GEOPHYSICS, VOL. 75, NO. 6 (NOVEMBER-DECEMBER 2010); P. F151–F171, 17 FIGS.
10.1190/1.3496902

Resolution and uncertainty in 1D CSEM inversion: A Bayesian approach and open-source implementation

James Gunning[1], Michael E. Glinsky[2], and John Hedditch[3]

ABSTRACT

Resolution and uncertainty in controlled-source electromagnetic (CSEM) inversion is most naturally approached using a Bayesian framework. Resolution can be inferred by hierarchical models with free parameters for effective correlation lengths ("Bayesian smoothing"), or model-choice frameworks applied to variable resolution spatial models (Bayesian splitting/merging). Typical 1D CSEM data can be modeled with quite parsimonious models, typically O(10) parameters per common midpoint. Efficient optimizations for the CSEM problem must address the challenges of poor scaling, strong nonlinearity, multimodality and the necessity of bound constraints. The posterior parameter uncertainties are frequently controlled by the nonlinearity, and linearized approaches to uncertainty usually are very poor. In Markov Chain Monte Carlo (MCMC) approaches, the nonlinearity and poor scaling make good mixing hard to achieve. A novel, approximate frequentist method we call the Bayesianized parametric bootstrap (sometimes called randomized maximum likelihood) is much more efficient than MCMC in this problem, considerably better than linearized analysis but tends to modestly overstate uncertainties. The software that implements these ideas for the 1D CSEM problem is made available under an open-source license agreement.

INTRODUCTION

In recent years, controlled-source electromagnetic (CSEM) or seabed logging (SBL) techniques have become a popular element of the hydrocarbon exploration toolkit. This method is designed to detect resistive anomalies in the marine subsurface which might be caused by hydrocarbon accumulations. Taken in conjunction with seismic data for geological and structural delineation, this is potentially a powerful discriminator between high- and low-gas saturations, because gas saturation controls resistivity in a far more linear fashion than it does seismic reflectivity in AVO studies. The CSEM technique is most useful when sufficient geological knowledge is available to exclude lithological causes of high-resistivity near anomalous zones, such as sequences of evaporites, volcanics, or carbonates.

Many articles have appeared outlining the general nature of the CSEM acquisition framework (Constable (2006), Constable and Srnka (2007), Tompkins and Srnka (2007). There are practical limitations on the suitability of the technique originating in basic physics principles, such as the impact of the airwave in shallower waters, the limitations on depth of penetration and detectability imposed by absorption, and the thermal noise of the transmitter-receiver system, frequency content restrictions from skin depths, etc. A large number of offshore petroleum prospects still fall within the domain of applicability of the technique.

In our view, two overriding factors limit the usefulness of the technique. First is that deeper penetrations require low frequencies, and the diffusive energy fronts do not justify sounding arrays with spacings very much smaller than the depth of interest, which automatically limits resolution. Second, the dynamic range of conductivity from seawater to resistive anomalies or deeper rocks usually is at least several decades. These large contrasts in resistivity make the changes in observed fields large and thus useful in an exploration context, but make the inverse problem very nonlinear. In an inverse-problem context, the subsurface response is very poorly modeled as a weak deviation from some agnostic reference model, so the Born approximation, so beloved and central to seismic imaging, is rarely very useful for real CSEM data. Because the forward model is strongly nonlinear in any resistivity parameter, the solution null space nearly always is multimodal, badly scaled, and contorted in shape. We agree strongly with Snieder (1998) that this makes these problems particularly difficult.

Meaningful 2D or 3D CSEM inversion thus is a difficult problem. The strong absorption induces a large dynamic range in the gradient

Manuscript received by the Editor 7 October 2009; revised manuscript received 8 July 2010; published online 11 November 2010.
[1]CSIRO CESRE, Ian Wark Laboratory, Bayview Ave., Clayton, Victoria, Australia. E-mail: James.Gunning@csiro.au.
[2]CSIRO CESRE, Kensington, Australia. E-mail: glinsky@qitech.biz.
[3]Google, 48 Pirrama Rd., Pyrmont, NSW, Australia. E-mail: j.hedditch@gmail.com.

or sensitivity matrices, and because this makes the problem very poorly scaled, nearly all inverse approaches require additional terms to improve the stability or conditioning of the matrices. For diverse reasons, the bulk of the inverse-theoretical work done in the EM community is not overtly statistical in nature, but rather approaches the stability problem using pragmatic Tikhonov-regularization methods. This introduces the awkward problem of how to estimate the free parameters in these regularizing operators and make meaningful statements about what these pieces imply about model resolution and uncertainty. Regularization, in our view, is an unsatisfactory framework for the problem of integrating other kinds of information, like rock-physics models, or data from seismic acquisition.

Most of these conceptual difficulties disappear if a more explicitly statistical approach is taken to the inverse problem. Evans and Stark (2002) put the case eloquently: "Describing inverse problems in statistical language permits a unified view of standard inversion techniques, and provides reasonable criteria for choosing among them." Sambridge et al. (2006) provide a theoretical framework strongly aligned with ours and offer a useful summary of the Bayesian approach to inverse problems and model selection.

Bayesian frameworks allow inverse problems to be stated as inference problems for the posterior distribution of a suite of model parameters and possible metaparameters, and questions about resolution or uncertainty are answerable directly from this posterior distribution. Two recent geophysical examples using empirical-Bayes ideas for metaparameter estimation are Malinverno and Parker (2006) and Mitsuhata (2004). Because such statements are conditional on the chosen model, a framework that enables sensible comparison of different models, or families of models — even of varying dimensionality — is very desirable (Hoeting et al., 1999). Bayesian approaches also are the most natural way to introduce knowledge from other data sources or professional expertise with its requisite precision and interdependencies, via additional likelihood terms or priors. Multidisciplinary information of this form is germane to earth resources delineation.

In this paper we show two new Bayesian approaches to the question of resolution and uncertainty for the CSEM problem and introduce the open-source reference code DeliveryCSEM, implementing these ideas for the 1D problem. This paper and the code implementation are confined to the isotropic case, though it is now recognized that modest electrical anisotropy is more common than not. The central ideas of this paper will extend readily to the anisotropic case, and the presentation is simplified when we need not carry the tensorial notational baggage along.

We do not wish to be misconstrued as advocating isotropic 1D inversions for problems that are clearly dominated by 3D effects or other forward-modeling issues. Nonetheless, for reasonably flat geometries without significant bathymetry issues, the 1D approach is a good first approximation to the 3D earth. Much can be learned about the limits of resolution and inversion uncertainties by a successful attack on the 1D problem.

We do not focus on the virtues or drawbacks of acquisitional details such as the number of frequencies to be measured, types of fields to be recorded, use of phase, or other similar details. Other papers, for example Key (2009), take up these issues. Our central themes are resolution and uncertainty via Bayesian approaches, so the bulk of this paper is devoted to these topics. The novel contributions of this paper are the application of model-selection, empirical-Bayes, and Bayesianized bootstrap ideas to CSEM applications.

The layout of this paper is as follows. In "Approaches to resolution issues," we introduce the central ideas needed for Bayesian approaches to resolution inference. In "Constrained Bayesian inversion," we present the machinery needed for resolution approaches based on variable correlated priors on a fixed grid. "Model hierarchies — splitting methods," shows how this machinery can be used to infer resolution via model choice, with the spatial correlations switched off and the Bayesian model choice operating over models of varying spatial discretization. The fundamental workhorse in both approaches is an efficient globalized, bound-constrained nonlinear least-squares optimization, so we visit several important topics in "Optimization details:" efficient bound-constrained Gauss-Newton and Marquardt techniques, multimodality and global optimization/enumeration, and mode distinguishability or connectivity. Two methods for uncertainty evaluation follow in the section titled "Approaches to inversion uncertainty," one fully Bayesian (MCMC), the other a faster, approximate technique we call the Bayesianized parametric bootstrap. Some "Example problems" are presented to illustrate all the various ideas, followed by a brief discussion of the "Software," and the "Conclusions."

APPROACHES TO RESOLUTION ISSUES

Resolution is most effectively understood as an interaction between the spatial representation (gridding) of an inversion model and the effective number of degrees of freedom which can be estimated meaningfully from the data. From this angle, there are two distinct approaches to resolution. First, if a somewhat fine spatial model m is supplemented by well-chosen metaparameters θ expressing effective spatial correlation, the resolution is embodied in the marginal distribution for the correlation parameters θ given the data. Overfitted, or excessively deconvolved, models correspond to low-probability regions of the correlation-parameter posterior marginal distribution(s). Second, resolution can be approached as a model-selection problem of choosing, among a family of models $k = 1 \ldots N$ of varying spatial discretization, the model or models having most posterior support in the data. Clearly the measure of support implied here must incorporate automatic penalties for overfitting, so the statistical significance of the models is the central issue.

Both of these approaches can be expressed in a Bayesian framework. We use the usual notation $L(d|m)$ for the likelihood of the data d (length n_d) under model m, and $p(m)$ for the prior probability of the parameters in model m. The likelihood often is the most contentious part of any Bayesian framework. It depends centrally on a model for the "effective" noise, which is defined as the difference between modeled and (processed) data. This difference clearly absorbs instrumental noise, external and cultural noise, and errors in the forward modeling assumptions. Rarely is it beyond dispute that the computer model adequately models the physics. One often works with the pragmatic assumption that the data are well-processed (mistakes/outliers removed etc.), the errors are zero-mean independent, and the dominant unknown is the variance of the error. For analytical convenience, Gaussian error models are most useful, so the likelihood often is of the form $L(d/m) \sim \exp\left(-\frac{1}{2}(d - f(m))^T C_D^{-1}(d - f(m))\right)$, with $f(m)$ the forward model for the data, and the unknown noise parameters σ_i buried in the matrix $C_D = \mathrm{diag}\{\sigma_i^2\}$. Some kind of dilution of this likelihood distribution might be required to model correlated or biased data.

To provide some context, the 1D forward CSEM problem considered here is a layer-based model, usually with the transmitter close (≈ 30 m) to the seafloor, receivers for electric or magnetic fields on

the seafloor, known resistivity through the seawater profile, and unknown resistivity in each of some n_{layers} layers under the mudline, terminating in a half-space. The forward problem and sensitivity matrix $\partial f_i/\partial m_j$ for this configuration is a well-studied problem (Constable et al., 1987; Key, 2009), with received fields a simple sum of Hankel transforms with kernels arising from reflectivity recursions running down the stack of layers. The measurements d_i are taken as electric or magnetic fields, unrolled over frequency and transmitter-receiver offset. Typically, the noise estimates σ_i initially are estimated at some fraction of the field amplitude, say 5%, so these have a large dynamic range. (The large range is required by the absorption of modeling errors as much as anything else). The acquisition usually attempts to keep the source dipole a constant height over the seafloor, and this can be used to advantage in splining the fast Hankel transforms in the forward model to retrieve fields at all offsets for a given frequency and transmitter height.

In the model selection problem, the central entity is the marginal model likelihood (MML), or evidence, obtained by integrating the Bayesian posterior density over the model parameters m in model k:

$$\pi_{\text{MML}}(k) = \int L(d|m)p(m)dm.$$

In general, the integral is quite difficult to perform, but approximations like the Laplace approximation are very effective if the posterior is modestly compact (Raftery, 1996). It is known that the Laplace approximation behaves asymptotically like the Bayes information criterion (BIC) (Denison et al., 2002), and thus has the required "Occamist" characteristic of favoring the simplest model that adequately explains the data.

It is less obvious how the notion of "simplicity" is quantified and induced in the context of single models with metaparameters. Although a strict Bayesian would confine the statement of "posterior knowledge" to the full posterior distribution, certain characteristics of this distribution usually are of significant interest as point estimates. In particular (1) the largest mode of the joint posterior distribution — usually called the maximum a posteriori point (MAP), and (2) the MAP point of particular marginal distributions, are of interest. Within the extremely common multiGaussian framework for noise and prior distributions, possibility (1) coincides with the minima of the negative log-posterior, a function which often closely resembles typical "objective" functions used in regularization approaches.

Statisticians of all flavors reflexively associate point-estimates with maxima of probability functions, and maximum-likelihood methods are virtually canonical in the statistical community. Under Gaussian error models, these invariably lead to least-squares minimization problems. For this reason, regularization approaches based on the optimization of penalized objective functions like

$$\chi^2(m,\mu) = \chi^2_{\text{misfit}}(m) + \mu\|Dm\|^2, \qquad (1)$$

where μ is a "free" parameter, and D is an operator whose null space does not overlap that of the forward model in $\chi^2_{\text{misfit}}(m)$, always seem philosophically unsatisfactory because the mathematical optimum clearly is at $\mu = 0$. Statisticians instinctively will feel that something is missing from the "objective function" that favors simplicity (large μ).

A well-known approach to this difficulty is Morozov's discrepancy principle (Hansen, 1998). Assuming the model is rich enough to potentially overfit the data, the multiplier μ can be set by minimizing equation 1 to a desired level of misfit, say, $\chi^2_{\text{misfit}}(m) \approx n_d$. It is difficult to make statements about a strongly nonlinear problem with great confidence, but we might take inspiration from what is known about the linear case: Hansen's discussions are quite extensive, and recommend, roughly, $\chi^2_{\text{misfit}} \approx n_d - n_p$, where there are n_p effective degrees of freedom. Essentially, the target value $n_d - n_p$ is based on the known frequentist result in linear regression that the (error-scaled) residual sum of squares has expectation $n_d - n_p$ if the regression model is the same as that producing the data, and the error variance is correct.

Use of the discrepancy principle is central to the well-known Occam code of Constable et al. (1987), but this framework does not yield a point estimate that is obviously the maximum of some distribution. A common criticism is that the technique is rather sensitive to the noise levels buried inside $\chi^2_{\text{misfit}}(m)$, and in practice these usually are poorly known (Farquharson and Oldenburg, 2004; Mitsuhata, 2004). Pessimistic estimates lead to oversmoothed solutions, and overoptimistic estimates might prevent convergence at all. It is common to see target values $\chi^2_{\text{misfit}} = n_d$ invoked, even for rather rich models, and Hansen has demonstrated this leads to oversmoothing in the linear context.

In a Bayesian approach, maximum likelihood estimation is possible for problems with smoothing contributions (μ), but it is necessary to treat the smoothing parameters as genuine metaparameters in a hierarchical framework. The normalization associated with the metaparameters then introduces the contributions that favor large values of the smoothing and compete with the data misfit terms. A Bayesian approach naturally will induce simplicity in the choice among models and in the inference of metaparameters (e.g., smoothing) within a model, so Occam's razor is a natural consequence. Thus we would see "variable-smoothing" type inversions as a special kind of Bayesian inversion, rather than a different approach. Parker (1994) has remarked that the Occam approach is "…lacking theoretical underpinnings, but… has been found to be remarkably effective in practice." We believe the Bayesian approach described in the following section, using spatial correlation as a metaparameter, supplies this missing theory.

Some known invariances for the 1D CSEM problem are useful to recall at this point. Loseth (2007) has shown that if a subsurface resistive layer is present against a more typical (say 1 Ω.m) conductive background, the dominant mode of energy transmission is a TM mode, with vertical electric field. His analytical approximations for the Hankel transforms show that this response is controlled by the resistivity-thickness product of the anomalous layer. We expect then that the response of a packet of layers thinner than the "natural" data resolution is controlled by the resistivity-thickness product of the effective medium formed by these layers. This forms a useful test case for many of the subsequent ideas.

