Mapping electron density in the ionosphere: a principal component MCMC algorithm

Eman Khorsheed,[*] Merrilee Hurn,[†] and Chris Jennison [‡]

November 26, 2008

Abstract

The outer layers of the Earth's atmosphere are known as the ionosphere, a plasma of free electrons and positively charged atomic ions. The electron density of the ionosphere varies considerably with time of day, season, geographical location and the sun's activity. Maps of electron density are useful because local changes in this density can produce inaccuracies in the Navy Navigation Satellite System (NNSS) and Global Positioning System (GPS). Satellite to ground-based receiver measurements produce tomographic information about the density in the form of path integrated snap-shots of the total electron content which must be inverted to generate maps. We propose a Bayesian approach to the inversion problem using spatial priors which allow us parsimoniously to include knowledge of how density varies with height. Less helpfully, this parameterisation does not lend itself well to standard Metropolis-Hastings algorithms and so we develop a much more efficient form of Markov chain Monte Carlo sampler using a transformation of variables based on a principal components analysis of initial output.

Keywords: Bayesian Modelling, Ionospheric Mapping, Inversion, Markov Chain Monte Carlo, Principal Components, Tomography.

1 Introduction

The Earth's atmosphere is categorised into five regions at increasing height from the Earth's surface, the troposphere, stratosphere, mesosphere, thermosphere and exosphere. The two outermost layers of the Earth's atmosphere, the thermosphere and exosphere, starting at about 75km from the Earth's surface are sufficiently thin that ultraviolet radiation causes them to be ionized; electrons are knocked out of atoms by photons, and

[*]Department of Mathematics, University of Bahrain, Sakhir, P.O.Box 32038, Kingdom of Bahrain

[†]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 Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK

the sparsity of the atmosphere allows them to live free for some time before recombining with a nearby positive ion. This plasma of free electrons and positively charged atomic ions is known as the ionosphere.

One noticeable effect of the ionosphere for us is on the transmission of radio waves, with the ionosphere acting almost as a mirror to bounce signals around the Earth. Less helpfully, the ionosphere also affects the radio wave signals of satellite based navigational systems with a negative effect on their accuracy. This effect could be assessed and accounted for using a map of electron density (which varies considerably including, but not exclusively, with altitude, time of day, season, geographical location and with the sun's activity). With a view to generating such maps, satellite to ground-based receiver systems can be used to provide tomographic information about the density in the form of path integrated snap-shots of the total electron content. These tomographic data must then be inverted to map the electron density.

In Section 2 we discuss the tomographic inversion problem based on NNSS data. We propose a Bayesian approach using priors which allow us to include knowledge of how electron density varies with height as well as spatial smoothness at fixed altitude. This parameterisation does not lend itself well to standard Metropolis-Hastings algorithms and so we develop a much more efficient transformation-of-variables modification linked to principal components in Section 3. Sections 4 and 5 present results, discussions and directions for further work. Finally an investigation of the stability of the principal components MCMC is given in the Appendix.

2 Inversion of ionospheric data

2.1 Data and existing methodology

The data which will be used here exploit the fact that the extent to which radio waves are affected by the electron density of the ionosphere depends on their frequency. By transmitting two satellite-to-ground receiver signals simultaneously at two different frequencies, it is possible to measure the Total Electron Content (TEC) along that particular satellite-to-ground path. The TEC is effectively a path integral of the electron density (and is measured in TECU, 1 TECU being 10^{16} electrons per m^2). Such data are believed by engineers to have very high accuracy, but they do only give tomographic projections of the electron density. Moreover, since the paths are between moving satellites and fixed ground-based receivers which must be in their line of sight, the geometry means a rather restricted set of projections in comparison to, say, Magnetic Resonance Imaging used in medicine.

Existing approaches to inverting TEC projections discretise space and correspondingly approximate the path integrals. Writing the set of TEC observations as a vector Y, the set of electron densities over space discretised into a set of voxels as a vector D, and forming a matrix W with ij^{th} entry the length of the i^{th}

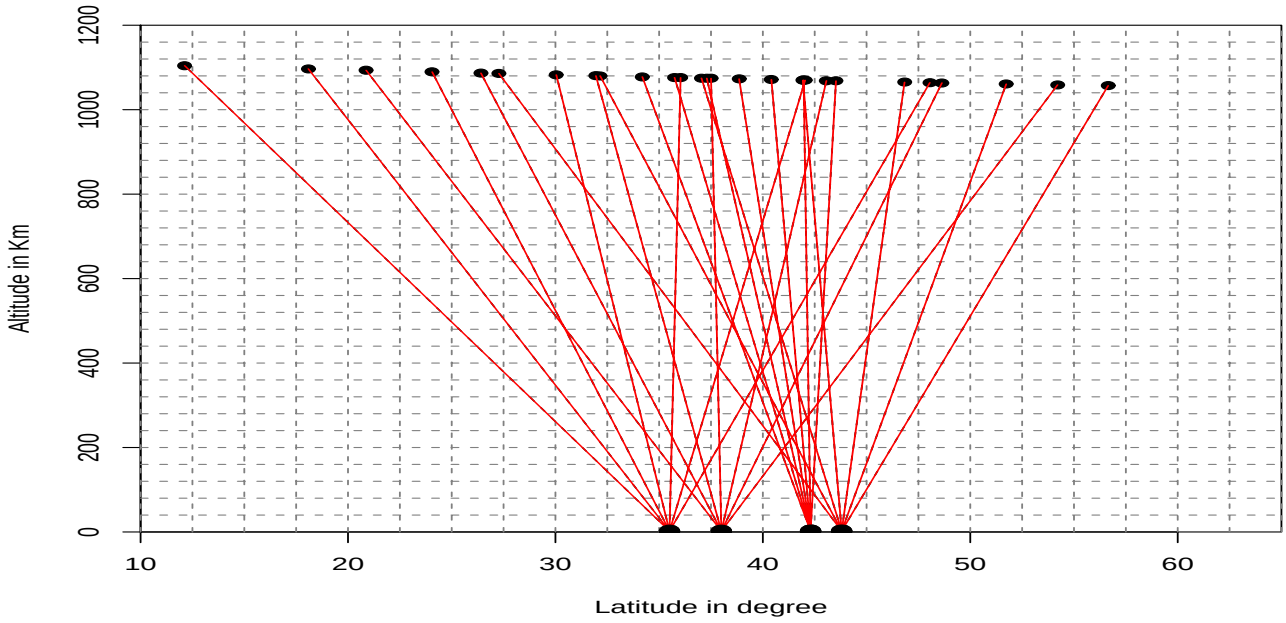

Figure 1: The approximate geometry of paths between a Navy Navigation Satellite System (NNSS) satellite and a chain of four ground based receivers (for clarity only a subset of intermediate satellite positions have been shown, there are 572 paths in total).

path across the j^{th} spatial voxel, the system to be inverted is written as

$$Y = WD. \tag{1}$$

Recovering the discretised electron density D is complicated because this is an under-determined system; there are usually large areas of the discretised ionosphere through which no paths pass. For example, Figure 1 shows the geometry of a sequence of measurements made from an orbiting NNSS satellite to four ground-based receivers. Various authors have proposed iterative algorithms which can be used for this type of inversion problem, see for example Gordon, Bender and Herman (1970) and Censor (1983). A non-iterative algorithm for two-dimensional imaging was proposed in 1992 by Fremouw, Secan and Howe, and extended in 2001 by Spencer and Mitchell. It is this last approach, known as MIDAS (Multi-Instrument Data Analysis System) which we will consider as the motivation for the work that follows. MIDAS transforms the voxel-based representation of the electron density into an alternative domain, partly using a Fourier basis, and performs a generalised inversion in this domain using a singular value decomposition. Although computationally fast, MIDAS does not provide interval estimates of the electron density and can also occasionally generate negative estimates of the density as there is no explicit positivity constraint imposed.

2.2 A Bayesian Modelling approach

We wish to use a Bayesian approach to explore the space of probable electron densities D which could have given rise to the observed TEC measurements Y. Although the measurement system is believed to be virtually noise-free, what little noise there is will be augmented by the modelling error of discretisation in formulating the data as linear combinations of average voxel densities, Equation (1). We shall make the assumption that the TEC data are conditionally independent given D and τ^2 with

$$Y_i|D, \tau^2 \sim N([WD]_i, 1/\tau^2), \;\; i = 1, \ldots, I \tag{2}$$

where the weights W are assumed known, τ^2 is the common precision, and the variance term τ^{-2} represents both measurement and discretisation error. This modelling assumption will be examined later. We assign a quite standard choice of prior for the precision τ^2 using Jeffreys' invariance rule

$$p_{\tau^2}(\tau^2) \propto \frac{1}{\tau^2}, \quad \tau^2 > 0. \tag{3}$$

One possible choice for a prior distribution for D which would reflect the fact that we expect the electron density to have smooth spatial variation would be an intrinsic Gaussian Markov Random Field (GMRF) model (Rue and Held (2005))

$$p_D(D) \propto \exp(-\beta_D \sum_{i \sim j} (D_i - D_j)^2) \tag{4}$$

where $i \sim j$ defines a nearest neighbour structure on the voxels, and β_D is a non-negative parameter controlling the degree of smoothness expected. It would not be uncommon to expand this model to incorporate different degrees of smoothness in different directions; in this example, we might expect different smoothness in the vertical and horizontal directions. However, we have more information about the electron density than this. Chapman (1931) considers the rate at which electrons are produced, which is proportional to the intensity of the ionising radiation and to the density of the gas being ionised. The competing effects of greater radiation intensity but lower gas density as altitude increases give rise to a unimodal profile of density with height which takes its name from him, the *Chapman profile*. For a single gas whose absorption coefficient is assumed constant with wavelength, the Chapman profile can be described by the equation

$$\begin{aligned} \gamma(z) &= \gamma_0 \exp(1 - z - e^{-z}) \\ &= \gamma_0 \exp(-\frac{z^2}{2!} + \frac{z^3}{3!} - \frac{z^4}{4!} + \ldots) \end{aligned} \tag{5}$$

where $\gamma(z)$ is the density at a height z measured relative to the height of maximum density ($z = 0$) where the density is γ_0. Figure 2 indicates the shape of this idealised profile. The reality of the ionosphere is more complicated, not least by the fact that multiple gases exist in different proportions at different altitudes.

