GEOPHYSICS, VOL. 80, NO. 4 (JULY-AUGUST 2015); P. S113–S126, 18 FIGS., 2 TABLES.
10.1190/GEO2014-0364.1

Iterative estimation of reflectivity and image texture: Least-squares migration with an empirical Bayes approach

S. Ahmad Zamanian[1], William L. Rodi[2], Jonathan A. Kane[3], and Michael C. Fehler[2]

ABSTRACT

In many geophysical inverse problems, smoothness assumptions on the underlying geology are used to mitigate the effects of nonuniqueness, poor data coverage, and noise in the data and to improve the quality of the inferred model parameters. Within a Bayesian inference framework, a priori assumptions about the probabilistic structure of the model parameters can impose such a smoothness constraint, analogous to regularization in a deterministic inverse problem. We have considered an empirical Bayes generalization of the Kirchhoff-based least-squares migration (LSM) problem. We have developed a novel methodology for estimation of the reflectivity model and regularization parameters, using a Bayesian statistical framework that treats both of these as random variables to be inferred from the data. Hence, rather than fixing the regularization parameters prior to inverting for the image, we allow the data to dictate where to regularize. Estimating these regularization parameters gives us information about the degree of conditional correlation (or lack thereof) between neighboring image parameters, and, subsequently, incorporating this information in the final model produces more clearly visible discontinuities in the estimated image. The inference framework is verified on 2D synthetic data sets, in which the empirical Bayes imaging results significantly outperform standard LSM images. We note that although we evaluated this method within the context of seismic imaging, it is in fact a general methodology that can be applied to any linear inverse problem in which there are spatially varying correlations in the model parameter space.

INTRODUCTION

Seismic imaging (also known as *migration*) refers to the process of creating an image of the earth's subsurface reflectivity from seismograms generated by sources and recorded by receivers located, typically, at or near the surface. Traditional migration methods for constructing the image generally involve operating on the seismic data with the adjoint of an assumed forward-modeling operator (Claerbout, 1992), possibly along with a modifying function that attempts to correct for amplitude loss due to geometric spreading, transmission, absorption, etc. (Bleistein, 1984; Hanitzsch et al., 1994). In recent years, attempts have been made to cast the imaging problem as a least-squares inverse problem (Nemeth et al., 1999; Duquet et al., 2000). This approach to imaging is conventionally referred to as *least-squares migration* (LSM). Early treatments of this approach can be found in LeBras and Clayton (1988) and Lambare et al. (1992). This paper will deal mainly with Kirchhoff-based LSM , which uses a ray-theoretic-based forward-modeling operator; its derivation and application are discussed in Nemeth et al. (1999) and Duquet et al. (2000). In our formulation, we make the additional simplifying assumption that there is a single reflectivity at each grid point independent of offset or angle of incidence (where the offset effects are instead captured in our forward model). Beyond Kirchhoff-based methods, LSM can also be applied with wave-equation-based forward modeling, as shown by Kühl and Sacchi (2003). In solving the least-squares inverse problem, it is common to include some form of regularization in the LSM cost function to penalize less smooth images. For example, Clapp (2005) describes two regularization schemes for LSM in which the image is constrained to be smooth either along geologic features predetermined by a seismic interpreter or along the ray-parameter axis. In these and other ap-

This paper was presented at the 2013 SEG Annual Meeting in Houston, Texas.
Manuscript received by the Editor 7 August 2014; revised manuscript received 24 January 2015; published online 10 June 2015.
[1]Formerly MIT Earth Resources Laboratory, Cambridge, Massachusetts, USA; presently Shell International E&P Inc., Shell Technology Center, Houston, Texas, USA. E-mail: zamanian@alum.mit.edu.
[2]MIT Earth Resources Laboratory, Cambridge, Massachusetts, USA. E-mail: rodi@mit.edu; fehler@mit.edu.
[3]Shell International E&P Inc., Cambridge, Massachusetts, USA. E-mail: jonathan.kane@shell.com.

plications of LSM, the regularization is chosen independently of the seismic data; i.e., it is a fixed input to the inversion procedure (as it is in the vast majority of geophysical applications of inversion). This, however, may result in suboptimal inversion results; overly strong regularization may result in oversmoothing the image, whereas weak regularization may not adequately penalize roughness in the image due to noise. Even if an appropriate regularization strength is determined, the true smoothness structure of the model need not be spatially uniform or even isotropic; for example, the true earth may typically contain many sharp discontinuities in which any form of smoothing would be undesirable.

In this paper, we propose a more general approach to LSM that solves for parameters defining the image regularization in conjunction with the optimal image itself. The approach is formulated within the framework of Bayesian inference, in which regularization can be accomplished with a prior probability density function (PDF) on the image parameters. We define a prior PDF with spatially varying smoothness properties and seek to jointly estimate its parameters along with the image. In particular, we use a variant of Bayesian inference known as *hierarchical Bayes*, which provides a rigorous mathematical framework for addressing the joint estimation of the image and regularization parameters. This should allow for preserving sharpness in the image at the true discontinuities while still smoothing the effects of noise.

Hierarchical Bayesian methods have been applied in geophysics in many settings. Examples include Malinverno and Briggs (2004), who apply a hierarchical Bayesian framework to 1D traveltime tomography, Buland and Omre (2003), who apply hierarchical Bayesian methods in amplitude variation with offset (AVO) inversion, and Bodin et al. (2012), who apply Bayesian techniques to determine group velocities for the Australian continent; in these three examples, the hierarchical framework is used to estimate the noise in the data. Another method, in the same spirit as hierarchical Bayes, is transdimensional Bayesian inference, in which the dimensionality of the model parameters is inferred from the data; this is applied by Bodin et al. (2012) and Bodin (2010), in the aforementioned study, by Malinverno (2002, 2000), who uses it to find optimal parameterizations of 1D density and resistivity models, and by Ray and Key (2012) and Ray et al. (2013) to invert marine controlled-source electromagnetic data. An important distinction among the work presented in this paper and these studies is that we are using the hierarchical Bayesian method to specifically infer spatially varying regularization parameters, rather than data noise or model parameterization.

In the next sections, we review Kirchhoff-based LSM and proceed to develop the hierarchical Bayesian framework and discuss the algorithms used to solve the inference problem. We conclude with two synthetic 2D data examples and a discussion of our results.

METHODOLOGY

Standard Kirchhoff-based least-squares migration framework

Kirchhoff modeling

The Kirchhoff modeling operator is a ray-based forward modeling operator that gives the seismic data as a linear function of the reflectivity model. In particular, to simulate the seismogram $d_{sr}(t)$ recorded at a seismic receiver r from a seismic source s, Kirchhoff forward modeling first generates a source-to-reflector-to-receiver traveltime (or two-way traveltime) field $\tau_{sr}(\mathbf{x})$ by using what is known as the *exploding reflector* concept. This concept refers to the treatment of each point in the reflectivity model as a *point source*. The two-way traveltime can be computed as the sum of the source-to-reflector and reflector-to-receiver traveltimes, as determined by ray tracing through a specified background velocity model of the subsurface. The ray tracer also computes the field of raypath lengths $R_s(\mathbf{x})$ and $R_r(\mathbf{x})$ and opening angles between the source and receiver rays at each reflection point $\theta_{sr}(\mathbf{x})$. Once these quantities have been computed, the synthetic data $\hat{d}_{sr}(t)$ are computed by superposition over reflector locations $\mathbf{x}$ of scaled and shifted versions of the source wavelet $w_s(t)$ (after applying a 90° phase shift to simulate the effects of 2D propagation). For each $\mathbf{x}$, the phase-shifted wavelet $\tilde{w}_s(t)$ is delayed by $\tau_{sr}(\mathbf{x})$ and scaled by the reflectivity value $m(\mathbf{x})$, an obliquity correction factor $\cos(\theta_{sr}(\mathbf{x})/2)$, and a geometric spreading correction (in 2D, $1/\sqrt{R_s(\mathbf{x})R_r(\mathbf{x})}$). Thus,

$$\hat{d}_{sr}(t) = \int_{\mathcal{X}} m(\mathbf{x}) \frac{\tilde{w}_s[t - \tau_{sr}(\mathbf{x})] \cos[\theta_{sr}(\mathbf{x})/2]}{\sqrt{R_s(\mathbf{x})R_r(\mathbf{x})}} \mathrm{d}\mathbf{x}, \quad (1)$$

where $\mathcal{X} \subset \mathbb{R}^2$ is the model domain. We note that the above Kirchhoff modeling operator is precisely the adjoint operator to the Kirchhoff migration operator, given by

$$\hat{m}(\mathbf{x}) = \sum_s \sum_r \int_t d_{sr}(t) \frac{\tilde{w}_s[t - \tau_{sr}(\mathbf{x})] \cos[\theta_{sr}(\mathbf{x})/2]}{\sqrt{R_s(\mathbf{x})R_r(\mathbf{x})}} \mathrm{d}t. \quad (2)$$

If we discretize time and space, we can represent our data and image as finite-dimensional vectors $\mathbf{d}$ and $\mathbf{m}$, where the dimension of $\mathbf{d}$ is the number of source-receiver pairs times the number of time samples, and where the dimension of $\mathbf{m}$ is the number of points in a spatial grid sampling the model domain. Then, replacing the integral in equation 1 with a summation, we can express the Kirchhoff modeling operator in matrix form:

$$\hat{\mathbf{d}} = \mathbf{A}\mathbf{m}. \quad (3)$$

In particular, the ith column of $\mathbf{A}$, corresponding to a point x_i in the model grid, will contain a sampled version of the source wavelet for each source-receiver pair, appropriately scaled or shifted, giving (in 2D)

$$A_{\mathrm{srt},i} = \frac{\tilde{w}_s[t - \tau_{sr}(x_i)] \cos[\theta_{sr}(x_i)/2]}{\sqrt{R_s(x_i)R_r(x_i)}} \ell^2, \quad (4)$$

where ℓ is the spatial discretization interval.

Standard least-squares migration framework

LSM attempts to solve the imaging problem by seeking the image $\mathbf{m}_{\mathrm{LS}}$ that minimizes the ℓ^2-norm of the residual (the difference between the observed data $\mathbf{d}$ and the modeled data $\hat{\mathbf{d}} = \mathbf{A}\mathbf{m}$). Without regularization, the LSM image is given by

$$\mathbf{m}_{\mathrm{LS}} = \arg\min_{\mathbf{m}} \|\mathbf{d} - \mathbf{A}\mathbf{m}\|_2^2, \quad (5)$$

where $\|\cdot\|_2$ denotes the ℓ^2-norm in the (discretized) data-space given by

$$\|\mathbf{d}\|_2^2 = \sum_s \sum_r \sum_t d_{sr}(t)^2. \tag{6}$$

To ensure well posedness of the LSM solution, regularization is often introduced by augmenting the LSM cost function with a term that penalizes differences between model parameters and an additional term that penalizes the model norm. This gives the regularized LSM image as

$$\mathbf{m}_{\text{RLS}} = \arg\min_{\mathbf{m}} \|\mathbf{d} - \mathbf{A}\mathbf{m}\|_2^2 + \lambda \left[\sum_{(i,j)\in\mathcal{E}} \beta_{ij}(m_i - m_j)^2 + \epsilon \sum_i m_i^2 \right] \tag{7}$$

$$= \arg\min_{\mathbf{m}} \|\mathbf{d} - \mathbf{A}\mathbf{m}\|_2^2 + \lambda \mathbf{m}^T(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})\mathbf{m}, \tag{8}$$

where $\beta_{ij} \in [0, 1]$ indicates how strongly to penalize the difference between m_i and m_j, $\mathcal{E}$ is the set of all pairs of image parameter indices whose difference we decide to potentially penalize, $\lambda > 0$ assigns the maximal weight given to penalizing these differences (and controls the trade-off between model smoothness and data fit), and $\epsilon > 0$ weights the penalty on the model norm. Equation 8 is equation 7 rewritten in compact matrix-vector notation, where $\mathbf{D}$ is a differencing operator defined by the vector $\boldsymbol{\beta} = \{\beta_{ij} : (i, j) \in \mathcal{E}\}$. Taking the derivative of the right side of equation 8 and setting it to zero yields the solution to the regularized LSM problem:

$$\mathbf{m}_{\text{RLS}} = [\mathbf{A}^T\mathbf{A} + \lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})]^{-1}\mathbf{A}^T\mathbf{d}. \tag{9}$$

Note that $\epsilon > 0$ ensures that the regularized LSM cost function is a positive-definite quadratic function of the image $\mathbf{m}$, and hence its minimizer is unique. It is worth noting that the LSM formulation leading to the minimum least-squares solution is one among many valid solution frameworks to the inverse problem, each of which may yield a different answer for the optimal model parameters.

