Geophysical Journal International

Geophys. J. Int. (2014) doi: 10.1093/gji/ggu079

Geophysical Journal International Advance Access published April 12, 2014

Receiver function deconvolution using transdimensional hierarchical Bayesian inference

J. M. Kolb[1,2] and V. Lekić[1]

[1]*Department of Geology, University of Maryland, College Park, MD* 20742, *USA. E-mail: jkolb1@terpmail.umd.edu*
[2]*Department of Geoscience, University of Calgary, Calgary, Canada*

Accepted 2014 March 3. Received 2014 February 4; in original form 2013 August 2

SUMMARY

Teleseismic waves can convert from shear to compressional (*Sp*) or compressional to shear (*Ps*) across impedance contrasts in the subsurface. Deconvolving the parent waveforms (*P* for *Ps* or *S* for *Sp*) from the daughter waveforms (*S* for *Ps* or *P* for *Sp*) generates receiver functions which can be used to analyse velocity structure beneath the receiver. Though a variety of deconvolution techniques have been developed, they are all adversely affected by background and signal-generated noise. In order to take into account the unknown noise characteristics, we propose a method based on transdimensional hierarchical Bayesian inference in which both the noise magnitude and noise spectral character are parameters in calculating the likelihood probability distribution. We use a reversible-jump implementation of a Markov chain Monte Carlo algorithm to find an ensemble of receiver functions whose relative fits to the data have been calculated while simultaneously inferring the values of the noise parameters. Our noise parametrization is determined from pre-event noise so that it approximates observed noise characteristics. We test the algorithm on synthetic waveforms contaminated with noise generated from a covariance matrix obtained from observed noise. We show that the method retrieves easily interpretable receiver functions even in the presence of high noise levels. We also show that we can obtain useful estimates of noise amplitude and frequency content. Analysis of the ensemble solutions produced by our method can be used to quantify the uncertainties associated with individual receiver functions as well as with individual features within them, providing an objective way for deciding which features warrant geological interpretation. This method should make possible more robust inferences on subsurface structure using receiver function analysis, especially in areas of poor data coverage or under noisy station conditions.

Key words: Time-series analysis; Inverse theory; Body Waves.

1 INTRODUCTION

Teleseismic waves can convert from shear to compressional (*Sp*) or compressional to shear (*Ps*) across impedance contrasts or due to the presence of seismic anisotropy (Park & Levin 2000). These conversions can be recorded at the surface by seismometers, and analysed to probe Earth structure beneath the receiver. The waveforms of the *Ps* or *Sp* can be used to infer velocity structure directly (Burdick & Langston 1977; Langston 1977); typically, however, the parent waveform (*P* for *Ps* or *S* for *Sp*) is deconvolved from the daughter waveform (*S* for *Ps* or *P* for *Sp*) to yield a receiver function (e.g. Vinnik 1977; Langston 1979), before structural inferences are made. The advantage of the additional deconvolution step stems from its ability to remove waveform complexity due to source-side structure and the source time function. Furthermore, receiver functions derived from multiple source–receiver pairs can be combined into a 3-D discontinuity model using either simple common conversion point stacking (e.g. Lekic *et al.* 2011; Levander & Miller 2012), or more sophisticated formal wavefield migration techniques (for a review, see Rondenay 2009), which is less straightforward to do with actual velocity profiles.

A variety of techniques have been developed to deconvolve the parent waveform from the daughter waveform including damped spectral division (e.g. Langston 1979; Ammon 1991; Bostock 1998), iterative time-domain deconvolution (Ligorría & Ammon 1999), the multitaper method (Park & Levin 2000; Helffrich 2006) and a form of Bayesian deconvolution (Yildirim *et al.* 2010).

Challenges common to all methods for obtaining receiver functions are background noise—which is loudest in the microseismic band (e.g. Peterson 1993)—and signal-generated noise, which cannot be modelled through a convolution operator common to all waveforms observed at a seismic station due to energy being singly or multiply scattered by structures that are not common to all paths. For an individual parent–daughter pair on a single path, signal-generated noise is not an issue, except for *Sp*, which arrives in the *P* coda. Noise is especially problematic for *Sp* receiver functions

because the parent *S* waves overlap in their frequency content with the primary and secondary microseism (e.g. Bromirski 2009) resulting in signal-to-noise ratios that are typically much smaller than in *Ps*. Creating a method that is robust with respect to this noise is important for enabling receiver function analyses for stations in noisy environments (e.g. ocean bottom, coastal, polar regions), temporary deployments and for inferring anisotropy through variations of receiver functions with backazimuth. The smaller bin size needed for stacking receiver functions by backazimuth requires reliably estimating receiver functions with many fewer waveforms.

Spectral division, which due to spectral holes and noise must be stabilized using either damping or a water-level (Clayton & Wiggins 1976), continues to be successfully applied to receiver function estimation (e.g. Abt *et al.* 2010; Schaeffer & Bostock 2010). However, introduction of damping/water-level typically generates side lobes, which complicate the interpretation of the resulting receiver functions. Another drawback is that reliable estimates of noise are required to choose an appropriate level of damping/water-level. Furthermore, when noise levels and damping/water-level are high, spectral division often does not yield reliable and interpretable receiver functions.

Frequency-domain deconvolution can be made more robust by the introduction of multitaper estimates of the spectra. While this enables spectral leakage to be reduced, it also decreases spectral resolution. The reduced spectral resolution obliterates receiver functions at large lag times (Helffrich 2006). Furthermore, receiver function amplitudes can be affected by the choice of the time–bandwidth product, the number of tapers, and by the fact that tapers reduce the amplitude for the majority of the time window (e.g. Shibutani *et al.* 2008).

An alternative, time-domain method is iterative time-domain deconvolution (Ligorría & Ammon 1999), which parametrizes the receiver function by a set of Gaussians of unknown amplitude but whose width and number has to be decided prior to the deconvolution. Also, the method can get stuck in local minima in the model space because once it places a Gaussian somewhere it cannot adjust that Gaussian's position later in the process.

Because estimation of receiver functions can be cast as an inverse problem, Bayesian approaches can be readily applied (Tarantola & Valette 1982). We believe that a Bayesian approach to deconvolution such as in Yildirim *et al.* (2010) is advantageous for generating robust receiver function estimates because of its ability to test different models and compare them probabilistically. In addition, in a Bayesian approach, the data uncertainty affects the posterior probability distribution and can be used to obtain more accurate results (e.g. Bodin *et al.* 2012).

It is well known that inferences of velocity profiles based on receiver functions are highly non-unique (e.g. Ammon *et al.* 1990). However, the step of creating the receiver function is also non-unique (e.g. Lavielle 1991), which is apparent from the fact that the convolution matrix is often singular or ill-conditioned. Therefore, methods that yield only a single deconvolution estimate are of limited utility. This aspect of the non-uniqueness may be particularly important when few waveforms are used under noisy conditions to estimate the receiver function.

We present a method based on transdimensional (see Sambridge *et al.* 2013) hierarchical (Malinverno & Briggs 2004) Bayesian inference in which both the noise magnitude and noise spectral character are parameters in calculating the likelihood probability distribution. In our method, we use a reversible-jump Markov chain Monte Carlo (henceforth, RJMCMC) algorithm—first proposed by Green (1995), applied to the analysis of mixtures with an unknown number of components by Richardson & Green (1997), and later adopted to geoscience by Malinverno (2002)—to find an ensemble of receiver functions whose relative fits to the data have been calculated while simultaneously inferring the values of the noise parameters. This RJMCMC algorithm is an extension of the Metropolis–Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970) in which the proposal distribution allows for jumps between subspaces of different dimensions, allowing for dimensionality to be inferred along with the model parameters. Motivated by the work of Piana Agostinetti & Malinverno (2010), who used RJMCMC to infer velocity structure from receiver functions, we treat the number of parameters defining the receiver function as an unknown to be estimated in the inversion. This is an appropriate choice since the complexity of the receiver function is not known *a priori* and will vary with geological setting. Yildirim *et al.* (2010) showed that including constraints on the sparsity of the receiver functions, which stems from the fact that it is a limited number of discrete impedance contrasts that produce *Sp* and *Ps* conversions, can improve stability and robustness of the deconvolution; this sparsity characteristic is not exploited by other deconvolution methods, except indirectly by iterative time-domain deconvolution. Therefore, we parametrize the receiver function as an unknown number of Gaussians of unknown amplitude and width.

Stawinski *et al.* (1998) performed transdimensional Bayesian deconvolution and Andrieu *et al.* (2001) performed transdimensional hierarchical Bayesian deconvolution (THBD; with hyperparameters accounting for noise), both in the field of nuclear imaging. Kang & Verotta (2007) also demonstrated transdimensional Bayesian deconvolution, but applied to pharmacokinetic data. Application to receiver function estimation requires, among other things, different proposal distributions for new models as well as a different noise parametrization. In addition, with advances in processor speed, changing multiple components of a Markov chain simultaneously can now be applied in this method, which due to a lower acceptance rate will require more iterations for the algorithm to converge as well as to sample the posterior distribution, but will reduce the likelihood of the algorithm getting trapped in local minima.

Using synthetic data with realistic noise characteristics, we show that the THBD method can accurately obtain a receiver function as well as estimate the noise parameters. Furthermore, we demonstrate that this new approach is far less susceptible to generating spurious features even at high noise levels due to the natural parsimony inherent in transdimensional Bayesian methods (e.g. Malinverno 2002). Finally, the method yields not only the most likely receiver function, but also enables its full uncertainty to be quantified by analysing the ensemble solution.

2 METHOD

We seek a method that can obtain reliable receiver functions in the presence of high noise levels, estimate uncertainty by creating an ensemble of solutions and estimate the characteristics of the noise. Piana Agostinetti & Malinverno (2010) presented a method based on transdimensional Bayesian inference, and Bodin *et al.* (2012) presented a method based on transdimensional hierarchical (with noise hyperparameters) Bayesian inference, both for inferring subreceiver velocity structure from receiver functions. Motivated by this idea, we develop a related method for generating receiver functions from parent and daughter waveforms. An overview of this method is shown in Fig. 1.

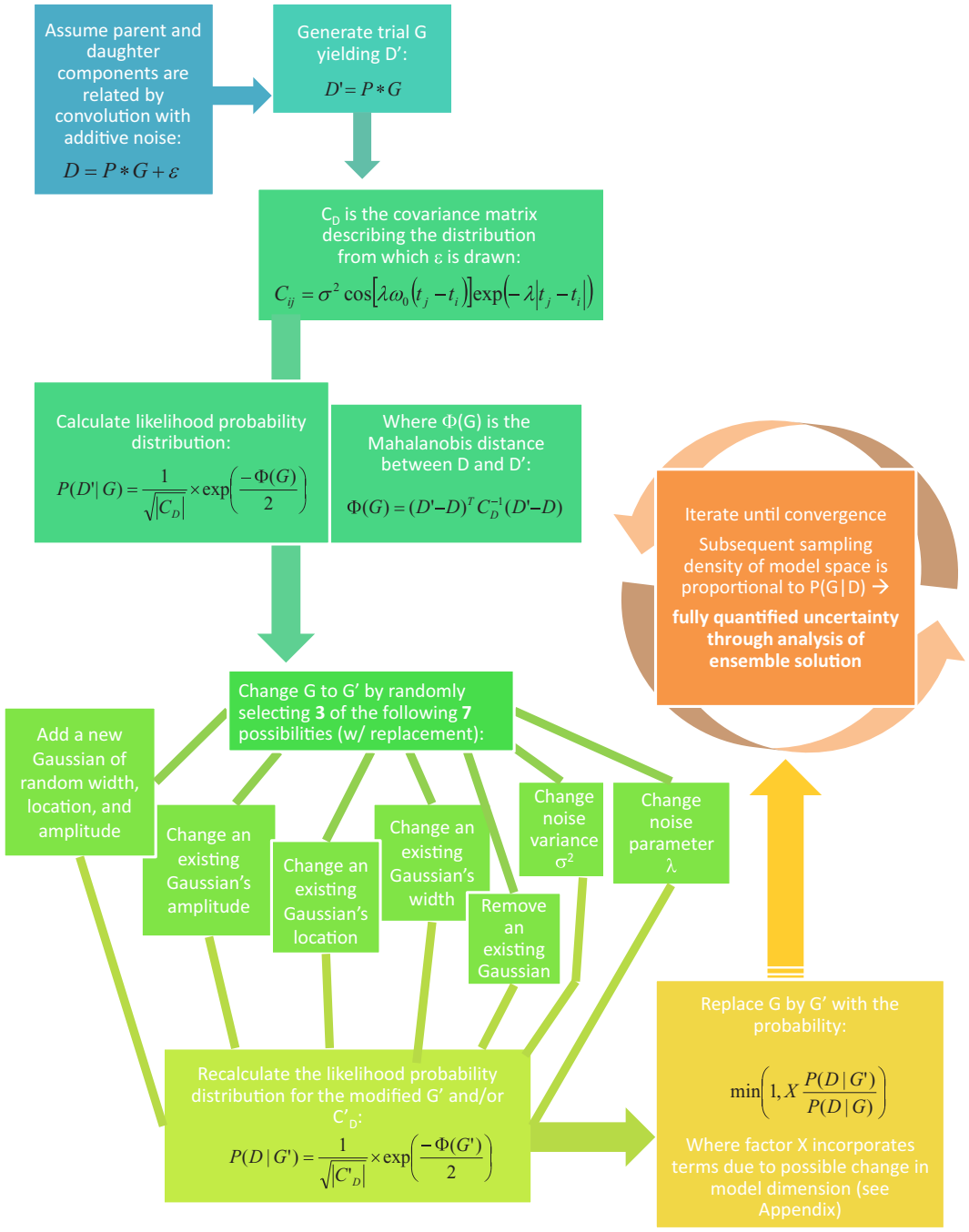

Figure 1. An overview of the reversible-jump Markov chain Monte Carlo algorithm used for our deconvolution. The $C_{D_{ij}}$ shown describes noise parametrization type 3. For parametrization types 1 and 2, see Section 3.