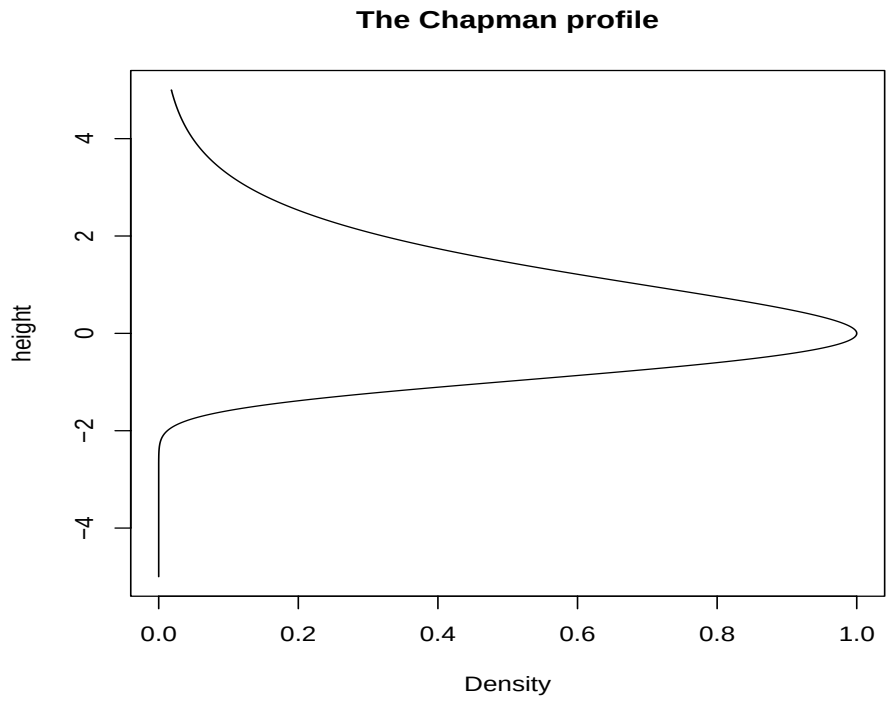

Figure 2: The shape of the Chapman profile.

We indirectly propose a prior for the densities D by proposing a prior for an approximation to the basic Chapman profile. Consider the n^{th} vertical column in the spatial discretisation bearing in mind Equation (5). We approximate the Chapman profile at this location by a scaled Normal probability density function, with electron density at height h given by

$$\tilde{\gamma}_n(h) = \frac{\gamma_n}{\sqrt{2\pi\sigma_n^2}} \exp(-\frac{(h-\mu_n)^2}{2\sigma_n^2}) \tag{6}$$

where μ_n represents the peak electron density, σ_n^2 represents a scaling parameter and γ_n represents the total electron count in that vertical column. This approximation is assumed to hold for all points on the ground lying in the n^{th} column, $n = 1, \ldots, N$, giving us N Chapman curve approximations. We relate the discretised densities D to the approximated Chapman curves $\{\tilde{\gamma}_n(h)\}$ by setting the density in each voxel to be constant and equal to the corresponding $\tilde{\gamma}_n(h)$ evaluated at h, the mid-height of that voxel. Notice that we have moved from $N \times N_h$ D variables (where N is the number of vertical columns, and N_h is the number of discretisation values in the height direction) to $3 \times N$ values of parameters $\{\mu_n, \sigma_n^2, \gamma_n\}$. The behaviour we would still like to capture though is that the electron density changes quite slowly and smoothly. We consider the priors

$$p(\{\mu_n, \sigma_n^2, \gamma_n\}) \quad = \quad p_\mu(\{\mu_n\}) p_{\sigma^2}(\{\sigma_n^2\}) p_\gamma(\{\gamma_n\})$$

$$
\begin{aligned}
p_\mu(\{\mu_n\}) &\propto \exp(-\beta_\mu \sum_{i\sim j}(\mu_i-\mu_j)^2) I_{[0<\{\mu_n\}<\mu_{\max}]} \\
p_{\sigma^2}(\{\sigma_n^2\}) &\propto \exp(-\beta_{\sigma^2} \sum_{i\sim j}(\sigma_i^2-\sigma_j^2)^2) I_{[0<\{\sigma_n^2\}<\sigma^2_{\max}]} \\
p_\gamma(\{\gamma_n\}) &\propto \exp(-\beta_\gamma \sum_{i\sim j}(\gamma_i-\gamma_j)^2) I_{[0<\{\gamma_n\}<\gamma_{\max}]}
\end{aligned} \tag{7}
$$

where $\sim$ indicates a common nearest neighbour structure and $\beta_\mu, \beta_{\sigma^2}, \beta_\gamma$ are non-negative parameters controlling the degree of smoothness. These parameters $\{\mu_n, \sigma_n^2, \gamma_n\}$ are all restricted to finite ranges which in practice can be chosen sufficiently wide that the posterior would have negligible mass outside them.

We are interested in the posterior distribution of the ionospheric parameters $\{\mu_n, \sigma_n^2, \gamma_n\}$ and the nuisance parameter τ^2 given the satellite-to-receiver observations $Y_1, \ldots, Y_I$:

$$
p(\{\mu_n, \sigma_n^2, \gamma_n\}, \tau^2 | y_1, \ldots, y_I) \propto p_\mu(\{\mu_n\}) p_{\sigma^2}(\{\sigma_n^2\}) p_\gamma(\{\gamma_n\}) p_{\tau^2}(\tau^2) \prod_{i=1}^{I} p(y_i | \{\mu_n, \sigma_n^2, \gamma_n\}, \tau^2) \tag{8}
$$

2.3 The application of standard MCMC

The posterior distribution defined by Equation (8) is rather intractable and we will need to resort to Markov chain Monte Carlo methods for inference (see Gilks, Richardson and Spiegelhalter (1996) for an overview). In order to test both the efficiency of a standard single-site MCMC implementation and the ability of the model to invert the signals $\{y_i\}$ to recover information about the electron density, we consider first a small simulated dataset. Figure 3 shows both the geometry of the test study and the shapes of the approximated Chapman profiles, Equation (6), which will be used to simulate a test data set. From these scaled curves, each voxel is assigned an electron density value equal to the corresponding $\tilde{\gamma}_n(h)$ evaluated at the height h which is the mid-height of that voxel. Data $Y_1, \ldots, Y_I$ are then simulated according to model Equation (2), where the weights W are the path lengths across each voxel, and τ^2 is taken to be 2000 reflecting the belief that measurement error is small. There are a couple of points to note. First, in this test example, there are ray paths between all satellite positions and receivers; in real data sets this is not necessarily the case because of the curvature of the Earth. Secondly, the level of discretisation is extremely coarse, especially in the vertical direction, with the number of D_i actually equaling the number of Chapman curve parameters $\{\mu_n, \sigma_n^2, \gamma_n\}$.

A standard application of MCMC updates each of the $\{\mu_n, \sigma_n^2, \gamma_n\}$ and τ^2 parameters in turn, using a Gibbs sampler for the updates of τ^2 and random walk Metropolis updates for all of the other parameters. The smoothing parameters $\beta_\mu, \beta_\gamma, \beta_{\sigma^2}$ are, for the moment, held fixed and the proposal distributions for the Metropolis random walk moves are tuned to have acceptance rates of 20-40%. Figure 4 shows trace plots of the first 150000 iterations of an MCMC run using this set-up together with the values used in the synthesis of the data. It is clear that convergence is slow, for example the precision parameter only really reaches the right region of the parameter space after about 75000 iterations. Before this point, the other parameters

