\,I}$. Of course, if the input models have sharp features at some particular redshift(s) $z > 6$, those features may not be recovered satisfactorily by including only first few terms.

A slightly more formal approach is to estimate M by minimizing the quantity risk, which is defined as (Wasserman et al. 2001)

$$R = \sum_{i=1}^{n_{bin}} \left(\delta N_i^{(M)} \right)^2 + \sum_{i=1}^{n_{bin}} \left\langle \left(\delta N_i^{(M)} \right)^2 \right\rangle. \tag{25}$$

The first term in the right-hand side (RHS) is the bias contribution which arises from neglecting the higher order terms, and the second term is the uncertainty given by Cramer–Rao bound which rises as higher order terms (i.e. those corresponding to smaller eigenvalues)
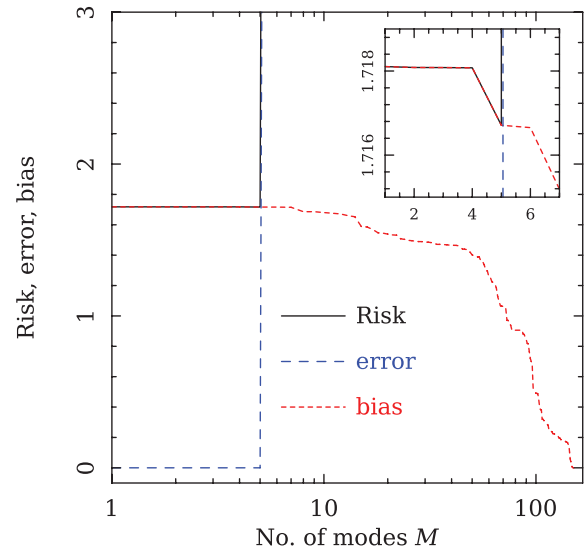
Figure 6. Dependence of risk, error and bias as defined in equation (25) on the number of modes M. The blow-up of a region around $M = 5$ is shown in the inset which shows that there is a clear minimum in the risk at $M = 5$.

are included

$$\left\langle \left(\delta N_i^{(M)} \right)^2 \right\rangle \geq \sum_{k=1}^{M} \frac{S_k^2(z_i)}{\lambda_k}. \tag{26}$$

However, the calculation of risk, as defined above, involves assumption of an 'underlying model', hence the determination of M using this method would be model dependent. Let us assume the underlying model to be the same as equation (24). Then the dependence of the risk on the number of modes M is shown in Fig. 6. In addition, we also show the plots of bias (first term of the RHS in equation 25) and the error (second term of the RHS in equation 25). It is clear that the value of error is small for lower M which is a direct consequence of small eigenvalues. The error shoots up drastically for $M > 5$ which is what we discussed in the previous section. On the other hand, the bias is higher for small M and decreases gradually as more and more terms in the summation are included. The risk, which is the sum of these two quantities, has a clear minimum at $M = 5$ (which is more clear from the inset in Fig. 6). Hence we conclude that $M = 5$ is the optimum value to be used.

The main conclusion of this section is that one needs five parameters to describe the reionization history which can be constrained with the data considered in this paper. Out of these five, four parameters are required to describe the emissivity at $z < 6$ where most of the data points exist; these parameters are the best-constrained ones. The fifth parameter characterizes the evolution of N_{ion} at $z > 6$ and is essentially determined by the *WMAP* constraints of τ_{el}. Inclusion of more parameters would lead to overfitting of the data and hence the constraints on the parameters would be highly uncertain.

4.3 Constraints on reionization history

The constraints on reionization are obtained by performing a Monte Carlo Markov Chain (MCMC) analysis over the parameter space of PCA amplitudes $\{m_1, \ldots, m_5\}$ and λ_0. The cosmological parameters are kept fixed to the *WMAP*7 best-fitting values. In order to carry out the analysis, we have developed a code based on the publicly

Table 3. The marginalized posterior probabilities with 95 per cent confidence limits errors of all free parameters (top six parameters) and derived parameters (from the seventh parameter down) for the reionization model with PCA.

Parameters	Mean value	95 per cent confidence limits
m_1	0.002	[−0.002, 0.018]
m_2	0.007	[−0.001, 0.032]
m_3	−0.003	[−0.012, 0.004]
m_4	0.003	[−0.003, 0.015]
m_5	−0.065	[−0.276, 0.003]
λ_0	4.450	[3.245, 5.906]
$z(Q_{H\,II} = 0.5)$	10.349	[9.528, 11.585]
$z(Q_{H\,II} = 0.99)$	8.357	[5.800, 10.270]

available COSMOMC[3] (Lewis & Bridle 2002) (which is widely used for running MCMC on CMB and other cosmological data sets). To get accurate results from MCMC, we ensure that the parameter chains contain enough independent samples over a sufficiently large volume of parameter space so that the density of the samples converges to the actual posterior probability distribution. We run a number of separate chains (varying between 5 and 10) until the Gelman and Rubin convergence statistics, R, corresponding to the ratio of the variance of parameters between chains to the variance within each chain, satisfies $R - 1 < 0.01$.