We start by assuming that the parent (P) and daughter (D) waveforms, sampled at times t_i are related by convolution with a receiver function (G)

$$D = P * G. \tag{1}$$

The challenge is to estimate G given P and D, both of which are contaminated by background and signal-generated noise of unknown amplitude and frequency content. We parametrize G as an unknown number of Gaussians of unknown width (w_i) and amplitude (a_i), centred at unknown lag times (c_i). We assume that the noise can be described by random sampling from a distribution defined by a covariance matrix C_{D}:

$$C_{\mathrm{D}_{ij}} = \sigma^2 R_{ij}, \tag{2}$$

where R_{ij} is a function of absolute lag time $|t_i - t_j|$ whose form depends on our choice of one of three different noise parametrizations. By parametrizing R as a Toeplitz matrix, we make the assumption that the noise is invariant with respect to time, within the duration of a single waveform pair; noise characteristics across waveform pairs, specified by different Toeplitz matrices, are allowed to vary. As we show in Section 3, the choice of parametrization will emerge as important because the covariance matrix is used in

calculating the likelihood of a particular receiver function. The first two types of noise parametrizations we explore were previously used by Bodin *et al.* (2012); in addition to these, we introduce a third parametrization consisting of a decaying sinusoidal function. This noise parametrization was motivated by covariance matrices estimated from actual pre-event noise observed at stations in a variety of tectonic settings. We analyse and discuss these different parametrizations in detail in Section 3.

We estimate the number of Gaussians, their locations, widths and amplitudes, as well as two additional parameters describing the noise amplitude and frequency content using an RJMCMC implementation of transdimensional hierarchical Bayesian inference, applied to receiver function inversion by Bodin *et al.* (2012).

To process our data, we start by bandpassing the parent and daughter waveforms to include signal between periods $T_{\min}$ and $T_{\max}$, and then decimate them to three times the Nyquist frequency. When estimating an *Sp* receiver function, we time-reverse both the *S* and *P* waveforms, in order to consider only positive lag times.

We start the Markov chain with an initial model, G, that contains no Gaussians, and with a starting σ equal to the standard deviation of the parent waveform. Note that though starting with an initial model closer to the 'true' model should cause the Markov chain to converge faster, we decided to start with a model without Gaussians to ensure that Gaussians in our result were placed by our algorithm and not simply due to our initial model. Also, by setting σ equal to the standard deviation of the parent, we assume that all of the waveform is noise, and reductions in the noise magnitude by the signal are indications that some of the waveform can be interpreted as a signal related through convolution. Note that if the actual noise level is low, our choice of a starting σ value will also slow the convergence rate. Finally, initializing the Markov chain with a single randomly generated Gaussian tends to produce larger misfits (for the first tested model) than starting with no Gaussians and does not cause the chain to converge any faster.

We first convolve G with the parent waveform and calculate the Mahalanobis distance between that convolution and D:

$$\Phi(G) = (P * G - D)^T C_{\mathrm{D}}^{-1} (P * G - D), \tag{3}$$

where C_{D} is the data covariance matrix with standard deviation σ and noise correlation parametrized as described in Section 3. In order to obtain $\Phi(G)$ with more efficiency, we solve for $C_{\mathrm{D}}^{-1}(P * G - D)$ as a system of linear equations, eliminating the need to calculate C_{D}^{-1} explicitly. Also, by calculating the LU factorization of C_{D}^{-1} and saving the factors, the process is sped up further, with the factorization only needing to be performed when the noise correlation is changed.

The Mahalanobis distance determines the likelihood probability of the observed daughter waveform given the model:

$$P(D|G) = \frac{1}{\sqrt{(2\pi)^n |C_{\mathrm{D}}|}} \mathrm{e}^{\frac{-\Phi(G)}{2}}, \tag{4}$$

where n is the number of points in the data vector. By Bayes' Theorem, this probability is proportional to the probability of the model, given the data, that is $P(G|D) \propto P(D|G)$.

At each step of the Markov chain, a new model G' is created by choosing, with replacement, three of the following seven possibilities:

(1) Creating a new Gaussian with random width, location and amplitude, according to probabilities listed in Table 1;
(2) Changing an existing Gaussian's amplitude;
(3) Changing an existing Gaussian's width;
(4) Changing an existing Gaussian's location;
(5) Changing the noise standard deviation;
(6) Changing the noise correlation parameter;
(7) Removing a Gaussian.

Table 1. Parameters defining the probability density functions (PDFs) from which the random model realizations/updates are drawn. The first three parameters have probabilities listed for creating a new Gaussian with uniform distributions between *min* and *max*. The last five parameters have probabilities listed for changing the value based on a normal distribution with mean μ and standard deviation θ. $T_{\mathrm{min,max}}$ are the low and high period corners of the bandpass applied to the data. σ_P denotes the standard deviation of the parent waveform. α is defined in the text.

Parameter	PDF type	min/μ	max/θ
Location (c_i)	Uniform	0 s	25 s
Width (w_i)	Uniform	$\frac{1}{10} T_{\min}$	$\frac{1}{10} T_{\max}$
Amplitude (a_i)	Uniform	-1.5α	1.5α
ΔLocation (c_i)	Gaussian	t	0.15
ΔWidth (w_i)	Gaussian	w	0.04
ΔAmplitude (a_i)	Gaussian	a	0.1α
$\Delta\sigma$ (noise)	Gaussian	σ	$0.0025\sigma_P$
$\Delta\lambda$ (noise)	Gaussian	λ	0.000125

If there are no Gaussians, option 1 is chosen. Otherwise, option 6 is given a 2.5 per cent probability of being chosen, with the remaining 97.5 per cent probability split between the other six options. This is done because changing the noise correlation necessitates updating the LU factorization of C_{D}^{-1} and is computationally expensive. Reducing the likelihood of selecting option 6 allows us to speed up the algorithm without imposing a fixed value for the noise correlation parameter.

After choosing the three options, if any of the Gaussians overlap—which we define as the area within one standard deviation of the centre of one Gaussian overlapping in time with the area within one standard deviation of the centre of another Gaussian—then that model is rejected and a new model is chosen. This incorporates our prior knowledge that the receiver function should be sparse. An alternative implementation would be to shift overlapping Gaussians so that they are no longer overlapping; we opt against this implementation because it would cause more models to be tested at that spot in the model space, which would violate the Metropolis algorithm.

Because of the transdimensional nature of our algorithm, initial models with no Gaussians can quickly add many Gaussians in order to model complexity without ever correctly modelling the largest Gaussians, or can model large Gaussians as a superposition of smaller nearby Gaussians. As the number of dimensions increases, changes to any single Gaussian become less likely. So, in order to more quickly converge to a well-fitting model, we place limits on how quickly the total number of Gaussians can increase during the burn-in period. This allows the algorithm more time to optimize the locations, widths and amplitudes of the existing Gaussians as the dimensionality and thus the model space increases. We impose a limit of k Gaussians in the first $1000k(k+1)$ iterations. For example, up to one Gaussian is allowed from 1 to 2000 iterations, up to two Gaussians are allowed from 2001 to 6000 iterations, up to three Gaussians are allowed from 6001 to 12 000 iterations, etc. This continues so that a maximum of 30 Gaussians are allowed until 9.3×10^5 iterations after which there is no limit to the number of Gaussians allowed. It should be noted, however, that

for computational convenience and to place limits on receiver function complexity, we set the prior on having more than 30 Gaussians to zero, effectively making our Gaussian limit scheme end with 30 Gaussians being allowed after 8.7×10^5 iterations.

The values of the relevant parameters are randomly drawn from probability density functions (PDFs) defined in Table 1. The factor of $\frac{1}{10}$ relating the minimum and maximum widths of the Gaussians to $T_{\min}$ and $T_{\max}$ is introduced so that the peak power of the narrowest and widest allowed Gaussians is approximately equal to the low ($T_{\min}$) and high ($T_{\max}$) period corners of the bandpass applied to the data. The α in Table 1 is an estimate of the magnitude of the largest Gaussian in the receiver function based on the cross-correlation of the parent and daughter waveforms and comes from eq. 12 in Kikuchi & Kanamori (1982), where x is D and w is the convolution of P and the minimum width Gaussian allowed in the algorithm. The likelihood probability given by eq. (4) is then calculated for the new model, G', and the chance that the model update is accepted is given by

$$\min\left(1, \frac{P(D|G')}{P(D|G)}\frac{(k+1)}{k'}\right), \tag{5}$$

where k is the number of Gaussians in the current model G and k' is the number of Gaussians in the proposed model G'. The extra ratio term is introduced by the prior, as explained in the Appendix.

The ratio of model probabilities determines the acceptance probability of the new model. We calculate the acceptance in log space, taking advantage of the fact that

$$\frac{1}{2}\log(|C_{\rm D}|) = n\log\sigma + \sum_i \log\left(R^{\rm U}_{ii}\right), \tag{6}$$

where $R^{\rm U}$ is the upper triangular Cholesky factor of R; this is a computationally efficient means of calculating the determinant of the covariance matrix.

Since the ensemble solution we obtain with THBD is only truly representative of the posterior probability density for the model parameters if convergence has been achieved, we run our chains at least this long before saving models. Slow convergence is a common problem in Markov chain Monte Carlo (MCMC) methods (Gilks *et al.* 1996) and a number of tools to analyse convergence has been proposed. Cowles & Carlin (1996) analysed 13 commonly used convergence diagnostics for MCMC algorithms, recommending using multiple tests to analyse convergence, but also noting that it is never possible to say with certainty that a finite sample from an MCMC algorithm has converged to the underlying stationary distribution. Therefore, we test convergence in a number of ways. First, we ensure that the misfit is and remains low and the calculated likelihood is and remains high. Fig. 2 shows the evolution of the model likelihood as a function of iteration for 32 parallel chains. Then, we compare the marginal probability densities obtained by a number of different MCMC chains; if all the chains have converged to the same solution, then these marginal probability densities will be similar. We confirm that this is indeed the case for the numbers of Gaussians, locations of Gaussians, amplitudes and widths, as shown in Fig. 3.

After a burn-in period lasting at least as long as the limit on the maximum number of Gaussians is in place, and convergence is achieved, the model is saved every 500th iteration for another 10^6 iterations. The purpose of saving only every 500th iteration is to reduce computer storage while allowing the algorithm enough iterations to create uncorrelated models in the model space. Typically, we find that the algorithm converges within $\sim 5 \times 10^5$ iterations, while the limit on the maximum number of Gaussians is still in place.

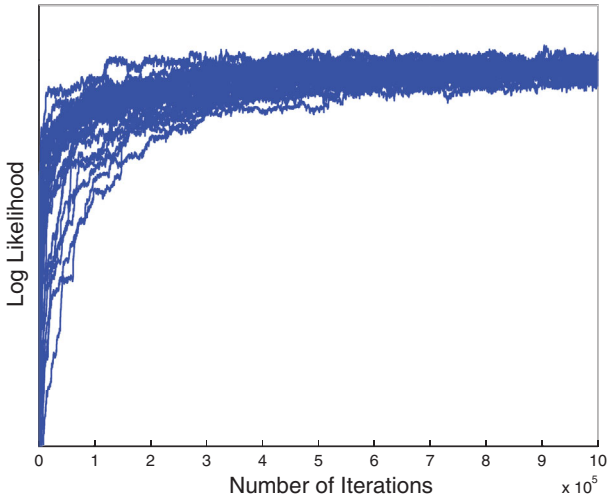

Figure 2. Likelihood progression by iteration for 32 Markov chains. The algorithm generally converges within $\sim 5 \times 10^5$ iterations for *Ps* waveform pairs.