CONSTRAINED BAYESIAN INVERSION FOR MODEL, NOISE, AND SPATIAL CORRELATION

Our inversion code can perform several flavors of inversion, all of which can be understood as special cases of the following general framework. We are interested in inverting for n_p model parameters $m_i = \log_{10}\rho_i$ (the layer resistivities are ρ_i), jointly with metaparameter-parameters describing spatial correlation structures (μ) or parameters of the noise distribution (σ_n). The full parameter vector is $\mathbf{M} = \{\mathbf{m},\mu,\sigma_n\}$.

A standard Bayesian approach to inversion (Tarantola, 1987),

based on a multiGaussian model of the errors and with a multiGaussian expression for the prior with prior mean $\mathbf{m}_p$ and covariance $C_p(\mu)$, yields a posterior density

$$\Pi(\mathbf{M}|\mathbf{d}) \sim \frac{e^{-(\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d(\sigma_n)^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m}))/2}}{(2\pi)^{n_d/2}|C_d(\sigma_n)|^{1/2}} \times \frac{e^{-(\mathbf{m}-\mathbf{m}_p)^T C_p(\mu)^{-1}(\mathbf{m}-\mathbf{m}_p)/2}}{(2\pi)^{n_p/2}|C_p(\mu)|^{1/2}}. \quad (2)$$

Here n_d is the number of measurements, and we will consider the particular case where $C_d(\sigma_n) = \sigma_n^2\,\mathrm{diag}\{\sigma_i^2\}$, the covariance matrix of the total error, is assumed diagonal and known up to the scalar σ_n^2. Similarly, the unknown metaparameter-parameters μ might appear in $C_p(\mu)$. For normalization and model-comparison purposes, the determinant terms and dependencies on n_p are important.

The first problem is which choice of prior is suitable for a particular model-layer resistivity. Typical CSEM hydrocarbon applications will occur in clastic-dominated areas, where shale abundances might be 80% or so. The model-layer resistivity is an "effective medium" property of a rock composite, whose (frequency) distribution is a complex function of rock-type abundances, the internal spatial arrangement of rock types, the internal variability within a rock type, and the effective-medium laws. In general, we should expect it to be a complex mixture distribution resulting from these factors. A rigorous calculation is doubtless rather subjective, but we can say a few definite things: It will have a heavy right tail, resulting from the lighter abundances of low-porosity facies, and it is reasonable to apply a strict lower bound, computed from the Hashin-Shtrikman lower bound on brine and shale-matrix mixtures via sensible upper-bounds on shale porosity (e.g., 50%). A typical, credible number is $\rho = 0.8\ \Omega\mathrm{m}\ (\log_{10}(\rho) = -0.1)$. A truncated Gaussian distribution for $m = \log_{10}(\rho)$ can be used to cover the prior support comfortably, has these required properties, and the added advantage of analytical convenience. If bounds are not applied, the logarithmic transform retains the advantage of guaranteeing positive resistivities.

Spatial smoothness-type beliefs about the model can be expressed by embedding spatial correlation into the multivariate prior distribution for the model parameters. We will use forms derived for the unbounded cases and impose constraints for the bounded case as required. A convenient form to work with is the Gaussian prior $p(\mathbf{m}) = N(\mathbf{m}_p, C_p)$, where $\mathbf{m}_p$ is a prior mean or prejudice about the subsurface structure. It is reasonable to suppose the prior marginal variance of any layer parameter (as imposed by the mixture distribution approximations above) to be independent of any vertical correlation. Thus it is simpler to specify C_p directly, rather than C_p^{-1}, as the diagonal elements contain the prior marginal variances. Specifically, if there are $i = 1 \ldots n_p$ layer parameters m_i, whose prior marginal standard deviations are set to a common value σ_p, the exponential correlation matrix $C_{p,ij} = \sigma_p^2 \exp(-\alpha|i-j|)$ is a convenient possible form for C_p, with a "lattice" correlation length $1/\alpha$.

To forestall confusion, we emphasize that we will make inferences about an effective correlation length $1/\alpha$ for the large-scale resistivity parameters m, as estimated by CSEM data solely, and not to be confused with correlation lengths inferred, for example, from wireline or core data. Although the correlation length might be argued to be an intrinsic geological property, a Bayes MAP estimate of this effective correlation length suggests the resolution characteristics of the measuring technique used to acquire the data.

Now C_p has a tridiagonal inverse which, for convenient comparison with other literature using the discrepancy principle, can be written in the form

$$C_p^{-1}(\mu) = \mu\partial^T\partial + \mathrm{diag}\{W_{p,1}^2, W_{p,2}^2, W_{p,2}^2, \ldots, W_{p,2}^2, W_{p,1}^2\},$$

where ∂ is the $n_p \times n_p$ finite-difference derivative matrix

$$\partial \equiv \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & \cdots \\ \cdots & \cdots & \ddots & \ddots & \cdots \end{pmatrix},$$

and the correlation length $1/\alpha$ is related to the "regularizing strength" μ by

$$\alpha(\mu) = \sinh^{-1}\left(\frac{1}{2\mu\sigma_p^2}\right).$$

(Further connections of the inverse covariance implied by the regularizing matrix ∂ with geostatistical ideas are drawn out in Kitanidis [1999].) Maintaining the prior standard deviation requires that the weights W_p vary with μ also

$$W_{p,1}^2 = \frac{1}{\sigma_p^2(1+e^{-\alpha})}, \quad (3)$$

$$W_{p,2}^2 = \frac{1-e^{-\alpha}}{\sigma_p^2(1+e^{-\alpha})}. \quad (4)$$

Clearly, α and μ are alternative ways to track the exponentially correlated prior; we will use the parameter μ henceforth as the metaparameter. Thus, if we define $\mathcal{W}_p(\mu) = \mathrm{diag}\{W_{p,1}^2, W_{p,2}^2, W_{p,2}^2, \ldots, W_{p,2}^2, W_{p,1}^2\}$, the inverse then is $C_p^{-1}(\mu) = \mu\partial^T\partial + \mathcal{W}_p(\mu)$.

In the absence of correlation ($\alpha \to \infty$, or $\mu = 0$), the $W_{p,i}$ are related to the prior marginal standard deviation σ_p by $W_{p,i} = 1/\sigma_p$. The determinant $|C_p| = \sigma_p^{2n_p}(1-e^{-2\alpha})^{n_p-1}$, with the property $|C_p| \to 0$ as $\alpha \to 0$, is helpful to know. The question of how to choose a suitable prior distribution for μ is rather tricky. Fortunately, the posterior distribution for μ is only very weakly influenced by the prior, so we use a flat prior on μ for simplicity.

The noise parameter σ_n is a global scalar correction term for the (white) Gaussian noise distribution, and we presume the error estimates σ_i in $C_d(\sigma_n) = \sigma_n^2\,\mathrm{diag}\{\sigma_i^2\}$ are sensible estimates based on preliminary data analysis, e.g., 5% of the expected field amplitude, down to some typical noise floor for the receivers. (Absolute noise floors are dependent on electronics design, possibly electrode chemistry, receiver motion, stacking and processing considerations, etc., and are typically around 10^{-15} V/Am^2 for E fields, 10^{-18} T/Am for B). This absorbs measurement and modeling errors. The additional term σ_n is an $O(1)$ correction parameter, corresponding fairly closely to the "unknown variance" parameter of traditional Bayesian regression treatments, e.g., Gelman et al. (1995). We will take the prior $P(\sigma_n)$ to be flat (constant) for simplicity.

There are two possible approaches to the inference problem at this point: pure maximum a posteriori or empirical Bayes. The general ideas are easier to see in the fully linear problem, which, for reasons of space, we have supplied in the supplementary material Appendix C of Gunning (2010). This material supplies some derivation details we skip in the following. The first and simplest idea is a pure "maximum a posteriori" approach, setting inferences at a global minimum of the negative log posterior of the full joint distribution in $\mathbf{m}, \mu, \sigma_n$.

This objective function in the optimization step could be written (dropping $n_d \log(2\pi)$ and $\log|\mathrm{diag}\{\sigma_i^2\}|$) as

$$\begin{aligned}-2\log(\Pi(\mathbf{m},\mu,\sigma_n|\mathbf{d})) &\equiv \chi^2 = (\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d(\sigma_n)^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m})) \\ &+ n_d\log(\sigma_n^2) + (\mathbf{m}-\mathbf{m}_p)^T(\mu\partial^T\partial+\mathcal{W}_p)(\mathbf{m}-\mathbf{m}_p) \\ &- \log(|\mu\partial^T\partial+\mathcal{W}_p|) + n_p\log(2\pi). \end{aligned} \quad (5)$$

Where the prior has weak influence and the degrees of freedom are few, this is a simple and effective approach. The estimates of μ will be biased if the data are too noisy, however, as shown in the supplementary Appendix C (Gunning, 2010).

The smoothing and noise parameters really are meta-parameter in a hierarchical construction. The empirical Bayes (EB) approach is to estimate these parameters at the maximum likelihood point of their marginal distribution, which is known to be less biased than the joint maximum-a posteriori method. The derivations for the EB case are somewhat messier, so we will show how things run for the joint maximum-a posteriori case first and summarize the EB results later.

In the joint maximum-a posteriori case, we minimize equation 5 by cyclically alternating minimizations on σ_n, μ, and $\mathbf{m}$, which is not inefficient if the three blocks are not strongly correlated in the posterior.[4] The minimization on σ_n involves only the first two terms and is trivially a standard ML variance estimate:

$$\sigma_n^2 = \frac{(\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m}))}{n_d}.$$

Substituting this again into equation 5, and dropping some constants, yields the reduced objective

$$\begin{aligned}\chi_J^2 &= n_d(1+\log((\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m}))/n_d)) \\ &+ (\mathbf{m}-\mathbf{m}_p)^T(\mu\partial^T\partial+\mathcal{W}_p)(\mathbf{m}-\mathbf{m}_p) \\ &- \log(|\mu\partial^T\partial+\mathcal{W}_p|). \end{aligned} \quad (6)$$

The optimization on μ then involves only the last two terms, a problem we can write as

$$\begin{aligned}\chi^2_{\mathrm{smooth}}(\mu) &= (\mathbf{m}-\mathbf{m}_p)^T(\mu\partial^T\partial+\mathcal{W}_p(\mu))(\mathbf{m}-\mathbf{m}_p) \\ &- \log(|\mu\partial^T\partial+\mathcal{W}_p(\mu)|).\end{aligned}$$

The determinant must be evaluated numerically in general (an $O(n_p)$ operation because $\partial^T\partial$ is tridiagonal), and this problem can be solved using any suitable 1D optimization routine, e.g., Brent's method (Press et al., 1992). We have found it prudent to step-limit the optimum found in this phase to within a trust region centered on the current value of μ, typically $\mu \pm 0.5$.

The final optimization in the cycle is for $\mathbf{m}$. For small changes in $\mathbf{m}$ about a current model $\mathbf{m}_0$, by linearizing the log( ) expression, the varying terms in equation 6 needed for the optimization can be written as

$$\begin{aligned}\chi_m^2 &= (\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m}))/\sigma_n^2 \\ &+ (\mathbf{m}-\mathbf{m}_p)^T(\mu\partial^T\partial+\mathcal{W}_p)(\mathbf{m}-\mathbf{m}_p). \end{aligned} \quad (7)$$

The Gauss-Newton step thus is the standard Bayesian update, with the data covariance merely adjusted by the current noise estimate σ_n^2. The full Newton update for this optimum, with the Jacobian $J_{ij} \equiv \partial F_i/\partial m_j$, is

$$\begin{aligned}\mathbf{m}' = &\left(\frac{1}{\sigma_n^2}J^T C_d^{-1}J + \mu\partial^T\partial + \mathcal{W}_p\right)^{-1} \\ &\times\left(\frac{1}{\sigma_n^2}J^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m})+J\mathbf{m}) + (\mu\partial^T\partial+\mathcal{W}_p)\mathbf{m}_p\right).\end{aligned}$$

Another important traditional form for the Newton step $\Delta\mathbf{m} \equiv \mathbf{m}' - \mathbf{m}$ is

$$\begin{aligned}\Delta\mathbf{m} = &\left(\underbrace{\frac{1}{\sigma_n^2}J^T C_d^{-1}J + \mu\partial^T\partial + \mathcal{W}_p}_{H}\right)^{-1} \\ &\times\underbrace{\left(\frac{1}{\sigma_n^2}J^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m})) + (\mu\partial^T\partial+\mathcal{W}_p)(\mathbf{m}_p-\mathbf{m})\right)}_{-\nabla\chi_m^2},\end{aligned} \quad (8)$$

with implied Hessian H and gradient $\nabla\chi_m^2$.

For the cases where no estimation of σ_n is desired, the same formalism applies, except the optimization on σ_n is omitted and $\sigma_n \to 1$ everywhere else. Similarly, if no optimization on μ is performed, μ is fixed at the desired value in all equations.

For the EB case, the derivations follow a similar spirit to supplementary Appendix C in Gunning (2010), save that one uses local linearization and the Laplace approximation in estimating the marginal distribution (marginal) for μ. The mode of the marginal for σ_n is straightforward, yielding the classical unbiased estimate

$$\sigma_n^2 = \frac{(\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m}))}{n_d - n_p},$$

and to a good approximation the marginal $\Pi(\mu,\sigma_n|d)$ for μ has an additional term in the optimization ($\Pi(\mu,\sigma_n|d) \sim \exp(-\chi^2_{\mathrm{smooth}}(\mu)/2)$):

$$\begin{aligned}\chi^2_{\mathrm{smooth}}(\mu) = &(\mathbf{m}-\mathbf{m}_p)^T(\mu\partial^T\partial+\mathcal{W}_p(\mu))(\mathbf{m}-\mathbf{m}_p) \\ &- \log(|\mu\partial^T\partial+\mathcal{W}_p(\mu)|) \\ &+ \log\left(\left|\frac{1}{\sigma_n^2}J^T C_d J + \mu\partial^T\partial+\mathcal{W}_p(\mu)\right|\right).\end{aligned} \quad (9)$$

Clearly, there is nothing particularly magical about the choice of the exponentially correlated prior. We have chosen it because the inverse (the "precision matrix") maps closely to the sorts of structures used in regularization approaches (i.e., the connection and differences are clear), and the determinant is simple. Other choices could be made, and blockwise forms arising from the use of "tear-surfaces" (discontinuities in the correlation) also would pass through the foregoing derivation simply.

Example of resolution via correlation meta-parameters: Bayesian smoothing

An example of how the empirical Bayes apparatus works, for fixed known noise, but unknown correlation parameter μ, is shown

[4]This is a good assumption for σ_n and $\mathbf{m}$ (a well known statistical phenomenon), but probably not for μ and $\mathbf{m}$: a joint Newton scheme would be much better for the latter pair.

in Figure 1. Synthetic data (inline $|E|$ field at 0.25, 0.75, 1.25 Hz, over offsets 1–12 km) for the depicted "truth-case" model are generated with varying noise levels by adding independent Gaussian noise deviates of the required standard deviation (e.g., $0.05|E|$ for 5% errors) to $|E|$. The uneven sampling (dropouts, etc.,) is inherited from a real data set "template," but the model and data all are synthetic. The inversion model is quite finely discretized, using layers of approximately 50 m to 100 m, and the marginal priors for each layer are set at $m_j \sim N(0,1)$.

MODEL HIERARCHIES — SPLITTING METHODS

Another approach to resolution is to perform model-selection on a set of models of increasing spatial resolution. Clearly, an exhaustive enumeration of a full suite of possible layer-grids, using, say, the theory of integer partitions based on some finer underlying lattice, will produce a huge (combinatorially large) number of possible models. These cannot be computed exhaustively, so some kind of heuristic is necessary for exploring model spaces. An obvious idea is some kind of recursive algorithm which either will adaptively refine a very coarse model, or remove detail from a fine model, such that resolution is created at the depths statistically justifiable from the data.