The mean values and the 95 per cent confidence limits on our parameters obtained from our analysis are shown in Table 3. Our fiducial model $m_1 = m_2 = m_3 = m_4 = m_5 = 0$ is included within the 95 per cent confidence limits of the parameters corresponding to the eigenmode amplitudes, however, the mean values show clear departures from the fiducial model. This implies that the model characterized by the mean values of parameters, loosely mentioned as the 'mean model' hereafter, is different from the fiducial one.

In order to see how different it is, we show the evolution of various quantities related to reionization is shown in Fig. 7. The solid lines represent the mean model while the shaded region corresponds to 95 per cent confidence limits. For comparison, we have also plotted the fiducial model (short-dashed) and the step model (long-dashed) which was introduced in Section 4.2. We find that the fiducial model is within the 95 per cent confidence limits for the whole redshift range, while the step model is within the 95 per cent confidence limits for $z < 10$. Also note that the fiducial model is actually near the edge of the shaded region, implying that there is a wide range of models allowed by the data which are characteristically different from the fiducial model.

The next point to note is that all the quantities are highly constrained at $z < 6$, which is expected as most of the observational information related to reionization exists only at those redshifts. The errors also decrease at $z > 12$ as there is practically no information in the PCA modes and hence all models converge towards the fiducial one. This implies that early stages of reionization are almost similar independent of the N_{ion} chosen. The most interesting information regarding reionization is concentrated within a redshift range $6 < z < 12$.

It is very clear from the plot of $N_{ion}(z)$ (top left-hand panel) that such quantity must necessarily increase from its constant value at $z < 6$. This rules out the possibility of reionization with a single stellar population having non-evolving IMF and/or star-forming

efficiency and/or escape fraction. The value of N_{ion} can be almost 40 times larger than its value at $z < 6$. Also note that N_{ion} need not be a monotonic function of z. For example, the mean model, which is constant for $z < 6$, shows an increase for $z > 6$ followed by a decrease at $z \approx 7$. The plot shows a subsequent increase around $z \approx 11$, however, one should remember that the information contained within eigenmodes are severely limited at these epochs.

From the plot of $\Gamma_{PI}(z)$ (top middle panel), we find that the mean model is consistent with the observational data at $z < 6$, as expected. The errors corresponding to 95 per cent confidence limits are also smaller at $z < 6$ for reasons discussed above. The photoionization rate for the fiducial model shows a smooth rise at $z > 6$ with a peak around $z \approx 10$, however, model described by the mean values of the parameters shows a much sharper rise and much prominent peak. The location of the peak is around $z \sim 6.5$. The highest value of Γ_{PI} allowed by the data can be as high as 10^{-10} s^{-1} (95 per cent confidence level), which is about 100 times the values typically observed at $z < 6$. The prominent peak-like structure is also present in the dN_{LL}/dz (top right-hand panel). Interestingly, the high-Γ_{PI} models predict that $dN_{LL}/dz \approx 0$ at $z \sim 6.5$, hence any sighting of LLS at these epochs would put more constraints on the models.

The limits on τ_{el} (bottom left-hand panel) are, as expected, similar to the *WMAP*7 constraints. We find that the mean τ_{el} is slightly higher than the best-fitting *WMAP*7 value because a wide range of models with early reionization is allowed by the data.

The constraints on the reionization history can be seen from the plot of $Q_{H\,II}(z)$ (bottom middle panel). The growth of $Q_{H\,II}$ for the fiducial model is somewhat gradual. On the other hand, the mean model, which is characterized by sharp peak structures in N_{ion} and Γ_{PI} at $z > 6$, shows a much faster rise in $Q_{H\,II}$ at initial stages, though the completion of reionization takes place only at $z \approx 6$. The shaded regions show that reionization can be complete as early as $z \approx 10.5$ (95 per cent confidence level). These models of early reionization are essentially characterized by high N_{ion} at $6.5 < z < 10$ (so that enough contribution to τ_{el} is achieved to match the *WMAP*7 constraints) followed by a sharp decrease at $z < 6.5$ so that the emissivity becomes low enough to match the photoionization rate obtained from Lyα forest data.