3 NOISE PARAMETRIZATION

In order to get reliable results with this algorithm, it is important that the covariance matrix $C_{\rm D}$ accurately describes the covariance of our data. Data residuals can be used to estimate the full data covariance matrix if array data—in which noise character is common across the array—are available (Holland *et al.* 2005; Dosso *et al.* 2006). However, because noise characteristics depend on time and location, this promising method is unavailable to us; nevertheless, it should be explored in situations where dense seismic arrays are operating.

Dettmer *et al.* (2012) propose a new parametric procedure for estimating the full data covariance matrix based on autoregressive error models of arbitrarily high order. This technique is very promising in MCMC applications, since it obviates the need for calculating the inverse and determinant of the data covariance matrix. However, without careful study of noise characteristics, this method runs the risk of misidentifying actual signals as noise, if the autoregressive order is chosen to be unjustifiably high. Therefore, in our study, we opt for a direct parametrization of $C_{\rm D}$ that approximates characteristics observed on actual pre-event time-series of noise at broad-band seismometers in a variety of different settings.

Bodin *et al.* (2012) suggest two parametrizations for the covariance matrix: type 1, an exponentially decaying correlation (eq. 7) and type 2, a Gaussian correlation (eq. 8). They also show that the type 2 parametrization is more realistic than type 1, because an exponentially decaying correlation tapers off quickly, producing noise that has a higher frequency content than observed in actual pre-event noise. On the other hand, the exponentially decaying correlation has the advantage that analytical forms for the inverse and determinant of its covariance matrix exist and can be quickly computed. A type 2 covariance matrix, however, is difficult to invert; Ababou *et al.* (1994) show that it is, in fact, one of the worst-conditioned covariance matrices possible due to its zero slope at $t = 0$.

To evaluate how appropriate these noise parametrizations are for representing actual noise, we analysed pre-event noise at ANMO. We created covariance matrices based on 250-s samples of noise before 5833 recorded earthquakes, by normalizing the noise time-series, placing them in a matrix, and then multiplying that matrix by its transpose. Then, each row of the resulting matrix is divided by the value of the diagonal entry. The first 50 s of the resulting covariance matrix can be seen in Fig. 4(d). We then averaged the values along

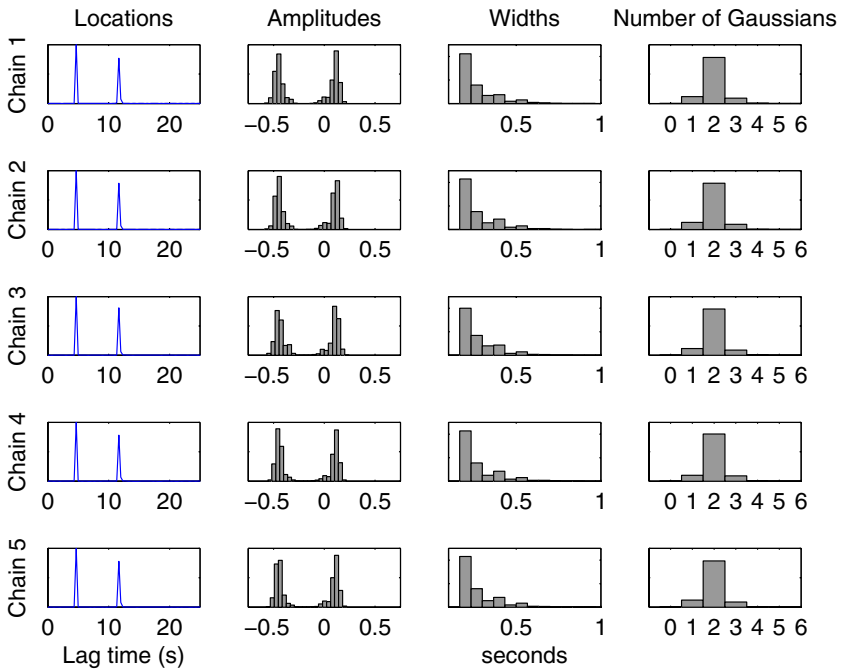

Figure 3. Histograms showing marginal densities of Gaussian locations, amplitudes, widths and number for five chains run in parallel on the same waveform pair. The fact that the parallel chains are sampling the same parameter values increases our confidence that the algorithm is successfully converging during the burn-in period.

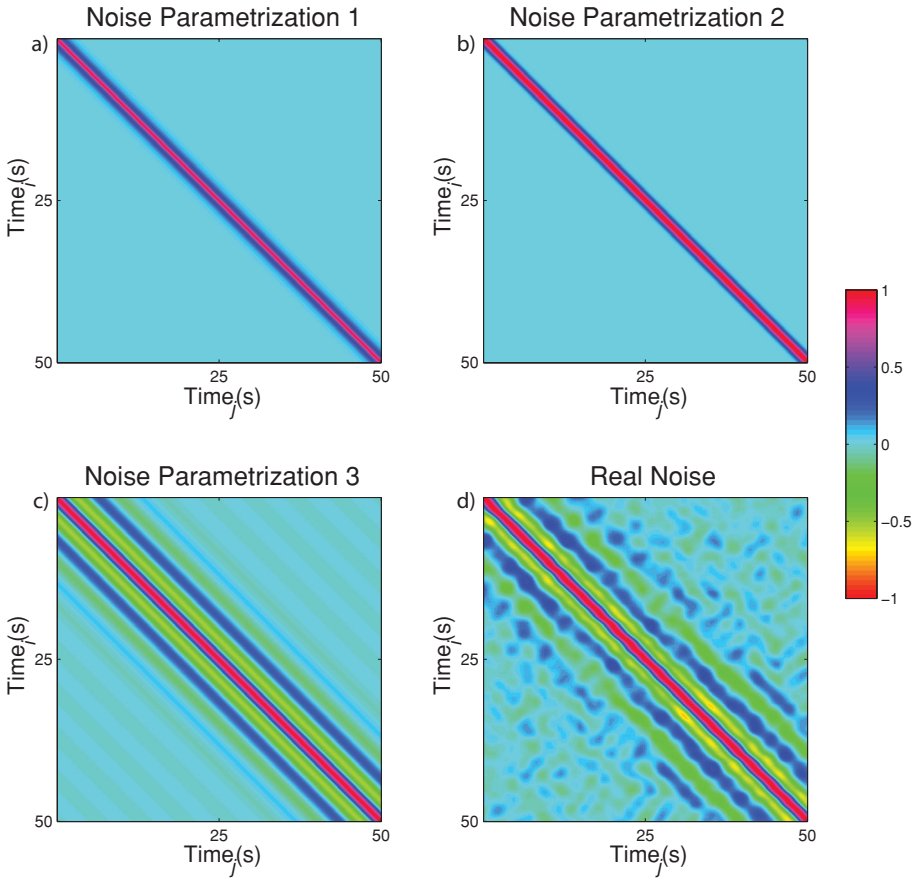

Figure 4. Covariance matrices from the three types of noise parametrization and actual noise. (a–c) Parametrizations 1–3. (d) Covariance matrix generated from pre-event noise at ANMO. The rate at which correlation decays with increasing lag times varies across the parametrizations; parametrization 3 (c) is most similar to the behaviour seen in real noise covariance (d), due to its ability to represent the microseism.

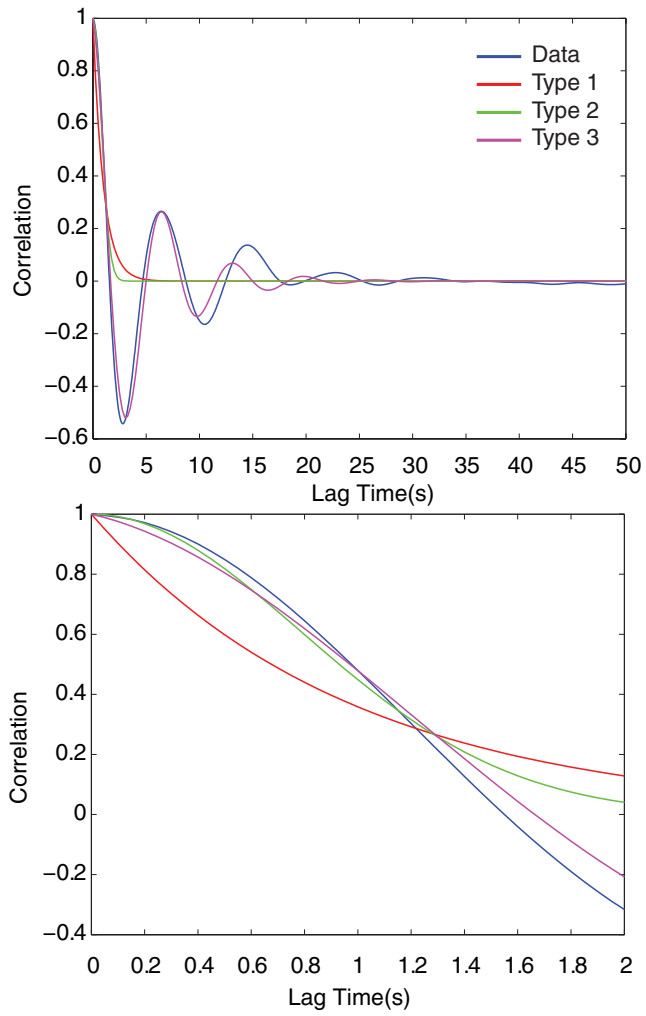

Figure 5. Noise correlation from actual noise versus parametrizations as a function of lag time. Parametrization 3 (with $\lambda = 0.2$ and $\omega_0 = 4.4$) fits the real noise correlation best for all but the smallest lag times, where parametrization 2 fits the real noise correlation slightly better. Parametrization 1 yields poor fits at all lag times.

the diagonal of our covariance matrix to create an estimate of the correlation as a function of lag time, which is shown in Fig. 5. It can be seen that the noise resembles a decaying sinusoidal function, whose periodic anticorrelations cannot be described by the first two types of noise parametrization. We verify that similar covariance matrices describe the noise at seismic stations in other geological settings, including islands, coastlines and Antarctica.

Based on this shape of the noise correlation, we propose a parametrization of noise, type 3, which is a decaying exponential multiplied by a cosine function. Given this, our three noise parametrizations are

Type 1:

$$R_{ij} = \mathrm{e}^{-\lambda|t_j - t_i|}, \tag{7}$$

Type 2:

$$R_{ij} = \mathrm{e}^{-\lambda^2|t_j - t_i|}, \tag{8}$$

Type 3:

$$R_{ij} = \mathrm{e}^{-\lambda|t_j - t_i|}\cos(\lambda\omega_0|t_j - t_i|). \tag{9}$$

In order to limit the number of parameters when using the third noise parametrization, we fixed ω_0, in essence scaling the decay to the oscillation rate of the covariance. Note that the R_{ij} appear in the definition of the covariance matrix in eq. (2).

For each noise parametrization, we found the parameters that best fit the actual noise, and plotted the modelled and observed correlations in Fig. 5. Parametrization type 2 has the best fit at very small lag times ($T < 0.6$ s), but in keeping close to the zero slope of the correlation observed at zero lag time, it yields ill-conditioned covariance matrices. Parametrization type 3 fits the observed correlation across a much wider range of lag times, and especially at long lag times. Parametrization type 1 does a poor job at all lag times.

In our problem, noise on the parent and daughter components will affect the misfit in different ways. If there is noise $\epsilon(t)$ on the parent component, noise $\eta(t)$ on the daughter component and the true receiver function is G_0, then the parent component we observe is

$$P_{\mathrm{obs}} = P + \epsilon, \tag{10}$$

and the observed daughter component is

$$D_{\mathrm{obs}} = P * G_0 + \eta. \tag{11}$$

Since we are estimating G from observed waveforms, that is, $D_{\mathrm{obs}} = P_{\mathrm{obs}} * G$, we are solving for G in

$$P * G_0 + \eta = (P + \epsilon) * G, \tag{12}$$

and the Mahalanobis distance (eq. 4) becomes

$$\Phi(G) = [(P + \epsilon) * G - (P * G_0 + \eta)]^T \times C_{\mathrm{D}}^{-1}[(P + \epsilon) * G - (P * G_0 + \eta)]. \tag{13}$$

Therefore, assuming G and G_0 are approximately equal, the effective noise is $\epsilon * G - \eta$. To ensure that the covariance matrix we modelled from noise on one of the components can accurately represent the overall effective noise, we compared the covariance matrix of 2000 random time-series $\epsilon * G - \eta$ (Fig. 6b) against our parametrization type 3 covariance matrix (Fig. 6a). One of these time-series can be seen in Fig. 6(c). We generated the random noise time-series η and ϵ from the parametrized covariance matrix by multiplying the Cholesky factor of the matrix with samples drawn at random from a standard normal distribution (Seydel 2009, p. 91). For G, we used a receiver function generated from our synthetic seismograms with no noise added. We find that because the receiver function has relatively low amplitudes and is convolved with the noise on the parent component, the overall contribution of the noise on the parent component is only a fraction of the contribution of the noise on the daughter component. This is why the covariance matrices in Fig. 6 are so similar, and justifies our approach of explicitly accounting only for noise on the daughter component.