Bayesian framework

Standard Bayesian formulation

The same solution to LSM can be derived from a Bayesian formulation of the imaging problem, wherein the image $\mathbf{m}$ and the data $\mathbf{d}$ are taken to be random vectors. In particular, we take $\mathbf{m}$ a priori to be Gaussian with zero mean and some covariance matrix $\mathbf{C}$ (i.e., $\mathbf{m} \sim \mathcal{N}(0, \mathbf{C})$) , so that the prior PDF $p(\mathbf{m})$ for $\mathbf{m}$ is given by

$$p(\mathbf{m}) \propto \exp\left\{-\frac{1}{2}\mathbf{m}^T\mathbf{C}^{-1}\mathbf{m}\right\}. \tag{10}$$

We choose to use a zero-mean prior due to the nature of the imaging problem: Because seismic reflectors result from discontinuities in the subsurface and are hence likely to be sparse, we assume a prior model involving no discontinuities. We model the seismic data as $\mathbf{d} = \mathbf{A}\mathbf{m} + \mathbf{n}$, where $\mathbf{A}$ is our Kirchhoff modeling operator and $\mathbf{n}$ is zero-mean Gaussian noise with some covariance matrix Σ (i.e., $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$). Thus, the conditional PDF for the data $\mathbf{d}$ given the model $\mathbf{m}$, known as the *model likelihood function*, will be

$$p(\mathbf{d}|\mathbf{m}) \propto \exp\left\{-\frac{1}{2}(\mathbf{d} - \mathbf{A}\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\mathbf{d} - \mathbf{A}\mathbf{m})\right\}, \tag{11}$$

i.e., $\mathbf{d}|\mathbf{m} \sim \mathcal{N}(\mathbf{A}\mathbf{m}, \boldsymbol{\Sigma})$. Other, non-Gaussian noise models (such as the Laplace PDF, resulting from a 1-norm in the exponential), could also be used to formulate our problem, which would result in a very different posterior PDF for the model. However, the Gaussian PDF is somewhat of a natural choice for our formulation for probabilistic reasons (given no other information about the noise other than its mean and covariance matrix, the Gaussian distribution is the distribution with maximum entropy) and for reasons of mathematical convenience (because it allows us to remain analogous to the least-squares formulation).

Applying Bayes' rule gives the posterior PDF for the model $\mathbf{m}$ conditioned on the data $\mathbf{d}$ as

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{m})p(\mathbf{d}|\mathbf{m})}{p(\mathbf{d})} \tag{12}$$

$$\propto \frac{1}{p(\mathbf{d})} \exp\left\{-\frac{1}{2}[\mathbf{m}^T\mathbf{C}^{-1}\mathbf{m} + (\mathbf{d} - \mathbf{A}\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\mathbf{d} - \mathbf{A}\mathbf{m})]\right\}. \tag{13}$$

Rearranging terms in equation 13 and dropping any multiplicative factors that do not depend on $\mathbf{m}$, we obtain

$$p(\mathbf{m}|\mathbf{d}) \propto \exp\left\{-\frac{1}{2}(\mathbf{m} - \boldsymbol{\mu}_{\text{post}})\boldsymbol{\Lambda}_{\text{post}}^{-1}(\mathbf{m} - \boldsymbol{\mu}_{\text{post}})^T\right\}, \tag{14}$$

where $\boldsymbol{\mu}_{\text{post}}$ is the posterior mean given by

$$\boldsymbol{\mu}_{\text{post}} = (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} + \mathbf{C}^{-1})^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{d} \tag{15}$$

and $\boldsymbol{\Lambda}_{\text{post}}$ is the posterior model covariance matrix given by

$$\boldsymbol{\Lambda}_{\text{post}} = (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} + \mathbf{C}^{-1})^{-1}; \tag{16}$$

that is, the posterior PDF for $\mathbf{m}$ conditioned on $\mathbf{d}$ is itself Gaussian: $\mathbf{m}|\mathbf{d} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Lambda}_{\text{post}})$.

The Bayesian maximum a posteriori (MAP) estimate m_{MAP} is the image that maximizes the posterior PDF (equation 13). It is clear from equation 15 that $m_{\text{MAP}} = \boldsymbol{\mu}_{\text{post}}$. Comparing to equation 9, we also see that $m_{\text{MAP}} = \mathbf{m}_{\text{RLS}}$ when we set the prior model and noise covariance matrices as

$$\mathbf{C} = \{\lambda[\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I}]\}^{-1} \tag{17}$$

and

$$\boldsymbol{\Sigma} = \mathbf{I}. \tag{18}$$

We note that although modeling the additive noise as white, as above, will cause the MAP solution to be equivalent to the regularized least-squares solution, we are free to design $\boldsymbol{\Sigma}$ according to a more realistic noise model; to emphasize this generality, we leave our formulation in terms of a general noise covariance matrix $\boldsymbol{\Sigma}$. When implementing our methodology on the synthetic data examples discussed in the "Results" section, we indeed make $\boldsymbol{\Sigma}$ propor-

tional to $\mathbf{I}$, but we test our algorithm against data containing correlated noise. As will be seen in the "Results" section, our algorithm is still able to handle correlated noise, despite assuming a white noise model.

Interpretation of the prior probability density function

The choice of $\boldsymbol{\beta}$ plays a key role in determining the spatial smoothness properties of the PDF on the model. In particular, $\boldsymbol{\beta}$ defines the prior model precision matrix $\mathbf{Q} = \lambda[\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I}]$, which induces a Markov random field (MRF) on the model. What this means in probabilistic terms is that $\boldsymbol{\beta}$ captures the prior conditional dependence structure of the image $\mathbf{m}$, such that $\beta_{ij} = 0$ implies that, prior to observing $\mathbf{d}$, m_i is conditionally independent of m_j when $\{m_k : k \neq i, j\}$ is given (a detailed review of probabilistic graphical models and MRFs can be found in Koller and Friedman, 2009). The MRF defined by $\boldsymbol{\beta}$ on a nine-pixel image is shown in the undirected graph of Figure 1, which depicts the conditional dependence that the β_{ij} (which parameterize the edges of the graph) impose on m_i and m_j (encoded in the vertices of the graph). For this reason, we sometimes refer to the elements of $\boldsymbol{\beta}$ as the edge strengths of the MRF and to $\mathbf{D}(\boldsymbol{\beta})$ as its weighted-graph Laplacian (weighted by $\boldsymbol{\beta}$).

Note that although Figure 1 shows edges connecting only nearest neighbors horizontally and vertically, this need not be the case. We can consider a situation in which each node shares an edge with all other nodes within a specified radius; the graphical model depicted in the figure results from using a radius of 1 node.

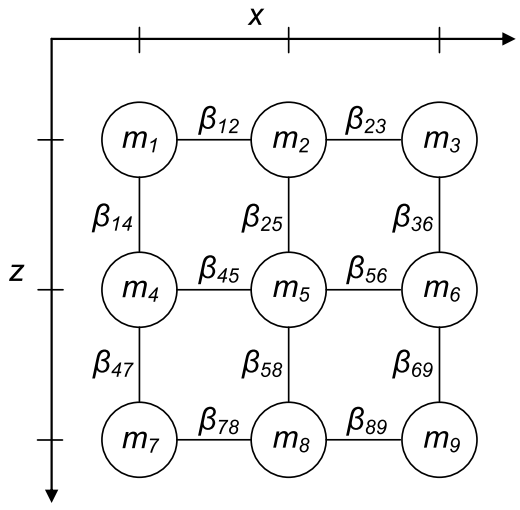

Figure 1. The MRF imposed on model $\mathbf{m}$ by fixing the spatially varying smoothness parameters $\boldsymbol{\beta}$ prior to observing the data $\mathbf{d}$, for a simple nine-pixel image. Note that we labeled the edges of the MRF with the β_{ij} because $\boldsymbol{\beta}$ parameterizes these edges.

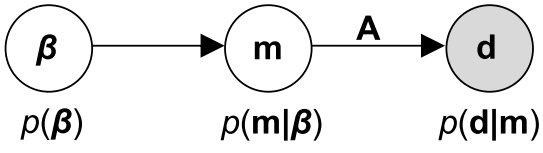

Figure 2. The directed graphical model capturing the relationship between the spatially varying smoothness parameters $\boldsymbol{\beta}$, the model $\mathbf{m}$, and the data $\mathbf{d}$. The node for $\mathbf{d}$ is shaded to indicate that $\mathbf{d}$ is an observed quantity that the posterior distributions of $\boldsymbol{\beta}$ and $\mathbf{m}$ are conditioned upon.

Hierarchical Bayesian formulation

Thus far, we have assumed that the parameters λ, ϵ, and $\boldsymbol{\beta}$, which determine the regularization in the LSM framework and the prior model covariance structure in the Bayesian framework, are known. We now describe how we can expand the Bayesian formulation to the problem of estimating these regularization parameters from the data $\mathbf{d}$, in addition to the image $\mathbf{m}$. We focus on the estimation of the edge strengths $\boldsymbol{\beta}$, which capture our belief about where we think the image should be smooth. Essentially, we would like to learn what the data have to say about the smoothness in the model parameters. By expanding our unknowns to include the edge strengths (so that they are now no longer considered part of a prior PDF), we can learn these edge strengths from the data by using the probabilistic relationship between the edge strengths, the model, and the data. This is depicted in the directed graphical model of Figure 2, which also illustrates the induced Markov chain structure between $\boldsymbol{\beta}$, $\mathbf{m}$, and $\mathbf{d}$.

