Resolution and uncertainty in 1D CSEM inversion: a Bayesian approach and open-source implementation[a]

[a]This is an updated manuscript, with additional Appendices, of the published manuscript (Gunning, 2010). Please cite as both (Gunning, 2010) and the website.

James Gunning[*] **and Michael E. Glinsky**[†] **and John Hedditch**[‡]
[*]*CSIRO CESRE, Ian Wark Laboratory, Bayview Ave. Clayton, Vic. 3168, Australia*
[†]*CSIRO CESRE, 26 Dick Perry Avenue, Kensington WA 6151, Australia*
[‡]*13 Mohr Drive, Denmark WA 6333, Australia*

(November 16, 2010)

ABSTRACT

We show that resolution and uncertainty in CSEM inversion is most naturally approached using a Bayesian framework. Resolution can be inferred by either hierarchical models with free parameters for *effective* correlation lengths ("Bayesian smoothing"), or model–choice frameworks applied to variable resolution spatial models ("Bayesian splitting/merging"). We find that typical 1D CSEM data can be modelled 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 linearised approaches to uncertainty are usually 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, is 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 may be due to hydrocarbon accumulations. Taken in conjunction with seismic data for geological and structural delineation, this tool is potentially a powerful discriminator between high and low gas saturations, since 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 to date outlining the general nature of the CSEM acquisition framework (Constable (2006); Tompkins and Srnka (2007); Constable and Srnka (2007)). There are practical limitations on the suitability of the technique originating in basic physics principles, such as the impact of the air wave 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. Notwithstanding these, a large number of offshore petroleum prospects in the world still fall within the domain of applicability of the technique.

In our view, two overriding factors limit the usefulness of the technique. The 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. The second is that the dynamic range of conductivity from seawater to resistive anomalies (or deeper rocks) is usually at least several decades. These large contrasts in resistivity make the changes in observed fields large and thus useful in an exploration context, but they also make the inverse problem very nonlinear. In an inverse–problem context, the subsurface response is very poorly modelled 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 parameters, the solution null space is nearly always multimodal, badly scaled, and contorted in shape. We agree strongly with Snieder (1998) that this character makes these problems particularly difficult.

Meaningful 2 or 3 dimensional CSEM inversion is thus a hard problem. The strong absorption induces a large dynamic range in the gradient or sensitivity matrices, and since 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 in turn 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 is also, in our view, 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 to the inverse problem is taken. 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 an inference problem for the posterior distribution of a suite of model parameters and possible meta–parameters, and questions about resolution or uncertainty are answerable directly from this posterior distribution. Two recent geophysical examples using empirical–Bayes ideas for meta–parameter estimation are Malinverno and Parker (2006), and Mitsuhata (2004). Since 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 are also the most natural way to introduce knowledge from other data sources or professional expertise, with its requisite precision and inter-dependencies, 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 recognised that modest electrical anisotropy is now 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 miscontrued as advocating isotropic 1D inversions for problems that are clearly dominated by 3D effects or other forward–modelling 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 learnt 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, like 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*: (1) efficient bound-constrained Gauss–Newton and Marquardt techniques (2) multimodality and global optimization/enumeration (3) mode distinguishability or connectivity. Two methods for uncertainty evaluation follow in *Approaches to Inversion Uncertainty*, one fully Bayesian (MCMC), the other a faster, approximate technique we call the Bayesianized parametric bootstrap. Some *Examples Problems* follow to illustrate all the various ideas, a brief discussion of the *Software*, and the usual *Conclusions.*

APPROACHES TO RESOLUTION ISSUES

Resolution is most effectively understood as an interaction between the spatial representation ("gridding") of a inversion model, and the effective number of degrees of freedom which can be meaningfully estimated 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 *meta*–parameters 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). Secondly, resolution can be approached as a model–selection problem of choosing, among a family of models $k = 1...N$ 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 is often 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 modelled and (processed) data. This difference clearly absorbs instrumental noise, external and cultural noise, and errors in the forward modelling 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 reasons of analytical convenience, Gaussian error models are most useful, so the likelihood is often of form $L(d|m) \sim \exp(-(d-f(m))^T C_D^{-1}(d-f(m)))$, with $f(m)$ the forward model for the data, and the unknown noise parameters σ_i buried in the matrix $C_D = \text{diag}\{\sigma_i^2\}$. Some kind of dilution of this likelihood distribution may be required to model correlated or biased data. See for example, the discussions in Appendix E.

To provide some context, the 1D forward CSEM problem considered herein is a layer based model, usually with transmitter close ($\approx 30m$) 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 (Key, 2009; Constable et al., 1987), 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 are initially *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 modelling 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:

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

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 favouring 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 meta–parameters. Although a strict Bayesian would confine the statement of "posterior knowledge" to the full posterior distribution, certain characteristics of this distribution are usually of significant interest as point estimates. In particular (1) the largest mode of the joint posterior distribution – usually called the maximum aposteriori (MAP) point, and (2) the MAP point of particular marginal distributions, are of interest. Within the extremely common multi–Gaussian 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 flavours 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

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

where μ is a "free" parameter, and D is an operator whose null space does not overlap that of the forward model in $\chi^2_{\mathrm{misfit}}(m)$, always seem philosophically unsatisfactory, since the mathematical optimum is clearly at $\mu = 0$. Statisticians will instinctively feel that something is missing from the "objective function" that favours 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 (1) to a desired level of misfit, say, $\chi^2_{\mathrm{misfit}}(m) \approx n_d$. It is difficult to make statements about a strongly nonlinear problem with great confidence, but we may take inspiration from what is known about the linear case: Hansen's discussions are quite extensive, and recommend, roughly, $\chi^2_{\mathrm{misfit}} \approx n_d - n_p$, 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 $\chi^2_{\mathrm{misfit}}(m)$, and in practice these are usually poorly known (Farquharson and Oldenburg, 2004; Mitsuhata, 2004). Pessimistic estimates lead to oversmoothed solutions, and overoptimistic ones may prevent convergence at all. It is also 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 meta–parameters in a hierarchical framework. The normalisation associated with the meta–parameters then introduces the contributions which favour large values of the smoothing, and compete with the data misfit terms. A Bayesian approach will naturally induce simplicity both in the choice among models, but also in the inference of meta–parameters (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 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 meta–parameter, 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 1Ω.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 will be controlled by the resistivity–thickness product of the effective medium formed by these layers. This forms a useful test cases for many of the subsequent ideas.

CONSTRAINED BAYESIAN INVERSION FOR MODEL, NOISE AND SPATIAL CORRELATION

Our inversion code can perform several flavours 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 meta–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 multi-Gaussian model of the errors and with a multi–Gaussian expression for the prior with prior mean $\mathbf{m}_p$ and covariance $C_p(\mu)$, yields a posterior density

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

Here n_d is the number of measurements, and we will consider the particular case where $C_d(\sigma_n) = \sigma_n^2 \text{diag}\{\sigma_i^2\}$, the covariance matrix of the total error, is assumed diagonal and known up to the scalar σ_n^2. Similarly, the unknown meta–parameters μ may 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 may be 80% or so. The model–layer resistivity will be an "effective medium" property of a rock composite, whose (frequency) distribution will be 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: (1) it will have a heavy right tail, resulting from the lighter abundances of low–porosity facies (2) 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 $\rho = 0.8\Omega\text{m}$ ($\log_{10}(\rho) = -0.1$). A truncated Gaussian distribution for $m = \log_{10}(\rho)$ can be used to cover the prior support comfortably, has both of these required properties, and has 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 $i = 1 \dots n_p$ 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 be making 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 from, e.g. 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, may be written in the form

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

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

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

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

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

(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:

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

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

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

In the absence of correlation ($\alpha \to \infty$, or $\mu = 0$), the $W_{p,i}$ are simply 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 global scalar correction term for the (white) Gaussian noise distribution, and we presume the error estimates σ_i in $C_d(\sigma_n) = \sigma_n^2\text{diag}\{\sigma_i^2\}$, 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 both measurement and modelling 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–aposteriori, or empirical Bayes. The general ideas are easier to see in the fully linear problem, which, for reasons of space, we have supplied in Appendix C. This material supplies also some derivation details we skip in the following. The first and simplest idea is a pure "maximum aposteriori" 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 may be written (dropping $n_d \log(2\pi)$ and $\log|\text{diag}\{\sigma_i^2\}|$) as

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

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 up if the data are too noisy, however, as shown in the supplementary Appendix C

The smoothing and noise parameters are really meta–parameters 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–aposteriori method. The derivations for the EB case are somewhat messier, so we will show how things run for the joint maximum–aposteriori case first, and merely summarise the EB results later.