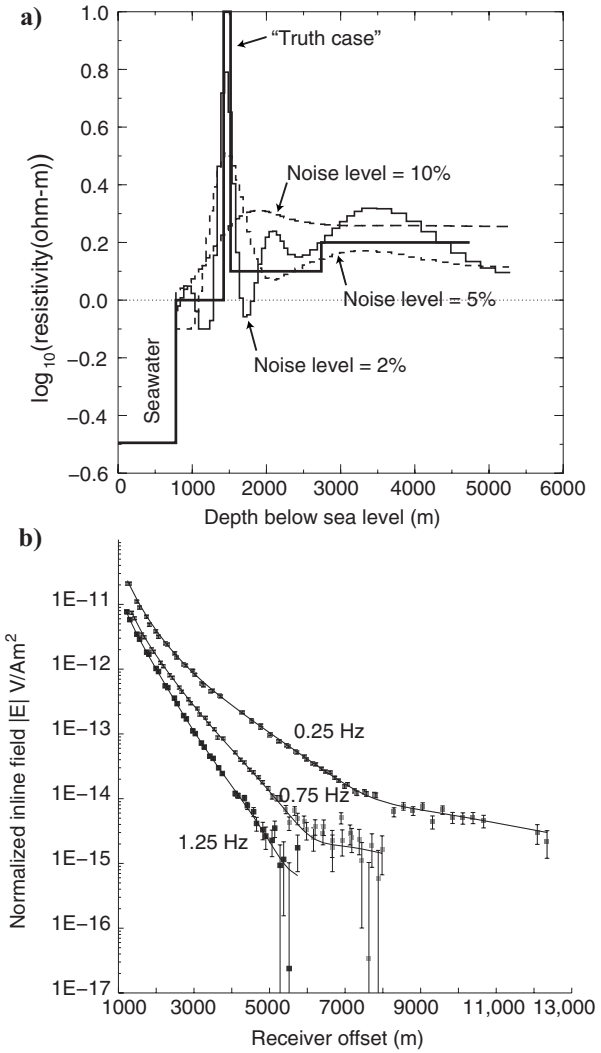

Figure 1. (a) Bayesian smoothing MAP inversions (μ as a meta-parameter) of CSEM data for the "truth-case" model shown, for noise levels 10%, 5%, and 2%. Though the termination at the optimum is not explicitly controlled by $\chi^2_{\mathrm{RMS}} \equiv [(\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m}))/n_d]^{1/2}$, χ^2_{RMS} values are typically $O(1)$ at the optimum; in this case, 1.21, 1.09, 1.02, respectively. Clearly resolution is strongly dependent on noise levels. (b) Typical data and fit at 5% noise. Note the error bars apply to $|E|$, not $\log_{10}|E|$, despite the scales.

We rank models on the basis of the marginal model likelihood (MML), obtained by integrating the Bayesian posterior density over the model parameters. For model k the MML is defined as

$$\pi(k) = \int L(d|m_k)p(m_k)dm_k.$$

The Laplace approximation for the MML (Raftery, 1996), for our CSEM problem, is

$$\begin{aligned}-\log(\pi(k)) = {} & \frac{1}{2}(\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m}))/\sigma_n^2 \\ & + \frac{1}{2}n_d \log(\sigma_n^2) + \frac{1}{2}(\mathbf{m}-\mathbf{m}_p)^T(\mu\partial^T\partial + \mathcal{W}_p) \\ & \times(\mathbf{m}-\mathbf{m}_p) - \frac{1}{2}\log(|\mu\partial^T\partial + \mathcal{W}_p|) \\ & + \frac{n_p}{2}\log(2\pi) + \frac{1}{2}(\log|H|), \end{aligned} \quad (10)$$

with all terms evaluated at the MAP point, the Hessian H as per equation 8, and the smoothing $\mu = 0$.

As a reference implementation, we have adopted a recursive greedy search algorithm based on successive refinement of an initial very coarse model. The algorithm proceeds as follows:

1) Compute the MAP solution and MML for a very coarse, sufficiently deep two-layer model (problem of dimension $n_p = 2$). This becomes the parent model.
2) Loop over all layers in the parent model, split each layer into two by turns to make child models, and invert for the MAP point and MML for each child model (n_p models of dimension $n_p + 1$ each). Record the best solution (favorite child) and best MML.
3) If the best child MML is an improvement on the parent's MML, embed the split, and iterate the process with the best child as the new parent. If no solution is better, terminate the algorithm on the n_p dimensional parent model.

In each case, default starting points for the optimization are obtained by injecting the parent MAP parameter values into the child parameter vector in a way that preserves the existing spatial distribution. Global inversion is very desirable for each candidate model, as superior solutions might not be in the basin of attraction of the starting point inherited from a parent.

These coarse models should require no spatial smoothing between layers, so in all the expressions above, $\partial = 0$ and the $\mathcal{W}_p$ will be calculated from the univariate prior variance.

Example of resolution via model-selection

A standard test problem in the CSEM literature is the canonical model (Constable, 2006), a 100 m-thick, 100 Ωm-reservoir buried 1 km deep in shales under deep water. An example of the evolution

of these split models for the canonical test model is shown in Figure 2.

It is clear that the reference algorithm above will arrive at relatively parsimonious models, but it is not clear that it always terminates at the simplest conceivable model. An alternative, more expensive algorithm based in splitting and merging can achieve the latter; an example is shown in Figure 10 later in the paper.

OPTIMIZATION DETAILS

Projected Newton or Marquardt methods with bound constraints

Experience shows that unconstrained inversion (very wide priors) often produces unphysically low values of resistivity in the shallower layers. Such values might occur not only at the final optimum, but during the optimization phase, and might allow the minimization to wander into an unwanted basin of attraction. Placing a sensible lower bound truncation in the prior distribution cures this problem, but introduces the problem of how to efficiently control the optimization in the presence of such bounds.

For badly scaled problems such as the CSEM problem we address, naive ideas easily can induce slow convergence, so some subtlety and care in implementation is required. We have implemented the projected Newton technique line-search described by Bertsekas (1982) and Kelley (1999) and a projected trust-region (Marquardt) method, adapted from Madsen et al. (2004). The implementation requires some care, so we make this available in Appendix D of Gunning (2010).

When optima occur at parameter boundaries, the Laplace approximation for the marginal model likelihood is certain to be less accurate, as the probability is truncated in at least one parameter. It is difficult to estimate the correction factors necessary, but the approximation will give at least an estimate of the order-of-magnitude of the integral.

Currently, all parameters ($\log_{10}(\rho)$) share the same bounds. Default bounds of $-0.1 < \log_{10}(\rho) < 4$ are applied, the lower corresponding to 0.8 Ωm, a respectable lower bound for shales based on Hashin-Shtrikman effective media theory. The bounds can be disabled or altered if desired.

Globalization — multiple start solutions

Virtually all modes of inversion except either very low-dimensional models or excessively oversmoothed finer models will suffer from multimodality. This is most obvious in dependence on the initial guesses in the optimization runs and algorithm dependence in the solutions found (e.g., the details of the line search). Reasonably rich models with weak smoothing usually have a significant number of local modes; some might be very poor fits, but several might be respectable.

The best strategy for dealing with this is to use models as parsimonious as the purpose of the study permits and attempt to enumerate and quantify as many local modes as possible. The code can be invoked with a suite of strategies, attempting multiple optimization passes at each point in the code where (by default) a single local optimization is performed (in addition to the default local-optimization pass). A variety of strategies is conceivable; we have implemented the following suite. (1) Default (and mandatory): use a starting point determined by the startup file. (2) Try N random starts in the hypercube $\hat{m}_i - 1 < m_i < \hat{m}_i + 1$, where $\hat{m}$ is the optima found by strategy (1). (3) Form starting points formed by flipping adjacent layer resistivities in the solution $\hat{m}$, pairwise, at layers where a reasonable contrast seems likely as judged by successive jumps in $\hat{m}_i$. The latter strategy is designed to (hopefully) lie in different basins of attraction to the existing $\hat{m}$. Some simple thought experiments and numerical experience show that the MAP solution for underresolved (fine-gridded) models tend to place all the required high resistivity in a single layer, so simple multimodality will exist in the precise location of that anomalous layer.

At the end of the mode enumeration, the code checks the modes for duplicates using some naive tests (e.g., Euclidean distance of MAP points less than some threshold) and sorts the modes by marginal model likelihood (usually very closely tied to rms misfit). Iteration, response, and model depth-profile files are written for each mode.

A typical example of distinct multiple modes is shown in Figure 3. These have the typical "layer-flipping" behavior mentioned before. Another useful function of the mode-enumeration facility is to check that the local modes occur at genuine optima of the -ve log posterior, not simply at points where the Newton scheme could make no further progress caused by coding errors, bad scaling, poor termination

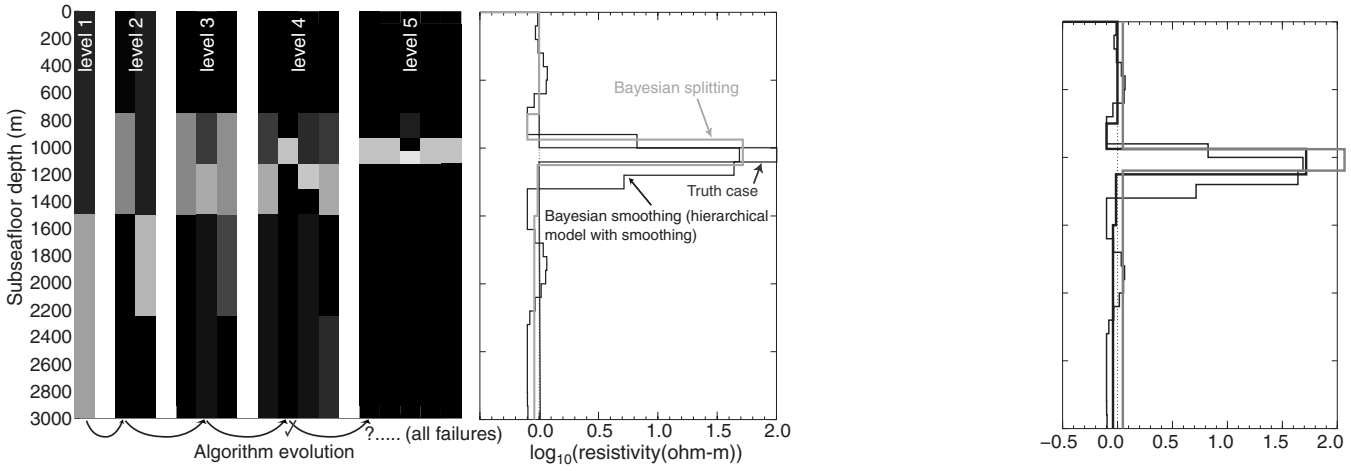

Figure 2. Canonical model under splitting: dark gray = truth case, light gray = final split model with minimal-log (MML) value, black = Bayesian smoothing MAP inversion on fine grid for comparison.

criterion, or other gremlins. Figure 3 shows a plot of the final objective function from 500 random starts of a typical problem, where the repeatable convergence to one of seven possible solutions is clearly evident. In this case, one mode clearly is very superior, and it is reassuring to see that it has an ample basin of attraction.

Mode uniqueness checks

An important consideration in any "mode enumeration" strategy is to avoid the double-counting of modes and understand the relation between modes. We know from simple thought-experiments that there can be distinct optima which are separated only by weak probability barriers in the posterior surface, and knowledge of these near-degeneracies might be useful in constructing MCMC strategies, among other reasons.

A particularly interesting question is how to construct the "lowest energy" path connecting two modes. This path should look like the gray geodesiclike path of Figure 4. This object might form a sort of backbone along which ridges of the posterior probability might form. One possible way to seek such paths is to minimize the path integral

$$\Lambda_{AB} = \int_A^B \chi^2(\mathbf{M}) d\ell \qquad (11)$$

along a smooth parametrized path from MAP point $\mathbf{M}_A$ (belonging to mode A) to a distinct mode MAP point $\mathbf{M}_B$. For χ^2, we would use the full Bayes -ve log posterior, equation 5, or at least the varying pieces of it. Algorithms to generate such paths are described in Appendix A, with some examples.

In summary, our findings are these: for many problems, we find the modes can be linked along paths whose probability barriers are very weak relative to the sampling fluctuations expected in the posterior. For certain near-degenerate cases, the paths correspond to sets of layers behaving as an effective medium with strictly known upscaling laws (e.g., responses depend only on a sum of resistivity-thickness products), but in general this is not the case. In such cases, sampling algorithms for the model uncertainty ought to be able to visit all the modes, and the chief challenge for such algorithms is the traversal of the twisting, steep-sided ridges of the posterior, not jumping between isolated modes per se.

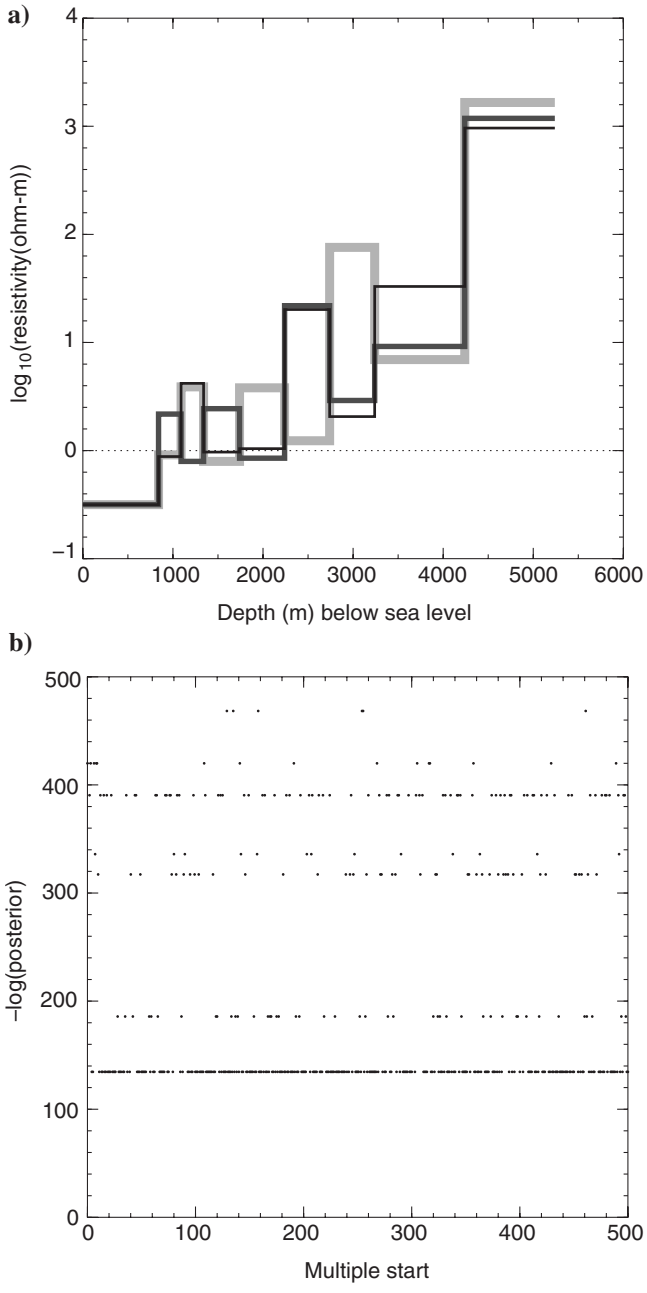

Figure 3. (a) Multiple local-mode MAP solution depth profiles of an eight-layer unsmoothed problem, shown as distinct curves. (b) –Log(posterior) of a large ensemble of random starts. Repeated convergence to particular optima is good evidence of sensible termination criteria.