To estimate $\boldsymbol{\beta}$ from $\mathbf{d}$, we consider $\boldsymbol{\beta}$ to be a random vector endowed with its own prior PDF $p(\boldsymbol{\beta})$. Accordingly, all probability distributions in the previous sections can be considered as conditional on $\boldsymbol{\beta}$. In particular, we now write the prior PDF on $m|\boldsymbol{\beta}$ as

$$p(\mathbf{m}|\boldsymbol{\beta}) = \frac{|\lambda[\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I}]|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{m}^T[\lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})]\mathbf{m}\right\}}{(2\pi)^{N/2}} \tag{19}$$

and the conditional PDF for $\mathbf{d}|\mathbf{m}, \boldsymbol{\beta}$ as

$$p(\mathbf{d}|\mathbf{m}, \boldsymbol{\beta}) = p(\mathbf{d}|\mathbf{m}) \tag{20}$$

$$= \frac{\exp\left\{-\frac{1}{2}(\mathbf{d} - \mathbf{A}\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\mathbf{d} - \mathbf{A}\mathbf{m})\right\}}{(2\pi)^{K/2}\boldsymbol{\Sigma}^{1/2}}, \tag{21}$$

where N is the number of model parameters (i.e., the dimension of $\mathbf{m}$) and K is the number of data points (the dimension of $\mathbf{d}$). We again apply Bayes' rule to obtain the joint posterior PDF for $\mathbf{m}$ and $\boldsymbol{\beta}$ given the data $\mathbf{d}$:

$$p(\mathbf{m}, \boldsymbol{\beta}|\mathbf{d}) = \frac{p(\boldsymbol{\beta})p(\mathbf{m}|\boldsymbol{\beta})p(\mathbf{d}|\mathbf{m}, \boldsymbol{\beta})}{p(\mathbf{d})} \tag{22}$$

$$= \frac{p(\boldsymbol{\beta})|\lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})|^{1/2}}{p(\mathbf{d})\,(2\pi)^{(N+K)/2}\boldsymbol{\Sigma}^{1/2}}$$

$$\exp\left\{-\frac{1}{2}[(\mathbf{d} - \mathbf{A}\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\mathbf{d} - \mathbf{A}\mathbf{m}) + \mathbf{m}^T(\lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I}))\mathbf{m}]\right\}. \tag{23}$$

To define $p(\boldsymbol{\beta})$, we endow each β_{ij} with a uniform prior PDF on the set $[0, 1]$ and we let the β_{ij} be mutually independent random variables, so that

$$p(\boldsymbol{\beta}) = \prod_{(i,j)\in\mathcal{E}} \mathbf{1}_{[0,1]}(\beta_{ij}), \quad \text{where} \quad \mathbf{1}_S(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases}. \tag{24}$$

We note that equation 23 is very similar to the posterior PDF in the nonhierarchical Bayesian setting (where $\boldsymbol{\beta}$ is fixed) with some important differences: First, equation 23 is now a function of $\mathbf{m}$ and $\boldsymbol{\beta}$, and second, outside the exponential of equation 23 is the determinant of $\mathbf{m}$'s prior precision matrix $\mathbf{Q}$ (which can no longer be dropped as a proportionality constant, because it depends on $\boldsymbol{\beta}$). Computing this determinant is expensive, with time complexity $\mathcal{O}(N^2)$ (because $\mathbf{Q}$ is a sparse matrix with bandwidth $N^{1/2}$); this reflects the additional computational cost of the hierarchical Bayesian approach.

Having obtained the joint posterior PDF $p(\mathbf{m}, \boldsymbol{\beta}|\mathbf{d})$, the task of estimating the best image remains. Here, we explore two estimation methodologies within the hierarchical Bayesian framework: the *hierarchical* Bayes solution and the *empirical* Bayes solution (Malinverno and Briggs, 2004). What is strictly known as the *hierarchical Bayes solution* is the full marginal posterior PDF of the image $p(\mathbf{m}|\mathbf{d})$ (marginalizing out $\boldsymbol{\beta}$ from the joint posterior PDF $p(\mathbf{m}, \boldsymbol{\beta}|\mathbf{d})$). Hence, we have for the hierarchical Bayes solution

$$p(\mathbf{m}|\mathbf{d}) = \int_{\mathcal{B}} p(\mathbf{m}, \boldsymbol{\beta}|\mathbf{d}) \mathrm{d}\boldsymbol{\beta}, \tag{25}$$

where $\mathcal{B}$ is the domain of admissible vectors $\boldsymbol{\beta}$. Unfortunately, the marginalization operation cannot be performed analytically and must be computed numerically. We may also consider the MAP estimates for the image that can be derived within the hierarchical Bayesian setting. The hierarchical Bayes MAP estimate $\mathbf{m}_{\mathrm{HB}}$ is the MAP estimate of $\mathbf{m}$ based on its marginal posterior PDF $p(\mathbf{m}|\mathbf{d})$:

$$\mathbf{m}_{\mathrm{HB}} = \arg \max_{\mathbf{m}} \int_{\mathcal{B}} p(\mathbf{m}, \boldsymbol{\beta}|\mathbf{d}) \mathrm{d}\boldsymbol{\beta}. \tag{26}$$

One can think of $\mathbf{m}_{\mathrm{HB}}$ as the single best image $\mathbf{m}$ over all choices of edge strengths $\boldsymbol{\beta}$. Although the posterior marginal PDF for the image (equation 25) is the complete solution to the Bayesian inference problem, several computational issues prevent its use in practice. First, due to the high dimensionality of $\mathcal{B}$ and the cost of evaluating the joint posterior PDF (which involves a matrix determinant; see equation 23), stochastic sampling from and direct marginalization of the joint posterior PDF are computationally intractable. Furthermore, even if we were able to evaluate the marginal posterior (equation 25), the high dimension of $\mathbf{m}$ would make it difficult to explore. One potential way to avoid the problem of high model dimensionality, although not pursued in this paper, is the transdimensional Bayesian approach (e.g., Bodin et al., 2012), in which the number of model parameters is also determined by the data.

A somewhat different solution for estimating the image is known as the *empirical Bayes* solution, which first looks for the best choice for $\boldsymbol{\beta}$, then, using that choice, finds the best image $\mathbf{m}_{\mathrm{EB}}$. If one takes the MAP estimate for $\boldsymbol{\beta}$, then we would have

$$\boldsymbol{\beta}_{\mathrm{MAP}} = \arg \max_{\boldsymbol{\beta}} \int_{\mathcal{M}} p(\mathbf{m}, \boldsymbol{\beta}|\mathbf{d}) \mathrm{d}\mathbf{m}, \tag{27}$$

where, it turns out, the marginalization over $\mathbf{m}$ can be performed analytically but the maximization over $\boldsymbol{\beta}$ must still be performed numerically. Given $\boldsymbol{\beta}_{\mathrm{MAP}}$, the empirical Bayes solution is taken as the MAP estimate with respect to $p(\mathbf{m}|\mathbf{d}, \boldsymbol{\beta}_{\mathrm{MAP}})$. The results of the previous sections then imply

$$\mathbf{m}_{\mathrm{EB}} = \{\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} + \lambda[\mathbf{D}(\boldsymbol{\beta}_{\mathrm{MAP}}) + \epsilon\mathbf{I}]\}^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{d}. \tag{28}$$

The empirical Bayes solution is within reach as long as we are able to compute $\boldsymbol{\beta}_{\mathrm{MAP}}$ by solving the marginal MAP problem of equation 27. To do so, we turn to the expectation-maximization (E-M) algorithm, which has direct application in solving such marginal MAP problems.

The expectation-maximization algorithm

The E-M algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) is a powerful and versatile algorithm for solving maximum-likelihood and MAP parameter estimation problems when a subset of the variables relevant to the parameter estimation is unobserved (referred to as *latent* variables). In the context of the seismic imaging problem we consider here, we view the image $\mathbf{m}$ as the latent variables. In the empirical Bayes approach, these variables must be marginalized from the joint posterior PDF on $\mathbf{m}$ and $\boldsymbol{\beta}$ when attempting to estimate the edge strengths $\boldsymbol{\beta}$. For our purposes, E-M can be thought of as a coordinate ascent algorithm for solving the marginal MAP optimization problem (equation 27), whereby subsequent estimations are performed between the latent variables ($\mathbf{m}$) and the parameters to be estimated ($\boldsymbol{\beta}$).

In Appendix A, we review the E-M algorithm by means of a brief derivation; similar derivations and a more thorough treatment of E-M can be found in Bishop (2006) or McLachlan and Krishnan (2008). We specialize the equations of the E-M algorithm to the LSM problem in Appendix B; this specialization yields the specific quantities that must be computed in each iteration of the E-M algorithm, particularly, the posterior model mean and the elements of the prior and posterior model covariance matrices that correspond to each β_{ij} (all conditioned upon the last iterates for the estimated $\boldsymbol{\beta}_{\mathrm{MAP}}$). Exact computation of these elements of the covariance matrices can be intractable because they would require direct matrix inversions; hence, in Appendices C and D, we discuss approximate methods that can be used to estimate these elements. Appendix E summarizes these developments with a pseudocode for our complete implementation of the E-M algorithm (using the above approximate methods).

As its output, our specific implementation of the E-M algorithm yields the MAP estimate of the edge strengths $\boldsymbol{\beta}_{\mathrm{MAP}}$ and the empirical Bayes MAP image $\mathbf{m}_{\mathrm{EB}}$; in addition, the posterior covariance matrix $\boldsymbol{\Lambda}$ (conditioned on $\boldsymbol{\beta}_{\mathrm{MAP}}$) is computed as a by-product of our implementation and quantifies the uncertainty associated with the empirical Bayes solution.

RESULTS

To validate our approach, we ran our inference algorithm on synthetic data sets. We present two test cases: the first case being a simple example in which the data arise from a small image consisting of three dipping reflectors separated by a weakly reflective fault and the second case being data simulated from the Marmousi model. Synthetic data were created using the same Kirchhoff modeling operator $\mathbf{A}$ that is used in the inference algorithms. To some-

what avoid the inverse crime, we add noise to the data according to two different noise models. We formulated our methodology assuming zero-mean white Gaussian noise, and this is the noise model we used for the three-layer synthetic model (with noise standard deviation equal to 10% of the maximum amplitude of the data) to test the results of our inversion procedure when the assumptions used in the formulation hold. In practice, however, noise will necessarily be band-limited and correlated; hence, for the Marmousi example, we used a more realistic noise model in which the noise is "colored" by the forward-modeling process. To be precise, we obtain colored noise by generating zero-mean white Gaussian noise in the image domain (having a standard deviation equal to 10% of the maximum amplitude of the true reflectivity model), then we pass this noise through the forward-modeling operator **A**, to obtain a colored zero-mean Gaussian noise vector that is added to the synthetic data.

In the first example of three dipping reflectors, the data are created from a single surface seismic source (at the center) and 50 equally spaced surface seismic receivers (with spacing of 50 m) using a homogeneous background velocity model (of 4000 m/s). The source wavelet is a 20-Hz Ricker wavelet; hence, the dominant wavelength is 200 m. The seismic traces are sampled at 1 ms, and the medium is sampled spatially at 50 m in the lateral and vertical directions. The entire medium has spatial dimensions of 2500 m × 2500 m; hence, $N_x = N_z = 50$ and the number of image parameters is $N = N_x N_z = 2500$. The purpose of testing our algorithm on such a small model is so that we can verify the performance of our algorithm in the absence of any approximations (i.e., in this case, we can directly compute the elements of the prior and posterior covariance matrices without the need of the approximations discussed in Appendices C and D). For this example, we used an MRF in which each node shares an edge with its four nearest neighbors, giving a total of 4900 edge strengths. Here, we ran 10 iterations of the E-M algorithm to obtain the MAP estimate of the edge strengths and the empirical Bayes image, in which each iteration of the E-M algorithm ran in approximately one minute on a quad-core Intel™ Xeon W3550 3.0 GHz processor. By contrast, a standard LSM run (using a fixed regularization scheme) took approximately 1 s on the same machine. Note that the vast increase in computational cost from that of the standard LSM to the empirical Bayes method is due to the fact that, in this example, we computed and stored the entire prior and posterior covariance matrices in each iteration, whereas on a larger problem, we would avoid directly inverting and storing matrices.