In the joint maximum–aposteriori 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*. The minimisation on σ_n involves only the first 2 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 (5), and dropping some constants, yields the reduced objective

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

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

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

The determinant must be evaluated numerically in general (an $O(n_p)$ operation since $\partial^T\partial$ is tri–diagonal), and this problem can be solved using any suitable one-dimensional 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 (6) needed for the optimization may be written as

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

The Gauss-Newton step here is thus 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

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

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

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

*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 with implied Hessian H and gradient $\nabla\chi^2_m$.

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

For the EB case, the derivations follow a similar spirit to supplementary Appendix C, 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 $(\Pi(\mu,\sigma_n|d) \sim \exp(-\chi^2_{\mathrm{smooth}}(\mu)/2)$:

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

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 block–wise forms arising from the use of "tear–surfaces" (discontinuities in the correlation) would also 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 in Figure 1. Synthetic data (inline $|E|$ field at 0.25,0.75,1.25 Hz, over offsets 1-12km) 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 are all synthetic. The inversion model is quite finely discretized, using layers of approximately 50m to 100m, 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 will not be able to be computed exhaustively, so some kind of heuristic for exploring model spaces is necessary. An obvious idea is some kind of recursive algorithm which will either adaptively refine a

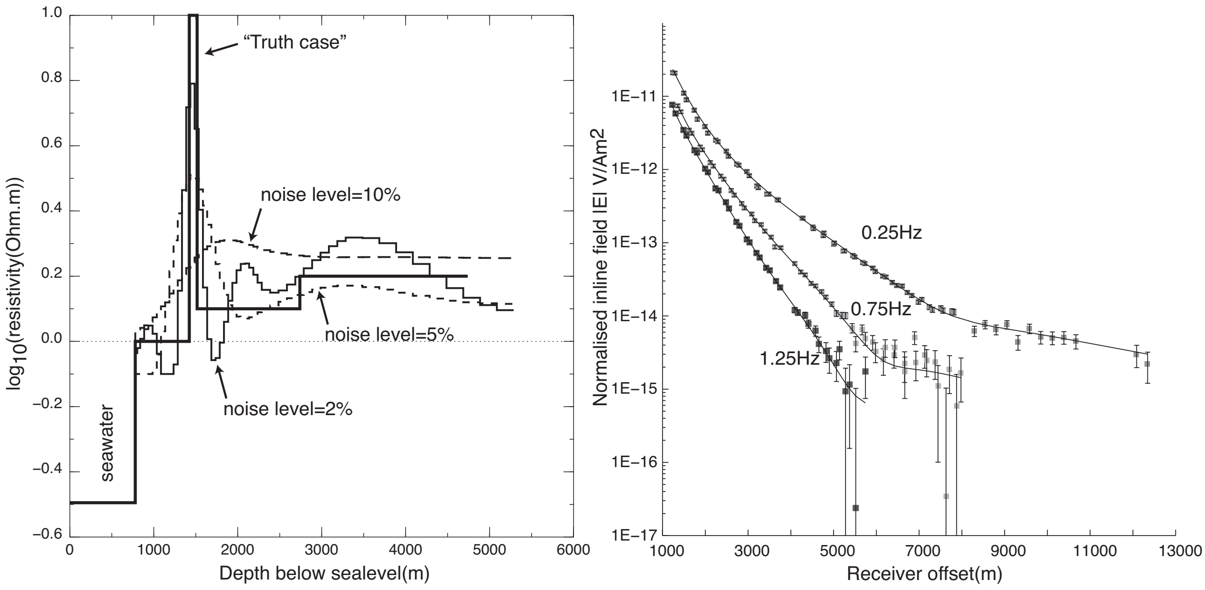

Figure 1: Left: "Bayesian smoothing" MAP inversions (μ as a meta–parameter) of CSEM data for the "truth case" model shown, for noise levels 10%,5% and 2%. Though the termination at the optimum is not explicitly controlled by $\chi^2_{\rm RMS} \equiv [(\mathbf{d}-\mathbf{F}(\mathbf{m}))^T C_d^{-1}(\mathbf{d}-\mathbf{F}(\mathbf{m}))/n_d]^{1/2}$, $\chi^2_{\rm RMS}$ 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. Right: typical data and fit at 5% noise. Note the error bars apply to $|E|$, not $\log_{10}|E|$, despite the scales.

very coarse model, or remove detail from a fine model, such that resolution is created at the depths statistically justifiable from the data.

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

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

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.

- Compute the MAP solution and MML for a very coarse, sufficiently deep 2 layer model (problem of dimension $n_p = 2$). This becomes the parent model.
- 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 ("favourite child") and best MML.
- 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 the way that preserves the existing spatial distribution. Global inversion is also very desirable for each candidate model, as superior solutions may 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 100m thick, 100Ωm reservoir buried 1km 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 may occur not only at the final optimum, but also during the optimization phase, and also may allow the minimisation 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 can easily induce slow convergence, so some subtlety and care in implementation is required. We have implemented both the projected Newton technique line–search described by Bertsekas (1982) and Kelley (1987), and also 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.