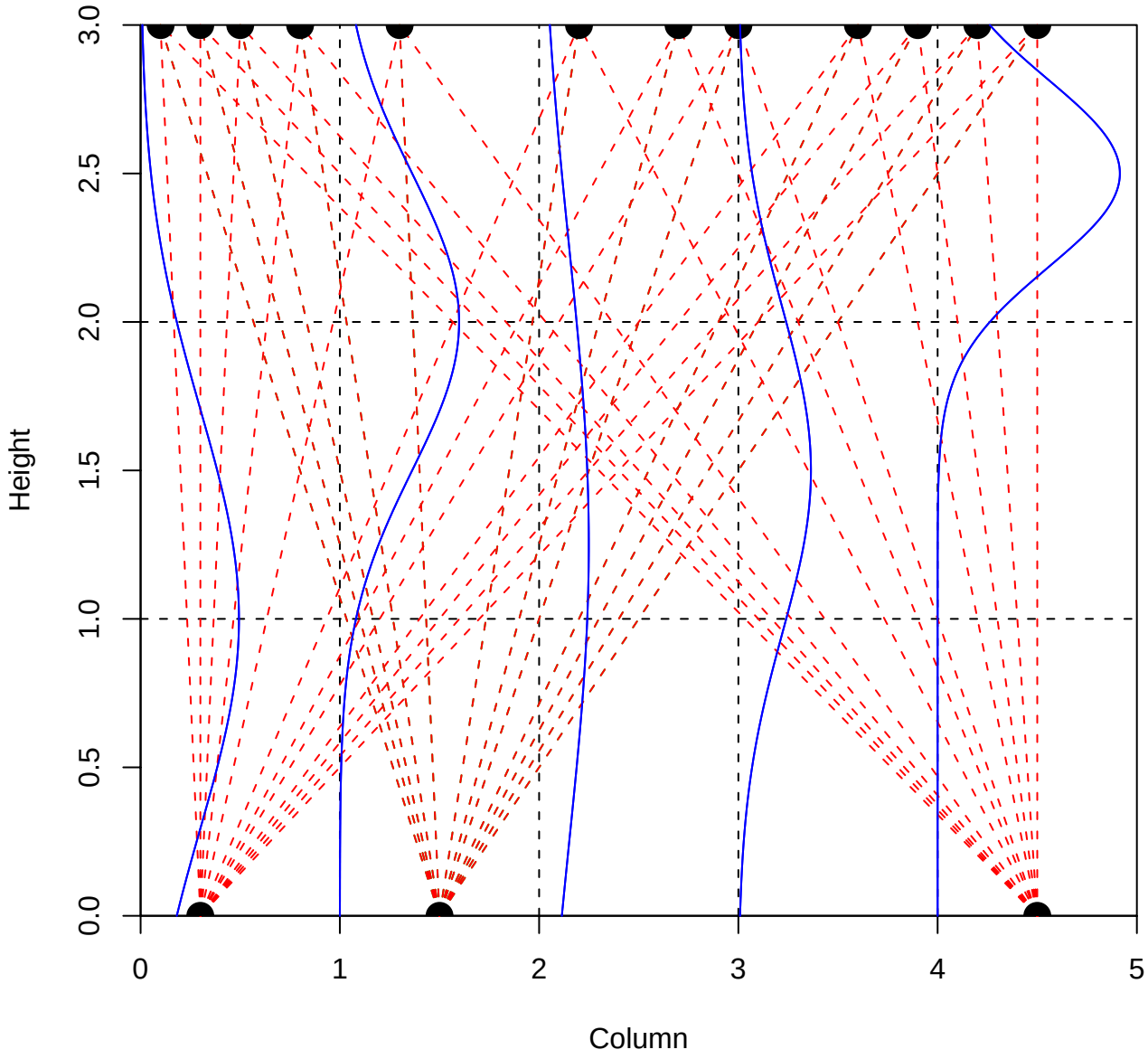

Figure 3: Physical set-up for the small simulated example. Red lines indicate the ray paths between 3 ground receivers and 12 satellite positions. The spatial discretisation is into five vertical sections and three horizontal ones. The solid blue lines indicate the shapes of the approximated Chapman profiles $\{\tilde{\gamma}_n(h)\}$, Equation (6), at the five vertical sections which will be used to simulate data. The curves do not indicate the values of the scalings $\{\gamma_n\}$ which will be, from left to right, 35, 30, 25, 20 and 29. D values for each voxel are evaluated at the mid-height point.

experience quite high variation, but after it this is damped down considerably. This is not perhaps surprising given the form of the likelihood, Equation (2), and the fact that the value of τ^2 used in the simulation of the data is very high, implying a variance for the data of 1/2000.

Unfortunately there is a more serious problem than the slow convergence. From Figure 4 it is clear that several of the Chapman parameters have settled into regions of the parameter space quite far from the values used to generate the data. While this could in some instances suggest that the posterior provides greater support for these non-synthesising values, in this case running the sampler with different starting values or different random number generator seeds can lead to the chains settling in different regions: The mixing is extremely poor. Figure 3 helps to explain why this is the case. Updating a single Chapman parameter alters all the density values D in the corresponding column in a non-linear fashion. In terms of the likelihood contribution to the posterior, what matters is that the weighted combination of these D values match the observed data values $Y_1, \ldots, Y_I$ well (bearing in mind how small the noise variance is). Once a combination of $\{\mu_n, \sigma_n^2, \gamma_n\}$ has been found which does give a reasonable match, it is hard to move away to any another well-matching set by single-site updates alone since this may involve moving through a sequence of far less likely combinations. One possibility here might be a block update of several parameters simultaneously, although it is not immediately clear how to block the parameters in order to maintain this "data-matching". The perhaps more intuitive solution is to consider a transformation of variables approach, proposing small changes to the D values themselves (and back-transforming to recover the new values of the $\{\mu_n, \sigma_n^2, \gamma_n\}$). (In this set-up of three horizontal divisions, each column has three D values and three Chapman parameters.) To evaluate the acceptance probability of a move of this type, the Jacobian of the transformation is required. Any perturbation of the D which does not give rise to a valid set of $\{\mu_n, \sigma_n^2, \gamma_n\}$ (for example a negative value of μ) will be rejected automatically as it will violate the range constraints incorporated into the priors.

Figure 5 gives the trace plots from an MCMC run using, alternately, the single-site moves and the moves in D space. While it still takes the algorithm a large number of iterations to reach the right area of the parameter space, the extreme "stickiness" of the single-site sampler is avoided. What the traces now reveal is part of the reason why this is a potentially hard inversion problem. For example, consider the turquoise trace plots which exhibit much higher variability than the other four sets of curves. This set of traces correspond to the fifth column in Figure 3. Looking at the ray paths passing through this column, it is clear that many of the rays are close to vertical or only have a short path across the top of the column. Moreover, the discretisation means that the D values are calculated using only the Chapman values at the mid-height point of the voxel. As this curve has a high μ value, the two lower D contributions are very close to zero, simply because of the shape of the Chapman curve. This effectively means that three parameter values $\mu_5, \sigma_5^2, \gamma_5$ can vary relatively widely while still maintaining a fairly constant set of D values.

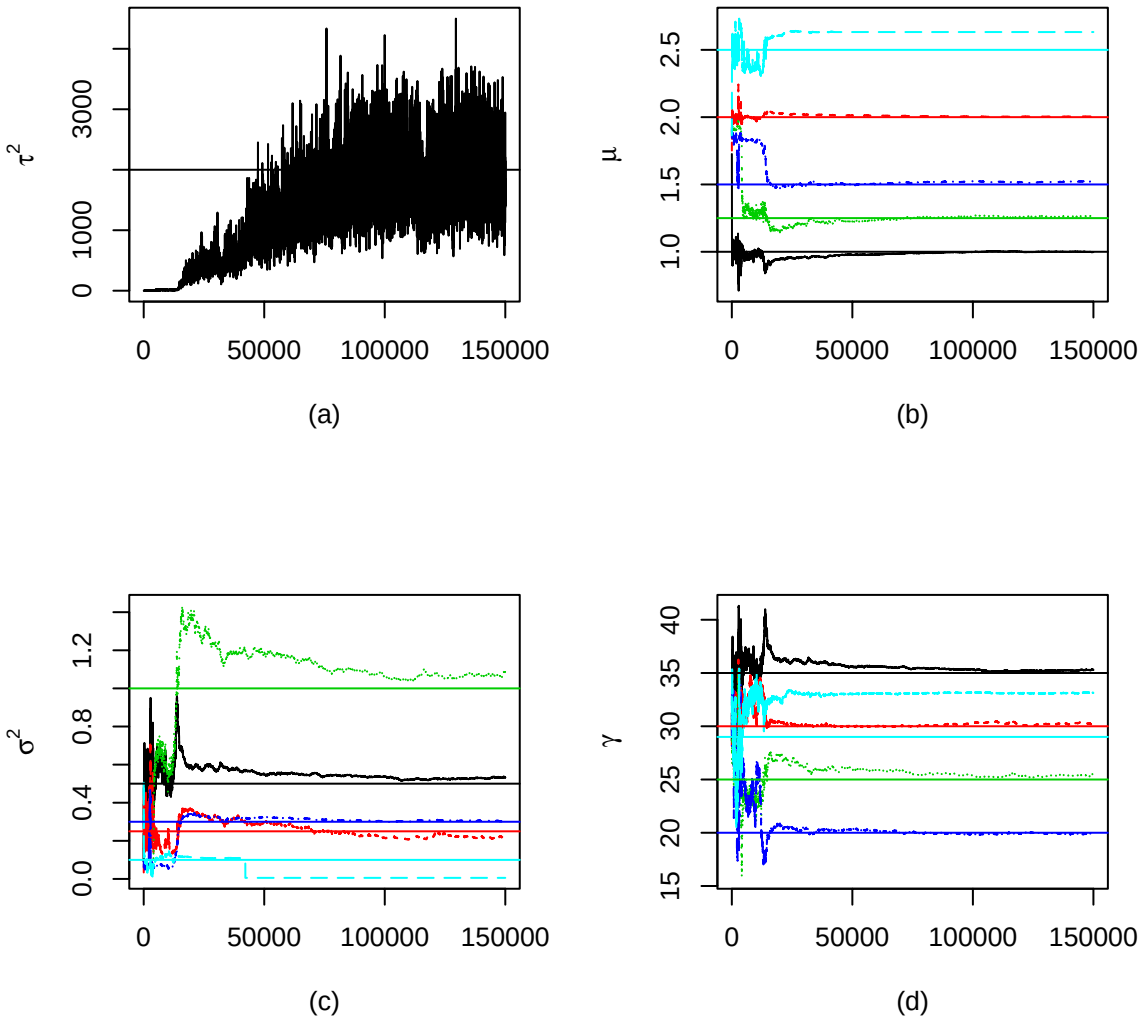

Figure 4: Trace plots of the precision τ^2 and sets of Chapman parameters $\{\mu_n\}$, $\{\sigma_n^2\}$, $\{\gamma_n\}$ for a single-site MCMC run. The values used to generate the simulated data are indicated as solid horizontal lines. The five columns left to right are represented in the plot by colours black, red, green, blue and turquoise respectively.