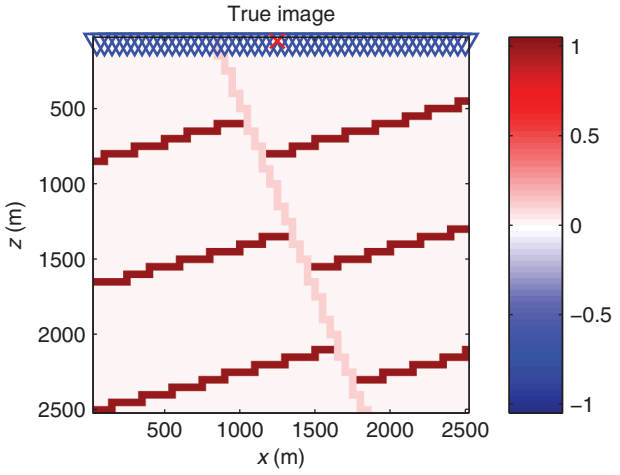

Figure 3. True image (normal incidence reflectivities) for the three-layer test case. The receiver locations are indicated by the blue inverted triangles, and the location of the single source is indicated by the red "x." The discretization interval used here is 50 m in the lateral and vertical directions. Note that the "staircased" appearance of the dipping reflectors is due to the spatial discretization interval.

For the case of the Marmousi model, we use a smoothed version of the true Marmousi velocity model (sampled at 24-m spacing) for our background velocity model in conjunction with the true (unsmoothed) reflectivity model to simulate the data. The data are created from a set of 20 evenly spaced surface sources firing into 20 receivers at the same location (resulting in 400 traces with different offsets), with 480-m spacing between stations, in which the source wavelet is a 25-Hz Ricker wavelet. We chose to use such a sparse data set (where standard methods often perform poorly) to show the strength of the empirical Bayes method in the face of poor data resolution. Because we have a larger model in this example, with 46,343 image parameters, we must resort to the approximate methods outlined in Appendices C and D to compute the quantities required by the E-M algorithm. Here, to capture the more complex dipping structures of the Marmousi model, we defined the MRF so that each node shares an edge with all nodes within a radius of $\sqrt{2}$ nodes (i.e., a node shares an edge with its four diagonal neighbors in addition to its four nearest neighbors, resulting in 183,862 edge strengths to be estimated). We note here that the choice of the radius defining the extent of the edge set $\boldsymbol{\mathcal{E}}$ is ad hoc; the practitioner of this method should pick an edge set large enough to be able to suitably capture the complexity of the structures expected to be seen, and often the appropriate radius is decided using a trial-and-error approach with a few different options (e.g., radii of 1, $\sqrt{2}$, 2, etc.). Although it may be possible to formulate a methodology for learning the best choice (e.g., within a hierarchical Bayesian setting) for the edge set $\boldsymbol{\mathcal{E}}$ of an MRF, this remains an open area of research within the field of probabilistic graphical models. In this example, the MAP estimate of the edge strengths and the empirical Bayes image were obtained with three iterations of the E-M algorithm, where each iteration of the E-M algorithm took approximately 33 min on a quad-core Intel Xeon W3550 3.0 GHz processor. By comparison, a single LSM run (with fixed regularization parameters) took approximately 2–3 min on the same machine. We note that each iteration of the E-M algorithm requires performing a standard LSM in addition to the computation required to approximate the elements of the posterior and prior covariance matrices. Applying these approximations, the increase in computation from standard LSM to the empirical Bayes method is significantly less than for the previous example.

Figures 3–10 show the results from the test case of the three dipping layer model. Figure 3 displays the true reflectivity model used to generate the synthetic data (shown in Figure 4). Performing a Kirchhoff migration on the data results in the image of Figure 5; here, the reflectors are imaged somewhat, but we also see heavy imaging artifacts (i.e., the migration smiles) due to the limited source-receiver geometry (where only a single source is being used). We observe that in the case of the unregularized LSM image (Figure 6), the reflectors are imaged, but unfortunately, the noise in the data is also imaged so strongly that the reflectors are nearly impossible to distinguish from the noise. We can improve on the

unregularized image by using a uniform regularization scheme (setting each $\beta_{ij} = 1$) to obtain the regularized LSM image of Figure 7; here, the use of regularization has filtered out the noise, but as a side effect, it has also smoothed out the reflectors. Regularizing instead using our estimate of the edge strengths (Figure 8), we obtain the empirical Bayes MAP image (Figure 9) significantly improving on the previous result. This is clear from a qualitative comparison between the images; we can see the reflectors imaged quite strongly with sharpness preserved at the reflectors, while the noise is filtered out elsewhere in the image. Additionally, the weakly reflective fault is also slightly imaged in the empirical Bayes MAP image, whereas it cannot be seen in the other images. We further note that the correlation of the empirical Bayes MAP image with the true image is significantly higher than the correlations of the other images with the true image. Examining the estimate of the edge strengths in Figure 8, we see that the edge strengths take on a pattern similar to what we expect: They are high where the image is constant, but they are close to zero where there are differences in the image (surrounding the reflectors). The uncertainty associated with the empirical Bayes estimate of the model parameters is given by the posterior model standard deviations (conditioned on $\boldsymbol{\beta}_{\mathrm{MAP}}$), which are shown in Figure 10 and were computed from the square root of the diagonal of the posterior model covariance matrix. As seen in the figure, the model uncertainty increases both in deeper parts of the model (as expected, due to weaker signal contribution) and at parts of the model adjacent to the reflectors, where the edge strengths are close to zero. A small edge strength means that very little smoothness is being enforced across the corresponding model parameters, thereby allowing for larger changes between these model parameters; this relaxation of constraints translates probabilistically to the increased model uncertainty seen in the figure.

Figures 11–18 show the results from the test case with the Marmousi model. The true reflectivity model is shown in Figure 11 and

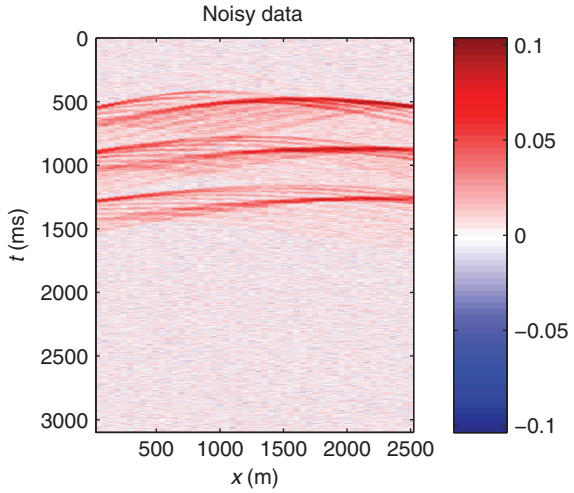

Figure 4. Noisy synthetic data for the three-layer test case.

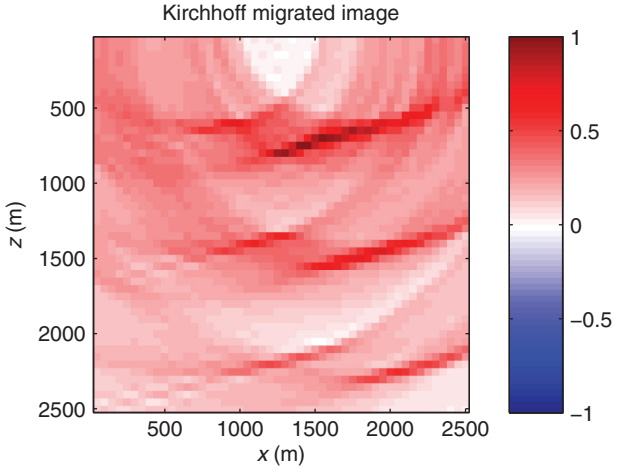

Figure 5. Kirchhoff-migrated image for the three-layer test case. Correlation with the true image = 0.4705. Note that the fault is not visible in this image and the reflectors are somewhat obscured by strong migration artifacts.

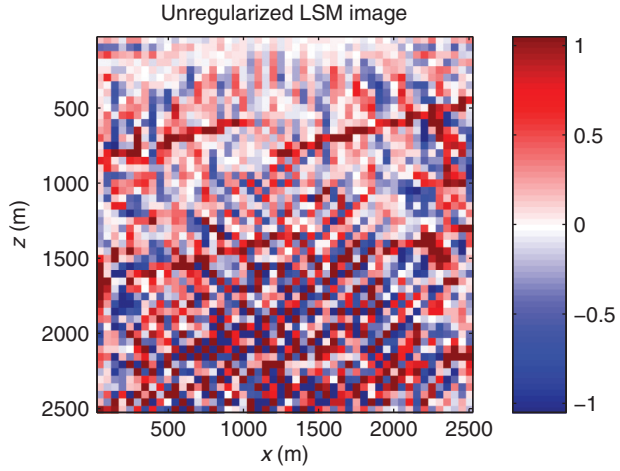

Figure 6. Unregularized LSM image (each $\beta_{ij} = 0$) for the three-layer test case. Correlation with the true image = 0.3649. Due to the lack of regularization, the noise in the image is so strong that it obscures the dipping reflectors and the fault.

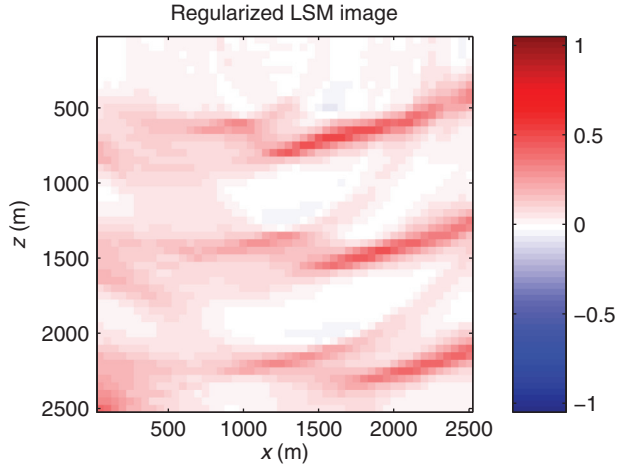

Figure 7. Uniformly regularized LSM image (each $\beta_{ij} = 1$) for the three-layer test case. Correlation with the true image = 0.5879. Here, due to the uniform regularization scheme, smoothness is enforced in the inversion even at places in which sharpness is desired (e.g., adjacent to the reflectors and fault). Hence, the fault is again not visible, and the reflectors have been obscured by the smoothing.

is used to generate the synthetic seismic data shown in Figure 12. We again observe the same features in the images as seen in the three-layer test case. Here, performing a Kirchhoff migration (Figure 13) results in many reflectors not being imaged correctly due to heavy acquisition artifacts and sparse sampling of the seismic wavefield. The unregularized LSM image (Figure 14) shows the reflectors along with a very strong noise component. Regularizing in a uniform fashion (by setting each $\beta_{ij} = 1$) results in the regularized image of Figure 15 in which the noise has been filtered out, but the

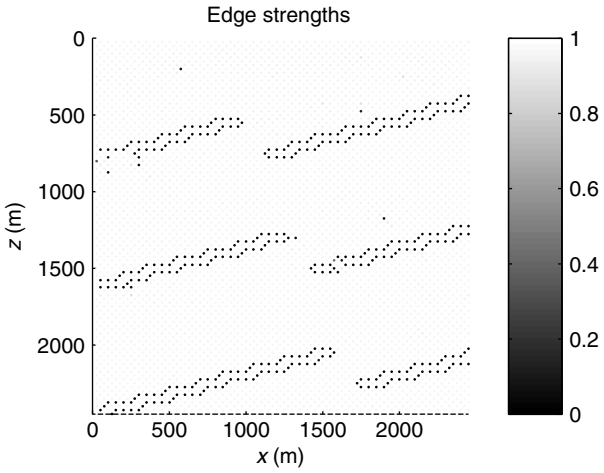

Figure 8. Edge strengths (or smoothness parameters) $\boldsymbol{\beta}$ estimated with the E-M algorithm for the three-layer test case. Note that the edge strengths spatially define the level of smoothness to enforce between adjacent pixels, and they have been plotted at the midpoint between the image pixels they "connect" (see the graph in Figure 1). The edge strengths are close to zero near the reflectors (to prevent smoothing out the true structure) but are close to one away from the reflectors and along the direction of the reflectors (allowing noise to be smoothed out).