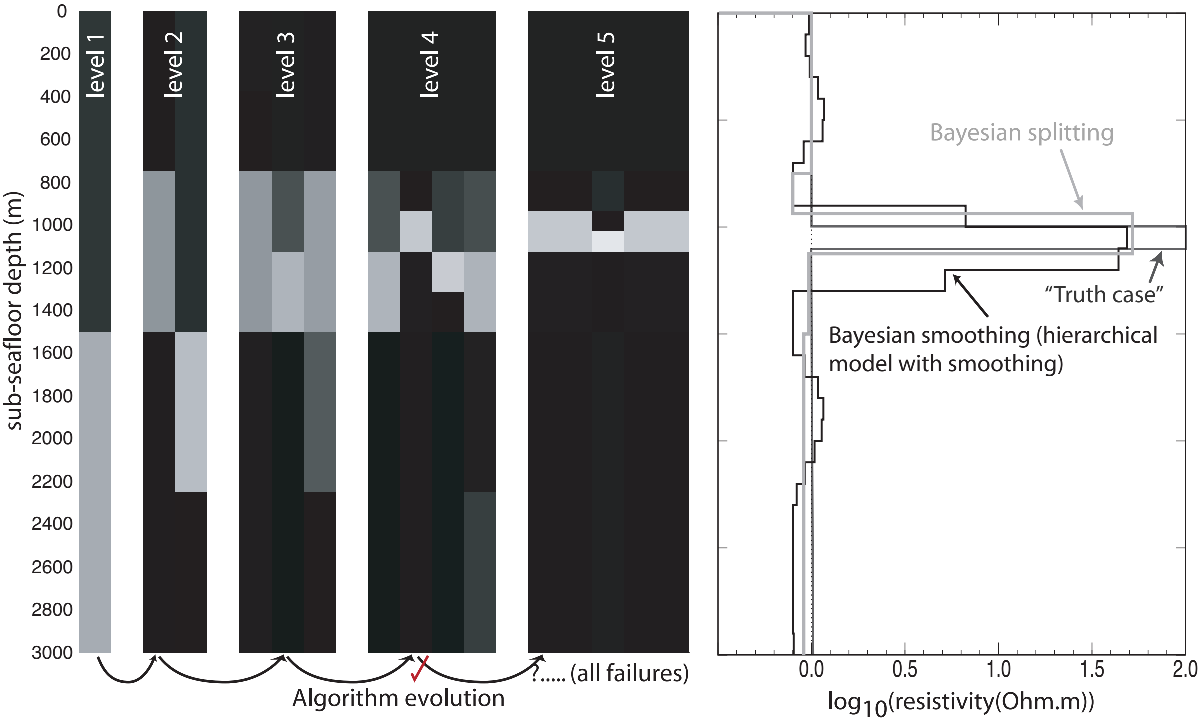

Figure 2: Canonical model under splitting: dark gray="truth case", light gray=final split model with minimal $-\log(\mathrm{MML})$ value, black="Bayesian smoothing" MAP inversion on fine grid for comparison.

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 0.8Ω-m, a respectable lower bound for shales based on Hashin-Shtrikman effective media theory. The bounds can be disabled or altered if desired.

Globalisation – multiple start solutions

Virtually all the modes of inversion except either very low dimensional models or excessively over–smoothed finer models will suffer from multi–modality. 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 of which may be very poor fits, but also several which may 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–optimisation pass). A variety of strategies are conceivable; we have implemented the following suite. a) Default (and mandatory): use a starting point determined by the startup file. b) 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 (a). (c) 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 shows 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" behaviour 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 due to either coding errors, bad scaling, poor termination 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 7 possible solutions is clearly evident. In this case, one mode is clearly very superior to the others, 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 also understand the relation between modes. We know from simple thought–experiments that there can be distinct optima which are separated by only weak probability barriers in the posterior surface, and knowledge of these near-degeneracies may clearly be useful in constructing MCMC strategies, among other reasons.

A particularly interesting question is that of how to construct the "lowest energy" path connecting two modes. This path should look like the gray "geodesic–like" 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 minimise the path integral

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

along a smooth parametrised 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.

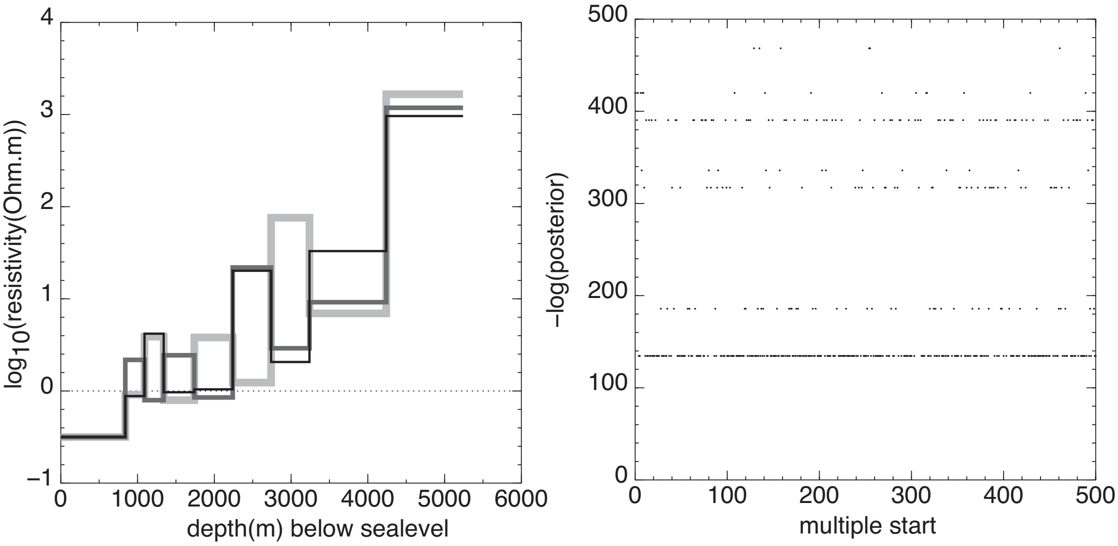

Figure 3: Left: Multiple local–mode MAP solution depth profiles of an 8–layer unsmoothed problem, shown as distinct curves. Right: -Log(posterior) of a large ensemble of random starts. Repeated convergence to particular optima is good evidence of sensible termination criteria.

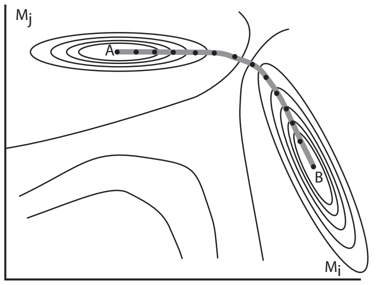

Figure 4: Optimal path connecting two modes A and B. The dots depict nodal points on a discretized approximation to the path, used in the optimisation algorithms detailed in the main text.

Algorithms to generate such paths are described in Appendix A, with some examples. In summary our findings are this. For very 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.

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 both 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 since the nonlinearity in CSEM is severe, the local 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, since the smoothing is really an artificial construct.

In this section, we confine the discussion to uncertainties within models, and present two canonical approaches to sampling. i) Markov–Chain Monte Carlo (MCMC) from the Bayesian point of view, and ii) 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 Markov–Chain Monte Carlo 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 "optimisation–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 (2003) 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 optimisation apparatus is explicitly used for each bootstrap sample. This has certain advantages for the CSEM problem, as the optimisation 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–parametrised models where the Bayesian prior is essential for stabilising the posterior. In the section *Bayesian Parametric Bootstrap* below, 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.

MCMC

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 optimisation and mode enumeration passes. For convenience, suppose the mode–enumeration has found a set of local optima $i = 1 \dots N_m$, which we characterise 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 normalised so $\sum_i \pi_N(i) = 1$. 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 reasonable 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 can potentially 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

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

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.

- Random jumps of form $q(m'|m) \sim N(m, \xi\hat{C}_1)$, where $\hat{C}$ is the linearised 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.
- "Layer–flip" type 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 $\min(1, \frac{\Pi(m')}{\Pi(m)} e^{-J_H})$ The jump sizes f_A, f_B are tunable parameters, typically $f_A \approx 0.4, f_B \approx 0.02$.

• '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 tunnelling 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 are very safe bets 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 efficiently implement a multi–component version for this problem. Some experiments with hybrid molecular–dynamics samplers (see Ch.9, Liu (2003)) 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 a 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 non-linear regression (i.e. minimise $\chi^2_{\text{misfit}} = (y - f(m))^T C_d^{-1}((y - f(m)))$, 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 $\hat{m}_i$ are then formed by nonlinear regressions of each resampled data set, i.e. minimising $\chi^2_{i,\text{misfit}} = (y_i - f(m))^T C_d^{-1}((y_i - f(m)))$. The statistics of the ensemble $\hat{m}_i$ are then 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 is typically 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, simply by treating the prior as additional "data". The upshot is that bootstrap model samples are then found by an optimization problem with *both* resampled synthetic data and resampled prior means $\mathbf{m}_p$. The distributional statement above also hold, 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 1km overburden shale (m_1), then two 50m reservoir layers (m_2, m_3), and shale underburden m_4. "Truth case" data are synthetically generated with the shale background 1Ωm ($m_1 = m_4 = 0$) and the reservoirs 100Ωm ($m_2 = m_3 = 2$). Since 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 may be subtle depths preferences.