APPROACHES TO INVERSION UNCERTAINTY

In Bayesian inversion, we emphasize that the full posterior distribution embodies all we can know about the model, and point estimates (e.g., MAP solutions) are very imperfect as tools for making decisions. Ideally, parameter inference from CSEM data should take into account model uncertainty and parameter uncertainty.

Within a model, typical approaches to parameter uncertainty will involve computing posterior covariance matrices from the inverse of the Hessian at MAP points. This is very useful, efficient, and usually satisfactory. But because the nonlinearity in CSEM is severe, the lo-

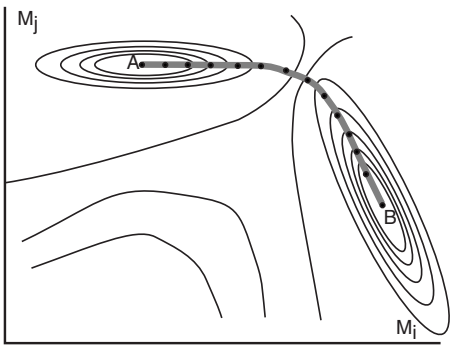

Figure 4. Optimal path connecting modes A and B. The dots depict nodal points on a discretized approximation to the path, used in the optimization algorithms detailed in the main text.

cal linearization is unreliable, and methods based on sampling must be adopted. Notwithstanding this, our implementation writes out linearized MAP posterior covariances ($\tilde{C} \equiv H^{-1}$), correlation-coefficient matrices ($\{\tilde{C}_{ij}/\sqrt{\tilde{C}_{ii}\tilde{C}_{jj}}\}$), and 1-sigma posterior marginal error bars ($\hat{m}_i \pm \sqrt{\tilde{C}_{ii}}$) for the inverted models, for comparison purposes. In the hierarchical Bayesian smoothing mode, the "smoothing-free" approximate covariance ($\tilde{C} \equiv (J^T C_d J/\sigma_n^2 + \sigma_p^{-2} I)^{-1}$) is used because the smoothing is an artificial construct.

In this section, we confine the discussion to uncertainties within models, and present two canonical approaches to sampling: 1) Markov Chain Monte Carlo (MCMC) from the Bayesian point of view, and 2) the frequentist parametric bootstrap method, adapted for the Bayesian framework we use. This latter technique has appeared in the hydrology/petroleum "history-matching" literature under the rubric "randomized maximum likelihood" (Kitanidis, 1995; Oliver et al., 1996), but we prefer the name "Bayesian parametric bootstrap."

The MCMC approach is the method of choice for fully Bayesian frameworks where little can be done analytically, and fast forward model evaluations are possible. It is the standard tool of choice for Bayesian statistical work. The validity of the MCMC algorithm rests critically on constructing a "model proposal" scheme which can visit all the parameter space efficiently and satisfies the requirements for reversibility. This is a very stringent requirement and greatly restricts the ability of these samplers to use "optimization-related" information to construct proposals. For posterior distributions that are very poorly scaled, distorted in shape, and modestly sharp in some dimensions, this makes the construction of good schemes very difficult. Liu (2001) is a good survey of the technique. The section MCMC below has the details of our implementation for 1D CSEM.

Frequentist statisticians are more used to dealing with uncertainty estimation using varieties of the bootstrap or jackknife (Efron and Tibshirani, 1994). These rely on performing separate parameter inferences for each member of a suite of synthetic data sets (generated from an initial best-fit model using the actual data), so the use of optimization apparatus is used explicitly for each bootstrap sample. This has certain advantages for the CSEM problem, as the optimization machinery in place is then able to help find good samples in the domain of support of the posterior. Bootstrap theory has foundations and justifications related to large n (number of data) expansions of the posterior (Hall, 1992) and can be expected to closely resemble Bayesian posteriors if the prior has weak influence (i.e., the likelihood swamps it). This latter is only partially true of the CSEM problem, especially in somewhat over-parametrized models where the Bayesian prior is essential for stabilizing the posterior.

In the section "Bayesian parametric bootstrap" and Appendix B, we show that the parametric bootstrap can be used in a Bayesian framework by treating the prior information as "effective observations" on the parameters. Clearly the number of "extra" data points generated in this way does not grow as we acquire more data, and if the forward model has implicit degeneracies (i.e., near rank-deficiency in the sensitivity), the "large n" assumptions of bootstrap are not strictly valid. Nonetheless, bootstrap theory has been shown to be remarkably effective even for few data, as some of the test examples show, and the ability to straightforwardly apply optimization techniques helps greatly in visiting a greater spread of parameter space.

Markov Chain Monte Carlo

The code incorporates a tentative implementation of an MCMC sampler suitable for sampling from low-dimensional models. It relies heavily on information collected during the optimization and mode enumeration passes. For convenience, suppose the mode enumeration has found a set of local optima $i = 1 \ldots N_m$, which we characterize by their MAP points $\hat{m}_i$, local approximate covariance (inverse Hessian) $\hat{C}_i$ and estimated relative probability $\pi_N(i)$ (we add the subscript N to indicate the $\pi(i)$ are normalized so $\Sigma_i \pi_N(i) = 1$. These are sorted by $\pi_N(i)$, so mode 1 is estimated to be most likely. The algorithm below is robust to the enumeration missing a mode as long as it is reasonably accessible by the random-walk proposals.

A Markov chain is a sequence of samples m_j whose overall equilibrium distribution approaches that of the Bayesian posterior $\Pi(m|y)$. All that is required is a proposal kernel $q(m'|m)$ for visiting a new state m' from an existing state m, which potentially can visit the entire support of the distribution (irreducibility), and a probability for accepting or rejecting a proposal. The art in MCMC implementation consists in constructing proposal schemes that rapidly move across the support of the posterior.

In fixed dimensions, the well-known Metropolis scheme uses an acceptance probability

$$\alpha = \min\left(1, \frac{\Pi(m'|y) q(m|m')}{\Pi(m|y) q(m'|m)}\right),$$

where $\Pi(m'|y)$ is the posterior density of model m, given data y, up to a fixed normalization constant. Models outside the bound constraints are assigned an extremely low probability.

At present, the sampler is implemented for known noise σ_n, and zero smoothing, so we use equation 2 with C_p a diagonal matrix populated from the user-specified prior variances.

The proposal kernel q is a random mixture of three types of proposal:

1) Random jumps of form $q(m'|m) \sim N(m, \xi \hat{C}_1)$, where $\hat{C}$ is the linearized posterior covariance (inverse Hessian) of the most likely mode, and ξ is a scaling parameter tuned such that the final acceptance rate from this kernel is about 0.25.
2) "Layer-flip" moves seeking to exploit the possibility of nearly constant resistivity-thickness product between adjacent layers. The scheme below is a random jump in m_j followed by a conditional random jump in m_{j+1}, designed so as to nearly conserve this property between layers j and $j+1$. Layers have thickness T_j, subsea depth d_j. At initialization, a set of candidate layers S_{LF} suitable for possible layer flipping is assembled. Currently, adjacent layers with $T_j < d_j/4$ form this set. If a layer-flip is chosen, the algorithm is:
   Choose $j \in S_{\mathrm{LF}}$ at random. All parameters but m_j, m_{j+1} will remain the same. Initialize $J_H = \infty$.
   Propose $m_j' = m_j + \delta m_j$, where $\delta m_j \sim N(0, f_A^2)$.
   If $m_j' \geq m_{L,j}$, compute $\xi = (T_j 10^{m_j} + T_{j+1} 10^{m_{j+1}} - T_j 10^{m_j'})/T_{j+1}$.
   If $(\xi > 0)$, propose $m_{j+1}' = \log_{10}(\xi) + \delta m_{j+1}$, where $\delta m_{j+1} \sim N(0, f_B^2)$ and compute $R = (T_j 10^{m_j'} + T_{j+1} 10^{m_{j+1}'} - T_j 10^{m_j})/T_{j+1}$. If $R > 0$, compute $J_H = (\delta m_{j+1}^2 - ((\log_{10}(R) - m_{j+1})/f_B)^2)$.
   Accept the proposal with probability $\min\left(1, \frac{\Pi(m')}{\Pi(m)} e^{-J_H}\right)$. The jump sizes f_A, f_B are tunable parameters, typically $f_A \approx 0.4$, $f_B \approx 0.02$.

3) "Mode jumps" from mode i into mode j of form

$$m' = m + \hat{m}_j - \hat{m}_i.$$

This proposal is made with probability $\pi_N(j)$, so $q(m'|m) = \pi_N(j)$, and the Metropolis equation requires the piece $q(m|m')/q(m'|m) = \pi_N(i)/\pi_N(j)$. This kernel is designed on the assumption that the random-walk part of the sampler will stay close to the mode MAP point relative to the separation between modes, that modes will have a similar shape (local covariance), and that no tunneling between modes will occur (so the "targeted" offset $\hat{m}_i - \hat{m}_j$ is useful). The mode weights $\pi_N(j)$ are used in the proposal so little time is spent constructing a jump to a mode that is very likely to be rejected. None of the assumptions just outlined is a very safe bet for the CSEM problem, unfortunately.

Although Chen et al. (2007) express enthusiasm for the slice sampler of Neal (2003), our impression is that the component-wise slice sampler has significant difficulties with highly correlated posteriors (as would any component-wise method), and it is not clear to us how to implement efficiently a multicomponent version for this problem. Some experiments with hybrid molecular-dynamics samplers (see Chapter 9, Liu (2001)) have produced indifferent results. The fundamental difficulty is that, for many problems, the posterior is very badly scaled (narrow in shallow parameters, wide in deep ones) and highly nonlinear for degenerate parameters: "steep-sided curving valley(s)" in parameter space. The scaled random-walk proposal works well for modestly poorly scaled problems, but only those that do not twist or snake. The fundamental difficulty is very strong, but twisting parameter correlations, and virtually all MCMC techniques we know of, have difficulties in this regime.

Bayesian parametric bootstrap (or Monte Carlo)

An alternative method for assessing inversion uncertainty is an older technique called Monte Carlo simulation, referred to in more modern literature as the parametric bootstrap. For overdetermined, stable inverse problems without any kind of Bayesian prior, the usual procedure is to estimate a maximum-likelihood model $\hat{m}$ by, say, nonlinear regression (i.e., minimize $\chi^2_{\text{misfit}} = (y - f(m))^T C_d^{-1}((y - f(m)))$, estimate the parameters of the noise distribution of $\epsilon = (y - f(m))$ (e.g., a noise variance), then simulate an ensemble of bootstrapped "synthetic" data sets $y_i = f(\hat{m}) + \epsilon_i$, with ϵ_i new samples from the noise distribution. A matching ensemble of bootstrapped parameter estimates $\hat{m}_i$ then are formed by nonlinear regressions of each resampled data set, i.e., minimizing $\chi^2_{i,\text{misfit}} = (y_i - f(m))^T C_d^{-1}((y_i - f(m)))$. The statistics of the ensemble $\hat{m}_i$ then are used for interval estimates, etc.

Appendix B reviews the known result from linear theory that if the noise model is correct and the noise variance unbiased, the mean bootstrap model is an unbiased estimator of the mean (in fact the ordinary least-squares estimate), and the ensemble average residual sum of squares (RSS) is χ^2_{n-p} distributed and has mean $n - p$. This result is what motivates suitable "target misfit" values in discrepancy principle approaches. Another important result is that the distribution of the RSS of the bootstrap residuals with respect to the original data set is χ^2_p, but offset to the right by the regression misfit $n - p$. This suggests the range of data misfits that should be encountered in the posterior distribution.

In Bayesian frameworks, the objective function (log-posterior) above typically is augmented with terms from the prior, usually to something like

$$\chi^2 = (y - f(\mathbf{m}))^T C_d^{-1}(y - f(\mathbf{m})) + (\mathbf{m} - \mathbf{m}_p)^T C_p^{-1}(\mathbf{m} - \mathbf{m}_p).$$

We show in Appendix B that the usual parametric bootstrap arrangement can be modified to work for this case, by treating the prior as additional "data." The upshot is that bootstrap model samples then are found by an optimization problem with resampled synthetic data and resampled prior means $\mathbf{m}_p$. The distributional statement above also holds, with the number of data n now taken as $n + p$. In short, a Bayes MAP model $\hat{m}$ is found using the real data y, and bootstrap samples are found by optimization with synthetic data drawn from $y_i \sim N(f(\hat{m}), C_d)$, and a synthetic prior from $\mathbf{m}_{p,i} \sim N(\hat{m}, C_p)$.

From the material and example shown in Appendix B, it emerges that the recentering of the prior mean, which is required in the fully linear case to achieve rigorous, unbiased sampling, has a strong effect in the nonlinear and multimodal case, effectively oversampling the posterior in the region close to the MAP estimate $\hat{m}$. To overcome this effect, at the price of some weak bias, we advocate a non-recentered version, using the same recipe as above, but drawing bootstrap prior means from $\mathbf{m}_{p,i} \sim N(\bar{m}, C_p)$. The example below illustrates how this helps for a CSEM problem with well-understood ambiguities.

Example: CSEM split-canonical model of underresolved layers

Here we examine parameter uncertainties using a test case we like to call the "split" canonical model: a 1 km overburden shale (m_1), then two 50-m reservoir layers (m_2), m_3), and shale underburden (m_4). "Truth-case" data are synthetically generated with the shale background 1 Ωm ($m_1 = m_4 = 0$) and the reservoirs 100 Ωm ($m_2 = m_3 = 2$). Because the reservoirs are thin relative to natural resolution, we expect the CSEM data to resolve only the total resistivity of the two reservoir layers, but there might be subtle depths preferences.

Samples drawn using the recentered bootstrap are shown in Figure 5. The spread of models is fairly wide, but there does appear to be a concentration of the anomaly in the deeper layer, parameter m_3. This requires a little explanation. First, in the Monte Carlo experiment where we generate synthetic data from the standard three-layer canonical model with Gaussian noise, and invert for bootstrap MAP split-canonical (four-layer) models using globalized mode-searching, about 75% of the time the "most-likely mode" places all the

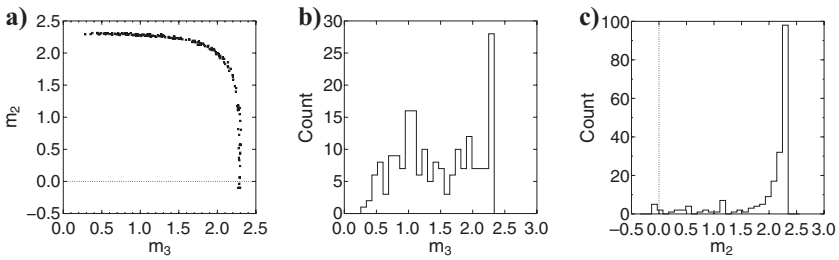

Figure 5. (a) Joint samples of m_2,m_3 from recentered parametric bootstrap on the split-reservoir canonical model. (b) Histogram of m_3 from the samples, and (c) histogram of m_2, from the samples. For discussion on asymmetry, see main text.

anomaly in the deeper thin layer,[5] so the layers obviously are thick enough to break the symmetry modestly. Second, the particular data used for the truth case produced a MAP solution $\hat{m} \approx (0.9,2.3)$, so the recentered bootstrap samples consequently are more concentrated in this region. The weak preference for the deep layer in the Monte Carlo experiment is of no great significance, but once the recentered bootstrap has been fired off with a MAP solution in a particular part of parameter space, bootstrap realizations clearly will be more sharply concentrated in that region than is desirable.