There are a number of model selection methods that are often used to identify preferred MCMC models, including the Akaike information criterion (Akaike 1974), the Bayesian information criterion (Schwarz 1978) and the deviance information criterion (Spiegelhalter *et al.* 2002). These methods reward goodness-of-fit, through the likelihood function, and penalize the number of parameters in a model. Therefore, in order to evaluate whether noise parametrization 3 that we present in this paper is indeed preferable over existing noise parametrizations, we compared the likelihoods and average number of Gaussians for parametrizations 1 and 3 (parametrization 2 is ill-conditioned and cannot be updated easily) when applied to synthetics with real noise added. We randomly selected 100-s segments of horizontal and vertical vectors of pre-event

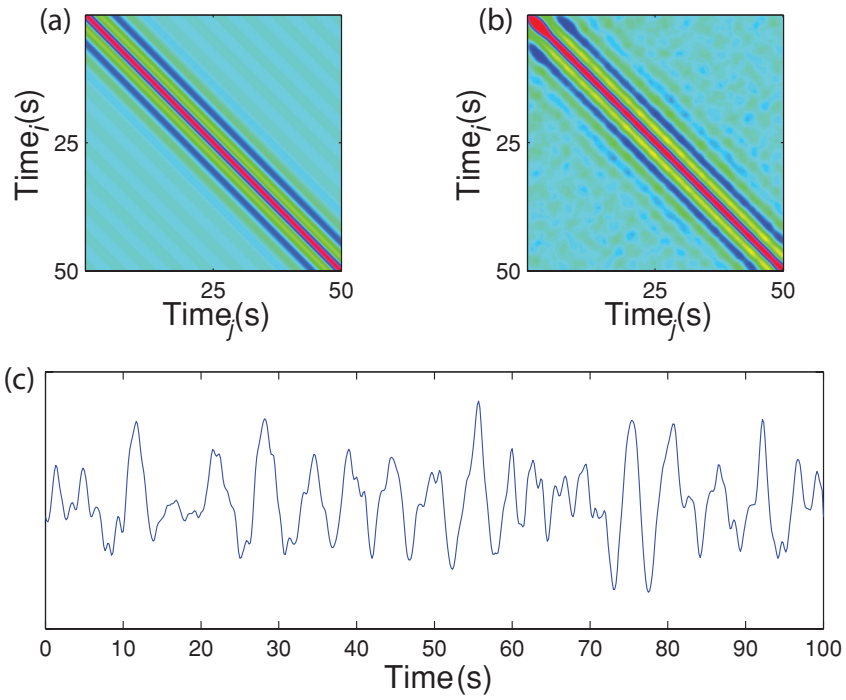

Figure 6. Test on overall noise. (a) The parametrization 3 covariance matrix. (b) The covariance matrix created from 2000 random samples of $\epsilon * G - \eta$, where ϵ is noise on the parent and η is noise on the daughter. (c) One of those random samples. The similarity of (a) and (b) justifies the use of parametrization 3 in modelling noise for the deconvolution problem.

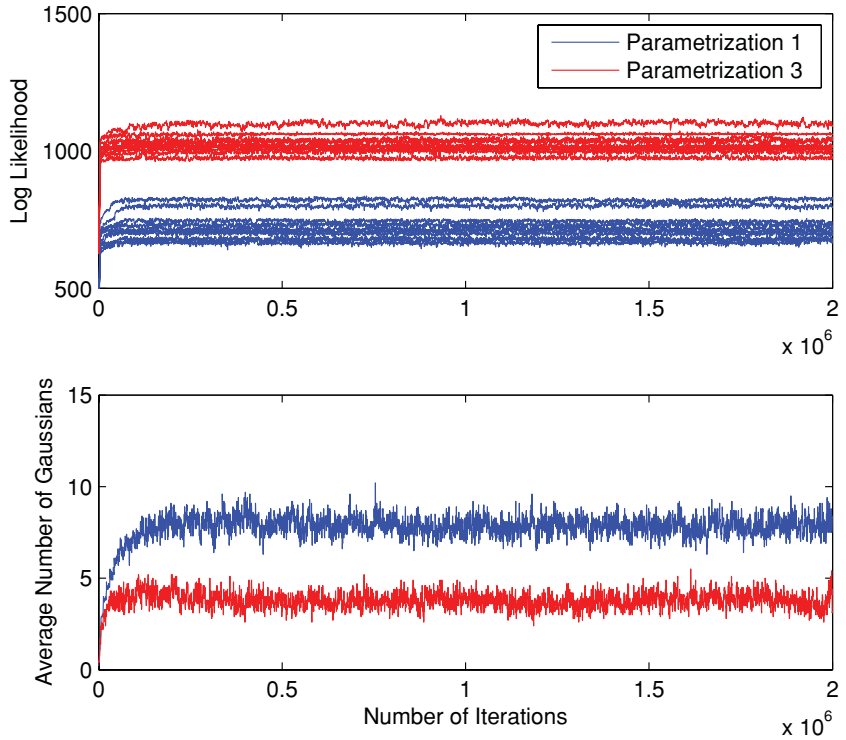

Figure 7. Log-likelihood and average number of Gaussians when using parametrization 1 (blue) versus parametrization 3 (red). Ten randomly selected noise pairs were added to a synthetic *Sp* waveform pair and THBD was performed on these waveform pairs using either parametrization 1 or parametrization 3. Parametrization 3 was not only able to produce models with a higher likelihood, but also was able to do so using fewer Gaussians.

noise at the Albuquerque, New Mexico station ANMO of the Global Seismic Network, and added them to synthetic *Sp* waveform pairs. These noisy synthetics were then deconvolved using THBD implemented with parametrization 1 in one case and parametrization 3 in the other. This process was repeated 10 times (using different randomly selected noise vectors each time). In Fig. 7, we plot the likelihoods and average number of Gaussians, as a function of iteration, for the two different noise parametrizations. We find

that parametrization 3 results in higher likelihoods and uses fewer Gaussians. Based on this observation, we conclude that regardless of which model selection method is used, parametrization 3 is a better model for our noisy data.

4 RESULTS

In order to test our algorithm, we created synthetic *Ps* and *Sp* seismograms [rotated into the *P-SV* system using a free-surface transform matrix (Kennett 1991)] from a model with three velocity layers—crust, mantle lithosphere and asthenosphere—using a propagator-matrix approach (Keith & Crampin 1977). The waveforms correspond to ray parameters of 0.0600 and 0.1129 s km^{-1}, respectively. We generated random noise time-series realizations from the covariance matrix of actual pre-event noise from ANMO as described in Section 3. We then added a different randomly generated noise time-series (of the same magnitude) to the parent and daughter components, and ran the deconvolution algorithm for 2×10^6 iterations, saving every 500th model after a burn-in period of 10^6 iterations; we tested running THBD for 10^7 iterations but this did not yield statistically different ensembles than THBD with 2×10^6 iterations. We fixed ω_0 at 4.4, the value which gave us the best fit to the covariance matrix obtained from real noise (see Fig. 5). In order to explore a range of noise levels that might be encountered with real data, we calculated this deconvolution for 32 increasing levels of noise, that is, increasing values of σ in eq. (2). For comparison, we also estimated the receiver functions using damped spectral division, in which the amplitude was normalized by calculating the deconvolution of the parent from itself using the same damping and then dividing the receiver function by the maximum value of that result. The results of the THBD and damped spectral division deconvolutions can be seen in Fig. 8.

The receiver functions generated from damped spectral division for *Ps* seismograms are shown in the left-hand panel of Fig. 8 and the receiver functions generated from our THBD method are shown in the right-hand panel. The amplitude of noise added to the seismograms increases from bottom to top with increasing waveform pair number. In the central panel are the parent and daughter waveform pairs corresponding to noise levels 4, 17 and 30. It can be seen that by adding noise generated from a covariance matrix created using real data to a synthetic seismogram one can get realistic-looking seismograms. The noise levels in the noisier seismograms are quite high, especially relative to the signal in the daughter components, and also overlap with signal in frequency content. This overlap can destroy information carried by the signal resulting in an inaccurate receiver function regardless of the method used. Good examples of this interference by noise are in waveform pairs 27, 28 and 29 where in both deconvolutions there are similar artefacts between 20 and 25 s. On both the damped spectral division and THBD receiver functions, the Moho conversion and the *PpPs* and *PpSs+PsPs* crustal multiples can be seen at approximately 4, 14 and 18 s, respectively. The LAB conversion at 9 s is harder to see in the receiver functions, especially when they are viewed individually. An advantage of our THBD method compared to damped spectral division is that there are many fewer artefacts which, without stacking, would be difficult to unambiguously interpret. Also, note that the true peaks are more compact in the THBD receiver functions—and closer to the delta functions expected in a receiver function given the layered model in which our synthetics are calculated—than in the damped spectral division receiver functions where they have longer period appearances as a result of the damping needed to stabilize the deconvolution.

The *Sp* receiver functions and seismograms in Fig. 9 are analogous to their *Ps* counterparts in Fig. 8. The receiver functions look simpler since they do not contain multiples, which arrive after the

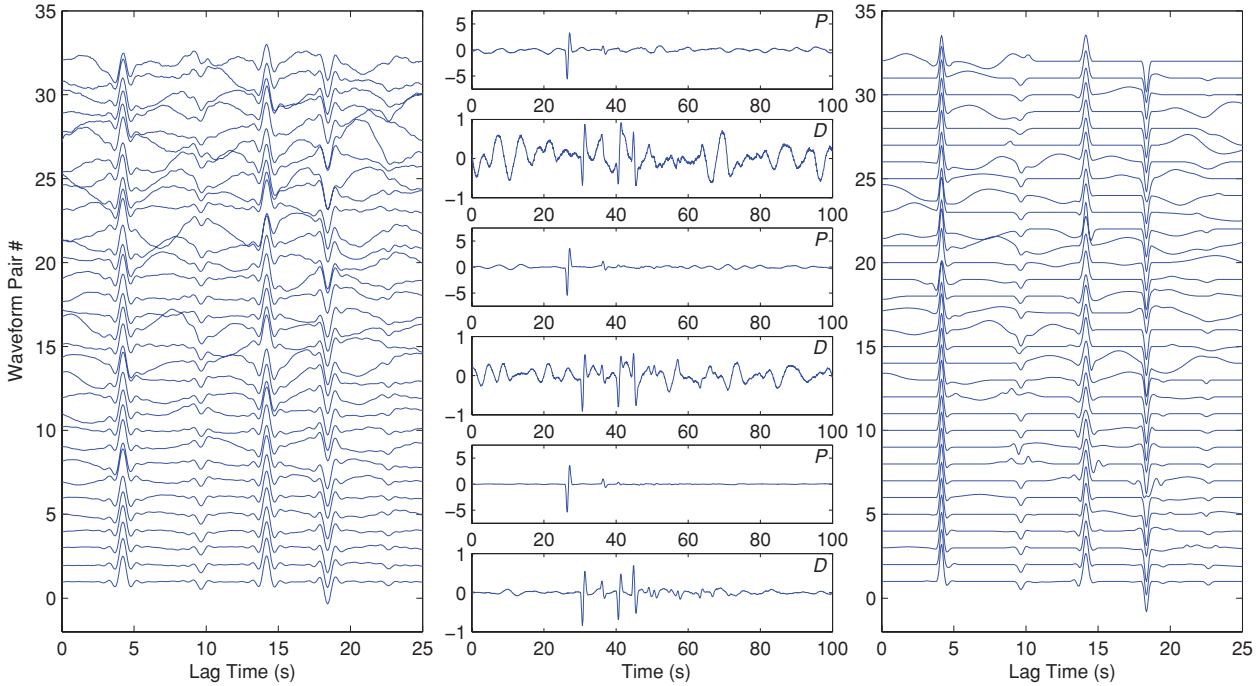

Figure 8. *Ps* waveforms and deconvolutions. Left-hand panel: the damped spectral division receiver functions for synthetic *Ps* seismograms with 32 increasing noise levels. Right-hand panel: the THBD receiver functions for the same seismograms. Central panels: the parent and daughter waveforms for waveform pairs 4, 17 and 30; note the increasing level of noise especially on the daughter component. The THBD method produces fewer artefacts than damped spectral division.