Samples drawn using the re–centered 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. Firstly, in the Monte Carlo experiment where we generate synthetic data from the standard 3–layer canonical model with Gaussian noise, and invert for bootstrap MAP split–canonical (4–layer) models using globalised mode–searching, about 75% of the time the "most–likely mode" places all the anomaly in the deeper thin layer [†], so the layers are obviously thick enough to break the symmetry modestly. Secondly, the particular data used for the "truth" case produced a MAP solution $\hat{m} \approx (0.9, 2.3)$, so the recentered bootstrap samples are consequently 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 re–centered bootstrap has been fired off with a MAP solution in a particular part of parameter space, bootstrap realisations will clearly 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, since 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 realisation. Figure 7 shows the comparable output using MCMC (with heavily decimated sampling output), showing heavier support in the corners of the distribution and also for smoother models.

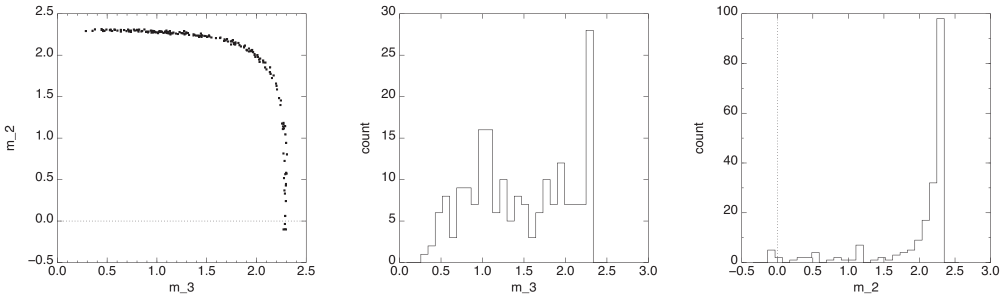

Figure 5: Left: joint samples of m_2, m_3 from *re–centered* parametric bootstrap on the split–reservoir canonical model. Middle and right: histograms of m_3 and m_2 from the samples, respectively. For discussion on asymmetry see main text.

For strongly non-linear 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 simply acquiring more data will not focus the posterior better: the model is intrinsically unresolvable, and only the uncertainty of the "effective medium"

[†] The bootstrap modes are also 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 $\hat{m}$ will be near either of these values.

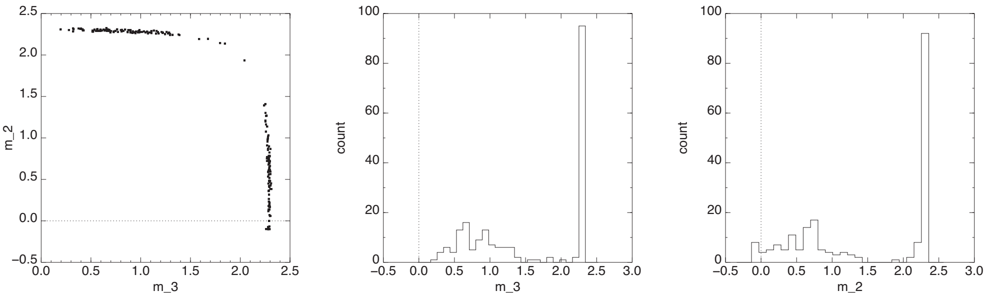

Figure 6: Left: joint samples of m_2, m_3 from *non–recentered* parametric bootstrap on the split–reservoir canonical model. Middle and right: histograms of m_3 and m_2 from the samples, respectively. For discussion on under–represented smooth models ($m_2 \approx m_3$) see main text.

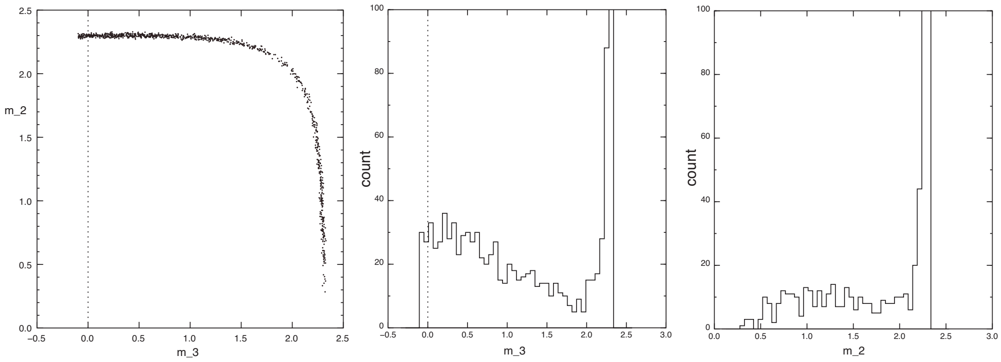

Figure 7: Left: joint samples of m_2, m_3 from MCMC sampling on the split–reservoir canonical model. Middle and right: histograms of m_3 and m_2 from the samples, respectively.

formed by m_1 and m_2 is reduced with more data.

Our recommendation at present is that the "non–recentered" 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. Since, in the CSEM case at present, the prior means are nearly always less than the MAP values, any biases are likely to reduce inferred resistivity values, which is a conservative tendency.

It is fairly likely that there exist adapted bootstrap techniques 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 1km deep in shale, in 1km of seawater, and extending from 10 to 450m in thickness; see Figure 8(a). 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 1Ωm, reservoir 100Ωm, and the data set is inline $|E|$ measurements at frequencies $f = .25, .5, .75, 1, 1.25, 1.5, 2$Hz, for offsets at 1km to 15km, on 500m spacings. Noise levels are taken as 5%, with a noise floor of 2.10^{-16}V/Am^2.

Figure 8(b) and (c) show MAP inversion images produced using"Bayesian smoothing" on two grids: (1) a regular 50m grid, and (2) a logarithmic grid (layer thicknesses increasing geometrically with depth). Both styles fit the data satisfactorily, so the inferred image is largely a function of the grid construction. Figure 8(d) is a plot of the MAP inverted reservoir thickness and resistivity–thickness (RTP) product, with error bars, based on a parametric study of a 3 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 also used to construct thickness and depth uncertainties for target layers. Discrete summations of the model probabilities ($\sim e^{\text{-MML}}$) over thicknesses/parameters not of interest is used to construct approximate marginal distributions for parameters of interest. Figure 8(d) 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 CMP5, where the "truth case" model was 135m thick (1000m 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 also the somewhat arbitrary extension of the 0.75Hz 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, the data is $|E|$ inline, from 1.2–12km. The true" model, data, and two styles of inversion are shown in Figure 10. Inversions have been run with both 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