The non-recentered bootstrap output is shown in Figure 6. Here there is a much better symmetry in where the anomaly is placed, but smoother models $m_2 \approx m_3$ are under-represented. This under-representation is caused by the modestly low probability of drawing models from the prior distribution close to this "knee" point in the maximum-likelihood surface because the MAP solution found by the bootstrap will be, roughly speaking, the closest point on the maximum-likelihood surface to the sample prior-mean for the realization. Figure 7 shows the comparable output using MCMC (with heavily decimated sampling output), showing heavier support in the corners of the distribution and for smoother models.

For strongly nonlinear models, empirical distributions produced by bootstrapping cannot be expected to yield the same results as procedures that correctly sample from the Bayesian posterior, such as MCMC. The theory is strong for the linear case, but the validity of the bootstrap procedure depends on being in an asymptotic regime with a large data-to-parameters ratio and a very focused (compact) likelihood, which means the linear approximation is respectably valid over the support of the posterior. The first example above represents a case where acquiring more data will not focus the posterior better; the model is intrinsically unresolvable, and only the uncertainty of the "effective medium" formed by m_1 and m_2 is reduced with more data.

Our recommendation at present is that the "nonrecentered" bootstrap be used, as it seems less likely to miss significant probability mass away from the mode belonging to the MAP solution $\hat{m}$ used as the basis for the bootstrap. Because, in the CSEM case at present, the prior means nearly always are less than the MAP values, any biases are likely to reduce inferred resistivity values, which is a conservative tendency.

It is fairly likely that adapted bootstrap techniques exist for multimodal target distributions, and that a good resampling scheme for multivariate Gaussian mixtures can be constructed. This requires further research.

EXAMPLE PROBLEMS

Thickness wedge model

Here we invert a known truth-case model, constructed as a resistive wedge buried 1 km deep in shale, in 1 km of seawater, and extending from 10 to 450 m in thickness; see Figure 8a. The wedge is presumed to be very gradual, so the 1D assumption is not violated; the wedge geometry is chosen specifically to illustrate resolution aspects. The underburden is also shale. The shale background is 1 Ωm, reservoir 100 Ωm, and the data set is inline $|E|$ measurements at frequencies f = .25,.5,.75,1,1.25,1.5,2 Hz, for offsets at 1 km to 15 km, on 500 -m spacings. Noise levels are taken as 5%, with a noise floor of 2.10^{-16} V/Am^2.

Figure 8b and c show MAP inversion images produced using Bayesian smoothing on two grids: a regular 50-m grid, and a logarithmic grid (layer thicknesses increasing geometrically with depth). Both styles fit the data satisfactorily, so the inferred image largely is a function of the grid construction. Figure 8d is a plot of the MAP inverted reservoir thickness and resistivity-thickness product (RTP), with error bars, based on a parametric study of a three-layer model, as follows. For low-dimensional models, the marginal model likelihood (MML) is a useful tool for examining model uncertainty involving depth and thickness of certain target layers. The code can be used to generate a model-study suite of inversions over a user-specified range of specified layer thicknesses in an arbitrary hypercube. The MAP model belonging to the maximum MML model chosen from this suite of models is what we describe as a "MML-based inversion." The MML outputs from this model study are used to construct thickness and depth uncertainties for target layers. Discrete summations of the model probabilities ($\sim e^{-\mathrm{MML}}$) over thicknesses/parameters not of interest are used to construct approximate marginal distributions for parameters of interest. Figure 8d is such an inversion result for the wedge model, using a parametric model study of the reservoir layer top-depth and thickness.

Figure 9 shows how the MML varies as the depth and thickness of a single-layer reservoir vary at location CMP 5, where the truth-case

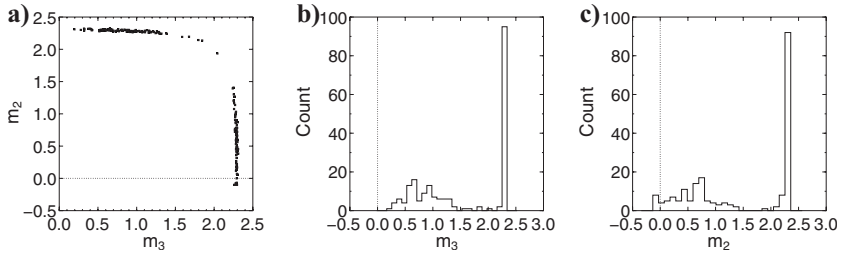

Figure 6. (a) Joint samples of m_2,m_3 from non-recentered parametric bootstrap on the split-reservoir canonical model. (b) Histogram of m_3 from the samples, and (c) histogram of m_2, from the samples. For discussion on under-represented smooth models ($m_2 \approx m_3$), see main text.

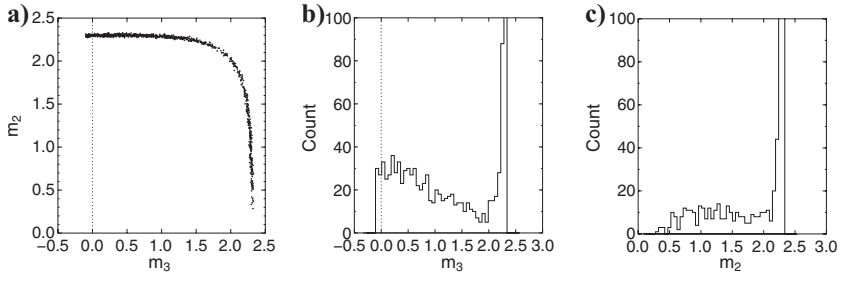

Figure 7. (a) Joint samples of m_2,m_3 from MCMC sampling on the split-reservoir canonical model. (b) Histograms of m_3 from the samples, and (c) histogram of m_2, from the samples.

[5]The bootstrap modes are very well-separated, focused clusters at $(m_2,m_3) \approx (0.4,2.5)$ and (2.5, 0.4), so we can expect that, for any data set, the MAP model $\hat{m}$ is near either of these values.

model was 135 m thick (1000 m deep). Thicker models have a slight tendency to image shallower. Though we do not show the details in the interests of brevity, under the Monte Carlo experiment of resampling the synthetic data and reconstructing the marginals each time via the parametric model study, the MAP estimate of depth and thickness can be shown to have low bias.

Bird model

This case is a surrogate for some field data, with subsurface target profiles approximating that of interest, and field data generated synthetically by adding independent Gaussian deviates to the truth-case data. The data sampling inherits some uneven spacing from the CMP processing on actual field data and the somewhat arbitrary extension

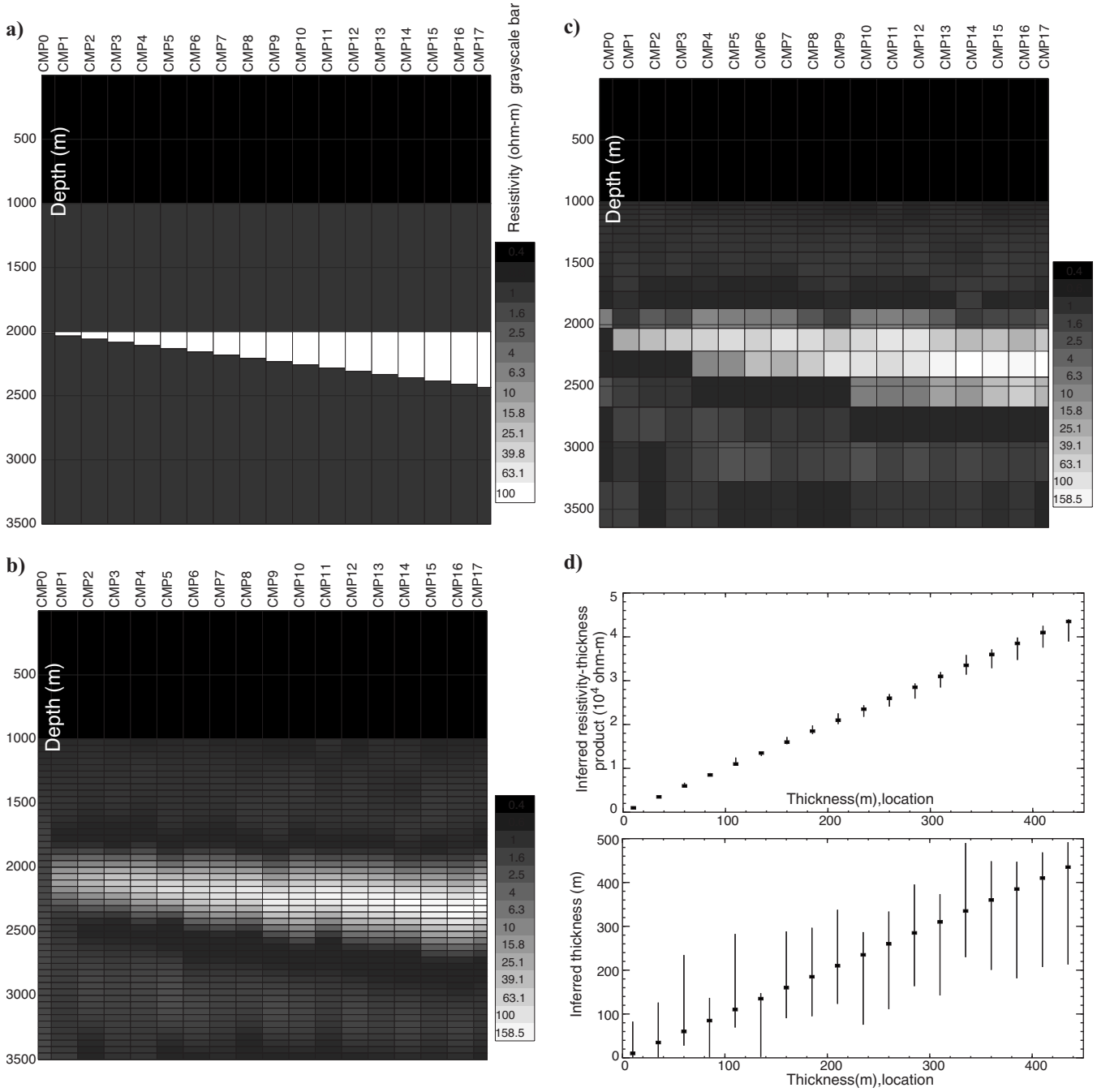

Figure 8. (a) Truth-case wedge model: 100 Ωm reservoir over 1 Ωm shale background. (b) MAP inversion image using Bayesian smoothing on a regular 50-m grid. (c) MAP inversion image using Bayesian smoothing on a logarithmic grid. (d) MML-based three-layer inversions for depth, thickness and resistivity, showing marginal-distribution 95% error-bars for thickness, and resistivity-thickness product (RTP). Clearly the RTP is much better identified by the data than thicknesses or resistivities.

of the 0.75 Hz and data near the noise floor. Here the error bars are 5% of $|E|$, thresholded at 2.10^{-16} V/Am^2. There are frequencies 0.25, 0.75, and 1.25 Hz, and the data is $|E|$ inline, from 1.2–12 km. The "true" model, data, and two styles of inversion are shown in Figure 10. Inversions have been run with Bayesian-smoothing and Bayesian model-selection styles, and both have similar "opinions" on the achievable resolution, and detect the two main anomalous (resistive) layers aside from basement. Some variation in the thickness of the final "GAP" lower-resistivity segment is observed (see Figure 10), but parametric variation of this thickness shows that it is very poorly resolved by the data (the MML shows support over about 1 km of thickness).

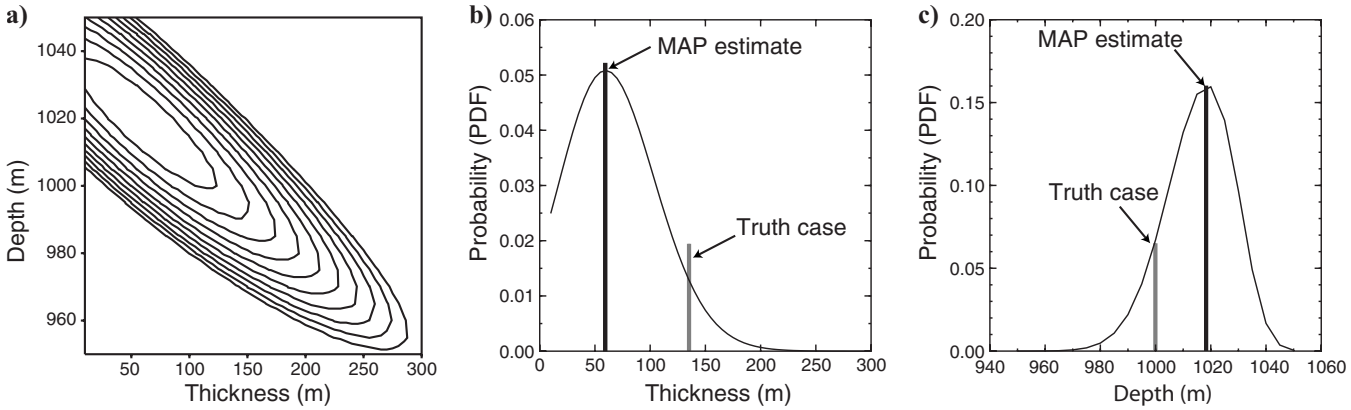

Figure 9. (a) Contour plot of marginal model likelihood of three-layer model fit to CMP5 data in wedge model, as a function of reservoir depth and thickness. Contours are at unit spacings of $\log_{10}$(MML), so three contours is about the conceivable span of model support in the data. Views (b) and (c) are the marginal distributions of thickness and depth of the reservoir layer associated with the MML measure, shown with the original truth-case model values from which the data were generated, and also MAP estimates of parameters. Note that no conclusions about bias could be drawn from these plots, as the marginal distributions have considerable stochastic uncertainty under resampling of the data.

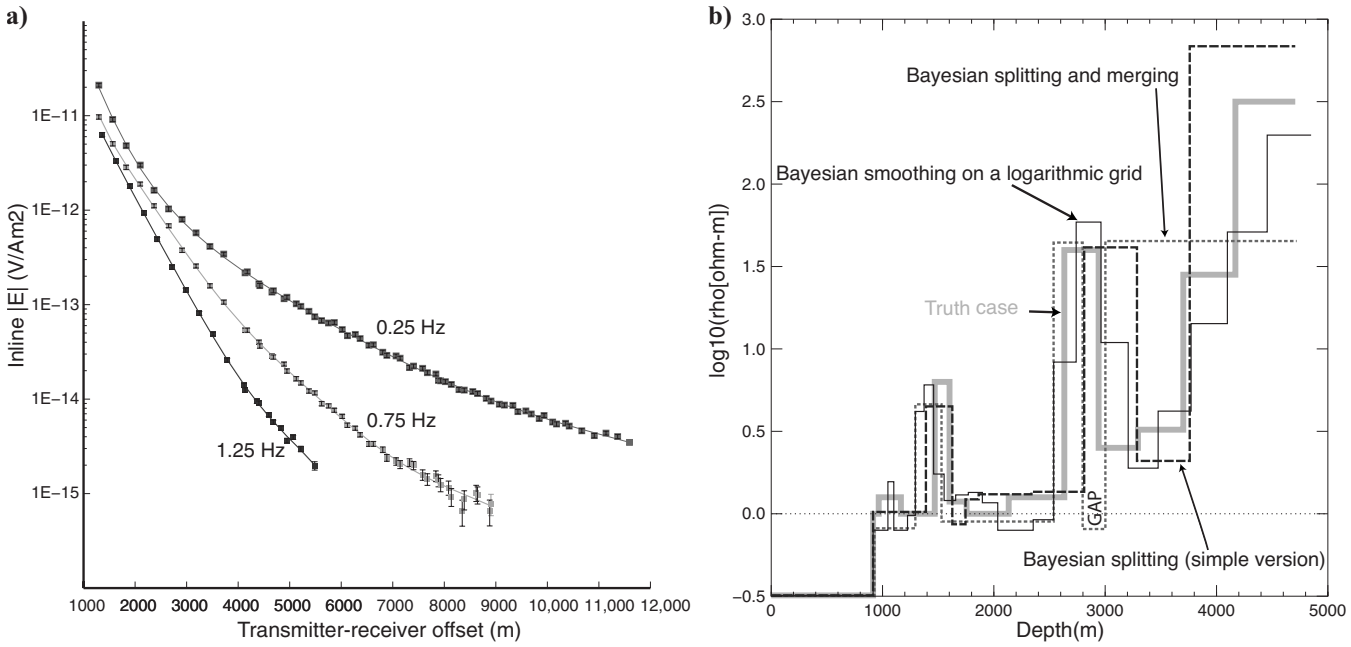

Figure 10. Bird model. (a) Three-frequency inline $|E|$ data used for inversion, with typical fitted model (Bayes-smoothing case). (b) Truth-case data (thick gray), and MAP inversions for three styles of "resolution-detecting" inversion (Bayesian-smoothing, splitting and split/merge). All three approaches fit the data at approximately $\chi^2_{\mathrm{RMS}} = 0.9$.