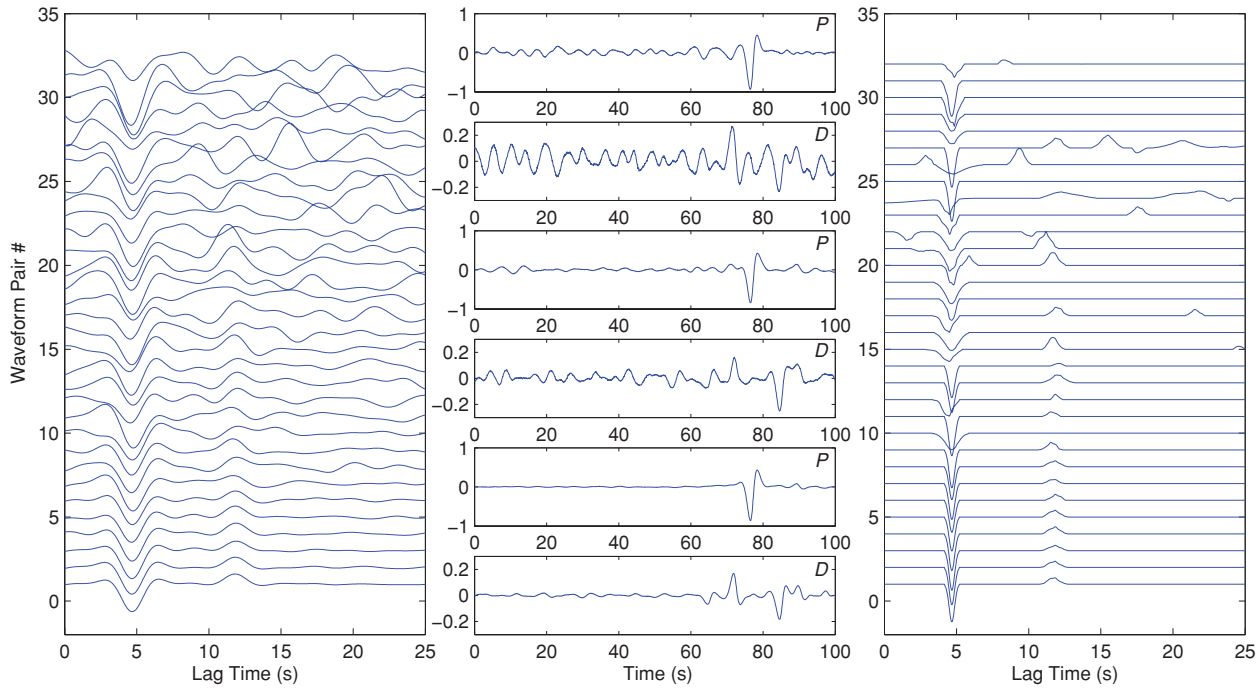

Figure 9. *Sp* waveforms and deconvolutions. Left-hand panel: the damped spectral division receiver functions for synthetic *Sp* seismograms with 32 increasing noise levels. Right-hand panel: the THBD receiver functions for the same seismograms. Central panels: the parent and daughter waveforms for waveform pairs 4, 17 and 30. At high noise levels, even the THBD method fails at retrieving the LAB phase, and information from multiple waveform pairs must be combined to retrieve it (see Fig. 11).

S. The absence of overprinting of direct conversions by multiples is a major advantage of the *Sp* method over the *Ps* one when investigating deep structure. In both the THBD method and damped spectral division, the Moho is visible at approximately 5 s, while the LAB at approximately 12 s is less reliably retrieved; at the highest noise levels, the LAB phase is indistinguishable from noise in the damped spectral division results and absent in the THBD results. A major difference, once again, is that there are many more artefacts in the damped spectral division receiver functions, which can only be diminished by averaging a substantial number of waveforms.

Some useful information contained in the ensemble solution produced by our THBD method is illustrated in Fig. 10, in which we analyse the results of the deconvolution for the *Sp* waveform pair 6 (Fig. 10a and see Fig. 9). Because the Metropolis algorithm yields an ensemble of models, and the probability that a parameter has a particular value is proportional to the number of models that have that value, we can plot the distributions of the parameter values across the ensemble models to visualize uncertainty on individual parameters (Figs 10b–d and f–i). The true (input) values of the receiver function parameters are denoted with red lines.

The histogram of the number of models with a Gaussian at a given lag time is plotted in Fig. 10(b), indicating the probability that the receiver function has a Gaussian at a particular lag time. The histogram shows that the majority of the models have two Gaussians (also shown in Fig. 10h), one at ∼5 s and the other at ∼12 s. The reason that the probability peak at 12 s has a lower amplitude than the 5-s peak, even though the majority of models have both Gaussians, is that there is less certainty on the location of the second Gaussian, broadening the peak and reducing its amplitude.

We can also quantify how well the amplitudes (Figs 10c and d) and widths (Figs 10f and g) of these two Gaussians are constrained by the waveform pair. In receiver function deconvolution, amplitudes and widths of the Gaussians are correlated so that the integral of a peak is proportional to the impedance contrast producing the *Ps* (*Sp*) conversion(s). The synthetic waveforms we used were generated through a model with instantaneous impedance contrasts that would manifest as instantaneous spikes in a receiver function estimated from truly broad-band data. Because we are using bandlimited signals, the spike widens. The 'true' values of amplitudes (red vertical lines) correspond to those associated with Gaussians of minimum allowable width. We can see that the Gaussian at 5 s has bimodal amplitude (Fig. 10c) and width (Fig. 10f) distributions. Though not bimodal, the distributions of amplitude (Fig. 10d) and width (Fig. 10g) of the Gaussian at 12 s is not normally distributed. This implies that uncertainty analyses that assume normally distributed errors on the receiver function parameters are inappropriate even in this relatively low-noise situation.

In Fig. 10(e), we present a visualization of the ensemble solution where all the models are binned by time and amplitude and the bins with more models in them are shaded brighter. The bimodal amplitude distribution for the first Gaussian in the ensemble solution can be seen by the dark spot (low probability) at about −0.3. The fuzziness in the second peak in the ensemble solution is an indication that there is some uncertainty in its location, width and amplitude.

The ensemble solution also enables us to quantify uncertainty on the recovery of the noise hyperparameters. For example, Fig. 10(i) shows the distribution of the noise magnitude (σ), which is approximately normally distributed around the value of 0.01. While the noise magnitude is estimated properly for lower noise levels, it should be noted that as noise levels increase, some of the noise may be interpreted as signal by the algorithm (especially if it has similar frequency content), reducing the estimated noise magnitude and narrowing the posterior distributions around incorrect parameter values (see Fig. 13 and associated discussion).

Real data sets often contain a number of noisy waveforms from which information is to be retrieved. In order to model this

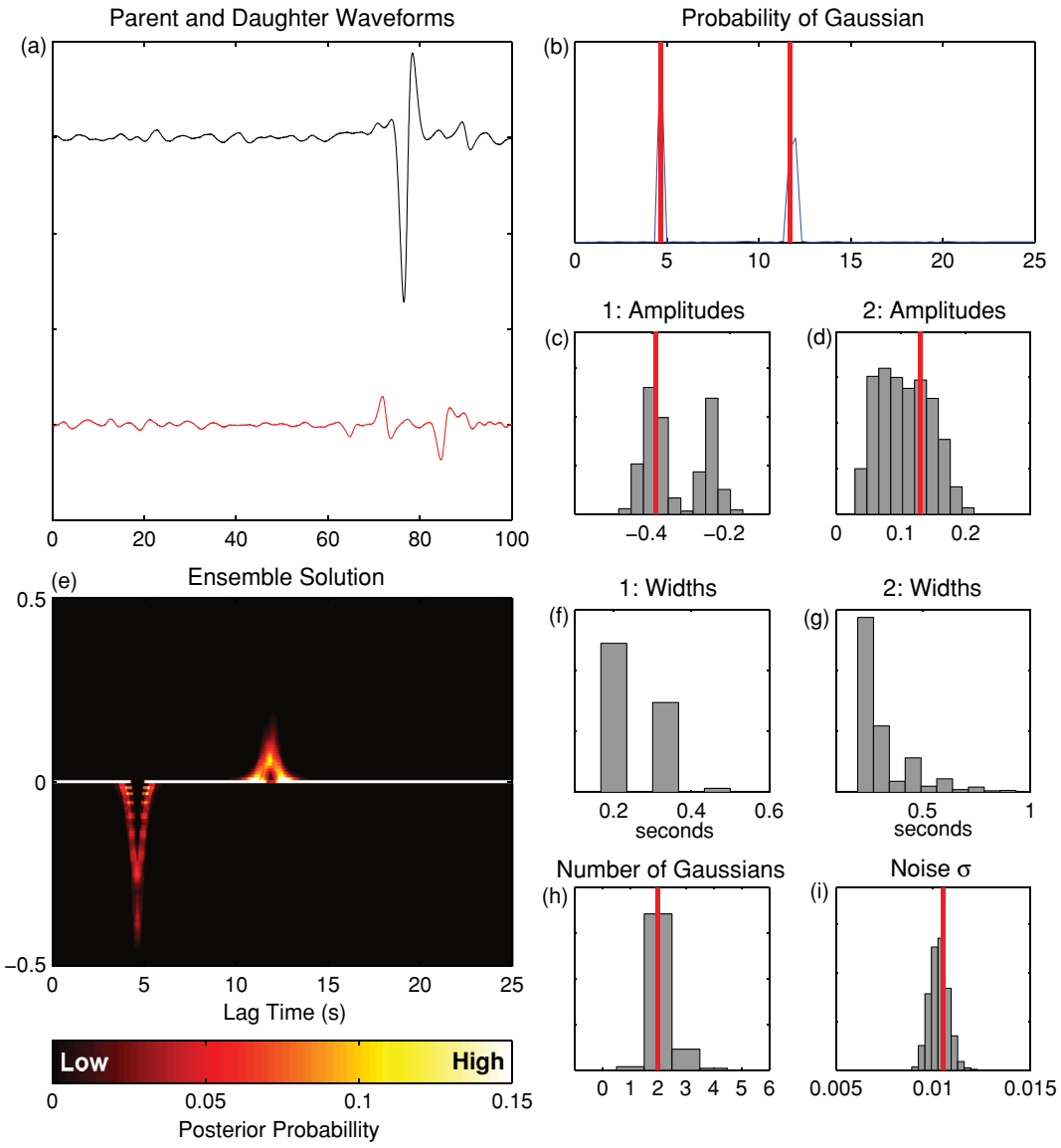

Figure 10. Analysis of a single deconvolution result. (a) Parent and daughter waveforms from *Sp* waveform pair 6. (b) The probability of a Gaussian being at a given time in the receiver function. (c and d) Histograms of the amplitudes of the two most common Gaussians (at ∼5 and 12 s). (e) Ensemble solution combining all of the models by binning them based on time and amplitude. (f and g) Histograms of the widths of the two most common Gaussians (at ∼5 and 12 s). (h) Histogram of the number of Gaussians. (i) Retrieved noise magnitudes (σ). True values are marked in red. For the amplitudes, the amplitude corresponding to the minimum allowable width is taken as the 'true' amplitude.

situation, we added high levels of noise to 20 *Sp* waveform pairs and deconvolved them using three methods. First, we used our THBD method to get 20 receiver function ensemble solutions and then averaged across them to obtain an overall receiver function (Fig. 11a). Next, we deconvolved the waveform pairs using damped spectral division following two approaches: (1) deconvolving the waveform pairs (damping each deconvolution individually) and averaging the results in the time domain (Fig. 11b), and (2) stacking the waveforms in the frequency domain with a single damping parameter and simultaneously deconvolving them (Fig. 11c). This figure shows that by combining the information from multiple noisy waveform pairs, we can retrieve a receiver function that contains Gaussians associated with both the Moho (red) and LAB (blue) conversions. Unlike the receiver functions obtained by the two implementations of damped spectral division (Figs 11b and c), the receiver function obtained with the THBD method does not show any artefacts that might mistakenly be interpreted as geological structures (impedance contrasts). This result demonstrates that our method is capable of obtaining easily interpretable receiver functions even from waveforms contaminated by high levels of noise, though doing so requires using information from multiple pairs of waveforms (see Fig. 10 for results obtained from only a single waveform pair).