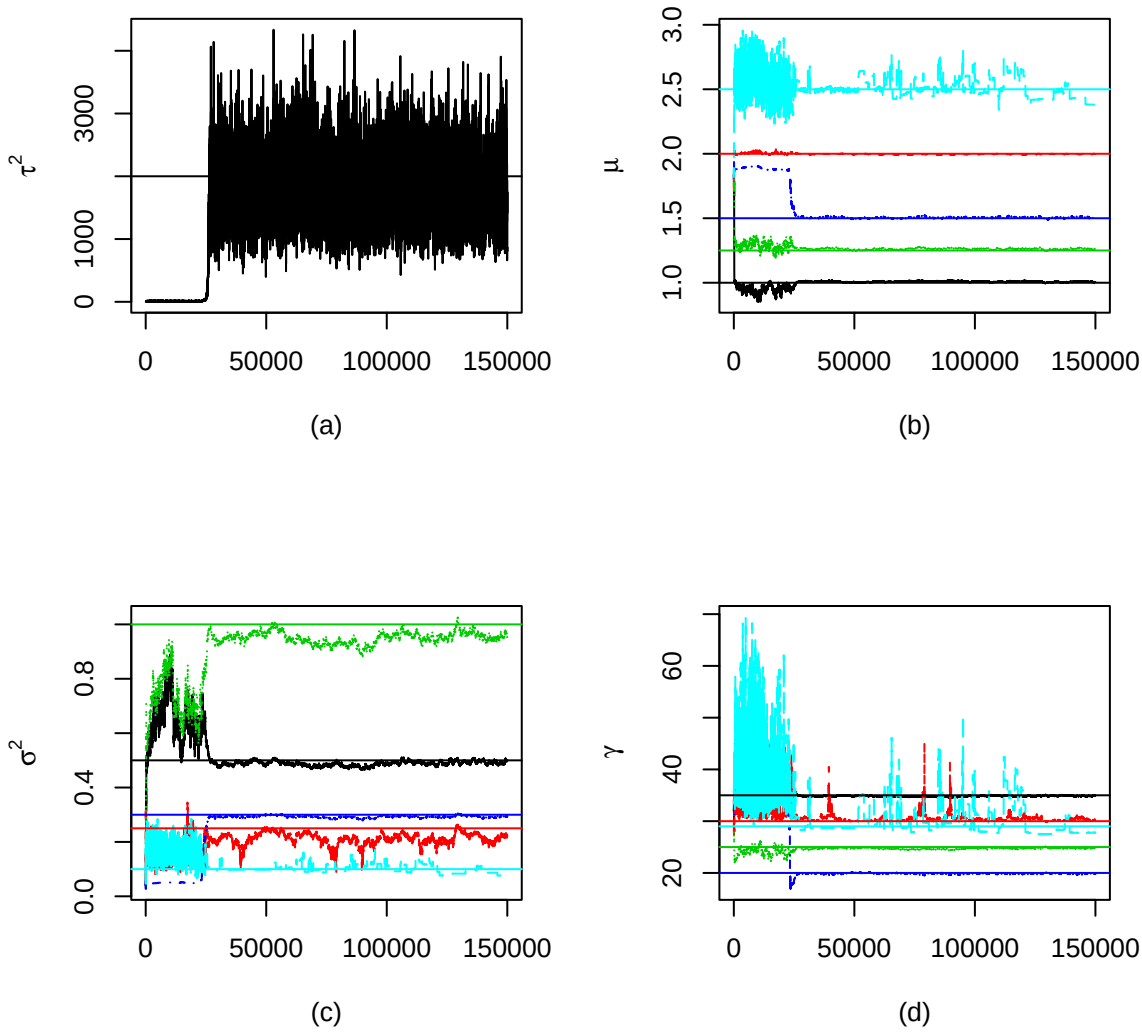

Figure 5: Trace plots of the precision τ^2 and sets of Chapman parameters $\{\mu_n\}$, $\{\sigma_n^2\}$, $\{\gamma_n\}$ for an MCMC run incorporating updates of the D values. The values used to generate the simulated data are indicated as solid horizontal lines. The five columns left to right are represented in the plot by colours black, red, green, blue and turquoise respectively.

It could be argued that the level of discretisation here is entirely artificial and certainly that only three levels of vertical discretisation is inadequate. The number of slices can certainly be increased to improve the sampling of the Chapman curve, and indeed experiments with this simulated data set show that this is helpful. However if the ray paths only intersect voxels where the electron density is very low, we would expect more uncertainty in the inversion for such columns. Considering the geometry for the NNSS data set in Figure 1, the sparsity of the ray-paths is apparent. Once we move to more than three horizontal sections, the MCMC move in D space is more complicated to define as there is no longer a dimension match between the number of voxels in a column and the corresponding number of Chapman parameters. Rather than concentrate on how this move can be generalised, we want to view this example of how a transformation of variables can improve MCMC performance to motivate a largely automatic route to improved mixing.

3 Principal Component MCMC

One of the well known introductory examples in MCMC for illustrating how single-site updates work is that of a bivariate distribution whose constituent variables are not aligned with the coordinate axes (see, for example, the first chapter of Gilks, Richardson and Spiegelhalter (1996) or Figure 6). Proposed moves are of one variable at a time and so are parallel to the coordinate axes. If the variables were uncorrelated, then this would be ideal. However, depending on the degree of correlation between the two variables, mixing can be very slow as the sampler slowly "tacks" up and down the area of high probability. The obvious solution would be to make a suitable simultaneous block update of both variables, enabling moves in the natural directions of the target distribution. The catch is that, in general, little is known in advance about the structure of the distribution and so it is not possible to specify what blocking of the components is suitable. However, if the variance matrix were known, then we could calculate the principal components of the distribution (see, for example, Chatfield and Collins (1980)). A principal component analysis will give us linear combinations of the original variables which are uncorrelated. Additionally the eigenvalues from the principal component analysis will give us the scale of variability in each of these new directions.

We propose to use the principal component idea to generate blockings of the original variables. Denoting the normalised eigenvectors and corresponding eigenvalues of the principal component analysis by $\mathbf{e_1}, \ldots, \mathbf{e_n}$ and $\lambda_1, \ldots, \lambda_n$ where n is the dimension of the original $\mathbf{X}$, the proposed block moves are

$$\mathbf{x}' = \mathbf{x} + z_i \mathbf{e_i}, \quad i = 1, \ldots, n \tag{9}$$

where $Z_i \sim N(0, \delta\lambda_i)$ and δ is a scalar used to control the acceptance rate. As the transformation of variables is linear and the proposal distribution is symmetric, the acceptance probabilities for these moves will simply take the form of the usual Metropolis rates.

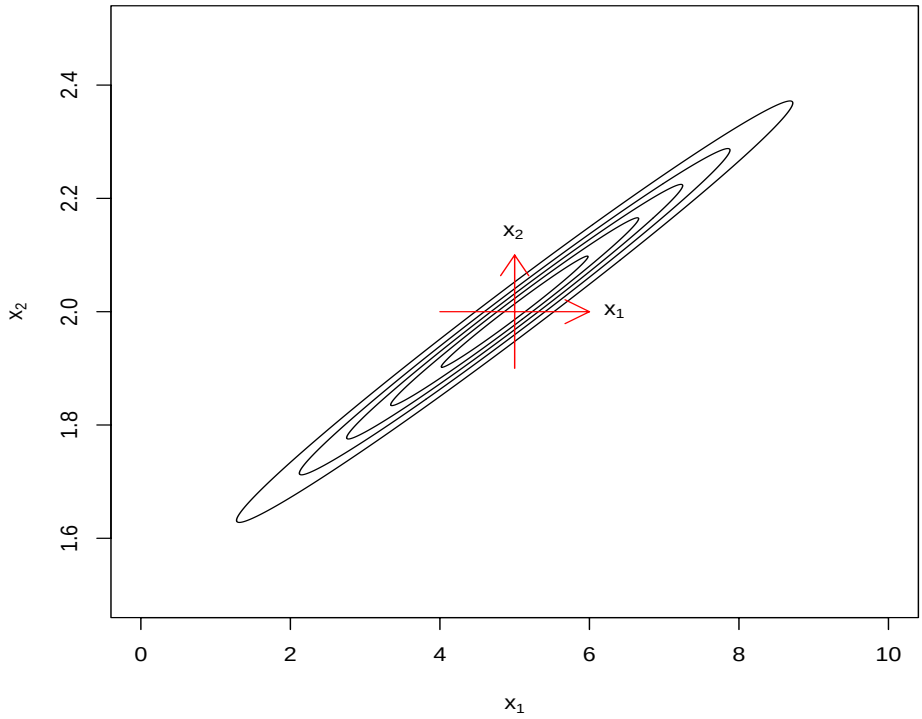

Figure 6: The contours of a distribution showing how the axes of the main variation are not aligned with the directions in which updates will be proposed by a single-site sampler.

Realistically of course, we do not know the variance of the distribution. Our suggestion is to use a preliminary run of a standard sampler to estimate either the covariance or the correlation matrix of the distribution (the correlation may be more appropriate in cases where the constituent components of $\mathbf{X}$ are on quite different scales). We might expect the idea of reparameterisation by principal components to give the best MCMC performance in situations where the target distribution is close to multivariate normal and where the variance structure is well estimated. In the Appendix, we investigate the effect of having little information from which to estimate the principal components. We also consider whether the reparameterisation remains of benefit when the target distribution is increasingly non-normal. Both sets of investigations show promising results for this methodology.