To examine inversion uncertainty, an unsmoothed inversion based on a $p = 18$ layer logarithmic grid was run, with model priors set at $N(0,1)$, and noise-variance σ_n an additional unknown. This inversion has modest uncertainty about which layer to place the two anomalies in, and the marginal posterior distributions in the anomalous layers clearly are multimodal. A typical example is shown in Figure 11.

This model is an interesting comparative test case for the posterior sampling techniques. We generate large bootstrap and MCMC ensembles and compute from these samples the P16, P50 and P84 quantiles (mean $\pm$ one std deviation for Gaussian deviates) of each layer parameter m_i. These quantiles and the truth-case model for both styles of calculation are shown in Figure 12. Neither method seems statistically anomalous in terms of mispredicting the actual model, but in general the bootstrapping interval estimates are a little wider, as suspected from the simple calculation for the split canonical model. Either method is very much preferable to linearized error analysis (using local mode Hessians); these are not shown. The MCMC calculation is at least 10 times the expense of the bootstrapping run in this case, as slow mixing is a controlling factor. The correlation test procedures of Raftery and Lewis (1996) have been used to estimate the adequacy of the final ensemble. The tendency of bootstrapping to undersample the smoother models makes certain bimodal distributions more accentuated, and hence some of the P50 quantiles are more volatile.

Another test of the sanity of the sampling procedures is statistical plots of the sample-log (posterior) distribution, relative to what might be expected from linear theory. From equation B-7, Appendix B, we expect the sampling distribution to "resemble" an offset χ^2_p distribution if the model were nearly linear. For the nonlinear case all

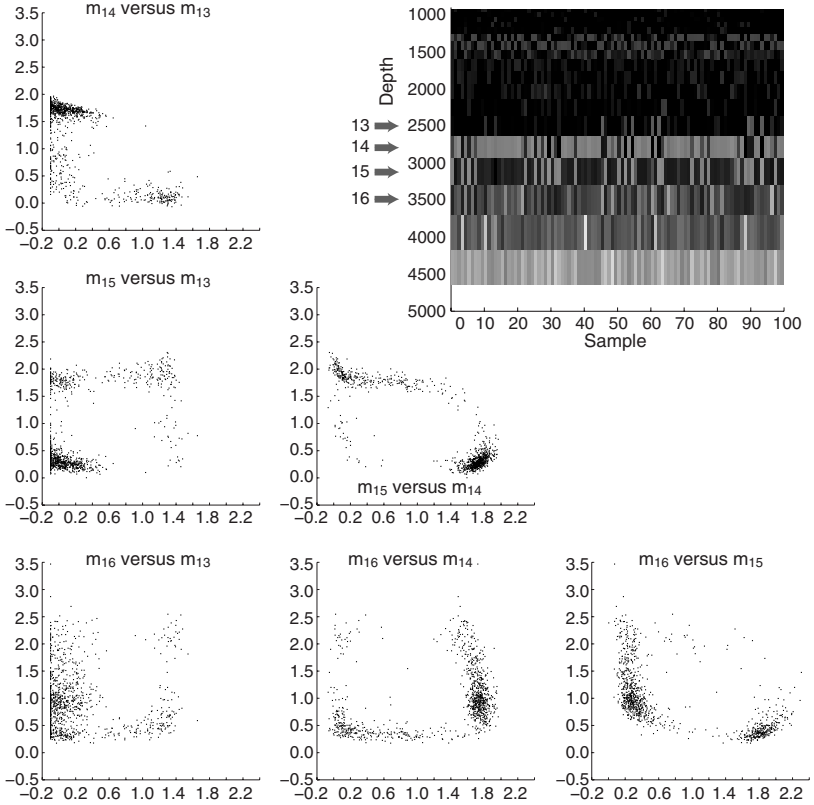

Figure 11. Bird model unsmoothed inversion uncertainty. Cross-scatterplots of bootstrap sample models in layers 13, 14, 15, and 16, where the deeper anomaly lies. Inset: Grayscale model depictions of the model parameters in depth, for 100 "realizations" from the posterior using bootstrapping. In these samples, the anomaly prefers to reside solely in one of two or three layers over a "background."

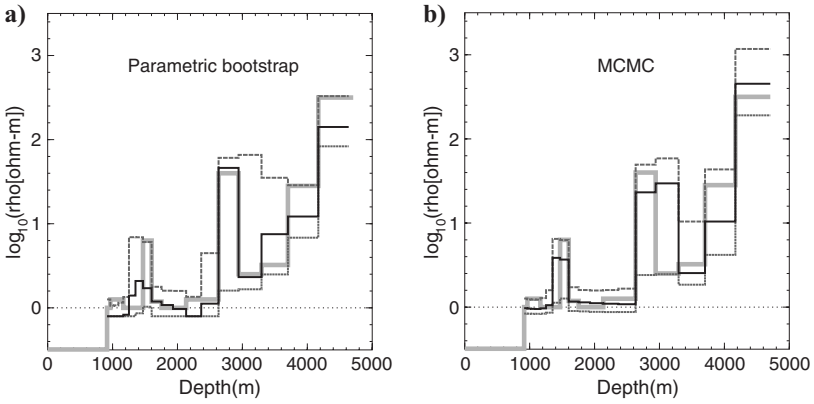

Figure 12. Bird model marginal model-interval estimates for an unsmoothed inversion. (a) P16 (thin gray, dashed), P50 (black), P84 (thin gray, dashed) quantiles computed from a bootstrap ensemble of 1000 models. "True" model shown in thick light gray. (b) Same graphs, generated from long, strict MCMC calculation.

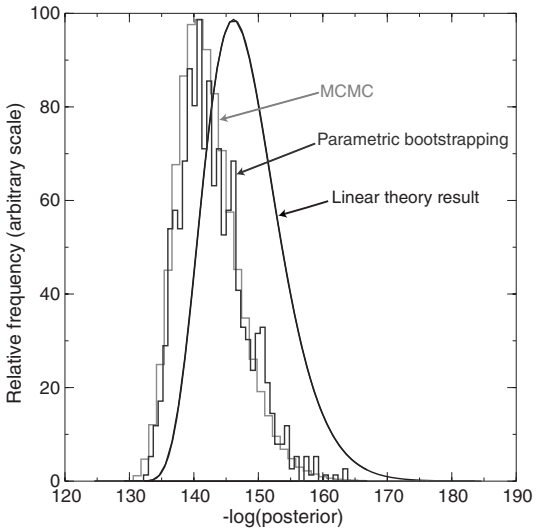

Figure 13. Bird model, posterior distribution of the -log(posterior) distribution (total prior and data misfit). Difference from the theoretical linear curve is a result of the nonlinearity, but MCMC and parametric bootstrapping generate comparable results.

bets are off, but we should expect a modest concurrence, and in particular we should expect an alternative scheme to MCMC to agree closely on this issue. See Figure 13.

One can conclude from this exercise that both sampling methods are good at generating plausible models (i.e., all fit the data within the "expected" variation), but the bootstrap models are more widely variable, i.e., tend to concentrate an undue fraction of the resistive anomaly in single layers. The bootstrap technique is very good at generating independent samples; even given the price of optimization for each sample, the overall optimization cost (say O(100) forward runs with sensitivity) still is less than the cost of progressing to a decorrelated state in the MCMC chain. However, the bootstrap does not visit the more remote portions of the posterior as well as MCMC and is overall a mildly biased sampler for this seriously nonlinear problem.

SOFTWARE

The open-source DeliveryCSEM code implementing these ideas is a companion software to the Delivery software used for seismic AVO inversion (Gunning and Glinsky, 2004). It is released under a General Public License-style license into the public domain and can be obtained at the Commonwealth Scientific and Research Organization (CSIRO) web site (Gunning, 2003). The bulk of the code is Java, but uses the public domain Scripps forward engines in FORTRAN (Dipole1D (Key, 2009), also seafloor.f and dependencies (Constable et al., 1987)) called through the Java Native Interface. Test examples and usage documents, etc., are at the web site.

CONCLUSIONS

We have presented two Bayesian approaches to resolution inference and uncertainty in CSEM inversion problems. Resolution can be inferred by either hierarchical models with free parameters for correlation lengths (Bayesian smoothing), or model-choice frameworks applied to variable resolution spatial models (Bayesian splitting/merging). Globalized optimization with bound constraints is an essential workhorse for either method. The smoothing methods tend to be faster, but the final models are not as parsimonious. Both methods offer a coherent alternative to regularization approaches, with more explicit control of the prior distribution, and a more intimate relationship to the large statistical literature on model inference using maximum likelihood or empirical Bayes methods.

Local linearization approaches to model uncertainty based on covariance matrices at modes are of very limited use and usually chronically underestimate uncertainty for models with multimodal or heavily skewed posterior marginal distributions. A reasonably efficient technique based on a Bayesianized version of the parametric bootstrap is much better, but likely to modestly overestimate uncertainties. Full MCMC sampling is possible for these problems but very expensive compared to either of the preceding techniques.

Software for performing these inversions is made available under an open-source license agreement, with reference implementations of all the main ideas described in this paper.

ACKNOWLEDGMENTS

We gratefully acknowledge the work of Kerry Key, Agusta Flosadottir, Alan Chave and Steve Constable in the Scripps forward modeling kernels we use. Helpful conversations with Kerry Key, Steve Constable, Niels Christensen and Henning Omre are appreciated, in addition to the suggestions and comments of Jinsong Chen and three other anonymous referees.

APPENDIX A

ALGORITHMS FOR FINDING MODE CONNECTIONS

One possible approach to finding a locally minimum path for the integral (equation 11) is by discretizing the integral using some quadrature scheme. In the following examples, neither free-noise nor smoothing is used, so it is sufficient to use the objective (with C_p diagonal)

$$-2\log(\Pi(\mathbf{m}|d)) \equiv \chi^2 = (\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m})) + (\mathbf{m}-\mathbf{m}_p)^T C_p^{-1}(\mathbf{m}-\mathbf{m}_p).$$

A very simple "midpoint" Euler scheme for equation 11 is

$$\Lambda_{AB} \approx \sum_{i=0}^{i=N+1} \frac{\chi^2(\mathbf{M}_i)+\chi^2(\mathbf{M}_{i+1})}{2}\|\mathbf{M}_{i+1}-\mathbf{M}_i\|, \qquad \text{(A-1)}$$

where $\mathbf{M}_0 = \mathbf{M}_A$, $\mathbf{M}_{N+1} = \mathbf{M}_B$, and $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_N$ are path "nodes" fairly evenly distributed along the path connecting A and B. We then minimize the sum for the joint parameters $\mathcal{M} = \{\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_N\}$ using standard optimization techniques. Start with an initial configuration of points $\mathbf{M}_i$ evenly distributed along the straight line connecting A and B. Efficient optimization will require, at least, $\nabla_{\mathcal{M}}\Lambda_{AB}$. Because the gradient $\nabla_{\mathbf{M}_i}\chi^2$ at the ith path-node already is coded and available, the bulk of the work is done. For completeness, the full joint gradient, in components, is

$$(\nabla\Lambda_{AB})_{ij} = \sum_{i=1}^{N}\{\nabla_\chi^2(\mathbf{M}_i)\}_j(\Delta\mathbf{M}_i+\Delta\mathbf{M}_{i-1}) + \sum_{i=1}^{N+1}(\chi^2(\mathbf{M}_{i-1}) + \chi^2(\mathbf{M}_i))\frac{\mathbf{M}_{ij}-\mathbf{M}_{i-1,j}}{\Delta\mathbf{M}_{i-1}} - \sum_{i=0}^{N}(\chi^2(\mathbf{M}_{i+1})$$

$$+ \chi^2(\mathbf{M}_i))\frac{\mathbf{M}_{i+1,j} - \mathbf{M}_{i,j}}{\Delta \mathbf{M}_i}. \qquad \text{(A-2)}$$

Here, $\Delta \mathbf{M}_i = \|\mathbf{M}_{i+1} - \mathbf{M}_i\|$ is the forward-difference path-segment length. With function and gradient now readily computable, the optimization now can proceed using standard efficient methods. At present, we use a Broyden-Fletcher-Goldfarb-Shanno (BFGS) (variable metric) scheme (Nocedal and Wright, 1999), based on unconstrained minimization (UNCMIN) (Schnabel et al., 1982; Verrill, 2005). A simple example with known degeneracy is shown in Figure A-1. It is helpful to introduce apparatus to ensure the node points in the discrete approximation to the path-integral remain equispaced. We define the segment lengths $\Delta M_i = \|\mathbf{M}_{i+1} - \mathbf{M}_i\|$, the mean segment length

$$\bar{m}_S = \frac{1}{N+1}\sum_{i=0}^{N} \|\Delta M_i\|,$$

and the additional penalty term to Λ_{AB}

$$\Lambda^*_{\mathrm{AB}} = \mathcal{A}\sum_{i=0}^{N} (\|\mathbf{M}_{i+1} - \mathbf{M}_i\| - \bar{m}_S)^2,$$

whose gradient has components

$$(\nabla \Lambda^*_{\mathrm{AB}})_{ij} = 2\mathcal{A}((M_{ij} - M_{i+1,j})(1 - \bar{m}_S/\Delta M_i) - (M_{i-1,j} - M_{i,j})(1 - \bar{m}_S/\Delta M_{i-1})).$$

$\mathcal{A}$ is chosen as a suitable scaling constant (e.g., $\mathcal{A} = (N+1)^2/\|\mathbf{M}_A - \mathbf{M}_B\|^2$). Local minimization of $\Lambda_{\mathrm{AB}} + \Lambda^*_{\mathrm{AB}}$ then will generate the maximum probability local path, with equispaced points. Note that because Λ^*_{AB} penalizes only the "segment-length variance," it should not compete with the principal term we wish to minimize.

It is possible to formulate the problem using Euler-Lagrange equations for the minimum path, which could be solved by, e.g., shooting. Experiments with the BFGS implementation scheme above indicate that the number of outer iterations required to stabilize (around 50) is likely to be comparable to the number of forward shoots likely to be needed in any Newton-like shooting scheme. A Runge-Kutta or similar scheme for the latter is likely to require about the same amount of work (e.g., a function and a gradient evaluated about every $\bar{m}_S$ in space), so overall, the computational costs of the two ideas are probable comparable.