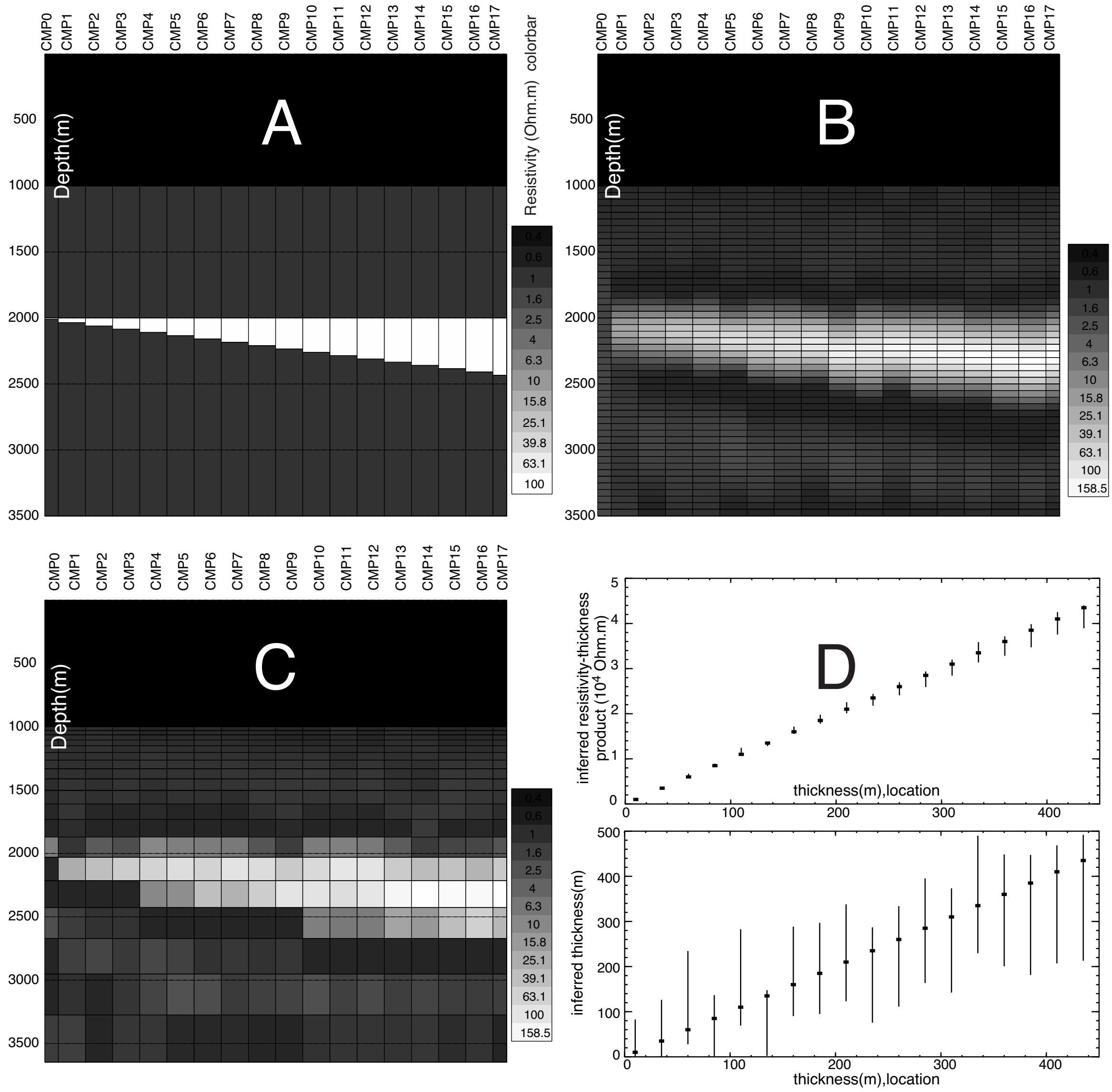

Figure 8: (a) Truth case wedge model: 100Ωm reservoir over 1Ωm shale background. (b) MAP inversion image using Bayesian smoothing on a regular 50m grid (c) MAP inversion image using Bayesian smoothing on a logarithmic grid (d) MML–based 3-layer inversions for depth, thickness and resistivity, showing marginal-distribution 95% error–bars for thickness, and resistivity–thickness (RTP) product. Clearly the RTP is much better identified by the data than thicknesses or resistivities.

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 1km of thickness).

To examine inversion uncertainty, an unsmoothed inversion based on an $p = 18$ layer

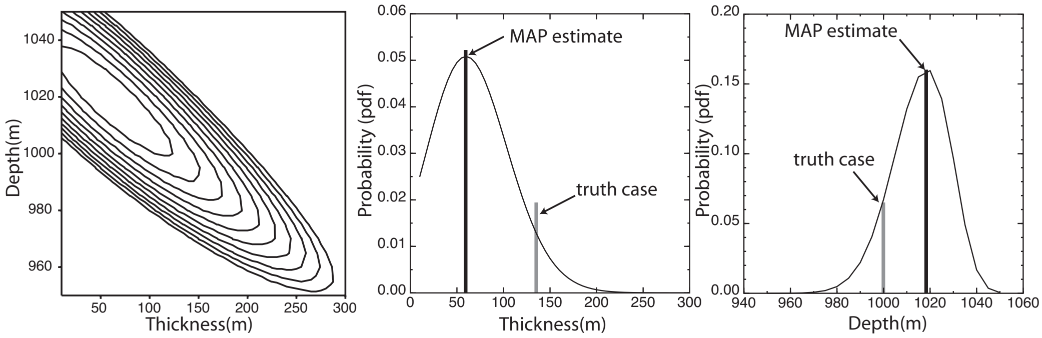

Figure 9: Left: Contour plot of marginal model likelihood of 3 layer model fit to CMP 5 data in wedge model, as a function of reservoir depth and thickness. Contours are at unit spacings of $\log_{10}$(MML), so 3 contours is about the conceivable span of model support in the data. Centre and right are 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 may be drawn from these plots, as the marginal distributions have considerable stochastic uncertainty under resampling of the data.

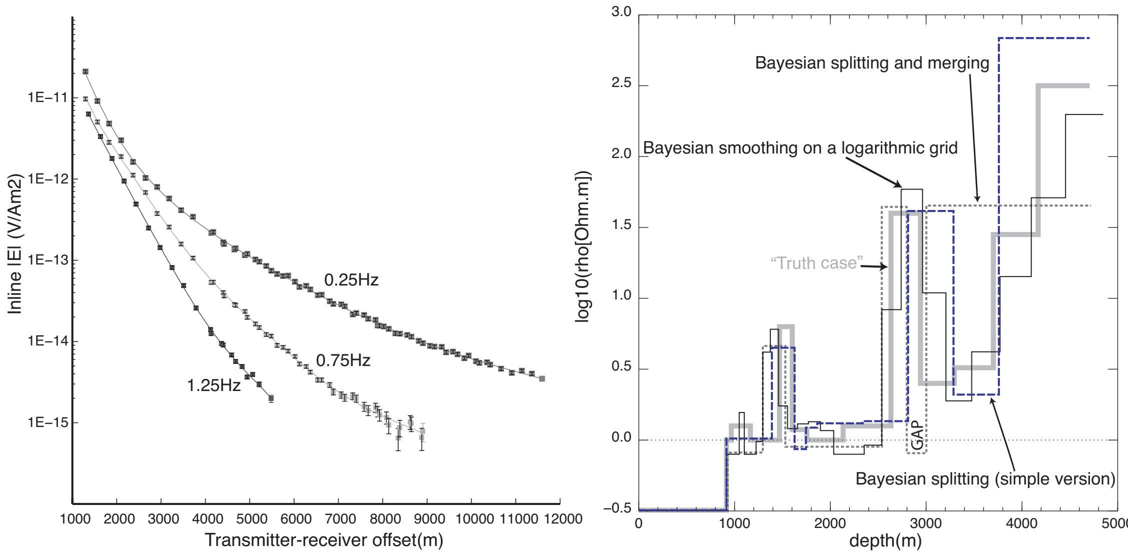

Figure 10: "Bird" model. Left: 3 frequency inline $|E|$ data used for inversion, with typical fitted model (Bayes–smoothing case). Right: Truth case data (thick gray), and MAP inversions for three styles of "resolution–detecting" inversion (Bayesian–smoothing, splitting and split/merge). All three approaches fit the data at approximately $\chi^2_{\mathrm{RMS}} = 0.9$.

logarithmic grid was run, with model priors set at $N(0,1)$, and noise–variance σ_n an additional unknown. This inversions has modest uncertainty about which layer to place the two anomalies in, and the marginal posterior distributions in the anomalous layers are clearly 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 ± one std deviation for Gaussian deviates) of each layer parameter m_i. These quantiles and the "truth case" model are shown in Figure 12 for both styles of calculation. 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 linearised 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 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) is still 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 GPL–style licence into the public domain, and may obtained at the CSIRO website (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 JNI. Test examples and usage documents etc are to found at the website.