The tests in Fig. 11 naturally lead to the idea of stacking observed waveforms of a given slowness and backazimuth and then using the stacked waveforms to perform a single deconvolution with the THBD method. This would be advantageous because it would reduce the computation time of the Markov chain, which is our method's main disadvantage. The problem with this idea, however, is that the noise characteristics—and, therefore, the hyperparameters describing the noise amplitude and frequency content—are unlikely to be identical across different pairs of waveforms, so that a stack of waveforms cannot be represented with a single set of

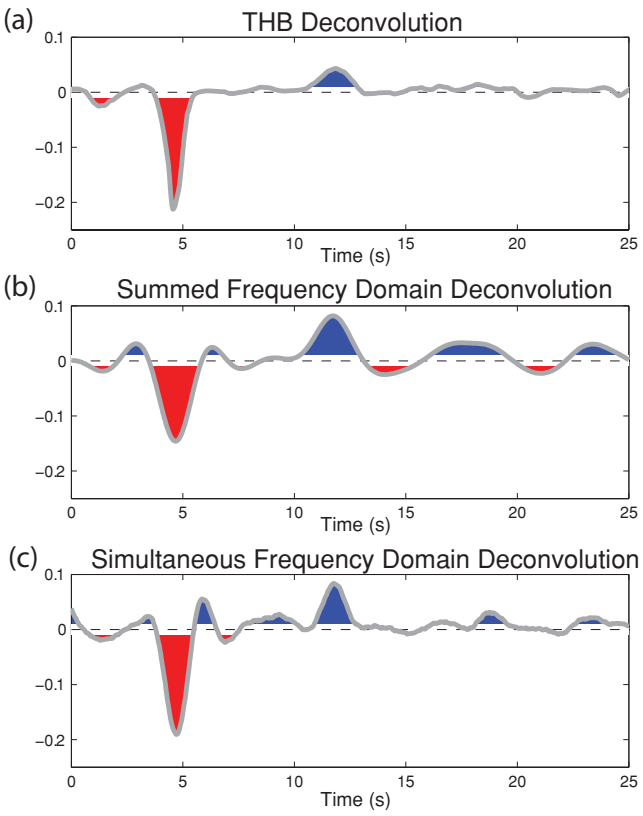

Figure 11. *Sp* receiver functions from 20 waveform pairs contaminated with high levels of noise. (a) Average receiver function generated using our THBD algorithm. (b) Receiver function generated by averaging individual frequency-domain receiver functions in the time domain. (c) Receiver function generated by simultaneous deconvolution of the waveforms in the frequency domain. Note the absence of artefacts in the THBD result, which yields more easily interpretable receiver functions.

noise hyperparameters. Instead, our preferred approach to stacking redundant waveforms is to first carry out the deconvolutions, determine their individual noise hyperparameters and then stack the ensemble solutions.

Another factor that often complicates receiver function analyses is the presence of spectral holes in the seismic waveforms, which can lead to instability in the deconvolution. In order to test how THBD copes with this effect, we convolved parent and daughter *Ps* waveforms with triangular source time functions of 5-s duration (whose Fourier transform is the square of the sinc function) before adding noise and performing deconvolution. The results of our deconvolution using three different methods—iterative time-domain, damped spectral division and THBD—are shown in Fig. 12. Even a cursory comparison of the receiver functions obtained from source time functions without (left) and with (right) spectral holes shows that both THBD and the traditional iterative time-domain deconvolution dramatically outperform damped spectral division in the presence of spectral holes added by the triangular source time function.

Having explored the ability of the THBD to obtain interpretable receiver functions, we proceed to quantify how well our method retrieves noise characteristics, given input noise spanning a range of amplitudes and four different frequency contents. To do this, we created random time-series of type 3 noise with ω_0 values of 2.2, 4.4 and 8.8, and noise levels (σ) from 0 to 0.18. An ω_0 value of 4.4 closely models real noise and is shown in Figs 4(c) and 5. An ω_0 value of 2.2 generates lower frequency noise and an ω_0 value of 8.8 generates higher frequency noise. We added realizations of each type of random noise to parent and daughter waveforms, and applied the deconvolution algorithm, with the ω_0 value fixed to that used in creating the added noise, in order to estimate the noise level and frequency content. We plot the input levels of noise (measured after bandpassing) against the corresponding retrieved levels of noise in Fig. 13, along with their best-fitting lines. We find that adding correlated noise causes the noise level to be systematically underestimated, that is, the slopes of the best-fitting lines in Fig. 13 are always less than 1. Yet, higher levels of noise are estimated for noisier records, making the retrieved noise level (σ) a useful parameter for comparing noisiness of different waveform pairs. This underestimation occurs because the misfit between the observed and predicted waveforms, taking into account noise, is

$$(P + \epsilon) * G - (P * G_0 + \eta). \tag{14}$$

If $G = G_0 + \delta$, where G_0 is the true model and δ is the difference between the true model and the predicted model due to the noise, then the misfit becomes

$$(P + \epsilon) * \delta + \epsilon * G_0 - \eta. \tag{15}$$

When the noise has a similar frequency content to that of the parent signal, a δ can be introduced which will reduce the misfit and thus also lead to a lower noise estimation. In fitting the noise, the δ will also lead to a greater difference between G and G_0, resulting in a less accurate deconvolution. To verify that this behaviour is giving rise to the results shown in Fig. 13, we repeat the analysis using white noise, which is uncorrelated by construction. Following the previous procedure, we apply the THBD algorithm to obtain noise-level estimates, but this time using parent and daughter waveforms contaminated by white noise of varying levels. We modify the algorithm to use a scalar matrix as the data covariance matrix C_D, instead of one based on the three noise parametrizations. In Fig. 13, we plot the retrieved noise levels against input levels, and find that the method obtains accurate estimates of the noise level. This is consistent with our explanation for why the method systematically underpredicts levels of correlated noise, since expression (15) is unlikely to be decreased by the introduction of δ when noise time-series ϵ and η are uncorrelated.

Finally, we turn our attention to quantifying our ability to estimate the correct value of the noise parameter λ using the THBD method. We do this by adding random realizations of noise time-series—once again generated from the covariance matrix of actual pre-event noise at ANMO (shown in Fig. 4d)—to 10 synthetic *Ps* waveform pairs. The amplitude of noise (σ) is kept the same for all noise realizations. We run the THBD algorithm on each noisy waveform pair, initializing each with a different starting λ value ranging from 0.05 to 0.5. In this run of the algorithm, we kept the probability of option 6 (changing the noise correlation parameter) being chosen equal to that of the other options. Doing so increases the computation time, but speeds up the convergence to an estimate of noise parameter λ, which may be necessary when the frequency content of pre-event noise is poorly known *a priori*. The results from this test are shown in Fig. 14. Regardless of their starting λ value, all the chains converge to a value of 0.2, which we previously found to best fit the covariance matrix at ANMO (see Fig. 6) that was used in generating the noise time-series for this test. This shows that our method can successfully estimate the frequency content of the noise, parametrized by λ, even from a grossly inaccurate starting guess.

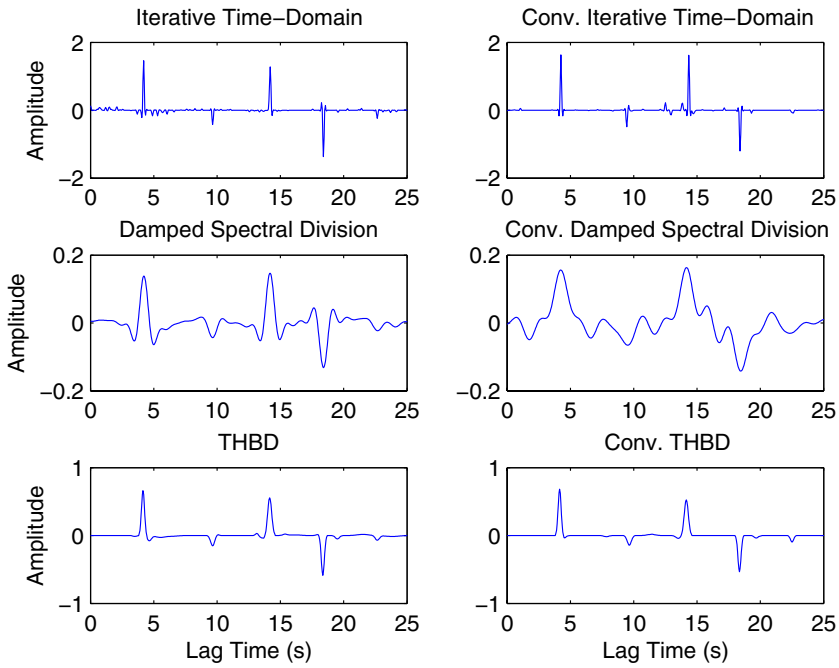

Figure 12. Spectral hole deconvolution test. The panels on the left-hand side show the results of three deconvolution methods on a waveform pair with a slight amount ($\sigma = 0.1$) of noise added. The panels on the right-hand side show the results of the same deconvolution test, but when the waveform pairs have been convolved with a 5-s triangular source time function before the addition of noise. The spectral holes (and destruction of high frequencies) introduced by this convolution have a larger effect on the damped spectral division than on the two time-domain methods (iterative time-domain and THBD).

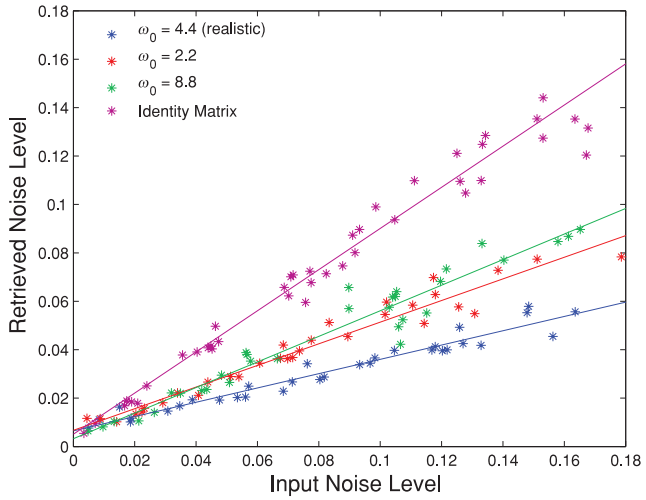

Figure 13. Noise retrievals. Noise levels (σ) were estimated by the THBD algorithm for varying input levels of noise given a realistic parametrization of noise frequency ($\omega_0 = 4.4$), lower frequency noise ($\omega_0 = 2.2$), higher frequency noise ($\omega_0 = 8.8$) and white noise (an identity covariance matrix times a scalar). When the noise has a similar frequency content to that of the signal, the noise level is systematically underestimated. Nevertheless, at a given value of ω_0, higher levels of noise are estimated for noisier records, making the retrieved noise level (σ) a useful parameter for comparing noisiness of different waveform pairs.

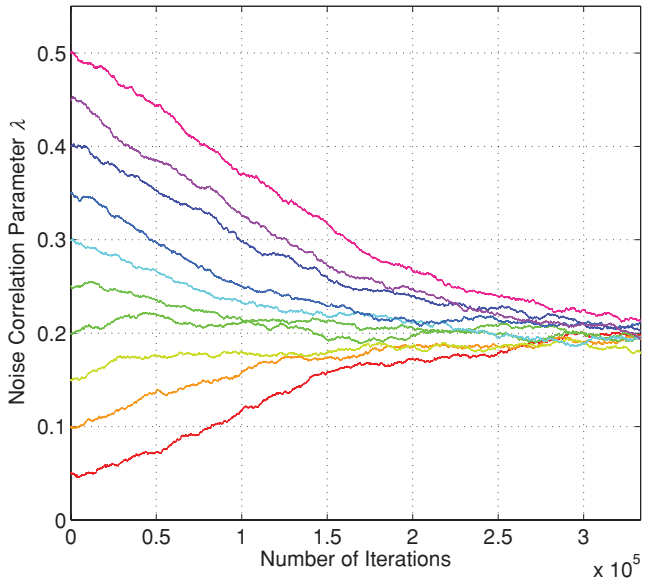

Figure 14. Noise correlation parameter λ by iteration. Markov chains with different starting values of λ are shown and all converge to ∼0.2, the same λ value we found best fit the covariance matrix that was used to generate this noise. This illustrates that our method can successfully estimate the frequency content of the noise, parametrized by λ, even from inaccurate starting guesses.