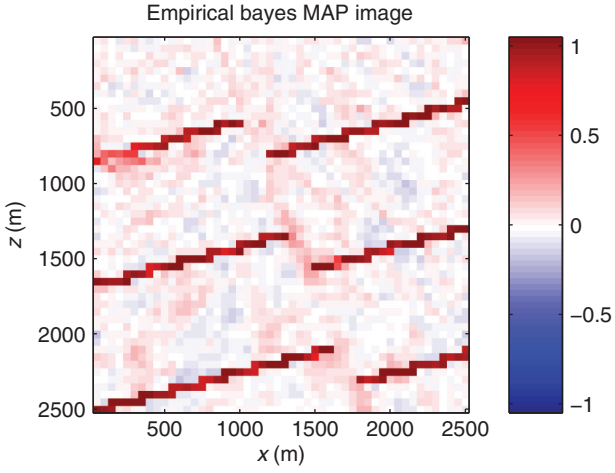

Figure 9. Empirical Bayes MAP image (computed after estimating $\boldsymbol{\beta}$) for the three-layer test case. Correlation with the true image = 0.9607. Here, due to adaptive regularization (tuned by the spatially varying parameter $\boldsymbol{\beta}$), smoothness is enforced away from the reflectors, thereby removing the noise, but sharpness is preserved at the reflectors. Additionally, now the weakly reflective fault is visible in the image.

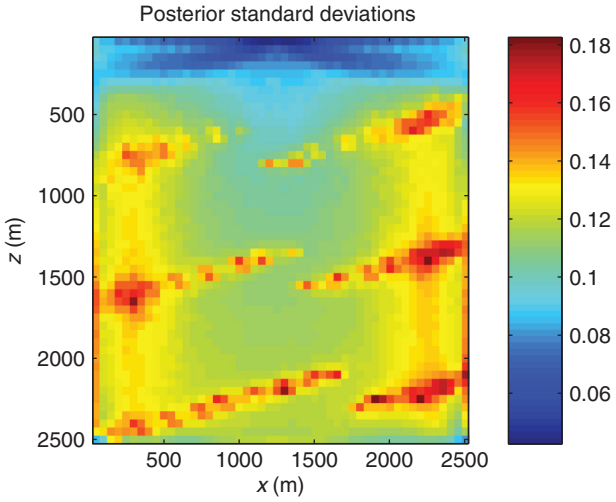

Figure 10. Posterior model standard deviation (conditioned on $\boldsymbol{\beta}_{\text{MAP}}$) obtained from the posterior model covariance matrix. Note that the uncertainty in the model parameter estimates increases in the deeper parts of the model and near the reflectors (where the edge strengths are closer to zero, allowing for larger changes in the model thereby resulting in increased model uncertainty).

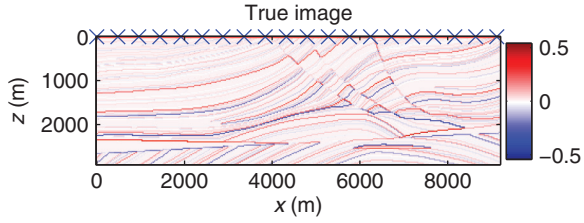

Figure 11. True image (normal incidence reflectivities) for the Marmousi test case. The source and receiver locations are indicated by the blue "x"s.

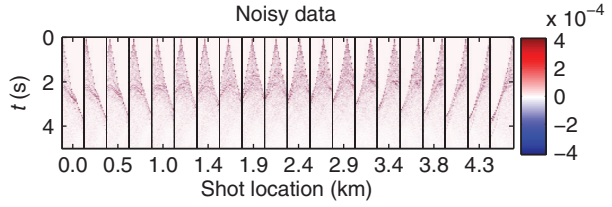

Figure 12. Noisy synthetic data for the Marmousi test case.

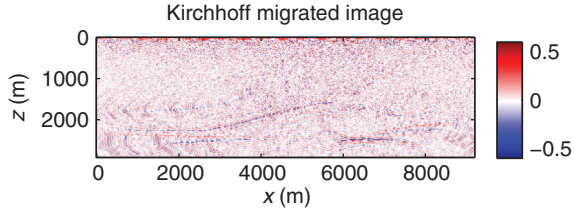

Figure 13. Kirchhoff-migrated image for the Marmousi test case. Correlation with the true image = 0.3439. Note that the sparsity of the data set does not allow for imaging of the shallow reflectors with Kirchhoff migration.

image is also overly smooth in some areas. Once again, using our algorithm to estimate the edge strengths (which are shown in Figure 16) results in the empirical Bayes MAP image of Figure 17. We notice the same qualitative improvements in the image as seen previously: The image remains sharp near the reflectors while smoothing out the noise away from the reflectors. As before, the correlation of the empirical Bayes MAP image with the true image is significantly higher than the correlations of the other images with the true image. The posterior model standard deviations are plotted in Figure 18, in which again we see increased model uncertainty in deeper parts of the model and below the strong reflectors (particularly those on the bottom left and right sides of the image, where the signal contribution is significantly weaker) as well as at points in the model near the reflectors because, as before, at these points the edge strengths are close to zero.

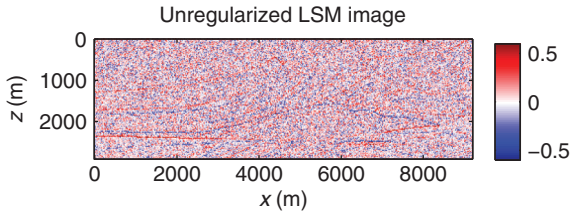

Figure 14. Unregularized LSM image (each $\beta_{ij} = 0$) for the Marmousi test case. Correlation with the true image = 0.4100. Here, the strong level of noise appearing in the image (due to lack of regularization) again obscures the reflectors.

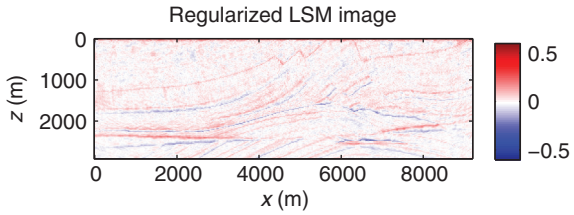

Figure 15. Uniformly regularized LSM image (each $\beta_{ij} = 1$) for the Marmousi test case. Correlation with true image = 0.4834. Here, the regularization removes the noise from the image, but it also smooths the image at the true reflectors.

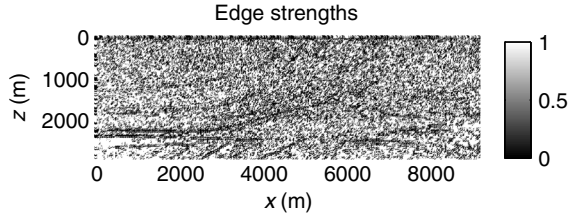

Figure 16. The edge strengths (or smoothness parameters) $\boldsymbol{\beta}$ estimated with the E-M algorithm for the Marmousi test case. Note that the edge strengths are close to zero near the reflectors (to prevent smoothing out the true structure) but are closer to one away from the reflectors (allowing noise to be smoothed out in these locations). The edge strengths are plotted at the midpoint between the image pixels they connect (see the graph in Figure 1).

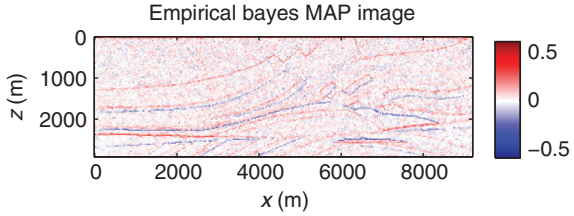

Figure 17. Empirical Bayes MAP image (computed after estimating $\boldsymbol{\beta}$) for the Marmousi test case. Correlation with the true image = 0.6909. Here, we see that regularizing with the optimal set of edge strengths allows for the smoothing out of noise in the image while preserving sharpness at the reflectors.

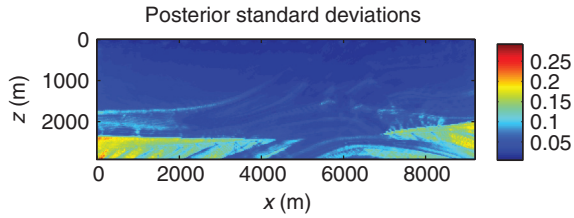

Figure 18. Posterior model standard deviation (conditioned on $\boldsymbol{\beta}_{\text{MAP}}$) obtained from the posterior model covariance matrix. Again, the uncertainty in the model parameter estimates increases in the deeper parts of the model (particularly below the strong reflectors on the lower left and right sides of the model), as well as near the reflectors (where the edge strengths are closer to zero).

CONCLUSIONS AND FUTURE WORK

Our study shows that the Bayesian framework provides a flexible methodology for estimating the image and smoothness parameters (or edge strengths) in an LSM setting. By estimating the spatially varying smoothness parameters, we are able to remove the effects of noise while, by and large, preserving sharpness at the reflectors in the image. The E-M algorithm, in particular, allowed us to solve the marginal MAP problem for estimating the smoothness parameters $\boldsymbol{\beta}$ (without having to explicitly integrate out or sample the high-dimensional model space from the posterior distribution to compute the marginal posterior PDF for $\boldsymbol{\beta}$).

We note that although our algorithm was presented within the context of the seismic imaging problem, the methodology we have developed is broadly applicable to many linear inverse problems in which the model parameters may exhibit spatially (or temporally) varying smoothness properties. The operator $\mathbf{A}$ (or, more generally, the conditional PDF for the data given the model $p(\mathbf{d}|\mathbf{m})$) would change if we were solving a different problem, but the methodology and algorithm described in this paper would still apply.

Although we have developed our algorithm in the setting of solving a linear inverse problem, an interesting direction for future work is to generalize this methodology to nonlinear inverse problems. This generalization is nontrivial, as the nonlinearity of the forward model will likely result in a posterior PDF that could be multimodal and not belong to a nice analytic family such as Gaussian or other exponential family PDFs. A second direction for future work is to explore alternative ways to parameterize the prior PDF on the image within the hierarchical Bayesian setting; for example, we may wish to apply a transdimensional Bayesian framework to the imaging problem using, for example, Voronoi cells. Another natural future direction is application of this methodology to a more realistic synthetic data set (or to a field data set), in which we expect

similar improvements in quality of the resulting image. We note here that in order for our method to be applied to a real data set, it will be necessary to use a more realistic forward model, such as a wave-equation-based forward modeling operator because the simplifying assumptions in our Kirchhoff operator do not correctly account for more complex physical phenomena such as AVO or attenuation effects. Because our modeling operator was formulated for a 2D case, we would also expect 3D out-of-plane effects (however, the Kirchhoff operator we have defined is readily generalizable to the 3D case). Having stated this, it is worth noting that as long as the forward modeling operator is linear in the image parameters (as is a typical assumption when inverting for reflectivity), the inversion methodology discussed herein can still be applied (because operator $\mathbf{A}$ can be any linear forward-modeling operator).

ACKNOWLEDGMENTS

This work was supported by Shell International E&P, Inc., through the MIT Energy Initiative and by the ERL Founding Member Consortium. The authors thank A. Chandran, V. Goh, K. Matson, and H. Kuehl of Shell for insightful discussions and feedback. We also wish to thank K. Marfurt, A. Malinverno, and A. Ray for providing valuable suggestions that helped improve the quality of this paper.