CONCLUSIONS

We have presented two Bayesian approaches to both 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"). Globalised 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 licence agreement, with reference implementations of all the main ideas described in this paper.

ACKNOWLEDGEMENTS

We gratefully acknowledge the work of Kerry Key, Agusta Flosadottir, Alan Chave and Steve Constable in the Scripps forward modelling 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 approaches to finding a locally minimum path for the integral (11) is by discretizing the integral using some quadrature scheme. In the following examples, neither free–noise nor smoothing are 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 (11) is

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

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 minimise 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 optimisation will require, at least, $\nabla_{\mathcal{M}}\Lambda_{AB}$. Since the gradient $\nabla_{\mathbf{M}_i}\chi^2$ at the ith path–node is already coded and available, the bulk of the work is done. For completeness, the full joint gradient, in components, is

$$\begin{aligned}(\nabla\Lambda_{AB})_{ij} &= \sum_{i=1}^{N}\{\nabla^2_\chi(\mathbf{M}_i)\}_j(\Delta\mathbf{M}_i+\Delta\mathbf{M}_{i-1}) \\ &\quad+\sum_{i=1}^{N+1}(\chi^2(\mathbf{M}_{i-1})+\chi^2(\mathbf{M}_i))\frac{\mathbf{M}_{ij}-\mathbf{M}_{i-1,j}}{\Delta\mathbf{M}_{i-1}} \\ &\quad-\sum_{i=0}^{N}(\chi^2(\mathbf{M}_{i+1})+\chi^2(\mathbf{M}_i))\frac{\mathbf{M}_{i+1,j}-\mathbf{M}_{i,j}}{\Delta\mathbf{M}_i}.\end{aligned} \quad \text{(A-2)}$$

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 optimisation can now proceed using standard efficient methods. At present, we use a BFGS (variable metric) scheme (Nocedal and Wright, 1999), based on UNCMIN (Koontz and Weiss, 1982; Verrill, 2005). A simple example with known degeneracy is shown in Figure 14.

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

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

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

It is also possible to formulate the problem using Euler–Lagrange equations for the minimum path, which may be solved by, e.g. shooting. Experiments with the BFGS implementation scheme above indicate that the number of outer iterations required to stabilise (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 meta–smoothing or noise parameters, so $\mathbf{M} = \mathbf{m}$. Multi-start optimization is enabled, using layer–flipping, and the code ends up collecting 8 modes. Figure 15 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 may say the modes are "statistically connectable". The modes found in this example easily satisfy this condition, as shown in Figure 16.

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

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

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 formulae 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, since $Q^2 = Q$, so has eigenvalues 1 or 0, and also 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 normalised residuals

$$\begin{aligned} e &= C_d^{-1/2}(y_u - \hat{y}_u) = y - \hat{y} \\ &= (I - Q) y. \end{aligned} \quad \text{(B-1)}$$

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

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

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

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

which might be, typically, around 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 $\chi^2_{\text{RMS}} = 1$ 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.

Bootstrap. A synthetic bootstrap data set for the linear problem can then 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 is then

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

which is obviously 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

$$\begin{aligned} e_i &= y - X\hat{m}_i \\ &= (I - Q)y + Q\epsilon_i. \end{aligned} \quad \text{(B-4)}$$

There are two kinds of ensemble distributions that are of interest here:

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

- The distribution of $e_i^T.e_i$ formed over the bootstrap samples only, i.e. for a given, fixed y. This is what is actually 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

$$\begin{aligned} e_i &= (I - Q)y + Q\epsilon_i \\ &= V^T((I - I_p)Vy + I_p V\epsilon_i) \\ &= V^T((I - I_p)Vy + I_p\epsilon_i'). \end{aligned} \quad \text{(B-5)}$$

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

$$\begin{aligned} e_i^T.e_i &= ||(I - I_p)Vy + I_p\epsilon_i'||_2 \\ &= ||(I - I_p)V_y||_2 + \sum_i^p \epsilon_i'^2, \\ &= ||y - X.\hat{m}||_2 + \sum_i^p \epsilon_i'^2, \end{aligned} \quad \text{(B-6)}$$

  or

$$e_i^T.e_i - ||y - X.\hat{m}||_2 \underset{\epsilon}{\sim} \chi^2_p, \quad \text{(B-7)}$$

  i.e. the sampling distribution of the data misfit $e_i^T.e_i$ is χ^2_p, 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^2_i whose mean is offset by p to the right of the "best fit" χ^2 in the regression.

In both 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 non–hierarchical 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. sec. 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 can then 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 use the known results from the previous section for maximum likelihood theory, with an 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 centred on the MAP estimate $f(\hat{m})$. For example, suppose a MAP estimate minimising equation (B-8) is $\hat{m}$. A bootstrap sample will then 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:

$$\begin{aligned}
\hat{m} &\equiv (X^T C_d^{-1} X + C_p^{-1})^{-1}(X^T C_d^{-1} y + C_p^{-1} \bar{m}) && \text{(B-9)} \\
Y_i &= \{y_i, \bar{m}_i\} = \{X.\hat{m} + \epsilon_i, \bar{m}_i\} \qquad \text{samples} && \text{(B-10)} \\
\hat{m}_i &= (X^T C_d^{-1} X + C_p^{-1})^{-1}(X^T C_d^{-1} y_i + C_p^{-1} \bar{m}_i) \qquad \text{MAP estimates} && \\
\langle \hat{m}_i \rangle &= (X^T C_d^{-1} X + C_p^{-1})^{-1}(X^T C_d^{-1} X \hat{m} + C_p^{-1} \hat{m}) && \\
&= \hat{m} && \text{(B-11)} \\
\mathrm{Cov}(\hat{m}_i) &= \langle (\hat{m}_i - \hat{m})(\hat{m}_i - \hat{m})^T \rangle = (X^T C_d^{-1} X + C_p^{-1})^{-1}. && \text{(B-12)}
\end{aligned}$$

The implications of this framework for the residual sum–of–squares can now be trivially inferred taking into account that there are now $n + p$ "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 will use also the *non–recentered* 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" 2–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 is thus confined to positive m_i. The Bayesian posterior is of form

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

For example, with $y = 20$, $\sigma = 1.0$, the posterior is focused on an arc, and Figure 17 shows both samples and an empirical marginal PDF of m_1 obtained using quadratures. For comparison, the marginal distribution obtained using the non–recentered parametric bootstrapping algorithm suggested above is also 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 minimising

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

Notice how the marginal distribution is subtly distorted, but it in general a reasonable approximation, especially since $n = 1$ 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 may widen interval estimates when we apply parametric bootstrapping to under–resolved CSEM models. Also, in this case, the recentered bootstrap will grossly underrepresent the frequency of large m_1 values.