5 CONCLUSION

We have developed and validated a method for deconvolution which retrieves noise levels and frequency character in addition to obtaining receiver functions with fewer artefacts even when the parent–daughter waveform pairs are contaminated by high levels of noise. This method of deconvolution is especially useful in receiver function analyses for noisy stations (e.g. ocean bottom, coastal, polar deployments) because of its treatment of noise level and frequency as hyperparameters. These noise hyperparameters affect the complexity of receiver functions, preventing the overfitting of noisy data, while simultaneously yielding an ensemble solution which fully quantifies associated uncertainties. The receiver functions ensemble solutions obtained by our THBD method can then be used

within a Bayesian framework (e.g. Piana Agostinetti & Malinverno 2010; Bodin *et al.* 2012), or a non-Bayesian framework, for example, a multistep modelling procedure (e.g. Tkalčić *et al.* 2012), and combined with constraints from seismic tomography. In this way, by treating our knowledge of the structure obtained by THBD as a random variable, uncertainties can be propagated from the receiver function to the final model of subsurface structure.

The THBD method's increased robustness in comparison to more traditional damped spectral division allows it to better constrain receiver functions from individual parent–daughter waveform pairs, which can improve structural inferences beneath stations with smaller numbers of recorded waveforms. This makes it particularly useful for constraining anisotropy beneath stations where there are only a small number of waveforms with a high signal-to-noise ratio available at a given backazimuth. Quantifying the reliability of features in receiver functions, which the ensemble solution produced by our method makes possible, is important to ensure that only those features that are robustly constrained are interpreted in geological analyses. In other words, the method can be used to ensure that features in the Earth whose existence is not required by the data are not mistakenly identified.

The THBD method's biggest drawback is that it is orders of magnitude slower than other commonly used methods of deconvolution. When run on an Intel i5-4570 3.2 GHz processor, the computation time for iterative time-domain deconvolution was $\sim 7 \times 10^3$ times longer than damped spectral division, while the THBD method was $\sim 1.2 \times 10^8$ times longer than damped spectral division. Of course, this comparison is unfavourable to THBD since it does not take into account the fact that THBD yields fully quantified uncertainties, which neither the iterative time-domain deconvolution nor damped spectral division do. Nevertheless, its relatively high computation times make THBD much better suited to situations where more robust receiver function is desired from a smaller data set. It should also be noted that by starting with a better initial model the burn-in time can be reduced, and that by further optimizing the code or coding and compiling the algorithm outside MATLAB, the time spent on each iteration can be reduced.

While we applied the THBD method to obtaining receiver functions, it can also be used for other deconvolution problems where the function being estimated can be parametrized by a series of Gaussians. In addition, if the features of the function being estimated via deconvolution are not suitably represented by a series of Gaussians, the method could be modified to use a parametrization that is better suited to the expected deconvolution result. For example, straightforward extensions of the method can be used to improve deconvolution in studies of *PP* and *SS* precursors (Schmerr & Garnero 2006), or for deconvolving geological structure out of signal in order to study source complexity.

ACKNOWLEDGEMENTS

The authors thank Hrvoje Tkalčić and an anonymous reviewer for constructive criticism that substantially improved this manuscript.

REFERENCES

Ababou, R., Bagtzoglou, A.C. & Wood, E.F., 1994. On the condition number of covariance matrices in kriging, estimation, and simulation of random fields, *Math. Geol.*, **26**(1), 99–133.

Abt, D., Fischer, K., French, S., Ford, H., Yuan, H. & Romanowicz, B., 2010. North American lithospheric discontinuity structure imaged by *Ps* and *Sp* receiver functions, *J. geophys. Res.*, **115,** B09301, doi:10.1029/2009JB006914.

Akaike, H., 1974. A new look at the statistical model identification, *IEEE Trans. Autom. Control*, **19**(6), 716–723.

Ammon, C.J., 1991. The isolation of receiver effects from teleseismic *P* waveforms, *Bull. seism. Soc. Am.*, **81**(6), 2504–2510.

Ammon, C.J., Randall, G.E. & Zandt, G., 1990. On the nonuniqueness of receiver function inversions, *J. geophys. Res.*, **95**(B10), 15 303–15 318.

Andrieu, C., Barat, É. & Doucet, A., 2001. Bayesian deconvolution of noisy filtered point processes, *IEEE Trans. Signal Process.*, **49**(1), 134–146.

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K. & Rawlinson, N., 2012. Transdimensional inversion of receiver functions and surface wave dispersion, *J. geophys. Res.*, **117,** doi:10.1029/2011JB008560.

Bostock, M., 1998. Mantle stratigraphy and evolution of the Slave province, *J. geophys. Res.*, **103**(B9), 21 183–21 200.

Bromirski, P.D., 2009. Earth vibrations, *Science*, **324**(5930), 1026–1027.

Burdick, L.J. & Langston, C.A., 1977. Modeling crustal structure through the use of converted phases in teleseismic body-wave forms, *Bull. seism. Soc. Am.*, **67**(3), 677–691.

Clayton, R.W. & Wiggins, R.A., 1976. Source shape estimation and deconvolution of teleseismic bodywaves, *Geophys. J. Int.*, **47**(1), 151–177.

Cowles, M.K. & Carlin, B.P., 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review, *J. Am. Stat. Assoc.*, **91**(434), 883–904.

Denison, D., Adams, N., Holmes, C. & Hand, D., 2002. Bayesian partition modelling, *Comput. Stat. Data Anal.*, **38**(4), 475–485.

Dettmer, J., Molnar, S., Steininger, G., Dosso, S.E. & Cassidy, J.F., 2012. Trans-dimensional inversion of microtremor array dispersion data with hierarchical autoregressive error models, *Geophys. J. Int.*, **188**(2), 719–734.

Dosso, S.E., Nielsen, P.L. & Wilmut, M.J., 2006. Data error covariance in matched-field geoacoustic inversion, *J. acoust. Soc. Am.*, **119,** 208–219.

Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., 1996. *Markov Chain Monte Carlo in Practice*, Vol. **2**, CRC Press.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**(4), 711–732.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**(1), 97–109.

Helffrich, G., 2006. Extended-time multitaper frequency domain cross-correlation receiver-function estimation, *Bull. seism. Soc. Am.*, **96**(1), 344–347.

Holland, C.W., Dettmer, J. & Dosso, S.E., 2005. Remote sensing of sediment density and velocity gradients in the transition layer, *J. acoust. Soc. Am.*, **118,** 163–177.

Jeffreys, H., 1939. *Theory of Probability*, Clarendon Press.

Kang, D. & Verotta, D., 2007. Reversible jump Markov chain Monte Carlo for deconvolution, *J. Pharmacokinet. Pharmacodyn.*, **34**(3), 263–287.

Keith, C.M. & Crampin, S., 1977. Seismic body waves in anisotropic media: synthetic seismograms, *Geophys. J. R. astr. Soc.*, **49**(1), 225–243.

Kennett, B., 1991. The removal of free surface interactions from three-component seismograms, *Geophys. J. Int.*, **104**(1), 153–163.

Kikuchi, M. & Kanamori, H., 1982. Inversion of complex body waves, *Bull. seism. Soc. Am.*, **72**(2), 491–506.

Langston, C.A., 1977. Corvallis, Oregon, crustal and upper mantle receiver structure from teleseismic *P* and *S* waves, *Bull. seism. Soc. Am.*, **67**(3), 713–724.

Langston, C.A., 1979. Structure under Mount Rainier, Washington, inferred from teleseismic body waves, *J. geophys. Res.*, **84**(B9), 4749–4762.

Lavielle, M., 1991. 2-D Bayesian deconvolution, *Geophysics*, **56**(12), 2008–2018.

Lekic, V., French, S.W. & Fischer, K.M., 2011. Lithospheric thinning beneath rifted regions of Southern California, *Science*, **334**(6057), 783–787.

Levander, A. & Miller, M.S., 2012. Evolutionary aspects of lithosphere discontinuity structure in the Western US, *Geochem. Geophys. Geosyst.*, **13**(7), doi:10.1029/2012GC004056.

Ligorría, J.P. & Ammon, C.J., 1999. Iterative deconvolution and receiver-function estimation, *Bull. seism. Soc. Am.*, **89**(5), 1395–1400.

Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.*, **151**(3), 675–688.
Malinverno, A. & Briggs, V.A., 2004. Expanded uncertainty quantification in inverse problems: hierarchical Bayes and empirical Bayes, *Geophysics*, **69**(4), 1005–1016.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E., 1953. Equation of state calculations by fast computing machines, *J. Chem. Phys.*, **21,** 1087–1092.
Park, J. & Levin, V., 2000. Receiver functions from multiple-taper spectral correlation estimates, *Bull. seism. Soc. Am.*, **90**(6), 1507–1520.
Peterson, J., 1993. *Observations and modeling of seismic background noise. Tech. rep.*, US Geological Survey, Albuquerque, New Mexico.
Piana Agostinetti, N. & Malinverno, A., 2010. Receiver function inversion by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.*, **181**(2), 858–872.
Richardson, S. & Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion), *J. R. Stat. Soc. B*, **59**(4), 731–792.
Rondenay, S., 2009. Upper mantle imaging with array recordings of converted and scattered teleseismic waves, *Surv. Geophys.*, **30**(4–5), 377–405.
Sambridge, M., Bodin, T., Gallagher, K. & Tkalčić, H., 2013. Transdimensional inference in the geosciences, *Phil. Trans. R. Soc. A*, **371**(20110547), doi:10.1098/rsta.2011.0547.
Schaeffer, A. & Bostock, M., 2010. A low-velocity zone atop the transition zone in northwestern Canada, *J. geophys. Res.*, **115**(B6), B06302, doi:10.1029/2009JB006856.
Schmerr, N. & Garnero, E., 2006. Investigation of upper mantle discontinuity structure beneath the central Pacific using *SS* precursors, *J. geophys. Res.*, **111**(B8), B08305, doi:10.1029/2005JB004197.
Schwarz, G., 1978. Estimating the dimension of a model, *Ann. Stat.*, **6**(2), 461–464.
Seydel, R., 2009. *Tools for Computational Finance*, Universitext (En ligne), Springer-Verlag Berlin Heidelberg.
Shibutani, T., Ueno, T. & Hirahara, K., 2008. Improvement in the extended-time multitaper receiver function estimation technique, *Bull. seism. Soc. Am.*, **98**(2), 812–816.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Van Der Linde, A., 2002. Bayesian measures of model complexity and fit, *J. R. Stat. Soc. B*, **64**(4), 583–639.
Stawinski, G., Doucet, A. & Duvaut, P., 1998. Reversible jump Markov chain Monte Carlo for Bayesian deconvolution of point sources, in *Proceedings of SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, International Society for Optics and Photonics, pp. 179–190.
Tarantola, A. & Valette, B., 1982. Inverse problems = quest for information, *J. geophys.*, **50**(3), 150–170.
Tkalčić, H., Rawlinson, N., Arroucau, P., Kumar, A. & Kennett, B.L., 2012. Multistep modelling of receiver-based seismic and ambient noise data from WOMBAT array: crustal structure beneath southeast Australia, *Geophys. J. Int.*, **189**(3), 1680–1700.
Vinnik, L., 1977. Detection of waves converted from *P* to *SV* in the mantle, *Phys. Earth planet. Inter.*, **15**(1), 39–45.
Yildirim, S., Cemgil, A., Aktar, M., Ozakin, Y. & Ertuzun, A., 2010. A Bayesian deconvolution approach for receiver function analysis, *IEEE Trans. Geosci. Remote Sens.*, **48**(12), 4151–4163.

APPENDIX: THE PRIOR AND ACCEPTANCE PROBABILITIES

A1 The prior

The prior model probability distribution can be separated into three terms

$$P(\mathbf{m}) = P(\mathbf{t}, \mathbf{a}, \mathbf{w}|k)P(k)P(\mathbf{n}), \tag{A1}$$

where $P(\mathbf{t}, \mathbf{a}, \mathbf{w}|k)$ is the prior on the centre times $\mathbf{c}$, widths $\mathbf{w}$ and amplitudes $\mathbf{a}$ of the Gaussians that make up our receiver function, $P(k)$ is the prior on the number of Gaussians and $P(\mathbf{h})$ is the prior on the N_{hp} noise hyperparameters (given our parametrization, $\mathbf{h} = [\sigma, \lambda]$). To minimize the prior information imposed on the deconvolution, we choose uniform distributions for the amplitudes and widths of individual Gaussians (i.e. a_i, w_i) as well as the hyperparameters

$$P(a_i|k) = \begin{cases} 1/(\Delta a) & a_{\min} < a_i < a_{\max} \\ 0 & \text{otherwise,} \end{cases} \tag{A2}$$

$$P(w_i|k) = \begin{cases} 1/(\Delta w) & w_{\min} < w_i < w_{\max} \\ 0 & \text{otherwise,} \end{cases} \tag{A3}$$

$$P(h_i|k) = \begin{cases} 1/(\Delta^j h) & {}^j h_{\min} < h_j < {}^j h_{\max} \\ 0 & \text{otherwise.} \end{cases} \tag{A4}$$

We use $(k+1)^{-1}$ for the prior on the number of Gaussians in order to represent complete ignorance of a positive quantity following the reasoning in Jeffreys (1939) and in order to have a finite prior when there are no Gaussians:

$$P(k) = \begin{cases} 1/(k+1) & 0 \le k \le 30 \\ 0 & \text{otherwise.} \end{cases} \tag{A5}$$

Although this prior does not normalize to unity, this does not represent a problem because only ratios of likelihoods will be used rather than the actual likelihoods to estimate our posteriors.