For the 3-level synthetic example of the previous section, we approximate the variance structure from a standard single-site update MCMC run. As the mixing is so bad for this MCMC, we abandon the first 100000 iterations to burn-in (see Figure 4). The scaling parameter δ is taken to be 2.4 for all the principal component moves, giving an acceptance rate of 0.07 for the first principal component and between 0.23

	μ_1	μ_2	μ_3	μ_4	μ_5
MCMC	4099	6256	4373	8008	7159
PCMCMC	234	198	227	186	6597
	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_5^2
MCMC	6129	7274	5416	6710	15444
PCMCMC	120	292	219	239	12201
	γ_1	γ_2	γ_3	γ_4	γ_5
MCMC	4371	7830	4471	7795	5455
PCMCMC	155	359	152	216	7850

Table 1: Comparison of the estimated integrated autocorrelation times using standard MCMC and PCMCMC for the 3-level synthetic example. In both cases, the estimates are from the final 50000 iterations.

and 0.42 for the others. Figure 7 shows the trace plots using PCMCMC corresponding to Figures 4 and 5; mixing and convergence are clearly improved over the standard single-site MCMC for all but the final column. To confirm this, Table 1 gives the estimated integrated autocorrelation times for the parameters of the five approximated Chapman profiles. What we see is that the PCMCMC has dramatically improved the performance for those parameters which exhibited some slow movement with the single-site sampler. However, for the fifth column the original MCMC run was particularly stuck with the parameters as good as fixed. As a result the principal component approach has essentially identified the three original parameters $\mu_5, \sigma_5^2, \gamma_5$ with the three eigenvectors with the three smallest eigenvalues. The fact that μ_5 is barely moving away from a value which is higher than the synthetic value, while σ_5^2 is barely moving away from a lower value than the synthetic one means that quite large changes in the scaling γ_5 have little effect on the D values. The lack of any improvement is easily spotted by a comparison of the MCMC and PCMCMC integrated autocorrelation times.

Overall the PCMCMC approach has dramatically improved the mixing behaviour for inverting the synthetic data. Although it is disappointing that this automatic reparameterisation has not performed as well as the tailored solution on the particularly difficult triplet of parameters, the reasons for this are well understood and the poor performance is easily noticed. Such pathological behaviour would not be expected in practice where discretisation levels are set at much more realistic levels. However it is always important to know the limitations of any algorithm and this experiment has shown us what we can expect of PCMCMC.

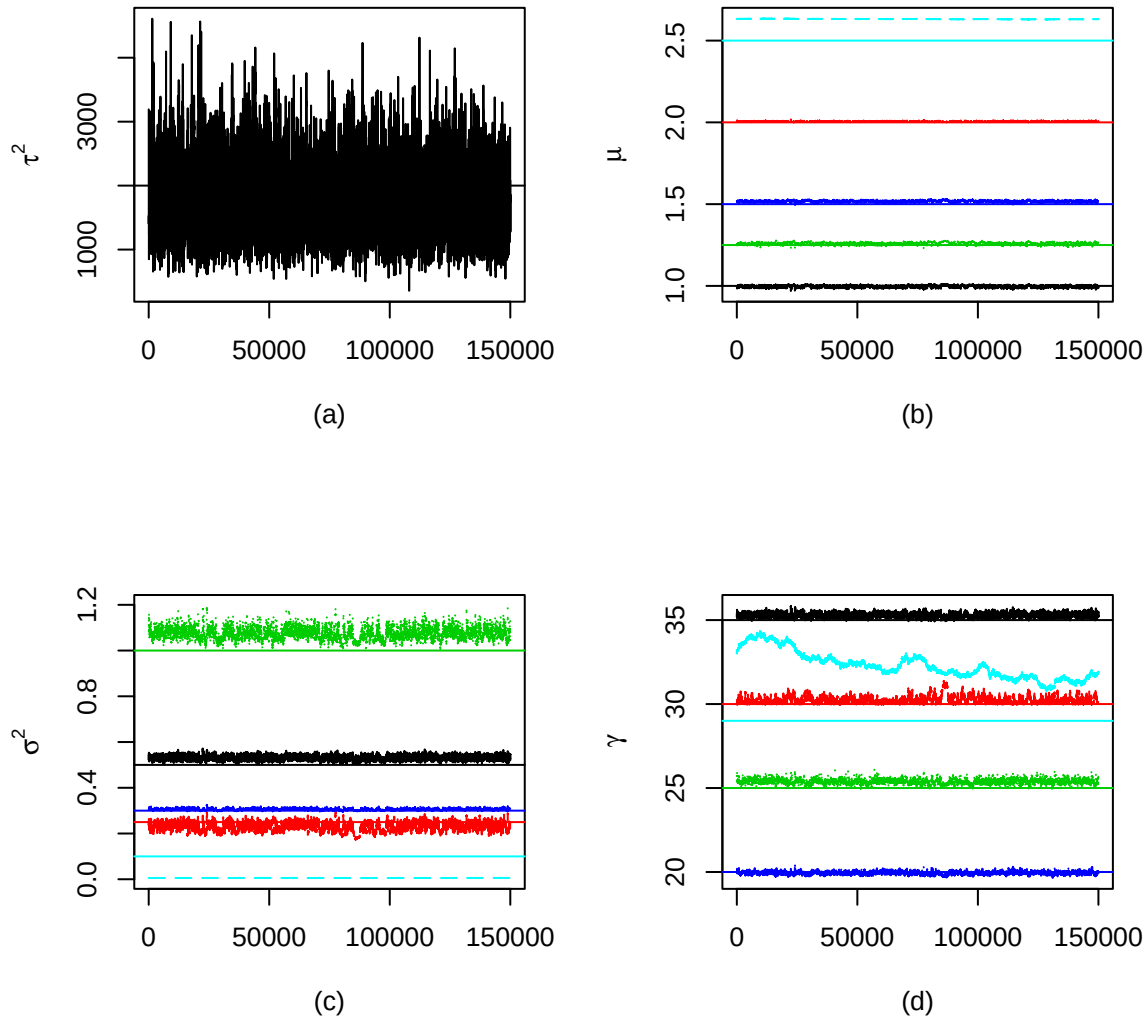

Figure 7: Trace plots of the precision τ^2 and sets of Chapman parameters $\{\mu_n\}$, $\{\sigma_n^2\}$, $\{\gamma_n\}$ using the PCMCMC algorithm. The values used to generate the simulated data are indicated as solid horizontal lines. The five columns left to right are represented in the plot by colours black, red, green, blue and turquoise respectively.

4 Applications of PCMCMC to ionospheric mapping

We are now is a position to fit the Bayesian model given by Equation (8) to data gathered in a satellite pass whose geometry is summarised in Figure 1. The data themselves are tomographic projections of the electron density along the satellite to receiver rays. There are various points to consider: the estimation of the nuisance parameters $\beta_\mu, \beta_\gamma, \beta_{\sigma^2}$, the choice of posterior summaries, the assessment of model fit.

We begin with the treatment of the nuisance parameters arising from the choice of the smoothing spatial priors, Equation (7). Perhaps the most satisfying approach would be to treat these parameters in a fully Bayesian manner but this is infeasible as the normalising constants of these priors, even when restricted to a finite support, are not tractable. Heikkinen and Högmander (1994) discuss a similar problem and possible alternative strategies in a spatial problem arising in ecology. We opt here for estimating the βs using maximum pseudo-likelihood (Besag (1975)). This technique can be implemented in alternation with updates of the main parameters during an initial MCMC run until the nuisance parameter estimates stabilise. Further, in this example, we opt to use only the middle third of the columns to estimate $\beta_\mu, \beta_\gamma, \beta_{\sigma^2}$; considering again Figure 1 we are aiming to include only columns which contribute to a significant number of the observations.

Once the maximum pseudo-likelihood estimates of the nuisance parameters have stabilised, the remainder of the initial MCMC run can be used for estimation of the variance with a view to finding approximate eigenvalues and eigenvectors. Some care needs to be taken over numerical stability in this application as the three parameters of the Chapman profile approximations are on vastly different scales. To compare the efficiency of the original MCMC to the PCMCMC algorithms, both samplers were run for 100000 iterations, estimating integrated autocorrelation times from the second half of both runs. Table 2 demonstrates the impressive scale of the improvement obtained from this computationally intensive stage.

The PCMCMC algorithm as described generates posterior samples of the parameters describing the approximate Chapman profiles, μ, γ, σ^2 as well as the noise parameter τ^2. However we may wish to make inferences about the discretised electron density itself, D, in which case each iteration's output should be transformed according to Equation (6). In either case, the important feature is that we can produce interval estimates of whatever quantity is of interest. What we present here are (non-simultaneous) 95% credible intervals for D. Figure 8 shows the MIDAS (Spencer and Mitchell 2001) solution to the inversion problem together with the posterior mean estimate of D and the upper and lower ends of the credible intervals for D. As black indicates low values of electron density and white high values, the image of the upper ends of the credible intervals shows a broader band of grey pixels than the image of the lower ends. Comparing the MIDAS and posterior mean estimates, a few points to note are, firstly, that MIDAS has produced a smoother estimate than the Bayesian reconstruction. This may in part be due to the way in which MIDAS works,

	MCMC	PCMCMC		MCMC	PCMCMC		MCMC	PCMCMC
μ_1	9505	7.7	σ_1^2	3552	6.1	γ_1	1396	5.3
μ_2	10185	7.8	σ_2^2	4089	6.6	γ_2	1285	5.4
μ_3	11036	8.3	σ_3^2	4572	6.9	γ_3	3281	5.6
μ_4	11916	8.8	σ_4^2	4640	7.1	γ_4	3886	6.1
μ_5	12842	8.6	σ_5^2	6957	7.4	γ_5	957	5.6
μ_6	13793	9.3	σ_6^2	6669	7.4	γ_6	730	5.3
μ_7	14413	9.1	σ_7^2	6884	7.4	γ_7	672	5.3
μ_8	15217	8.7	σ_8^2	7662	7.9	γ_8	518	5.9
μ_9	15829	8.4	σ_9^2	11381	8.6	γ_9	339	4.7
μ_{10}	15053	8.4	σ_{10}^2	13750	8.9	γ_{10}	735 5	6.6
μ_{11}	15040	8.4	σ_{11}^2	13565	9.2	γ_{11}	11994	7.7
μ_{12}	15203	8.5	σ_{12}^2	13295	10.0	γ_{12}	12374	7.9
μ_{13}	15149	8.6	σ_{13}^2	12289	10.3	γ_{13}	11258	8.2
μ_{14}	14762	8.6	σ_{14}^2	11188	10.4	γ_{14}	10468	7.7
μ_{15}	13268	8.2	σ_{15}^2	10007	10.3	γ_{15}	2375	6.2
μ_{16}	11851	7.4	σ_{16}^2	9868	9.6	γ_{16}	3337	5.4
μ_{17}	9604	6.9	σ_{17}^2	7834	9.1	γ_{17}	1719	5.8
μ_{18}	7621	6.5	σ_{18}^2	5919	8.6	γ_{18}	385	5.8
μ_{19}	6228	6.3	σ_{19}^2	4719	8.4	γ_{19}	538	6.1
μ_{20}	5269	6.2	σ_{20}^2	4017	8.1	γ_{20}	704	5.8
μ_{21}	4528	6.0	σ_{21}^2	3505	7.7	γ_{21}	774	5.6
μ_{22}	3463	5.8	σ_{22}^2	1775	7.5	γ_{22}	740	5.6