APPENDIX A
REVIEW OF THE E-M ALGORITHM

Here, we review the E-M algorithm by means of a brief derivation. Similar derivations and a more thorough treatment of E-M can be found in Bishop (2006) or McLachlan and Krishnan (2008). Recall that the goal of the E-M algorithm is to compute the MAP estimate for $\boldsymbol{\beta}$, such that

$$\boldsymbol{\beta}_{\text{MAP}} = \arg\max_{\boldsymbol{\beta}} p(\boldsymbol{\beta}|\mathbf{d}). \tag{A-1}$$

To derive the E-M algorithm, we start by noting that maximizing a PDF is equivalent to maximizing its logarithm, and we define our objective function as the log marginal posterior:

$$\ell(\boldsymbol{\beta}) = \log\, p(\boldsymbol{\beta}|\mathbf{d}). \tag{A-2}$$

We note that we are only passing from the marginal posterior PDF to its logarithm for the purpose of maximizing the PDF with respect to $\boldsymbol{\beta}$ to obtain its MAP estimate. Once we have obtained the optimal $\boldsymbol{\beta}_{\text{MAP}}$, we are able to use this to compute the posterior PDF (conditioned on $\boldsymbol{\beta}_{\text{MAP}}$) and take the posterior model mean and covariance matrix to find the best model and quantify its uncertainty. Rearranging terms in the joint posterior PDF, we can rewrite the MAP objective function as

$$\ell(\boldsymbol{\beta}) = \log \int_{\mathcal{M}} p(\mathbf{m}, \boldsymbol{\beta}|\mathbf{d})\mathrm{d}\mathbf{m} \tag{A-3}$$

$$= \log \int_{\mathcal{M}} \frac{p(\mathbf{m}, \boldsymbol{\beta}, \mathbf{d})}{p(\mathbf{d})}\mathrm{d}\mathbf{m} \tag{A-4}$$

$$= \log \int_{\mathcal{M}} \frac{p(\boldsymbol{\beta})p(\mathbf{m}, \mathbf{d}|\boldsymbol{\beta})}{p(\mathbf{d})}\mathrm{d}\mathbf{m} \tag{A-5}$$

$$= \log \int_{\mathcal{M}} p(\mathbf{m}, \mathbf{d}|\boldsymbol{\beta})\mathrm{d}\mathbf{m} + \log\, p(\boldsymbol{\beta}) - \log\, p(\mathbf{d}). \tag{A-6}$$

Here, we introduce a proxy PDF on the image, $q(\mathbf{m}|\mathbf{d})$, where we can choose q to be any PDF we like as long as it has the same support as $p(\mathbf{m})$ and where we have made explicit that q can depend on the data $\mathbf{d}$. Dividing and multiplying by q, we have

$$\ell(\boldsymbol{\beta}) = \log \int_{\mathcal{M}} \frac{q(\mathbf{m}|\mathbf{d})}{q(\mathbf{m}|\mathbf{d})} p(\mathbf{m}, \mathbf{d}|\boldsymbol{\beta})\mathrm{d}\mathbf{m} + \log\, p(\boldsymbol{\beta}) - \log\, p(\mathbf{d}) \tag{A-7}$$

$$= \log\, \mathbb{E}_{q(\mathbf{m}|\mathbf{d})}\left[\frac{p(\mathbf{m}, \mathbf{d}|\boldsymbol{\beta})}{q(\mathbf{m}|\mathbf{d})}\right] + \log\, p(\boldsymbol{\beta}) - \log\, p(\mathbf{d}), \tag{A-8}$$

where the integral in equation A-7 has been recognized as the expected value with respect to q (denoted by $\mathbb{E}_q$) to arrive at equation A-8. Now, by Jensen's inequality (McLachlan and Krishnan, 2008) and the concavity of the log function, we have

$$\ell(\boldsymbol{\beta}) \geq \mathbb{E}_{q(\mathbf{m}|\mathbf{d})}\left\{\log\left[\frac{p(\mathbf{m}, \mathbf{d}|\boldsymbol{\beta})}{q(\mathbf{m}|\mathbf{d})}\right]\right\} + \log\, p(\boldsymbol{\beta}) - \log\, p(\mathbf{d}) \tag{A-9}$$

$$= \hat{\ell}(q, \boldsymbol{\beta}). \tag{A-10}$$

We see that the function $\hat{\ell}(q, \boldsymbol{\beta})$ is a lower bound on the original objective function $\ell(\boldsymbol{\beta})$. The E-M algorithm maximizes this lower bound according to the following coordinate ascent scheme, starting with an initial guess $\hat{\boldsymbol{\beta}}^{(0)}$ and iterated for $t = 0, 1, 2, \ldots$:

$$\textbf{E-Step}: \hat{q}^{(t+1)} = \arg\max_{q} \hat{\ell}(q, \hat{\boldsymbol{\beta}}^{(t)}) \tag{A-11}$$

$$\textbf{M-Step}: \hat{\boldsymbol{\beta}}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \hat{\ell}(\hat{q}^{(t+1)}, \boldsymbol{\beta}). \tag{A-12}$$

It turns out that the E-step can be solved analytically. Let us propose a candidate solution $\tilde{q}$ as the Bayesian posterior of $\mathbf{m}$ conditioned on $\mathbf{d}$ and the last iterate $\hat{\boldsymbol{\beta}}^{(t)}$ of $\boldsymbol{\beta}$:

$$\tilde{q}(\mathbf{m}|\mathbf{d}) = p(\mathbf{m}|\mathbf{d}, \hat{\boldsymbol{\beta}}^{(t)}). \tag{A-13}$$

Then, if we plug $\tilde{q}$ into the E-step objective function (equation A-11), we have

$$\hat{\ell}(\tilde{q}, \hat{\boldsymbol{\beta}}^{(t)}) = \mathbb{E}_{p(\mathbf{m}|\mathbf{d}, \hat{\boldsymbol{\beta}}^{(t)})}\left\{\log\left[\frac{p(\mathbf{m}, \mathbf{d}|\hat{\boldsymbol{\beta}}^{(t)})}{p(\mathbf{m}|\mathbf{d}, \hat{\boldsymbol{\beta}}^{(t)})}\right]\right\} + \log\, p(\hat{\boldsymbol{\beta}}^{(t)}) - \log\, p(\mathbf{d}). \tag{A-14}$$

Recognizing the quotient in equation A-14 as $p(\mathbf{d}|\hat{\boldsymbol{\beta}}^{(t)})$, and because the expectation of $p(\mathbf{d}|\hat{\boldsymbol{\beta}}^{(t)})$ is just itself, we have

$$\hat{\ell}(\tilde{q},\hat{\boldsymbol{\beta}}^{(t)}) = \log\, p(\mathbf{d}|\hat{\boldsymbol{\beta}}^{(t)}) + \log\, p(\hat{\boldsymbol{\beta}}^{(t)}) - \log\, p(\mathbf{d}) \quad \text{(A-15)}$$

$$= \log\, \frac{p(\mathbf{d}|\hat{\boldsymbol{\beta}}^{(t)})p(\hat{\boldsymbol{\beta}}^{(t)})}{p(\mathbf{d})} = \log\, p(\hat{\boldsymbol{\beta}}^{(t)}|\mathbf{d}) \quad \text{(A-16)}$$

$$= \ell(\hat{\boldsymbol{\beta}}^{(t)}) \quad \text{(A-17)}$$

$$(\text{by eq. A} - 10) \geq \hat{\ell}(q,\hat{\boldsymbol{\beta}}^{(t)}) \quad \forall\, q. \quad \text{(A-18)}$$

Because $\hat{\ell}(q,\hat{\boldsymbol{\beta}}^{(t)}) \leq \ell(\hat{\boldsymbol{\beta}}^{(t)})$ for any q, it is clear that the candidate solution $\tilde{q}$ solves the E-step; i.e.,

$$q^{(t+1)} = p(\mathbf{m}|\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)}). \quad \text{(A-19)}$$

Now, coming to the M-step, we can simplify its objective function by dropping all terms that do not depend on $\boldsymbol{\beta}$. Thus, plugging into equation A-12 and employing equation A-19, we can write

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \{\log\, p(\boldsymbol{\beta}) + \mathbb{E}_{p(\mathbf{m}|\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)})}[\log\, p(\mathbf{m},\mathbf{d}|\boldsymbol{\beta})]\}. \quad \text{(A-20)}$$

Because we were able to solve the E-step analytically, the E-M algorithm reduces to iterating the single step given by equation A-20. We do not actually need to compute the Bayesian posterior in the E-step, but need only take the *expectation with respect to it* (which is why the E-step is so named). It can be shown that, under very mild conditions, the E-M algorithm (via iteration of equation A-20) does indeed converge to a (local) maximum of the original marginal MAP problem of equation 27 (McLachlan and Krishnan, 2008).

APPENDIX B

APPLICATION OF E-M TO LSM

We now proceed to apply the E-M algorithm to our LSM problem. For notational convenience, we can rewrite the E-M algorithm of equation A-20 in terms of the E-M objective function $\phi^{(t)}(\boldsymbol{\beta})$ given by

$$\phi^{(t)}(\boldsymbol{\beta}) = \log\, p(\boldsymbol{\beta}) + \mathbb{E}_{p(\mathbf{m}|\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)})}[\log\, p(\mathbf{m},\mathbf{d}|\boldsymbol{\beta})], \quad \text{(B-1)}$$

so the E-M iteration becomes

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \phi^{(t)}(\boldsymbol{\beta}). \quad \text{(B-2)}$$

For every iteration of E-M, we perform the maximization of $\phi^{(t)}$ via a gradient ascent scheme, for which we must compute the gradient of $\phi^{(t)}$.

To derive the exact form of $\phi^{(t)}$ and its gradient, we substitute our distributions into the E-M objective function. From equation 24, we have

$$\log\, p(\boldsymbol{\beta}) = \begin{cases} 0 & \text{if } \beta_{ij} \in [0,1], \quad \forall\, (i,j) \in E \\ -\infty & \text{otherwise} \end{cases}, \quad \text{(B-3)}$$

which means the prior PDF on $\boldsymbol{\beta}$ restricts us to consider only $\beta_{ij} \in [0,1]$. From equations 19 and 20, we have

$$\log\, p(\mathbf{m},\mathbf{d}|\boldsymbol{\beta}) = \frac{1}{2}\{\log\, \det\, [\lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})] - \mathbf{m}^T[\lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})]\mathbf{m} - (\mathbf{d} - \mathbf{A}\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\mathbf{d} - \mathbf{A}\mathbf{m})\} - Z, \quad \text{(B-4)}$$

where Z is a normalization constant given by

$$Z = \frac{(N+K)\log\, 2\pi + \log\, \Sigma}{2}. \quad \text{(B-5)}$$

Inserting these into equation B-1 yields (when every $\beta_{ij} \in [0,1]$)

$$\phi^{(t)}(\boldsymbol{\beta}) = \frac{1}{2}\mathbb{E}_{p(\mathbf{m}|\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)})}\{\log\, \det\, [\lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})] - \mathbf{m}^T[\lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})]\mathbf{m} - (d - \mathbf{A}\mathbf{m})^T\Sigma^{-1}(\mathbf{d} - \mathbf{A}\mathbf{m})\} - Z. \quad \text{(B-6)}$$

The log determinant term in equation B-6 only depends on $\boldsymbol{\beta}$ and is not affected by the expectation with respect to $\mathbf{m}$. Now, we can rewrite the second term in the above expectation as

$$\mathbf{m}^T\{\lambda[\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I}]\}\mathbf{m} = \lambda \text{tr}\{[\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I}]\mathbf{m}\mathbf{m}^T\}, \quad \text{(B-7)}$$

so

$$\mathbb{E}_{p(\mathbf{m}|\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)})}\{\mathbf{m}^T[\lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})]\mathbf{m}\} = \lambda \text{tr}\{[\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I}]\mathbb{E}_{p(\mathbf{m}|\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)})}[\mathbf{m}\mathbf{m}^T]\}. \quad \text{(B-8)}$$