Similar conclusions can be obtained from the plot of $x_{H\,I}(z)$ (bottom right-hand panel). In general, the models allowed by the 95 per cent confidence limits are consistent with the available data points (shown by points with error bars). Models of very early reionization (i.e. those with high N_{ion} at $6.5 < z < 10$) show sharp decrease in $x_{H\,I}$ at $z \approx 10$ and it can become as low as 10^{-6} at $z \approx 6.5$. However, the neutral fraction has to increase sharply again at $z < 6.5$ (corresponding to sharp decrease in N_{ion}) so as to match the Lyα forest constraints. Thus the evolution of $x_{H\,I}$ is not monotonic for these models. On the other hand, models with relatively smoothly evolving N_{ion} (ones similar to the fiducial model) show gradual decrease in $x_{H\,I}$ between $6 < z < 10$ and it smoothly matches the Lyα forest data. The evolution of the neutral fraction is thus monotonic in such models with smoothly evolving N_{ion}.

If we now go back to the lower portion of Table 3, we find that reionization is 50 per cent complete between redshifts 9.6 and 12.0 (95 per cent confidence level), while it is almost (99 per cent) complete between redshifts 5.8 and 10.6 (95 per cent confidence level). Note that the lower limit on the redshift of reionization (5.8) is imposed as a prior on the parameters.

Thus, the PCA shows that a wide range of reionization histories is still allowed by the data. Reionization can be quite early or can be gradual and late, depending on the behaviour of $N_{ion}(z)$. Hence,
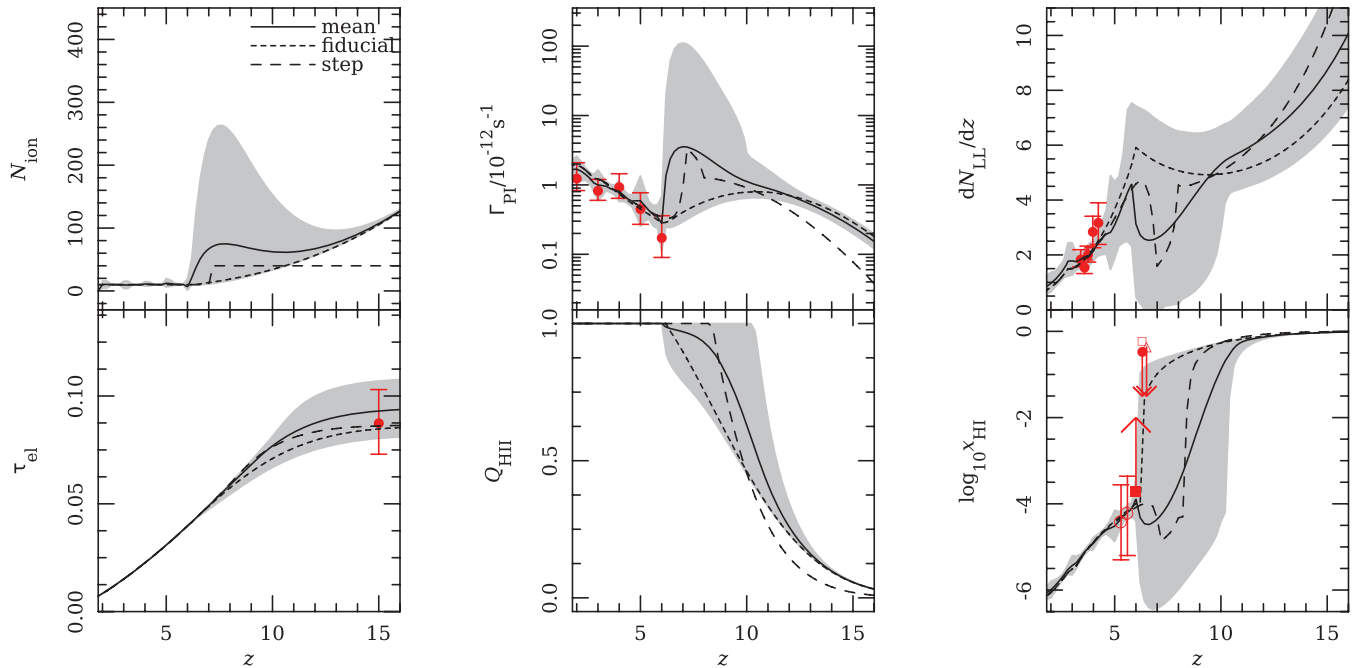
Figure 7. The marginalized posterior distribution of various quantities related to reionization history obtained from the PCA. The different quantities in the panels are identical to those in Fig. 1. The solid lines correspond to the model described by mean values of the parameters while the shaded regions correspond to 2σ limits. In addition, we show the properties of the fiducial model (short-dashed lines) and the step model used in Section 4.2 (long-dashed lines). The points with error bars denote the observational data points which are identical to those in Fig. 1.