MORE COMPLEX EXAMPLE

Here we consider an 18-layer logarithmic-gridded model with $n = 138$ data for inline $|E|$. The code is run in naive style, with no metasmoothing or noise parameters, so $\mathbf{M} = \mathbf{m}$. Multistart optimization is enabled, using layer-flipping, and the code ends up collecting eight modes. Figure A-2 shows a scatterplot of the path linking modes 1 and 3, for layers 4, 5, 6, 12, 13, 14. The inset "morph" figure shows how the model evolves from model 1 into model 3 along the path. The layers chosen for the scatterplot are those undergoing significant changes.

A question of great importance is whether the modes are "statistically interconnected" at the level of noise specified by the inversion. A rough guess at this can be inferred by assigning the most likely mode MAP point as the offset in an offset χ^2_p distribution (see the regression discussion in Appendix B, and equation B-7). Random samples from the posterior should spread out with χ^2 values no higher than the support of the offset χ^2_p distribution. If this latter comfortably covers the probability barriers separating modes, then we might say the modes are "statistically connectable." The modes found in this example easily satisfy this condition, as shown in Figure A-3.

It is important to point out that these "connecting links" are not trivial entities in general. They do not arise in the general case from straight-line interpolation of mode points in either (transformed) $\log(\rho)$ space or the untransformed space of resistivities. Such straight-line trajectories usually encounter enormous probability barriers caused by serious data misfits.

APPENDIX B

CLASSICAL REGRESSION RESULTS, BOOTSTRAP, AND BAYESIANIZED BOOTSTRAP

Here we wish to motivate the Bayesian parametric bootstrap by revisiting some known results from classical linear regression and bootstrap theory.

Suppose that, in truth, the n data are generated by a linear model in p parameters

$$y_u = X_u . m + \epsilon_u$$

where X_u is $n \times p$, the noise $\epsilon_u \sim N(0, C_d)$, and usually C_d is a diagonal matrix of noise variances. The suffix u denotes "unscaled" variables. The least-squares estimate of m is

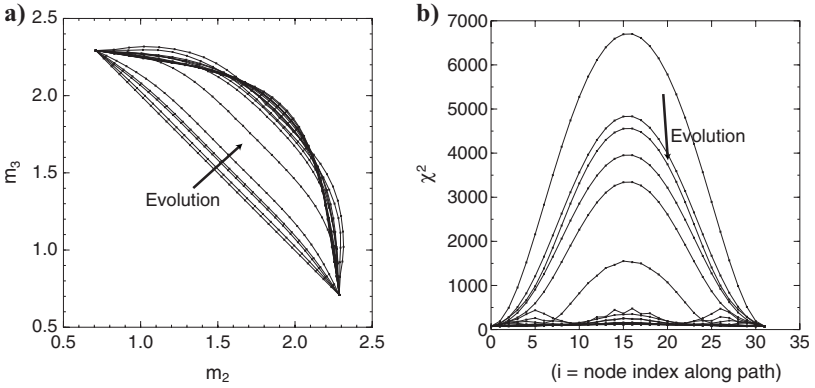

Figure A-1. Evolution of "mode-linking" path under optimization, for the simple experiment of the "split-canonical model," where the 100-m resistive layer (buried 1 km deep) of the canonical model is replaced by two 50-m layers (parameters m_2,m_3). Two modes can be found by the "layer-flipping" strategy of the global optimizer, corresponding (roughly) to placing the resistive anomaly in each thin layer solely. (a) The $\log_{10}(\rho)$ parameters of each are plotted as the path evolves. The optimal path is close to that describing a conserved resistivity times thickness sum over the two layers. (b) χ^2 cross-sections of the posterior surface as the optimal path evolves. Clearly the early paths are extremely improbable ways to connect models.

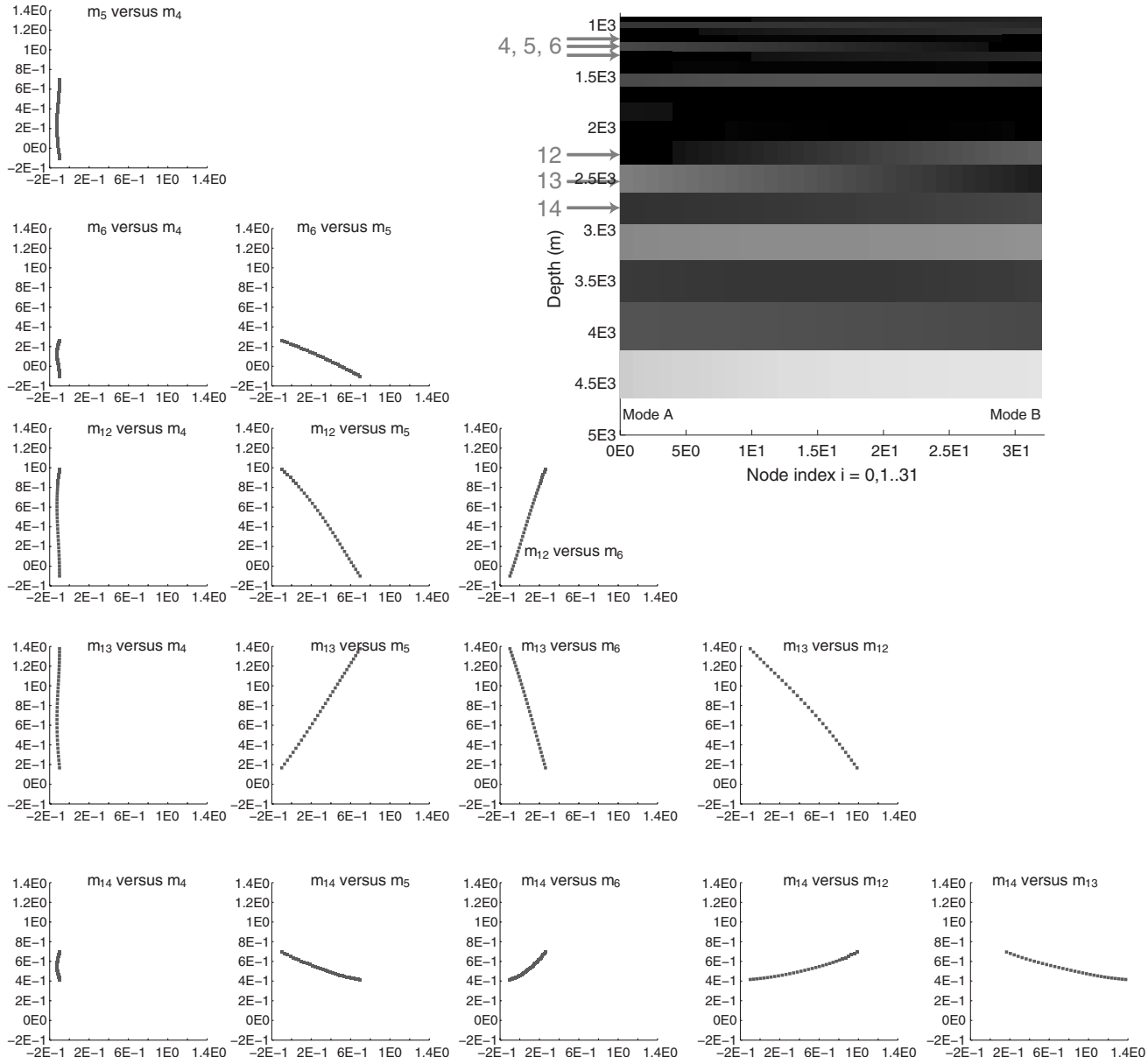

Figure A-2. Scatterplot along "mode-linking" path between two endpoint-modes. The inset shows grayscale morphing of mode A into mode B along this path, with the exchange of resistivity quite obvious (light shades = resistive). Crossplots for the interesting layers 4, 5, 6, 12, 13, 14 (arrows) are shown.

$$\hat{m} = (X_u^T C_d^{-1} X_u)^{-1} X_u^T C_d^{-1} y_u,$$

from which the "predicted" data are

$$\hat{y}_u = X_u . \hat{m} = X_u (X_u^T C_d^{-1} X_u)^{-1} X_u^T C_d^{-1} y_u.$$

For the algebra that follows, it is most simple to think in terms of scaled data $y \equiv C_d^{-1/2} y_u$, a scaled design matrix $X \equiv C_d^{-1/2} X_u$, and standard normal noise $\epsilon \equiv C_d^{-1/2} \epsilon_u \sim N(0,I)$, in terms of which the formulas read

$$\hat{m} = (X^T X)^{-1} X^T y,$$

$$\hat{y} = X (X^T X)^{-1} X^T y.$$

Here, the overall coefficient matrix $Q = X(X^T X)^{-1} X^T$ is known as a "hat" matrix (it "puts the hat" on y). Q has some important properties. It is symmetric and idempotent because $Q^2 = Q$, so has eigenvalues 1 or 0, and has the same rank as X, i.e., possessing p eigenvalues 1, the remainder 0. It therefore follows that $\text{rank}(I - Q) = n - p$, which is of use in the below. Another standard result we need is that if $\mathbf{z} \sim N(0,I)$, and A is a fixed symmetric idempotent matrix of rank k, then $\mathbf{z}^T A \mathbf{z}$ is distributed as χ_k^2, which has mean k.

We are interested in the normalized residuals

$$e = C_d^{-1/2} (y_u - \hat{y}_u) = y - \hat{y} = (I - Q) y. \qquad \text{(B-1)}$$

These have expectation $\langle e \rangle = 0$ if the model is true (because $(I - Q)\ X = 0$). Another important quantity is the residual-sum-of squares $\chi^2_{\text{RSS}} = e^T e$, with expectation

$$\langle \chi^2_{\text{RSS}} \rangle = \langle e^T e \rangle = \langle y^T (I - Q)^T (I - Q) y \rangle$$
$$= \langle \epsilon^T (I - Q) \epsilon \rangle = n - p \qquad \text{(B-2)}$$

after a few lines of algebra. Clearly, χ^2_{RSS} is distributed as χ^2 with $n - p$ degrees of freedom. It follows, in connection with the "discrepancy" principle used in the Occam style inversions, that if the noise estimates are correct and Gaussian, the "target value" of $\chi^2_{\text{RMS}} = \sqrt{e^T e / n}$ ought to be

$$\chi^2_{\text{RMS}} = \sqrt{(n - p)/n}, \qquad \text{(B-3)}$$

which might be, typically, about 0.9. See also the discussions in Hansen (1998). Roughly, this means we expect the regression to fit within $n - p$ "standard predictive errors." (Hansen has also a discussion of how setting $\chi^2_{\text{RMS}} \to 0 = 1$ tends to produce oversmoothing.) In the thought-experiment of making a very rich model with $p \to n$, we get $\chi^2_{\text{RMS}} \to 0$, which is a standard symptom of complete overfitting.

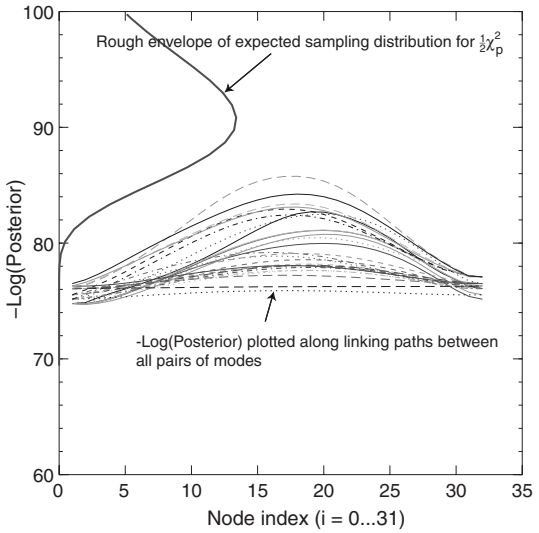

Figure A-3. Plots of the -ve log-posterior (omitting constant terms) along the $8 \times 7/2 = 28$ possible links among eight different modes, along the minimum integral path. On the left is a profile of the associated approximate offset $-\chi^2_p$ sampling distribution attached to the most likely mode. All these interconnecting paths appear reasonably accessible to the sampler. Note that an $\approx 20\%$ correction to the noise level has been used to adjust the vertical scale.

BOOTSTRAP

A synthetic bootstrap data set for the linear problem then can be sampled as

$$y_{u,i} = X_u \hat{m} + \epsilon_{u,i},$$

or equivalently

$$y_i = X \hat{m} + \epsilon_i.$$

The LS bootstrap model $\hat{m}_i$ estimated from this sample then is

$$\hat{m}_i = (X^T X)^{-1} X^T y_i = \hat{m} + (X^T X)^{-1} X \epsilon_i,$$

which obviously is unbiased ($\langle \hat{m}_i \rangle_\epsilon = \hat{m}$). The predictive accuracy of this sample model with respect to the original data set is of interest. Consider the predictive residuals

$$e_i = y - X \hat{m}_i = (I - Q) y + Q \epsilon_i. \qquad \text{(B-4)}$$

There are two kinds of ensemble distributions of interest:

1) The distribution of e_i formed by sampling over the data set y and the bootstrap variables ϵ_i, which will denote with y,ϵ subscripts. Because these are distinct spaces, it then is trivial to show that $\langle e_i \rangle_{y,\epsilon} = 0$ and that the bootstrap prediction residual sum of squares $e_i^T e_i \sim \chi^2_n$ i.e., $\langle e_i^T e_i \rangle_{y,\epsilon} = n$.

2) The distribution of $\overset{y,\epsilon}{e_i^T} . e_i$ formed over the bootstrap samples only, i.e., for a given, fixed y. This is what is handled in practice, and identical to the negative log-posterior term in a MCMC approach. We use the eigendecomposition $Q = V^T I_p V$, where V is orthogonal, I_p is a diagonal matrix of p leading ones, and so

$$e_i = (I - Q) y + Q \epsilon_i = V^T ((I - I_p) V y + I_p V \epsilon_i)$$
$$= V^T ((I - I_p) V y + I_p \epsilon_i'). \qquad \text{(B-5)}$$

where $\epsilon_i' \equiv V \epsilon_i$ also is $N(0,I)$ (i.e., standard normal). Thus

$$e_i^T . e_i = \| (I - I_p) V y + I_p \epsilon_i' \|_2 = \| (I - I_p) V_y \|_2 + \sum_i^p \epsilon_i'^2,$$

$$= \| y - X . \hat{m} \|_2 + \sum_i^p \epsilon_i'^2, \qquad \text{(B-6)}$$

or

$$e_i^T.e_i - \|y - X.\hat{m}\|_2 \underset{\epsilon}{\sim} \chi_p^2, \tag{B-7}$$

i.e., the sampling distribution of the data misfit $e_i^T.e_i$ is χ_p^2, but offset to the right by the minimum misfit found in the regression. Roughly speaking, we then expect the bootstrap "samples" of the model to generate an original-data misfit distribution χ_i^2 whose mean is offset by p to the right of the "best fit" χ^2 in the regression.

In classical and empirical Bayes methods, the noise level (if uncertain, as is usually the case) would be estimated such that $\|y - X.\hat{m}\|_2 = n - p$, so the mean of the data misfit $e_i^T.e_i$ under both kinds of ensemble averages is n.