Table 2: Comparison of the estimated integrated autocorrelation times using standard MCMC and PCMCMC for the ionospheric data set.

obtaining a least-squares fit to the data using a basis function of smoothly varying images. Secondly, both MIDAS and the Bayesian approach identify an area of high electron density in the left of the region. This is all the more surprising given that few rays pass directly through this region of high density (see the ray paths in Figure 1) although rays do pass through the columns in which it is situated. As with our experiments in earlier sections, we would expect a reasonable amount of uncertainty associated with such a feature; visually the credible intervals for D do indicate greatest width around this feature. On the far right of the image, there are also some columns through which no ray passes at all and so the only information we have about the values in these columns comes through our prior expectation of smoothness. The mean estimates here are essentially flat left to right (ie there is little change from column to column), as we might have expected. The widths of the corresponding credible intervals are largely governed by the strengths of the smoothing parameters of the priors.

We also need to consider the question of model fit and appropriateness of the assumptions made in formulating the likelihood. Figure 9 gives residual versus fitted value plots from both the MIDAS fit to the data and the Bayesian model (using residuals calculated by averaging the iteration-wise residuals). In the latter case, the fitted value of the noise variance in the model is $1/\tau^2 = 0.7$. As there is certainly residual structure in both cases, we have further identified the residuals associated with the four receivers using different colours. In the case of the MIDAS reconstruction, the greatest discrepancies are for the two left-hand receivers from Figure 1 (denoted in green and blue in the residual plot), ie those which are more affected by the area of high electron density. Generally the Bayesian model is able to provide a comparable or better fit to the data than the MIDAS algorithm including for these two receivers.

We also note though that while there is no apparent dependence of the residuals on the receiver or the size of the fitted value, there is a marked cyclical structure for all four receivers which calls into question our assumption of conditional independence of the observations. The data acquisition process of the satellite passing over the receivers does mean that the observations are not made simultaneously, a fact which we have ignored in our modelling. Successive observations, which are more closely spaced than indicated in Figure 1, follow rather similar routes (especially when considered in the discretised pixel representation) but through an ionosphere which has moved on in time (for the NNSS satellites, the orbiting time is about 20 minutes). There are also approximations to the geometry arising from flattening an inherently 3-dimensional problem into the two dimensions portrayed in Figure 1.

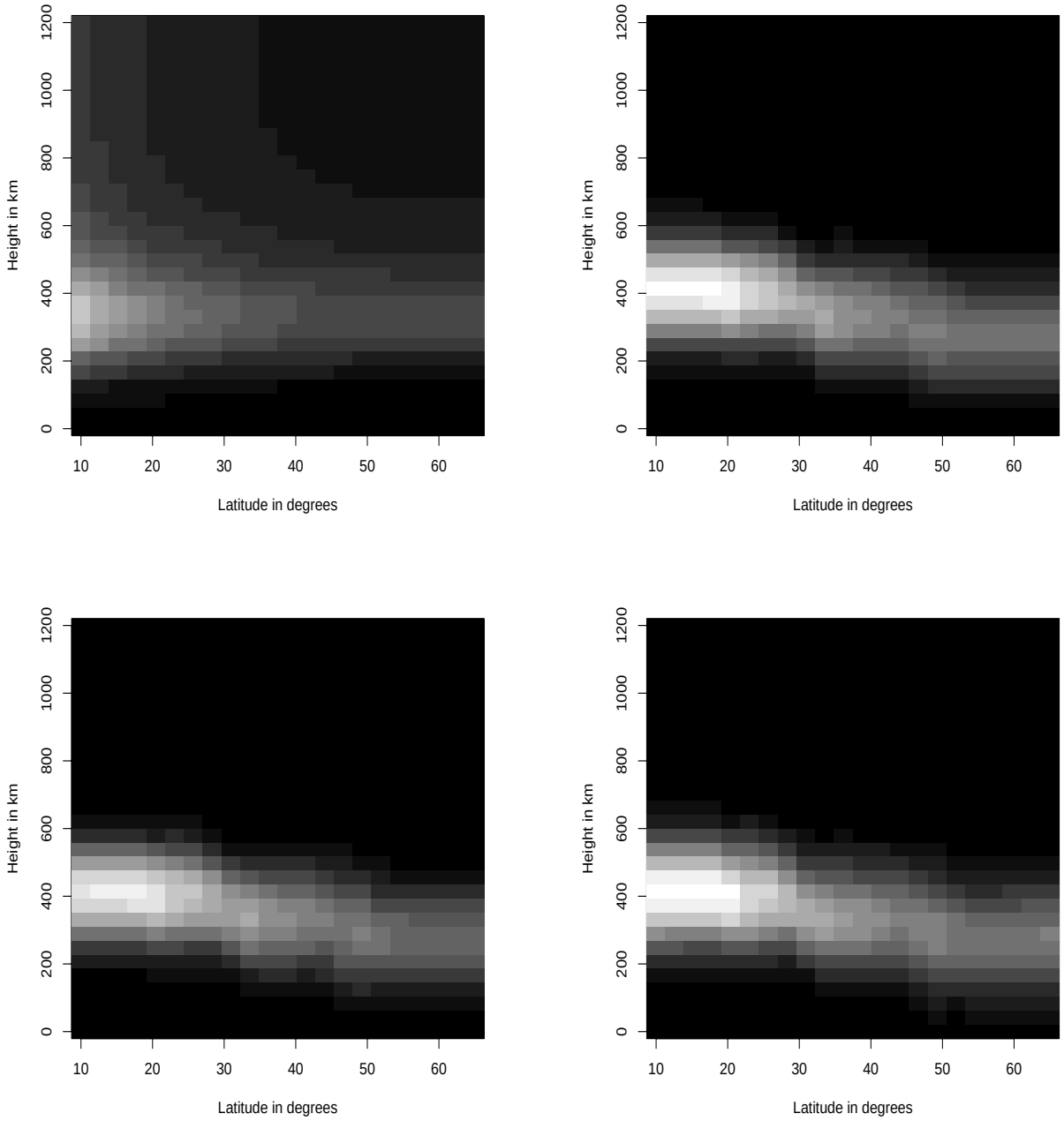

Figure 8: The ionosphere maps of free electron densities in units of $10^{11} el/m^2$ obtained from MIDAS (upper left), the posterior mean (upper right), the lower end of 95% non-simultaneous credible intervals (bottom left), the upper end of 95% non-simultaneous credible intervals (bottom right).

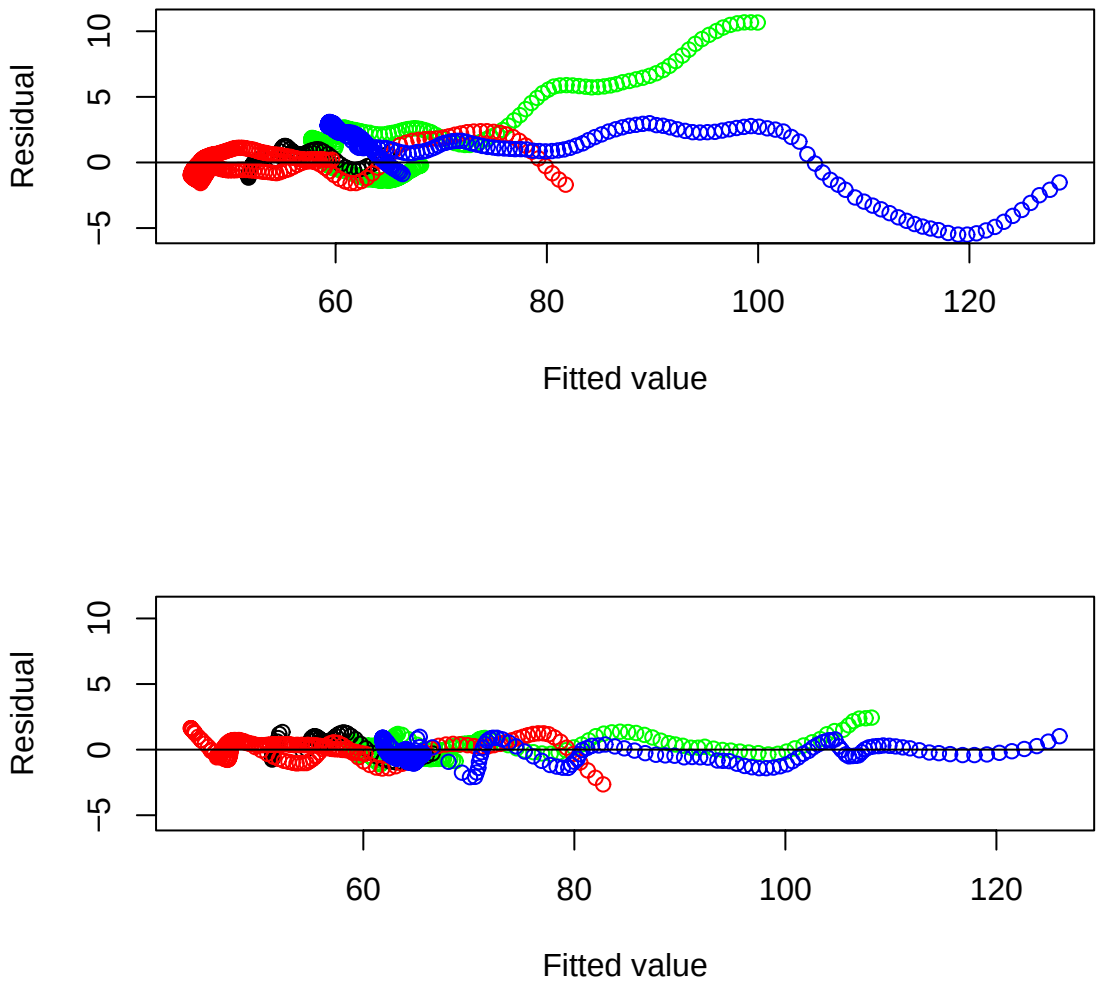

Figure 9: Residuals plots (with residuals defined as data minus the fitted value). The four different colours indicate the four different receivers (ordering from left to right in Figure 1: blue, green, red, black). Top panel: residuals from the MIDAS fit. Bottom panel: residuals from the Bayesian fit.