if one considers only the data we have used, it is practically impossible to put any sensible constraints on chemical feedback and/or the evolution of star-forming efficiencies and/or escape fractions. While this might seem somewhat disappointing at the moment, one can hope for much better constraints in near future when the magnitude of data sets are going to rise manifold. In fact, in order to keep the analysis simple, we have not used all the data sets available. For example, the constraints on the Lyα and Lyβ transmitted fluxes now extend beyond $z = 6$ and possibly could constrain the models much more. However, our numerical code takes significantly more time while calculating the transmitted fluxes and also there remain uncertainties in the theoretical modelling of the IGM at such redshifts (like the distribution of baryonic matter and the scatter in the temperature–density relation); hence we have worked simply with the constraints on Γ_{PI}. Similarly the distribution of LLS at $z > 6$ could also be important in ruling out some of the allowed models. At present, there exists a data point at $z \approx 6$ which put limits $dN_{LL}/dz = 8.91 \pm 3.49$. On the other hand, very high emissivity models predict $dN_{LL}/dz \approx 0$ at $z \sim 6.5$. Hence constraints on LLS distribution at $z \sim 6.5$ can be helpful in shrinking the allowed parameter space significantly.

We should also mention that the constraints obtained through the PCA are widely different from those obtained using the chemical feedback model of Section 2 involving Population II and Population III stars. The model in Section 2 uses a particular prescription for chemical feedback and assumes constant $N_{ion,II}, N_{ion,III}$, which results in an effective $N_{ion}(z)$ which is smoothly evolving and monotonically increasing with z. On the other hand, the models allowed by the PCA do not have any physical constraint regarding how N_{ion} should evolve. It turns out that in absence of any physical motivation, current data do allow for non-monotonic N_{ion} which may contain sharp features. Hence it is not surprising that the shapes of

the allowed models are quite different from the chemical feedback models.

5 DISCUSSION AND SUMMARY

In this work, we have used a semi-analytical model (Choudhury & Ferrara 2005; Choudhury & Ferrara 2006b) to study the observational constraints on reionization. Assuming that reionization at $z > 6$ is primarily driven by stellar sources, we have developed a formalism based on PCA to model the unknown function $N_{ion}(z)$, the number of photons in the IGM per baryon in collapsed objects. We have used three different sets of data points, namely, the photoionization rates Γ_{PI} obtained from Lyα forest Gunn–Peterson optical depth, *WMAP*7 data on electron scattering optical depth τ_{el} and the redshift distribution of Lyman-limit systems dN_{LL}/dz at $z \sim 3.5$.

The main findings of our analysis are as follows.

(i) The elements of the Fisher information matrix have larger values for $z < 6$ where most of the data points are. There is hardly any information at $z > 14$, implying that no information on star formation and/or chemical feedback can be obtained at these redshifts using the available three data sets.

(ii) To model $N_{ion}(z)$ over the range $2 < z < 14$ it is necessary to include five modes. Using a larger number of modes improves the agreement but at the cost of increasing errors.

(iii) One may not be able to recover the actual form of $N_{ion}(z)$ using only these five modes, however, the recovery of Γ_{PI} and dN_{LL}/dz at $z < 10$ is quite satisfactory and that of Q_{HII}, x_{HI} is excellent.

(iv) It is not possible to match available reionization data with a constant N_{ion} over the whole redshift range, i.e. N_{ion} must increase at $z > 6$. This is a signature of either of a changing IMF

induced by chemical feedback and/or evolution in the star-forming efficiency and/or photon escape fraction of galaxies. The data allow for non-monotonic $N_{ion}(z)$ (and consequently of x_{H_I}). In particular, reionization histories could show sharp features around $z \approx 7$.

(v) The PCA implies that reionization must be 99 per cent completed between $5.8 < z < 10.3$ (95 per cent confidence level) and is expected to be 50 per cent complete at $z \approx 9.5$–12.