The expected value on the right side of equation B-8 is just the noncentral second moment matrix of $\mathbf{m}$, as determined by the posterior PDF $p(\mathbf{m}|\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)})$, given by

$$\mathbb{E}_{p(\mathbf{m}|\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)})}[\mathbf{m}\mathbf{m}^T] = \boldsymbol{\Lambda}^{(t)} + \mu^{(t)}\mu^{(t)T}, \quad \text{(B-9)}$$

and where $\mu^{(t)}$ and $\boldsymbol{\Lambda}^{(t)}$ are the posterior mean and covariance matrix, respectively, when conditioning on $\mathbf{d}$ and $\hat{\boldsymbol{\beta}}^{(t)}$, given by

$$\mu^{(t)} = \{\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} + \lambda[\mathbf{D}(\hat{\boldsymbol{\beta}}^{(t)}) + \epsilon\mathbf{I}]\}^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{d} \quad \text{(B-10)}$$

and

$$\boldsymbol{\Lambda}^{(t)} = \{\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} + \lambda[\mathbf{D}(\hat{\boldsymbol{\beta}}^{(t)}) + \epsilon\mathbf{I}]\}^{-1}. \quad \text{(B-11)}$$

We further note that the $\epsilon\mathbf{I}\mathbb{E}[\mathbf{m}\mathbf{m}^T]$ term in equation B-8 does not depend on the variable $\boldsymbol{\beta}$, which is being optimized and hence can be dropped from the E-M objective function $\phi^{(t)}(\boldsymbol{\beta})$. Similarly, the third and fourth terms in equation B-6 also do not depend on $\boldsymbol{\beta}$ and can be neglected. Combining the above and rearranging terms, we can rewrite the E-M objective function as

$$\phi^{(t)}(\boldsymbol{\beta}) = \frac{1}{2}\{\log\ \det\ [\lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})] - \lambda \mathrm{tr}[\mathbf{D}(\boldsymbol{\beta})\boldsymbol{\Lambda}^{(t)}] - \lambda\mu^{(t)T}\mathbf{D}(\boldsymbol{\beta})\mu^{(t)}\}. \tag{B-12}$$

To compute $\nabla\phi^{(t)}$, the gradient of $\phi^{(t)}$ with respect to $\boldsymbol{\beta}$, we first note that the $\boldsymbol{\beta}$-weighted graph Laplacian matrix $\mathbf{D}(\boldsymbol{\beta})$ is a linear function of $\boldsymbol{\beta}$, particularly

$$\mathbf{D}(\boldsymbol{\beta}) = \sum_{(i,j)\in\mathcal{E}} \beta_{ij} P^{ij}, \tag{B-13}$$

where the entries of P^{ij} are

$$P^{ij}_{kl} = \begin{cases} 1 & \text{if } kl = ii \text{ or } jj \\ -1 & \text{if } kl = ij \text{ or } ji \\ 0 & \text{otherwise} \end{cases}. \tag{B-14}$$

We also note that $\frac{\partial}{\partial\beta_{ij}} \log\det\{\lambda[\mathbf{D}(\boldsymbol{\beta})+\epsilon\mathbf{I}]\} = \mathrm{tr}\{[\lambda(\mathbf{D}(\boldsymbol{\beta})+\epsilon\mathbf{I})]^{-1} \frac{\partial[\lambda(\mathbf{D}(\boldsymbol{\beta})+\epsilon\mathbf{I})]}{\partial\beta ij}\}$.

Letting $\mathbf{C}(\boldsymbol{\beta}) = \{\lambda[\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I}]\}^{-1}$ denote the prior covariance matrix of the image (when conditioning on $\boldsymbol{\beta}$), to compute $\nabla\phi^{(t)}$, we have

$$\frac{\partial}{\partial\beta_{ij}}\phi^{(t)}(\boldsymbol{\beta}) = \frac{\lambda}{2}\{\mathrm{tr}[\mathbf{C}(\boldsymbol{\beta})P^{ij}] - \mathrm{tr}[\boldsymbol{\Lambda}^{(t)}P^{ij}] - \mu^{(t)T}P^{ij}\mu^{(t)}\} \tag{B-15}$$

$$= \frac{\lambda}{2}[C(\boldsymbol{\beta})_{ii} + C(\boldsymbol{\beta})_{jj} - 2C(\boldsymbol{\beta})_{ij} - (\Lambda^{(t)}_{ii} + \Lambda^{(t)}_{jj} - 2\Lambda^{(t)}_{ij}) - (\mu^{(t)}_i - \mu^{(t)}_j)^2]. \tag{B-16}$$

We constrain each β_{ij} to the interval $[0, 1]$ by introducing proxy variables γ_{ij} that we map to the β_{ij} using a sigmoidal function. In particular, we set

$$\beta_{ij} = \frac{\arctan(\gamma_{ij})}{\pi} + \frac{1}{2}, \tag{B-17}$$

so that although γ_{ij} is free to take any value in $\mathbb{R}$, β_{ij} remains within $[0, 1]$. We can then compute $\nabla\phi^{(t)}(\boldsymbol{\gamma})$, the gradient of $\phi^{(t)}$ with respect to $\boldsymbol{\gamma}$, by

$$\frac{\partial}{\partial\gamma_{ij}}\phi^{(t)}(\boldsymbol{\gamma}) = \frac{\partial\phi^{(t)}}{\partial\beta_{ij}}\frac{\partial\beta_{ij}}{\partial\gamma_{ij}} \tag{B-18}$$

$$= \frac{1}{\pi(1+\gamma^2_{ij})}\frac{\partial\phi^{(t)}}{\partial\beta_{ij}}. \tag{B-19}$$

Unfortunately, direct computation of the gradient would require matrix inversions to compute the prior and posterior model covariance matrices. To avoid this, noting that we only need the node- and edgewise elements of these covariance matrices, we develop approximate techniques for obtaining these quantities in Appendices C and D.

APPENDIX C

PERTURBATION-OPTIMIZATION SAMPLING OF GAUSSIAN DISTRIBUTIONS

Because application of the prior model covariance matrix is relatively cheap (as we will discuss), we can estimate its elements by sampling from its associated Gaussian probability distribution and approximate these elements from the samples. Thus, to approximate the prior covariance matrix $\mathbf{C}(\boldsymbol{\beta})$, we generate L samples $\mathbf{m}^{(1)}, \ldots, \mathbf{m}^{(L)}$, of the underlying Gaussian prior PDF of $\mathbf{m}|\boldsymbol{\beta}$ and approximate $\mathbf{C}$ as

$$\mathbf{C}(\boldsymbol{\beta}) \approx \frac{1}{L}\sum_{\ell=1}^{L}\mathbf{m}^{(\ell)}\mathbf{m}^{(\ell)T}. \tag{C-1}$$

To sample from $\mathcal{N}(\mathbf{0}, \mathbf{C})$, we first note that the precision matrix $\mathbf{Q} = \lambda(\mathbf{D}(\boldsymbol{\beta}) + \epsilon\mathbf{I})$ can be rewritten as

$$\mathbf{Q} = \lambda[F^T B(\boldsymbol{\beta})F + \epsilon\mathbf{I}], \tag{C-2}$$

where F is a first-differencing matrix (having a number of rows equal to $|\mathcal{E}|$, the number of edges in $\mathcal{E}$, and a number of columns equal to N, the number of image parameters) and $B(\boldsymbol{\beta})$ is an $|\mathcal{E}|$-×-$|\mathcal{E}|$ diagonal matrix, with the β_{ij} on its diagonal. Referred to as perturbation-optimization (P-O) sampling by Orieux et al. (2012), a straightforward sampling algorithm (that avoids the need for Cholesky factorization of the precision matrix) is available when the precision matrix can be expressed in the form

$$\mathbf{Q} = \sum_{t=1}^{T} M_t^T R_t^{-1} M_t \tag{C-3}$$

and sampling from $\mathcal{N}(\mathbf{0}, R_t)$ is feasible (which is certainly true in our case because we have diagonal R_t matrices). The sampling algorithm given in algorithm 1.

The proof that $\hat{\mathbf{m}}$ is a sample from $\mathcal{N}(0, \mathbf{Q}^{-1})$ is straightforward and given in Orieux et al. (2012). The optimization step requires solving the linear system

Algorithm 1. Perturbation-optimization algorithm for sampling from $\mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1})$ (Orieux et al., 2012).

1) Perturbation step: Generate independent vectors $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, R_t)$ for $t = 1, \ldots, T$
2) Optimization step: Compute $\hat{\mathbf{m}}$ as the minimizer of $J(\mathbf{m}) = \sum_{t=1}^{T}(\boldsymbol{\eta}_t - M_t\mathbf{m})^T R_t^{-1}(\boldsymbol{\eta}_t - M_t\mathbf{m})$

Return $\hat{\mathbf{m}}$ as the sample from $\mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1})$.

$$\mathbf{Q}\hat{\mathbf{m}} = \sum_{t=1}^{T} M_t^T R_t^{-1} \boldsymbol{\eta}_t, \tag{C-4}$$

which, in our case, is very fast ($\mathcal{O}(kN)$ using an iterative solver with k steps) due to the sparsity of F.

APPENDIX D

BLOCK DIAGONAL APPROXIMATIONS

Although the sampling approach of Appendix C can also be used to approximate the elements of the posterior covariance matrix $\boldsymbol{\Lambda}^{(t)}$, in practice, generating a reasonably large number of samples from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{(t)})$ is not feasible due to the increased cost of solving a system involving the posterior precision matrix $\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} + \lambda[\mathbf{D}(\hat{\boldsymbol{\beta}}^{(t)}) + \epsilon\mathbf{I}]$ (we would need to perform a regularized LSM inversion for *each sample* when using the P-O approach).

To estimate the node and edgewise elements of $\Lambda^{(t)}$, we note that when using the Kirchhoff operator $\mathbf{A}$, there is a closed-form expression for the elements of the posterior precision matrix $\boldsymbol{\Lambda}^{(t)-1}$ (combining equations B-11 and 4). With this in mind, we can estimate elements of $\boldsymbol{\Lambda}^{(t)}$ by considering a block diagonal approximation to the precision matrix. In particular, we can construct an M-×-M partition of the posterior precision matrix corresponding to an image point and its $M-1$ nearest neighbors in space within some radius (we used a 49-pixel neighborhood to perform this approximation), then we approximate the covariance matrix at image point i, $\boldsymbol{\Lambda}_{ii}^{(t)}$, from the inverse of this M-×-M partition matrix. The off-diagonal elements $\boldsymbol{\Lambda}_{ij}^{(t)}$ (for each edge $(i,j) \in \mathcal{E}$) are similarly estimated from the same matrix inverse by taking the elements corresponding to covariance between m_i and m_j (however, care must be taken to ensure that the M-×-M partition of the precision matrix is large enough to sufficiently "surround" the image point i and all its neighbors j with which it shares an edge). This approximation will perform reasonably well as long as the posterior precision matrix decays spatially (in the model domain) as we move away from the diagonal (as is the case here).

Algorithm 2. Expectation-maximization algorithm for least-squares migration.

Initialize each $\gamma_{ij}^{(0)} = 0$, so that $\hat{\beta}_{ij}^{(0)} = 0.5$.

Specify step-size α for gradient ascent.

Set $t = 0$. Iterate on t:

1) Compute $\boldsymbol{\mu}^{(t)}$ via equation B-10.

2) Compute $\boldsymbol{\Lambda}_{ii}^{(t)}$ $(\forall\, i \in \mathcal{V})$ and $\boldsymbol{\Lambda}_{ij}^{(t)}$ $(\forall (i,j) \in \mathcal{E})$ via the block diagonal approximation.