5 Discussion

We have seen that a rather simple Bayesian model is able to invert the type of tomographic data arising in studies of the ionosphere, improving upon an existing method in terms of data fit. A large attraction of any statistical approach is the generation of interval estimates but in this case the non-simultaneous credible intervals for D should be treated with some caution as the residuals suggests a more complex structure than was assumed in this initial modelling, possibly due to an unacknowledged temporal aspect. Since for this application, the ultimate goal is to generate 4-dimensional time + space electron density maps, the temporal structure of the data is the next obvious question to be tackled.

In order to get reasonable mixing of an MCMC algorithm, we have needed to find transformations of the original parameterisation. The principal component based approach proposed here is computationally demanding in set-up, but relatively robust at least for unimodal distributions and surprisingly effective. It also has the benefit of being non-specific to this particular application and might therefore prove useful in other hard sampling problems. In some cases, the orientation of the distribution to be sampled from may vary across the parameter space. It will then be desirable to replace the global set of principal component directions by local values. These may be computed by fitting a local multivariate normal approximation to the target distribution, using numerical approximations for the second derivatives of the log density. One option is to use a preliminary exploration to define sets of directions to work with in different parts of the sample space. Alternatively, a new set of directions can be generated for each move: local fitting of an approximating normal distribution is performed to create a proposal and, again, in constructing the reverse step that appears in the Metropolis-Hastings formula for the acceptance probability. Our initial investigations of highly curved distributions, such as those with high values of a in Appendix 2, indicate the computational effort in repeated computation of principal component directions is justified if the target distribution is essentially confined to a very thin region of varying orientation.

Appendix: PCMCMC stability

To be confident in the ability of the proposed PCMCMC algorithm to improve mixing, we must investigate how it behaves when the motivating assumptions (i.e. sampling along the principal directions of a multivariate normal) break down. Given that in practice the principal directions must be estimated from what may be a poorly mixed MCMC sample, we will first consider sensitivity to inaccuracies here. We will then consider what loss of performance there is as the distribution moves away from the assumption of normality.

A.1 Sensitivity to badly estimated principal components

One of the reasons why we resort to MCMC is that the distribution of interest is too complicated to handle analytically. It is unrealistic to imagine that we might know the covariance structure exactly, and we would only perhaps think of transforming variables in exactly those cases where an MCMC sampler failed to mix well in the original parameterisation. We need to investigate how PCMCMC performs if the principal components are badly estimated.

We will consider sampling from the highly correlated bivariate normal

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.0 & 0.999 \\ 0.999 & 1.0 \end{bmatrix} \right) \tag{10}$$

using n steps of a Metropolis algorithm started at the mean of the distribution (to avoid issues of burn in) and with a proposal standard deviation of 0.1 in both directions giving acceptance rates of about 45%. Figure 10 shows sample paths for chains of lengths between $n = 20$ and $n = 10000$ overlaid with the standardised eigenvectors estimated from the corresponding sample covariance matrices. Based on the $n = 10000$ run, the integrated autocorrelation times for X_1 and X_2 for this Metropolis sampler are 4756 and 5023.

Table 3 gives the corresponding estimated correlation and eigenvalues (the true values for the eigenvalues and eigenvectors are 1.999 and 0.001, and $(1/\sqrt{2}, 1/\sqrt{2})$ and $(1/\sqrt{2}, -1/\sqrt{2})$). While the eigenvectors stabilise even for quite small n, the eigenvalues take much longer, largely because the Metropolis sampler mixes so badly that the variances of X_1 and X_2 are severely under-estimated as can be seen from the scales of Figure 10. Based on these sets of estimated principal components, PCMCMC is run for 10000 iterations, initialised at the mean of the Metropolis run and using 2.4 times the square root of the eigenvalues for a proposal standard deviation (with a view to achieving an acceptance rate of about 45%). Table 3 shows the statistics of these runs. For $n = 20$, $n = 50$ and $n = 100$ the Metropolis under-estimation of the variances leads to high acceptance rates, and correspondingly high integrated autocorrelation times. Although the ergodic averages for $n = 20$ seem quite far from zero, they are in fact within 2 estimated standard errors once the integrated autocorrelation time has been taken into account. However, even for $n = 20$ the integrated autocorrelation times are a tenth those of the Metropolis sampler, and this drops to a thousandth for $n = 10000$. In this example, it seems that remarkably few steps of the original sampler are required in order to generate an efficient PCMCMC sampler. In the case of normal target distribution, it seems almost impossible to do worse by considering this approach.

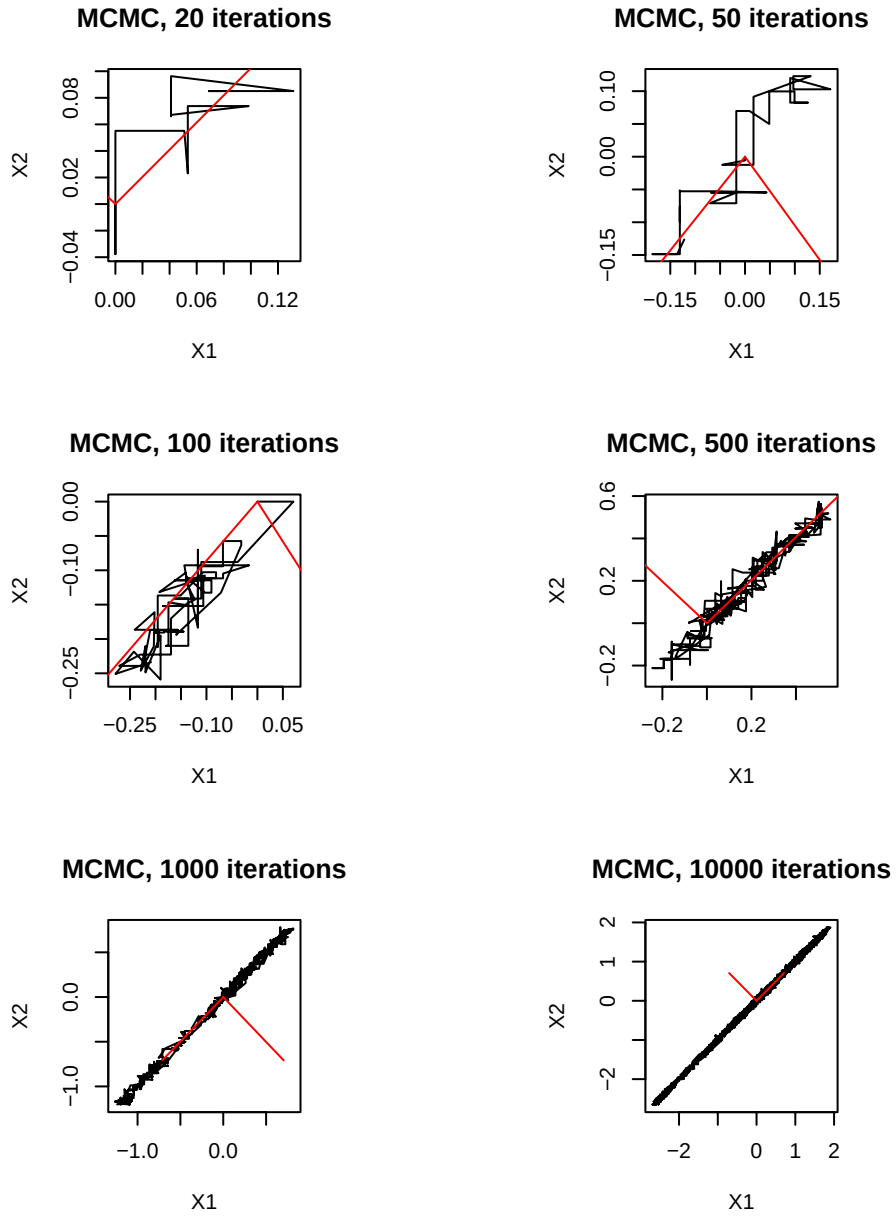

Figure 10: The sample paths of a Metropolis sampler run for n iterations for the distribution given in Equation (10). The estimated unit eigenvectors are overlaid in red.