The reason that we limit the number of Gaussians to 30 in our prior is that we do not want unnecessary complexity in our results when the noise level is low. When there are reasonable noise levels such as in real data, the number of Gaussians found by the code is reduced and this maximum number of Gaussians does not appear to affect the results.

Because the widths and amplitudes of the Gaussians are taken to be independent of one another, as are the values of the hyperparameters, the priors are given by

$$P(\mathbf{w}|k) = \prod_{i=1}^{k} P(w_i|k), \tag{A6}$$

$$P(\mathbf{a}|k) = \prod_{i=1}^{k} P(a_i|k), \tag{A7}$$

$$P(\mathbf{h}|k) = \prod_{i=j}^{N_{hp}} P(h_j|k). \tag{A8}$$

We set up the problem so that the centre times (c_i) of the Gaussians can only take on discrete values among the N_t at which the seismic time-series are sampled. However, since a new Gaussian of width w_{k+1} cannot be placed within $w_{k+1} + w$ of existing Gaussians with width w, the effective number of possible points is $N = N_t - N_b$, where N_b is the number of points blocked due to Gaussians already in place. Therefore, giving an equal probability to each possible configuration of Gaussian placements, the prior on $\mathbf{t}$ is

$$P(\mathbf{t}|k) = \frac{k!(N-k)!}{N!}. \tag{A9}$$

Combining these mutually independent terms, we have for the model prior probability distribution

$$P(\mathbf{m}) = \frac{k!(N-k)!}{\Delta k N! (\Delta a \Delta w)^k \prod_j^{N_{hp}} \Delta^j h}. \tag{A10}$$

It should be understood, that $P(\mathbf{m})$ is null when the prior limits on any of the parameters or hyperparameters are exceeded.

A2 Proposal distributions

At each step of the algorithm, we choose three of the following seven possibilities:

(1) Change the amplitude (a') of an existing Gaussian i whose amplitude is a. The new amplitude is chosen from the probability distribution

$$q_a(a_i'|a_i) = \frac{1}{\theta_a\sqrt{2\pi}}\exp\left\{-\frac{(a_i'-a_i)^2}{2\theta_a^2}\right\}, \tag{A11}$$

where θ_a^2 is the variance of the distribution.

(2) Change the width (w') of an existing Gaussian i whose width is w. The new width is chosen from the probability distribution

$$q_w(w_i'|w_i) = \frac{1}{\theta_w\sqrt{2\pi}}\exp\left\{-\frac{(w_i'-w_i)^2}{2\theta_w^2}\right\}, \tag{A12}$$

where θ_w^2 is the variance of the distribution.

(3) Change the location (c') of an existing Gaussian i whose current location is c. The new location is chosen from the probability distribution

$$q_c(c_i'|c_i) = \frac{1}{\theta_c\sqrt{2\pi}}\exp\left\{-\frac{(c_i'-c_i)^2}{2\theta_c^2}\right\}, \tag{A13}$$

where θ_c^2 is the variance of the distribution.

(4) Change the noise amplitude (σ') whose current value is σ. The new noise amplitude is chosen from the probability distribution

$$q_\sigma(\sigma'|\sigma) = \frac{1}{\theta_\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\sigma'-\sigma)^2}{2\theta_\sigma^2}\right\}, \tag{A14}$$

where θ_σ^2 is the variance of the distribution.

(5) Change the noise correlation timescale (λ') whose current value is λ. The new correlation timescale is chosen from the probability distribution

$$q_\lambda(\lambda'|\lambda) = \frac{1}{\theta_\lambda\sqrt{2\pi}}\exp\left\{-\frac{(\lambda'-\lambda)^2}{2\theta_\lambda^2}\right\}, \tag{A15}$$

where θ_λ^2 is the variance of the distribution.

(6) Create a new Gaussian, whose amplitude a_{k+1} and width w_{k+1} are drawn randomly from the prior distribution, and whose position is selected from the discrete set of possible centre times further than w_i from existing c_i. We find that this choice on minimum c_i spacing works well for our purposes, though it does introduce a slight dependence of c_{k+1} on existing c_i and w_i.

(7) Delete one Gaussian, with equal likelihood of picking any of the existing ones.

The values of θ_a, θ_w, θ_c, θ_σ and θ_λ are given in Table 1.

A3 The acceptance probability

In order for our method to converge to the transdimensional posterior distribution $P(\mathbf{m}|\mathbf{d}_{\rm obs})$, the probability of accepting a jump from model $\mathbf{m}$ to the proposed model $\mathbf{m}'$ has to be

$$\alpha(\mathbf{m}'|\mathbf{m}) = \min\left[1, \frac{P(\mathbf{d_{obs}}|\mathbf{m}')}{P(\mathbf{d_{obs}}|\mathbf{m})}\frac{P(\mathbf{m}')}{P(\mathbf{m})}\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}|\mathbf{J}|\right], \tag{A16}$$

where the first ratio gives voice to the data (through the Mahalanobis distance between prediction and observation), the second ratio gives voice to the prior information on the model distributions (see Appendix Section A1), the third ratio accounts for potential differences in probability of going from model $\mathbf{m}$ to model $\mathbf{m}'$ and reversing this jump, and the final term is the determinant of the Jacobian matrix for the transformation.

A3.1 Proposal ratios

An examination of the proposal probabilities associated with choices 2, 3, 4, 5 and 6—that is, all choices except those involving the birth and death of a Gaussian—involve symmetric distributions, so that

$$q_a(a_i'|a_i) = q_a(a_i|a_i'), \tag{A17}$$

$$q_w(w_i'|w_i) = q_w(w_i|w_i'), \tag{A18}$$

$$q_c(c_i'|c_i) = q_c(c_i|c_i'), \tag{A19}$$

$$q_{h_{\sigma,\lambda}}(h'_{\sigma,\lambda}|h_{\sigma,\lambda}) = q_{h_{\sigma,\lambda}}(h_{\sigma,\lambda}|h'_{\sigma,\lambda}). \tag{A20}$$

This means that the contributions of these terms to the ratio of $q(\mathbf{m}|\mathbf{m}')$ to $q(\mathbf{m}'|\mathbf{m})$ will cancel, being equal for the forward and reverse steps. The only non-trivial terms are, therefore, the ones associated with the birth and death of Gaussians.

Birth

When a Gaussian is born, the dimension of $\mathbf{w}'$, $\mathbf{a}'$ and $\mathbf{c}'$ increases from k to $k+1$. Since the choices of the width w_{k+1} and amplitude a_{k+1} of the new Gaussian are independent of one another and of the Gaussian location, the proposal ratio can be separated as

$$\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})} = \frac{q(\mathbf{c}|\mathbf{m}')}{q(\mathbf{c}'|\mathbf{m})}\frac{q(\mathbf{w}|\mathbf{m}')}{q(\mathbf{w}'|\mathbf{m})}\frac{q(\mathbf{a}|\mathbf{m}')}{q(\mathbf{a}'|\mathbf{m})}. \tag{A21}$$

When a new Gaussian is created, its amplitude a_{k+1} and width w_{k+1} are chosen from uniform distributions that do not change with current model location, and are therefore independent of the model $\mathbf{m}$. Therefore,

$$q(\mathbf{w}'|\mathbf{m}) = q(w_{k+1}) = 1/(\Delta w) \tag{A22}$$

and

$$q(\mathbf{a}'|\mathbf{m}) = q(a_{k+1}) = 1/(\Delta a). \tag{A23}$$

The probability of birth at a specific centre time c'_{k+1} is

$$q(\mathbf{c}'|\mathbf{m}) = 1/(N-k), \tag{A24}$$

where N is used instead of N_t due to the prohibition of placing a new Gaussian with width w_{k+1} within $w_{k+1}+w$ of existing Gaussians with centre times c and widths w. The reverse of birth is death, and the probabilities that removing a width and amplitude associated with killing a Gaussian is equal to unity

$$q(\mathbf{w}|\mathbf{m}') = q(\mathbf{a}|\mathbf{m}') = 1. \tag{A25}$$

The probability of deleting a Gaussian at position c'_{k+1} is

$$q(\mathbf{c}|\mathbf{m}') = 1/(k+1), \tag{A26}$$

since the chance of killing any particular Gaussian is equal to that of killing any other Gaussian, and there are $k+1$ Gaussians to choose from. Combining terms, we get that the reverse to forward proposal ratio associated with birth is given by

$$\left[\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}\right]_{\rm birth} = \frac{\Delta a\,\Delta w(N-k)}{(k+1)}. \tag{A27}$$

Death

Following the logic stated in the birth subsection, when a Gaussian is killed the probability a width or amplitude is removed is unity:

$$q(\mathbf{w}'|\mathbf{m}) = q(\mathbf{a}'|\mathbf{m}) = 1, \tag{A28}$$

and the probability a Gaussian centred at time c_k' is killed is

$$q(\mathbf{c}'|\mathbf{m}) = 1/k. \tag{A29}$$

The reverse of death is birth and the amplitudes are once again independent of the model, giving

$$q(\mathbf{w}|\mathbf{m}') = 1/(\Delta w) \tag{A30}$$

and

$$q(\mathbf{a}|\mathbf{m}') = 1/(\Delta a). \tag{A31}$$

The probability of birthing a model at time c_k' is

$$q(\mathbf{c}|\mathbf{m}') = 1/(N - k + 1), \tag{A32}$$

because there will be one more position to add a Gaussian in the reverse birth step. Combining terms, we get that the reverse to forward proposal ratio associated with death is given by

$$\left[\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}\right]_{\text{death}} = \frac{k}{\Delta a \Delta w (N - k + 1)}. \tag{A33}$$

A3.2 The Jacobian

Following the discussion in Bodin *et al.* (2012), the Jacobian only needs to be calculated when the proposed model $\mathbf{m}'$ involves a change in dimension with respect to the current model $\mathbf{m}$. In other words, we need only worry about the birth and death of Gaussians. When a Gaussian is born, the transformation going from $\mathbf{m}$ to $\mathbf{m}'$ involves a discrete transformation of Gaussian position and continuous transformation of its width and amplitude. The Jacobian term associated with discrete transformations is equal to unity (Denison *et al.* 2002), while that associated with the transformation of the widths and amplitudes is also unity since all widths and amplitudes are drawn from the same distribution and do not involve changes in scale. Therefore, the Jacobian terms are unity and can be neglected for both birth and death steps.

A3.3 Putting it together

For birth, combining terms and simplifying, we have

$$\frac{P(\mathbf{m}')}{P(\mathbf{m})}\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}|\mathbf{J}| = \frac{(k+1)}{(k+2)}, \tag{A34}$$

which is a fortunate result that greatly simplifies the implementation of the deconvolution algorithm that we propose, since the acceptance ratio is independent of the proposal ratios and only depends on the prior in terms of the ratio of the prior of the number of Gaussians, that is,

$$\alpha(\mathbf{m}'|\mathbf{m})_{\text{birth}} = \min\left[1, \frac{P(\mathbf{d_{obs}}|\mathbf{m}')}{P(\mathbf{d_{obs}}|\mathbf{m})}\frac{(k+1)}{(k+2)}\right]. \tag{A35}$$

For death, we have

$$\frac{P(\mathbf{m}')}{P(\mathbf{m})}\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}|\mathbf{J}| = \frac{(k+1)}{k}. \tag{A36}$$

So, we can once again simplify the acceptance ratio since it is independent of the proposal ratios and only dependent on the prior in terms of the ratio of the number of Gaussians, resulting in

$$\alpha(\mathbf{m}'|\mathbf{m})_{\text{death}} = \min\left[1, \frac{P(\mathbf{d_{obs}}|\mathbf{m}')}{P(\mathbf{d_{obs}}|\mathbf{m})}\frac{(k+1)}{k}\right]. \tag{A37}$$