BOOTSTRAP FOR BAYESIAN FRAMEWORKS

For ill-conditioned problems with regularization modifications, or in Bayesian frameworks, the data-misfit objective function (log-posterior) above is modified with terms containing "prior" beliefs about the model mean $\bar{m}$. Typically, for a nonhierarchical model, with a Gaussian prior $m \sim N(\bar{m}, C_p)$, and Gaussian likelihood, we have

$$\chi^2 = (y - f(m))^T C_d^{-1}((y - f(m))) + (m - \bar{m})C_p^{-1}(m - \bar{m}). \tag{B-8}$$

The extra term can be interpreted as "extra" data points (e.g., section 8.9 of Gelman et al. (1995)) as follows. Form a new data vector $Y = \{y, \bar{m}\}$, with observational model $F = \{f(m), m\}$ and augmented noise covariance

$$\mathcal{C}_d = \begin{pmatrix} C_d & 0 \\ 0 & C_p \end{pmatrix}.$$

The log–posterior then can be written as

$$\chi^2 = (Y - F(m))^T \mathcal{C}_d^{-1}(Y - F(m)),$$

and the local linearization of the forward model F at any point will produce a Jacobian that looks like

$$X_u = \begin{pmatrix} J \\ I \end{pmatrix}.$$

Thus we can use the known results from the previous section for maximum likelihood theory, with a total of $n + p$ data points, and an augmented-data vector to consider for the residuals.

The upshot is that the "prior mean" used in each bootstrap optimization must be a sample from the prior distribution, centered on the MAP estimate using the real data, just as the real data are resampled with errors $N(0, C_d)$ and centered on the MAP estimate $f(\hat{m})$. For example, suppose a MAP estimate minimizing equation B-8 is $\hat{m}$. A bootstrap sample then will be $Y_i \equiv \{f(\hat{m}) + \epsilon_i, \bar{m}_i\}$, with $\epsilon_i \sim N(0, C_d)$, $\bar{m}_i \sim N(\hat{m}, C_p)$. For the linear case $F(m) = X_u m$, the proofs are trivial:

$$\hat{m} \equiv (X^T C_d^{-1} X + C_p^{-1})^{-1}(X^T C_d^{-1} y + C_p^{-1}\bar{m}) \tag{B-9}$$

$$Y_i = \{y_i, \bar{m}_i\} = \{X.\hat{m} + \epsilon_i, \bar{m}_i\} \quad \text{samples} \tag{B-10}$$

$$\begin{aligned} \hat{m}_i &= (X^T C_d^{-1} X + C_p^{-1})^{-1}(X^T C_d^{-1} y_i \\ &\quad + C_p^{-1}\bar{m}_i) \quad \text{MAP estimates} \\ \langle \hat{m}_i \rangle &= (X^T C_d^{-1} X + C_p^{-1})^{-1}(X^T C_d^{-1} X\hat{m} + C_p^{-1}\hat{m}) \\ &= \hat{m} \end{aligned} \tag{B-11}$$

$$\begin{aligned} \mathrm{Cov}(\hat{m}_i) &= \langle (\hat{m}_i - \hat{m})(\hat{m}_i - \hat{m})^T \rangle \\ &= (X^T C_d^{-1} X + C_p^{-1})^{-1}. \end{aligned} \tag{B-12}$$

The implications of this framework for the residual sum-of-squares now can be trivially inferred taking into account that there are $n + p$ "data" points and p parameters. Specifically, for the best-fit (MAP) model, we expect

$$\chi^2_{\text{data + prior}} = \langle (y - f(m))^T C_d^{-1}((y - f(m))) + (m - \bar{m})C_p^{-1}(m - \bar{m}) \rangle = (n + p) - p = n,$$

and for bootstrap samples,

$$\chi^2_{\text{data + prior}} = (y - f(m))^T C_d^{-1}((y - f(m))) + (m - \bar{m})C_p^{-1}(m - \bar{m}) \sim \chi^2_{n+p}.$$

In summary, the implied suggested recipe for the nonlinear CSEM problem, which we call the recentered Bayesian bootstrap, is (1) invert with the true data and actual prior $N(\bar{m}, C_p)$ to get the MAP model $\hat{m}$, (2) resample with Gaussian noise of correct variance added to the synthetic data produced by the MAP model $\hat{m}$, and use a Bayesian prior sampled from the centered Gaussian $N(\hat{m}, C_p)$ when inverting for the bootstrap samples.

We will see below that recentering the mean of the prior has strong implications for multimodal models. At the risk of incurring some bias, we also will use the nonrecentered Bayesian bootstrap, which is the same recipe above except that the prior samples are drawn from the original mean $N(\bar{m}, C_p)$. The reasons this more defensive strategy is useful will become clear in the simple examples below.

SIMPLE EXAMPLES

1) Analytical toy substitute for underresolved layers

Consider the nonlinear "degenerate sum-resistivity" two-parameter problem with $n = 1$ data point y, and predictive model $y = 10^{m_1} + 10^{m_2}$, measurement error $\epsilon \sim N(0, \sigma^2)$, and Gaussian prior $\mathbf{m} \sim N(0, I)H(m_1)H(m_2)$. ($H(x)$ is the Heaviside function, $H(x) = 1, x \geq 0, 0$ otherwise.) The model thus is confined to positive m_i.

The Bayesian posterior is of form

$$\pi(m_1, m_2|y) \sim \exp(-(10^{m_1} + 10^{m_2} - y)^2/2\sigma^2) \times \exp(-(m_1^2 + m_2^2)/2)H(m_1)H(m_2).$$

For example, with $y = 20$, $\sigma = 1.0$, the posterior is focused on an arc, and Figure B-1 shows both samples and an empirical marginal probability density function (PDF) of m_1 obtained using quadratures. For comparison, the marginal distribution obtained using the non-recentered parametric bootstrapping algorithm suggested

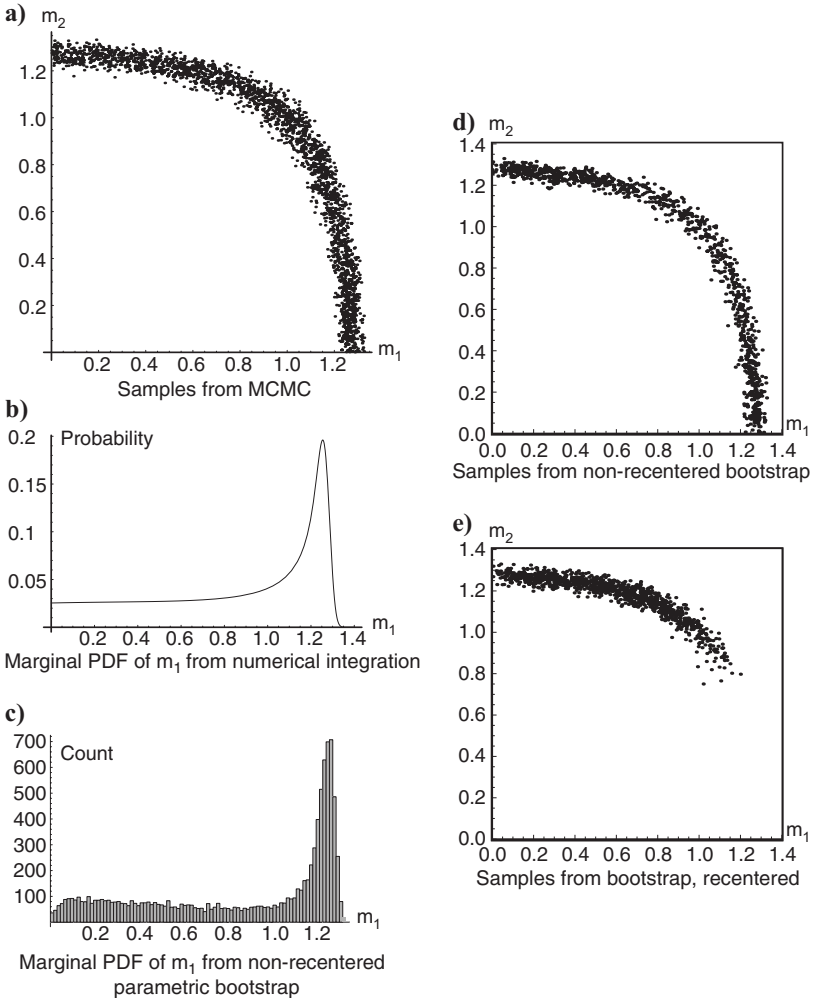

Figure B-1. (a) Joint samples of m_1,m_2 rigorously from MCMC. Inset (b) shows the exact marginal of m_1 from numerical integration, and (c) the approximate marginal of m_1 from non-recentered parametric bootstrapping. (d) Samples generated by the non-recentered bootstrap, and (e) samples from the recentered bootstrap, where the MAP model on the original "data" is at about (0.08, 1.27). The recentered bootstrap models are the suite of points on the maximum likelihood surface ($y_i = 10^{m_1} + 10^{m_2}$) nearest to independent draws from the recentered prior $N((0.08,1.27),I)$, and clearly such an ensemble undersamples the region with large m_1 values compared to MCMC.

above is shown. Specifically, the latter is: sample $\bar{m}_i \sim N(0,I)H(m_1)H(m_2)$ and error $\epsilon_i \sim N(0,\sigma^2)$, then estimate bootstrap samples $m_{i,1}$,$m_{i,2}$ by numerically minimizing

$$\chi_i^2 = (10^{m_1} + 10^{m_2} - (y + \epsilon_i))^2/2\sigma^2 + [(m_1 - \bar{m}_{i,1})^2 + (m_2 - \bar{m}_{i,2})^2]/2, \quad m_1,m_2 \geq 0.$$

Notice how the marginal distribution is distorted subtly, but is in general a reasonable approximation, especially because $n = 1$ and bootstrapping has origins as an asymptotic technique for large n (but remember that adding more data does not cure model-degeneracy stemming from the physics). The most obvious effect is the lower incidence of "smooth" solutions $m_1,m_2 \approx 1$ compared to the true posterior: Speculatively, this might widen interval estimates when we apply parametric bootstrapping to underresolved CSEM models. In this case, the recentered bootstrap will grossly underrepresent the frequency of large m_1 values.

REFERENCES

Bertsekas, D. P., 1982, Projected Newton methods for optimization problems with simple constraints: SIAM, Journal on Control and Optimization, **20**, 221–246, doi: 10.1137/0320018.

Chen, J., G. M. Hoversten, D. Vasco, Y. Rubin, and Z. Hou, 2007, A Bayesian model for gas saturation estimation using marine seismic AVA and CSEM data: Geophysics, **72**, no. 2, WA85–WA95, doi: 10.1190/1.2435082.

Constable, S., 2006, Marine electromagnetic methods — A new tool for offshore exploration: The Leading Edge, **25**, 438–444, doi: 10.1190/1.2193225.

Constable, S., and L. J. Srnka, 2007, An introduction to marine controlled-source electromagnetic methods for hydrocarbon exploration: Geophysics, **72**, no. 2, WA3–WA12, doi: 10.1190/1.2432483.

Constable, S. C., R. L. Parker, and C. G. Constable, 1987, Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data: Geophysics, **52**, 289–300, doi: 10.1190/1.1442303.

Denison, D. G. T., C. C. Holmes, B. K. Mallick, and A. F. M. Smith, 2002, Bayesian methods for nonlinear classification and regression: Wiley.

Efron, B., and R. Tibshirani, 1994, An introduction to the bootstrap: Chapman and Hall.

Evans, S. N., and P. B. Stark, 2002, Inverse problems as statistics: Inverse Problems, **18**, R55–R97, doi: 10.1088/0266-5611/18/4/201.

Farquharson, C. G., and D. W. Oldenburg, 2004, A comparison of automatic techniques for estimating the regularization parameter in non-linear inverse problems: Geophysical Journal International, **156**, 411–425, doi: 10.1111/j.1365-246X.2004.02190.x.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 1995, Bayesian data analysis: Chapman and Hall.

Gunning, J., 2003, Delivery website: http://www.csiro.au/products/Delivery.html. accessed 2 January 2010.

———, 2010, Supplementary material to Resolution and uncertainty in 1D CSEM inversion: a Bayesian approach and open-source implementation, http://www.csiro.au/products/Delivery.html, accessed 2 February 2010.

Gunning, J., and M. E. Glinsky, 2004, Delivery: an open-source model-based Bayesian seismic inversion program: Computers & Geosciences, **30**, 619–636, doi: 10.1016/j.cageo.2003.10.013.

Hall, P., 1992, The bootstrap and Edgeworth expansion: Springer.

Hansen, P. C., 1998, Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion: SIAM.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999, Bayesian model averaging: a tutorial: Statistical Science, **14**, 382–417.

Kelley, C. T., 1999, Iterative methods for optimization: SIAM.
Key, K., 2009, 1D inversion of multicomponent, multifrequency marine CSEM data: Methodology and synthetic studies for resolving thin resistive layers: Geophysics, **74**, no. 2, F9–F20, doi: 10.1190/1.3058434.
Kitanidis, P., 1995, Quasi-linear geostatistical theory for inversing: Water Resources Research, **31**, 2411–2419, doi: 10.1029/95WR01945.
Kitanidis, P. K., 1999, Generalized covariance functions associated with the Laplace equation and their use in interpolation and inverse problems: Water Resources Research, **35**, 1361–1367, doi: 10.1029/1999WR900026.
Liu, J. S., 2001, Monte Carlo strategies in scientific computing: Springer.
Løseth, L., 2007, Modeling of controlled source electromagnetic data: Ph.D. thesis, Department of Physics, Norwegian University of Science and Technology.
Madsen, K., H. B. Nielsen, and O. Tingleff, 2004, Methods for non-linear least squares problems (second edition.): Informatics and Mathematical Modeling, Technical University of Denmark, DTU.
Malinverno, A., and R. L. Parker, 2006, Two ways to quantify uncertainty in geophysical inverse problems: Geophysics, **71**, no. 3, W15–W27, doi: 10.1190/1.2194516.
Mitsuhata, Y., 2004, Adjustment of regularization in ill-posed linear inverse problems by the empirical Bayes approach: Geophysical Prospecting, **52**, 213–239, doi: 10.1111/j.1365-2478.2004.00412.x.
Neal, R. M., 2003, Slice sampling: Annals of Statistics, **31**, no. 3, 705–767, doi: 10.1214/aos/1056562461.
Nocedal, J., and S. J. Wright, 1999, Numerical optimization: Springer.
Oliver, D. S., N. He, and A. C. Reynolds, 1996, Conditioning permeability fields to pressure data: Presented at the fifth European Conference on the Mathematics of Oil Recovery.
Parker, R. L., 1994, Geophysical Inverse Theory: Princeton University Press.
Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. T. Flannery, 1992, Numerical recipes in C: The art of scientific computing, second edition: Cambridge University Press.
Raftery, A. E., 1996, Hypothesis testing and model selection, *in* Markov Chain Monte Carlo in Practice: Chapman and Hall.
Raftery, A. E., and S. M. Lewis, 1996, Implementing MCMC, *in* Markov Chain Monte Carlo in Practice: Chapman and Hall.
Sambridge, M., K. Gallagher, A. Jackson, and P. Rickwood, 2006, Trans-dimensional inverse problems, model comparison and the evidence: Geophysical Journal International, **167**, 528–542, doi: 10.1111/j.1365-246X.2006.03155.x.
Schnabel, R., J. Koontz, and B. Weiss, 1982, A modular system of algorithms for unconstrained minimization: Technical report CU-CS-240-82, Computer Science Department, University of Colorado at Boulder.
Snieder, R., 1998, The role of nonlinearity in inverse problems: Inverse Problems, **14**, 387–404, doi: 10.1088/0266-5611/14/3/003.
Tarantola, A., 1987, Inverse problem theory: Methods for data fitting and model parameter estimation: Elsevier, Amsterdam.
Tompkins, M. J., and L. J. Srnka, 2007, Marine controlled-source electromagnetic methods — Introduction: Geophysics, **72**, no. 2, WA1–WA2, doi: 10.1190/1.2557289.
Verrill, S., 2005, Java translation of UNCMIN. http://www1.fpl.fs.fed.us/optimization.html, accessed 2 February 2010.