	$\hat{\rho}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	Accept. rates	$\bar{X_1}$	$V(X_1)$	$\widehat{\tau^2}$ (X_1)	$\bar{X_2}$	$V(X_2)$	$\widehat{\tau^2}$ (X_2)
$n = 20$	0.736	0.0023	0.0003	97%, 60%	0.301	0.90	562	0.302	0.90	569
$n = 50$	0.890	0.0144	0.0008	91%, 48%	-0.043	0.97	121	-0.044	0.97	125
$n = 100$	0.861	0.0067	0.0005	85%, 56%	0.263	0.99	306	0.263	0.98	316
$n = 500$	0.977	0.0800	0.0009	84%, 46%	0.041	0.95	23	0.042	0.95	23
$n = 1000$	0.997	0.7184	0.0010	61%, 44%	-0.001	1.03	6	-0.002	1.03	6
$n = 10000$	0.999	2.268	0.0010	45%, 46%	-0.014	0.94	4	-0.014	0.94	4

Table 3: The performance of a PCMCMC sampler for the distribution given by Equation (10) when the eigen structure is estimated based on a Metropolis run of length n (see Figure 10). Estimated ρ, and estimated eigenvalues from the MCMC chain, acceptance rates for the moves in the two eigen directions, ergodic averages, variances and estimated integrated autocorrelation times for X_1 and X_2 from the PCMCMC chain.

A.2 Sensitivity to Non-normality

We now turn to the effect of non-normality on PCMCMC. Suppose that X_1 and X_2 are distributed

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 5 \\ 2 \end{bmatrix}, \begin{bmatrix} 2.0^2 & 0.396 \\ 0.396 & 0.2^2 \end{bmatrix} \right) \tag{11}$$

and that we construct a new pair of random variables by

$$\begin{aligned} Z_1 &= X_1 + a(X_2 - 2)^2 \\ Z_2 &= X_2, \end{aligned} \tag{12}$$

where a is a constant. We will consider this distribution for $a = 4, 40$ and 400, demonstrating increasing deviation from normality as a increases (although the distribution remains unimodal). A standard Metropolis sampler is started at $[6, 1]^T$ and run using normal proposals with standard deviations 1 and 0.1 in the Z_1 and Z_2 directions for 200,000 iterations with the first 50,000 iterations discarded as burn-in. Principal components are then calculated (using the correlation matrix), and PCMCMC again run for 200,000 iterations with the first 50,000 iterations discarded as burn-in and a scaling of 2.4 for the step size. Figure 11 shows the MCMC paths themselves after burn-in; the differences between the exploration of the distribution using the standard Metropolis sampler and PCMCMC are very clear using this representation. As a increases, both samplers struggle, but PCMCMC performs better of the two in terms of moving into the tails. The difficulty in mixing is also clear in Table 4 which gives estimated integrated autocorrelation times for the

		Acceptance rates	$\mathsf{E}(Z_1)$	$\bar{Z}_1$	$\widehat{\tau^2}(Z_1)$	$\mathsf{E}(Z_2)$	$\bar{Z}_2$	$\widehat{\tau^2}(Z_2)$
$a = 4$	MCMC	32%, 33%	5.16	5.095	489	2	1.994	447
	PCMCMC	40%, 37%		5.177	7.5		2.003	6.7
$a = 40$	MCMC	32%, 34%	6.6	6.49	1678	2	2.003	637
	PCMCMC	13%, 14%		6.550	112		1.998	67
$a = 400$	MCMC	33%, 12%	21	9.091	8064	2	2.011	589
	PCMCMC	5%, 7%		14.912	1364		2.028	371

Table 4: Comparison of the performance of a Metropolis sampler and a PCMCMC sampler for the distribution given by Equation (12) with $a = 4$, $a = 40$ and $a = 400$: Acceptance rates, theoretical means, ergodic averages and estimated integrated autocorrelation times.

Z_1 and Z_2 chains (estimated using the observed rather than known theoretical means). As the distribution moves further from normal, the advantage of PCMCMC over the Metropolis sampler decreases from a dramatic improvement in integrated autocorrelation time when $a = 4$ to a modest gain when $a = 400$. In this example, the non-normality of the target distribution has definitely not eliminated the benefits of PCMCMC.

Acknowledgements

We are grateful to Cathryn Mitchell and Paul Spencer of the Department of Electrical Engineering at the University of Bath for providing data, computational support and background information about the application. This research was supported by the Bath Institute for Complex Systems (EPSRC grant GR/S86525/01) and Eman Khorsheed was funded by the University of Bahrain.

References

[1] Besag, J. (1975), "Statistical analysis of non-lattice data", *The Statistician*, 24, 179–195.

[2] Censor, Y. (1983), "Finite series expansion reconstruction techniques", *Proceedings of the Institute of Electrical and Electronic Engineers*, 71, 409–419.

[3] Chapman, S. (1931), "The absorption and dissociative or ionising effect of monochromatic radiation in an atmosphere on a rotating Earth", *Proceedings of the Physical Society*, 43, 26–45.

[4] Chatfield, C. and Collins, A.J. (1980), *Introduction to Multivariate Analysis*. Chapman and Hall: London.

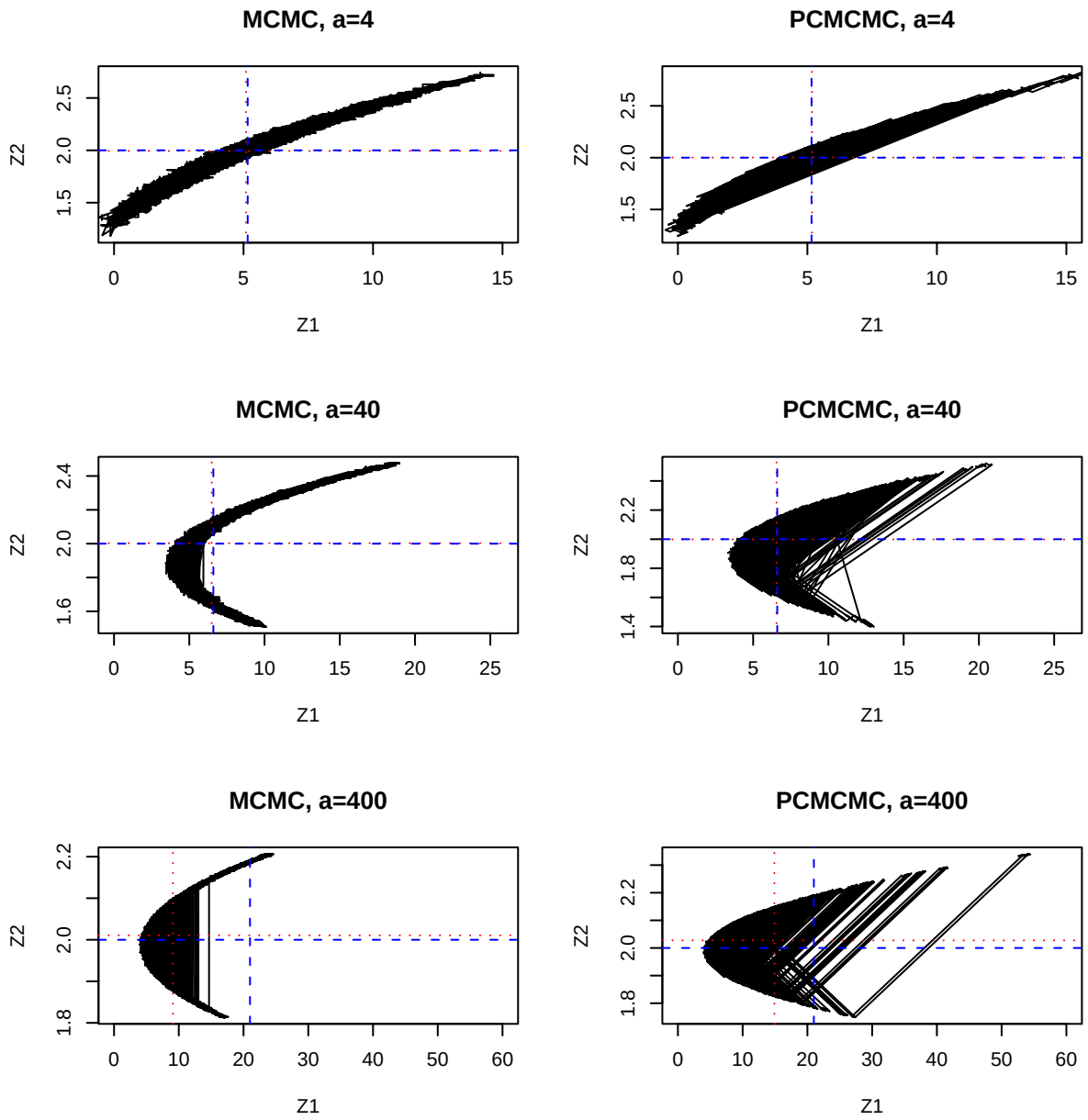

Figure 11: The sample paths of a Metropolis sampler and a PCMCMC sampler for the distribution given in Equation (12) with $a = 4$, $a = 40$ and $a = 400$. The theoretical mean in indicted in blue, and the ergodic averages in red.

[5] Fremouw, E.J., Secan, J.A. and Howe, B.M. (1992), “Application of stochastic inverse theory to ionospheric tomography”, *Radio Science*, 27, 721–732.

[6] Gilks, W., Richardson, S. and Spiegelhalter, D. (1996), *Markov chain Monte Carlo in practice*. Chapman and Hall: London.

[7] Gordon, R., Bender, R. and Herman, G.T. (1970), “Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography”, *Journal of Theoretical Biology*, 29, 471–481.

[8] Heikkinen, J. and Högmander, H. (1994), “Fully Bayesian approach to image restoration with an application in biogeography”, *Applied Statistics, Journal of the Royal Statistical Society, Series C*, 43, 569–582.

[9] Rue, H., and Held, L. (2005), *Gaussian Markov random fields: Theory and application.* Chapman and Hall/CRC: Boca Raton.

[10] Spencer, P.S.J. and Mitchell, C.N., (2001), “Multi-instrument data analysis system”, *Proceedings of the International Beacon satellite Symposium*, 4–6.