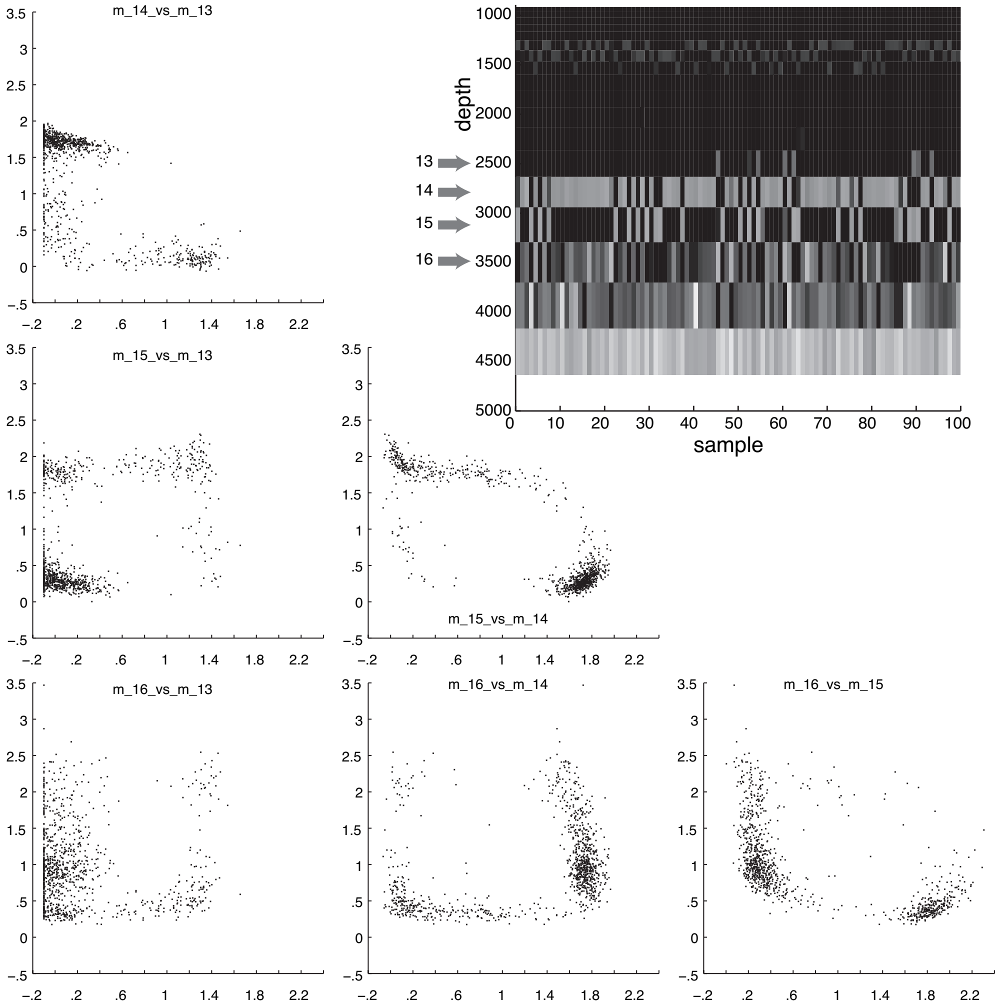

Figure 11: “Bird” model unsmoothed inversion uncertainty. Left: cross–scatterplots of bootstrap sample models in layers 13,14,15 and 16, where the deeper anomaly lies. Inset: grayscale model depictions of the model parameters in depth, for 100 “realisations” from the posterior using bootstrapping. In these samples, the anomaly prefers to reside solely in one of 2 or 3 layers over a “background”.

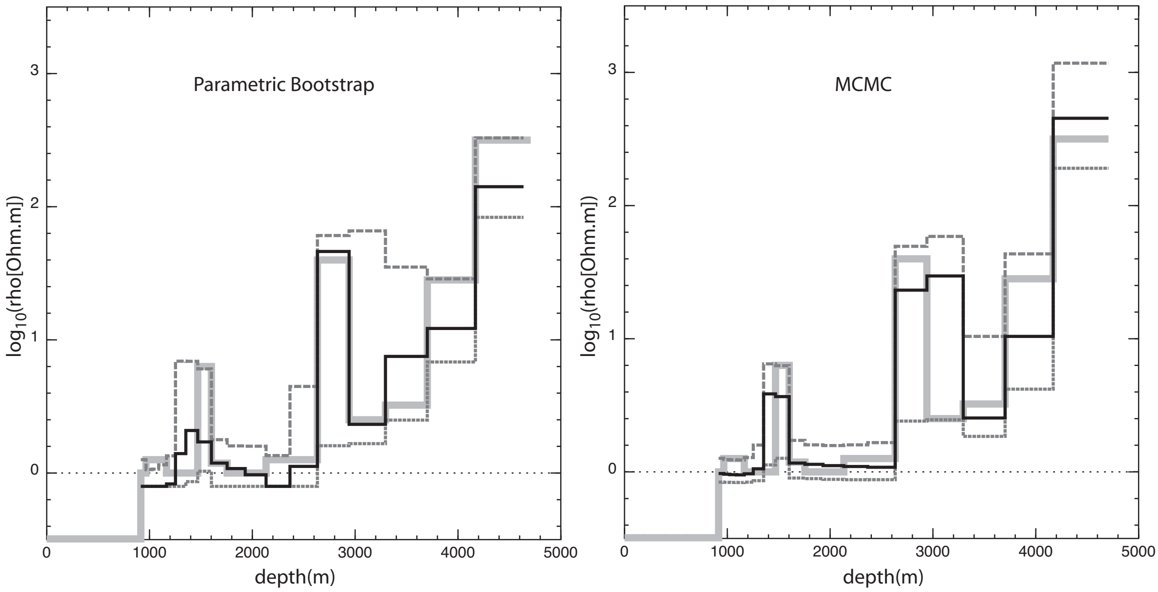

Figure 12: "Bird" model marginal model–interval estimates for an unsmoothed inversion. Left: P16(thin gray, dashed), P50(black), P84(thin gray, dashed) quantiles computed from a bootstrap ensemble of 1000 models. "True" model shown in thick light gray. Right: same graphs, generated from long, strict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

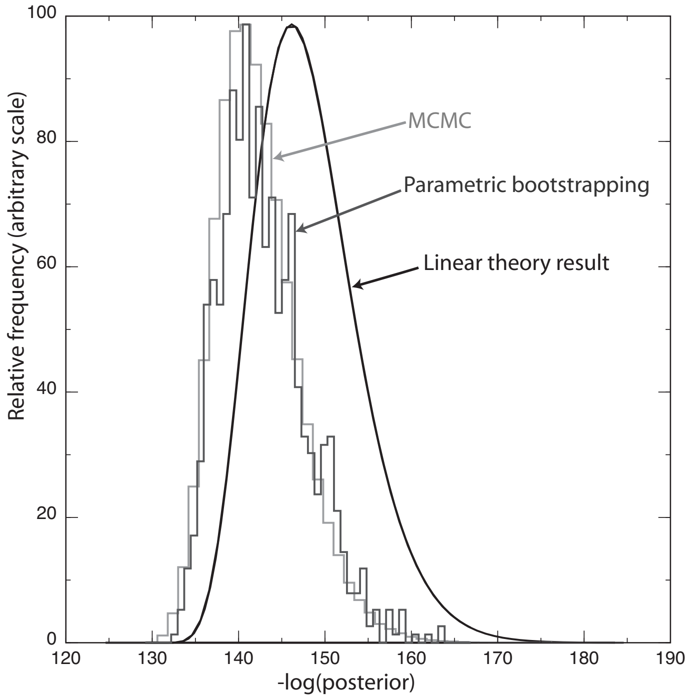

Figure 13: “Bird” model, posterior distribution of the -log(posterior) distribution (total prior and data misfit). Difference from the theoretical linear curve are due to the nonlinearity, but both MCMC and parametric bootstrapping generate comparable results.