Our analysis provides the widest possible range in reionization histories (shown in Fig. 7) allowed by available data sets. It is, in some sense, unfortunate that there still exists a wide range of reionization scenarios that are allowed by the data. While the constraints at $z < 6$ are quite tight, one requires additional data points at $z > 6$ to improve constraints on models of feedback and reionization. The most obvious addition would, of course, be observation of Gunn–Peterson trough in more QSOs at higher redshifts. In parallel, it is expected that observations of gamma-ray bursts (GRBs) and Lyα emitters could constrain x_{H_I} at $z > 6$, which again would result in improved constraints. Finally, observations of large-scale EE polarization signal by future CMB probes, like *Planck*,[4] would be extremely important in probing the evolution of N_{ion} at $z > 6$. Since the constraints obtained from the data are still unsatisfactory, there remains ample scope for developing physically motivated theoretical models which can match a wide variety of available data. This, in turn, requires significant improvement in our understanding of processes like chemical feedback and also the evolution of star-forming efficiencies and escape fraction.

ACKNOWLEDGMENTS

REFERENCES

Barkana R., Loeb A., 2001, Phys. Rep., 349, 125
Bolton J. S., Haehnelt M. G., 2007, MNRAS, 382, 325
Burigana C., Popa L. A., Salvaterra R., Schneider R., Choudhury T. R., Ferrara A., 2008, MNRAS, 385, 404

Chiu W. A., Fan X., Ostriker J. P., 2003, ApJ, 599, 759
Choudhury T. R., 2009, Curr. Sci., 97, 841
Choudhury T. R., Ferrara A., 2005, MNRAS, 361, 577
Choudhury T. R., Ferrara A., 2006a, in Fabbri R., ed., Cosmic Polarization. Research Signpost, Trivandrum, India, p. 205
Choudhury T. R., Ferrara A., 2006b, MNRAS, 371, L55
Choudhury T. R., Ferrara A., 2007, MNRAS, 380, L6
Clarkson C., Zunckel C., 2010, Phys. Rev. Lett., 104, 211301
Crittenden R. G., Pogosian L., Zhao G., 2009, J. Cosmology Astropart. Phys., 12, 25
Efstathiou G., 2002, MNRAS, 332, 193
Efstathiou G., Bond J. R., 1999, MNRAS, 304, 75
Fan X. et al., 2006a, AJ, 131, 1203
Fan X., Carilli C. L., Keating B., 2006b, ARA&A, 44, 415
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Gallerani S., Choudhury T. R., Ferrara A., 2006, MNRAS, 370, 1401
Gallerani S., Ferrara A., Fan X., Choudhury T. R., 2008a, MNRAS, 386, 359
Gallerani S., Salvaterra R., Ferrara A., Choudhury T. R., 2008b, MNRAS, 388, L84
Hopkins P. F., Richards G. T., Hernquist L., 2007, ApJ, 654, 731
Hu W., Holder G. P., 2003, Phys. Rev. D, 68, 023001
Huterer D., Cooray A., 2005, Phys. Rev. D, 71, 023506
Huterer D., Starkman G., 2003, Phys. Rev. Lett., 90, 031301
Kadota K., Dodelson S., Hu W., Stewart E. D., 2005, Phys. Rev. D, 72, 023510
Kashikawa N. et al., 2006, ApJ, 648, 7
Lacey C., Cole S., 1993, MNRAS, 262, 627
Larson D. et al., 2011, ApJ, 192, 16
Leach S., 2006, MNRAS, 372, 646
Lewis A., Bridle S., 2002, Phys. Rev. D, 66, 103511
Loeb A., Barkana R., 2001, ARA&A, 39, 19
Miralda-Escudé J., Haehnelt M., Rees M. J., 2000, ApJ, 530, 1
Mortonson M. J., Hu W., 2008, ApJ, 672, 737
Munshi D., Kilbinger M., 2006, A&A, 452, 63
Press W. H., Schechter P., 1974, ApJ, 187, 425
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in FORTRAN. The Art of Scientific Computing, 2nd edn. Cambridge Univ. Press, Cambridge
Pritchard J. R., Loeb A., Wyithe J. S. B., 2010, MNRAS, 408, 57
Prochaska J. X., O'Meara J. M., Worseck G., 2010, ApJ, 718, 392
Schneider R., Salvaterra R., Ferrara A., Ciardi B., 2006, MNRAS, 369, 825
Songaila A., Cowie L. L., 2010, ApJ, 721, 1448
Totani T., Kawai N., Kosugi G., Aoki K., Yamada T., Iye M., Ohta K., Hattori T., 2006, PASJ, 58, 485
Wasserman L. et al., 2001, ArXiv Astrophys. e-prints (astro-ph/0112050)
Wyithe J. S. B., Loeb A., 2003, ApJ, 586, 693

[4] http://www.esa.int/SPECIALS/Planck/index.html
[5] http://cluster.hri.res.in/index.html

This paper has been typeset from a TeX/LaTeX file prepared by the author.