3) Initialize $\tilde{\boldsymbol{\gamma}}^{(0)} = \gamma^{(t)}$. Set $s = 0$ and iterate on s to perform gradient ascent on $\boldsymbol{\gamma}$:

   a) Generate samples from $\mathcal{N}\{\mathbf{0}, \mathbf{C}[\boldsymbol{\beta}(\tilde{\boldsymbol{\gamma}}^{(s)})]\}$ via P-O sampling.

   b) Estimate $C[\boldsymbol{\beta}(\tilde{\boldsymbol{\gamma}}^{(s)})]_{ii}$ $(\forall\, i \in \mathcal{V})$ and $C[\boldsymbol{\beta}(\tilde{\boldsymbol{\gamma}}^{(s)})]_{ij}$ $(\forall (i,j) \in \mathcal{E})$ via equation C-1.

   c) Compute $\nabla\phi^{(t)}(\tilde{\boldsymbol{\gamma}}^{(s)})$ via equation B-19.

   d) Update $\tilde{\boldsymbol{\gamma}}^{(s+1)} = \tilde{\boldsymbol{\gamma}}^{(s)} + \alpha\nabla\phi^{(t)}(\tilde{\boldsymbol{\gamma}}^{(s)})$.

4) Update $\boldsymbol{\gamma}^{(t+1)} = \tilde{\boldsymbol{\gamma}}^{(s+1)}$.

5) Update $\hat{\boldsymbol{\beta}}^{(t+1)}$ via equation B-17 using $\boldsymbol{\gamma}^{(t+1)}$.

Upon termination, return:

$\boldsymbol{\beta}_{\text{MAP}} = \hat{\boldsymbol{\beta}}^{(t+1)}$,

$\mathbf{m}_{\text{EB}} = \{\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} + \lambda[\mathbf{D}(\boldsymbol{\beta}_{\text{MAP}}) + \epsilon\mathbf{I}]\}^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{d}$.

APPENDIX E

SUMMARY OF E-M ALGORITHM

To implement the approximations of Appendices C and D to calculate $\nabla\phi^{(t)}$, we need to approximate the entries of $\boldsymbol{\Lambda}^{(t)}$ only once per E-M iteration (because $\boldsymbol{\Lambda}^{(t)}$ does not vary with $\boldsymbol{\beta}$). However, we would need to reapproximate the entries of $\mathbf{C}(\boldsymbol{\beta})$ with the sampling algorithm in each iteration of the first-order gradient-ascent method (which must be rerun in each iteration of the E-M algorithm). We now summarize our above developments for applying the E-M algorithm to obtain the empirical Bayes estimate of the image in LSM in the algorithm 2.

NOMENCLATURE

$m(\mathbf{x})$	=	image/reflectivity parameters (i.e., model parameters) at subsurface points $\mathbf{x}$
$\hat{m}(\mathbf{x})$	=	migrated image at subsurface points $\mathbf{x}$
$\mathbf{m}$	=	discretized model vector
N	=	number of discretized model parameters
$\mathcal{X}$	=	spatial domain of the model
$d_{sr}(t)$	=	measured seismic data (pressure) from source s to receiver r measured at time t
$\hat{d}_{sr}(t)$	=	forward-modeled data from source s to receiver r measured at time t
$\mathbf{d}$	=	discretized data vector
K	=	number of discretized data points
$\tau_{sr}(\mathbf{x})$	=	two-way traveltime field from a source s to subsurface point $\mathbf{x}$ to receiver r
$\theta_{sr}(\mathbf{x})$	=	opening angle for rays traveling from a source s to subsurface point $\mathbf{x}$ and from $\mathbf{x}$ to receiver r
$w_s(t)$	=	source wavelet at source s and time t
$R_s(x), R_r(x)$	=	raypath length from source s or receiver r to subsurface point $\mathbf{x}$
$\mathbf{A}$	=	forward-modeling operator
$\mathbf{m}_{\text{LS}}$	=	unregularized LSM image
$\mathbf{m}_{\text{RLS}}$	=	regularized LSM image
$\boldsymbol{\beta}$	=	vector of spatially varying edge strengths/smoothness parameters ($\beta_{ij} \in [0,1]$ indicates how strongly to penalize differences between m_i and m_j)
$\mathcal{V}$	=	the vertex set of the MRF on which $\mathbf{m}$ is defined: the set of all indices into $\mathbf{m}$
$\mathcal{E}$	=	the edge set of the MRF on which $\mathbf{m}$ is defined: set of pairs of image parameter indices (i,j) indicating where smoothness constraints are allowed (i.e., the set of pairs of indices in which the β_{ij} are defined)
$\mathbf{D}(\boldsymbol{\beta})$	=	differencing matrix defined by $\boldsymbol{\beta}$

λ	=	overall model smoothness regularization parameter: assigns maximal weight given to penalizing differences in model parameters
ϵ	=	regularization parameter that weights penalty on the model norm
$p(\cdot)$	=	PDF
$\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$	=	Gaussian PDF with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$
$\mathbf{C}$	=	prior model covariance matrix (given the regularization parameters)
$\mathbf{Q}$	=	prior model precision matrix (inverse covariance matrix)
$\boldsymbol{\Sigma}$	=	noise covariance matrix
$\boldsymbol{\mu}_{\text{post}}$	=	posterior mean of model
$\boldsymbol{\Lambda}_{\text{post}}$	=	posterior model covariance matrix
$\mathbf{m}_{\text{MAP}}$	=	MAP estimate of model in the nonhierarchical framework
$\mathcal{B}, \mathcal{M}$	=	space of admissible edge strength $\boldsymbol{\beta}$ and model $\mathbf{m}$ vectors
$\mathbf{m}_{\text{HB}}$	=	hierarchical Bayes MAP estimate of model
$\mathbf{m}_{\text{EB}}$	=	empirical Bayes MAP estimate of model
$\boldsymbol{\beta}_{\text{MAP}}$	=	MAP estimate for edge strengths/smoothness parameters $\boldsymbol{\beta}$
$\ell(\boldsymbol{\beta})$	=	log-likelihood function for $\boldsymbol{\beta}$ (maximized to get $\boldsymbol{\beta}_{\text{MAP}}$)
$\hat{\ell}(q, \boldsymbol{\beta})$	=	lower bound on log-likelihood function that is maximized by the E-M algorithm
$\mathbb{E}_p$	=	expectation operator with respect to the distribution p (the subscript is omitted when the distribution is clear from the context)
$q(\mathbf{m}\|\mathbf{d})$	=	proxy model PDF that is maximized for (along with $\boldsymbol{\beta}$) in the E-M algorithm
$\phi^{(t)}(\boldsymbol{\beta})$	=	objective function for $\boldsymbol{\beta}$ at the tth iteration of the E-M algorithm
$\hat{\boldsymbol{\beta}}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Lambda}^{(t)}$	=	tth iterates of $\boldsymbol{\beta}$ and the posterior mean and covariance matrix (given $\hat{\boldsymbol{\beta}}^{(t)}$) computed in the E-M algorithm
γ_{ij}	=	proxy variables for the β_{ij} used to map β_{ij} to $[0, 1]$
P^{ij}	=	partition of the differencing matrix $\mathbf{D}(\boldsymbol{\beta})$ corresponding to β_{ij}
$F, B(\boldsymbol{\beta})$	=	first-differencing and diagonal matrices used to expand $\mathbf{D}(\boldsymbol{\beta})$
M_t, R_t	=	factors used to expand the precision matrix $\mathbf{Q}$ in the P-O sampling algorithm
η_t	=	samples of $\mathcal{N}(0, R_t)$ generated during the P-O sampling algorithm
$J(\mathbf{m})$	=	cost function to be minimized in the P-O sampling algorithm.

REFERENCES

Bishop, C. M., 2006, Pattern recognition and machine learning: Springer-Verlag, Information Science and Statistics.

Bleistein, N., 1984, Mathematical methods for wave phenomena: Academic Press, Computer Science and Applied Mathematics.

Bodin, T., 2010, Transdimensional approaches to geophysical inverse problems: Ph.D. thesis, The Australian National University.

Bodin, T., M. Sambridge, N. Rawlinson, and P. Arroucau, 2012, Transdimensional tomography with unknown data noise: Geophysical Journal International, **189**, 1536–1556, doi: 10.1111/j.1365-246X.2012.05414.x.

Buland, A., and H. Omre, 2003, Joint AVO inversion, wavelet estimation and noise-level estimation using a spatially coupled hierarchical Bayesian model: Geophysical Prospecting, **51**, 531–550, doi: 10.1046/j.1365-2478.2003.00390.x.

Claerbout, J., 1992, Earth soundings analysis: Processing versus inversion: Blackwell Scientific/Stanford Exploration Project.

Clapp, M., 2005, Imaging under salt: Illumination compensation by regularized inversion: Ph.D. thesis, Stanford University.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm: Journal of the Royal Statistical Society Series B (Methodological), **39**, 1–38.

Duquet, B., K. J. Marfurt, and J. A. Dellinger, 2000, Kirchhoff modeling, inversion for reflectivity, and subsurface illumination: Geophysics, **65**, 1195–1209, doi: 10.1190/1.1444812.

Hanitzsch, C., J. Schleicher, and P. Hubral, 1994, True-amplitude migration of 2D synthetic data: Geophysical Prospecting, **42**, 445–462, doi: 10.1111/j.1365-2478.1994.tb00220.x.

Koller, D., and N. Friedman, 2009, Probabilistic graphical models: Principles and techniques: MIT Press.

Kühl, H., and M. D. Sacchi, 2003, Least-squares wave-equation migration for AVP/AVA inversion: Geophysics, **68**, 262–273, doi: 10.1190/1.1543212.

Lambare, G., J. Virieux, R. Madariaga, and S. Jin, 1992, Iterative asymptotic inversion in the acoustic approximation: Geophysics, **57**, 1138–1154, doi: 10.1190/1.1443328.

LeBras, R., and R. W. Clayton, 1988, An iterative inversion of back-scattered acoustic waves: Geophysics, **53**, 501–508, doi: 10.1190/1.1442481.

Malinverno, A., 2000, A Bayesian criterion for simplicity in inverse problem parametrization: Geophysical Journal International, **140**, 267–285, doi: 10.1046/j.1365-246x.2000.00008.x.

Malinverno, A., 2002, Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem: Geophysical Journal International, **151**, 675–688, doi: 10.1046/j.1365-246X.2002.01847.x.

Malinverno, A., and V. A. Briggs, 2004, Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes: Geophysics, **69**, 1005–1016, doi: 10.1190/1.1778243.

McLachlan, G. J., and T. Krishnan, 2008, The EM algorithm and extensions, 2nd ed.: Wiley, Probability and Statistics.

Nemeth, T., C. Wu, and G. T. Schuster, 1999, Least-squares migration of incomplete reflection data: Geophysics, **64**, 208–221, doi: 10.1190/1.1444517.

Orieux, F., O. Feron, and J. F. Giovannelli, 2012, Sampling high-dimensional Gaussian distributions for general linear inverse problems: IEEE Signal Processing Letters, **19**, 251–254, doi: 10.1109/LSP.2012.2189104.

Ray, A., D. L. Alumbaugh, M. Hoversten, and K. Key, 2013, Robust and accelerated Bayesian inversion of marine controlled-source electromagnetic data using parallel tempering: Geophysics, **78**, no. 6, E271–E280, doi: 10.1190/geo2013-0128.1.

Ray, A., and K. Key, 2012, Bayesian inversion of marine CSEM data with a trans-dimensional self parametrizing algorithm: Geophysical Journal International, **191**, 1135–1151, doi: 10.1111/j.1365-246X.2012.05677.x.