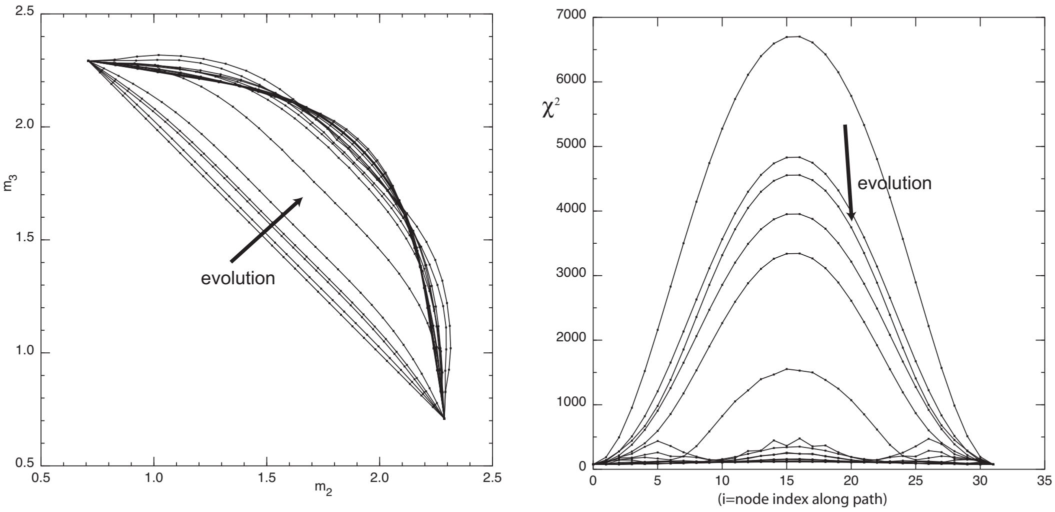

Figure 14: Evolution of "mode-linking" path under optimization, for the simple experiment of the 'split–canonical model', where the 100m resistive layer (buried 1km deep) of the canonical model is replaced by two 50m layers (parameters m_2, m_3). Two modes can be found by the "layer–flipping" strategy of the global optimizer, corresponding (roughly) to placing the resistive anomaly in each thin layer solely. Left inset: the $\log_{10}(\rho)$ parameters of each are plotted as the path evolves. The optimal path will be close to that describing a conserved resistivity times thickness sum over the two layers. Right inset: χ^2 cross–sections of the posterior surface as the optimal–path evolves. Clearly the early paths are extremely improbable ways to connect models.

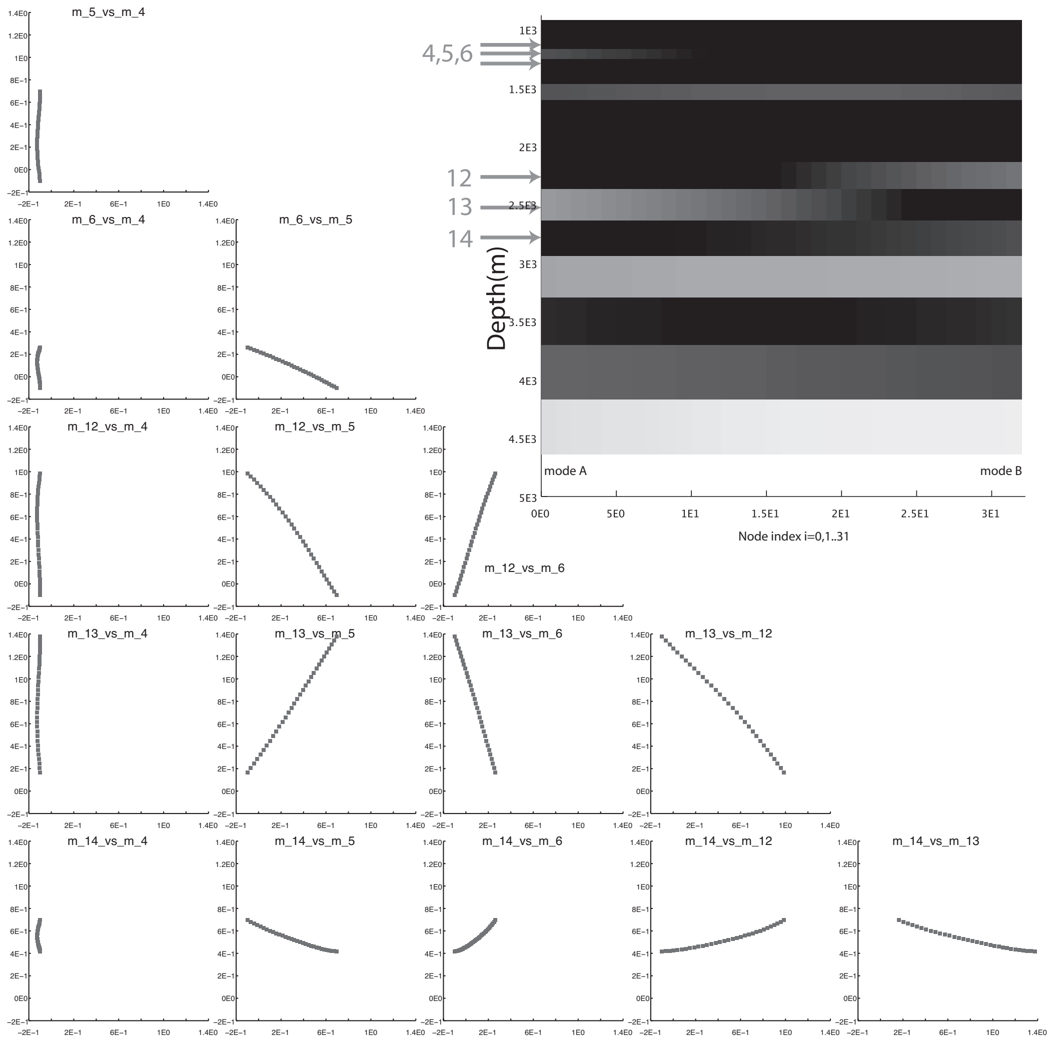

Figure 15: Scatterplot along "mode-linking" path between two endpoint–modes. The inset shows grayscale morphing of mode A into mode B along this path, with the exchange of resistivity quite obvious (light shades = resistive). Crossplots for the interesting layers 4, 5, 6, 12, 13, 14 (red arrows) are shown.

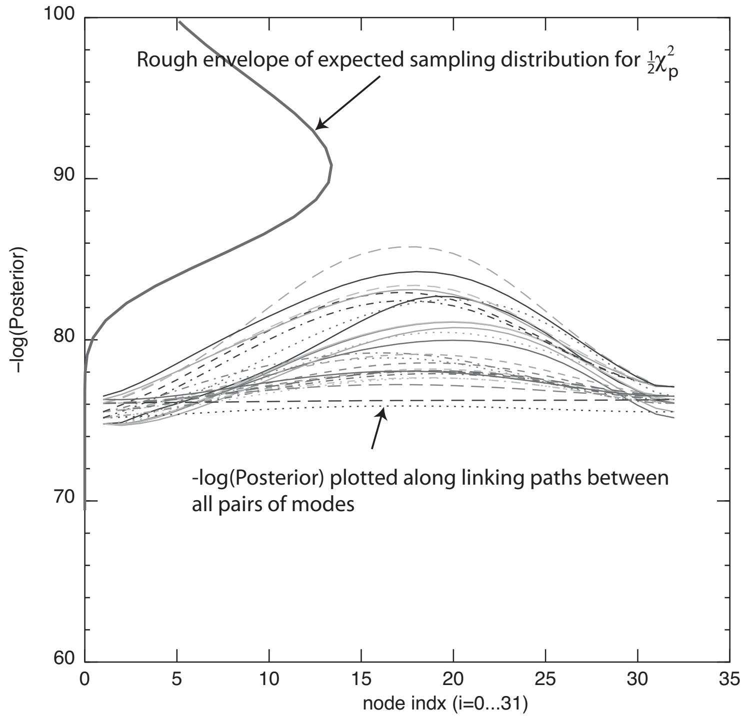

Figure 16: Plots of the -ve log–posterior (omitting constant terms) along the $8 \times 7/2 = 28$ possible links among 8 different modes, along the minimum integral path. On the left is a profile of the associated approximate offset–χ^2_p sampling distribution attached to the most likely mode. All these 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.

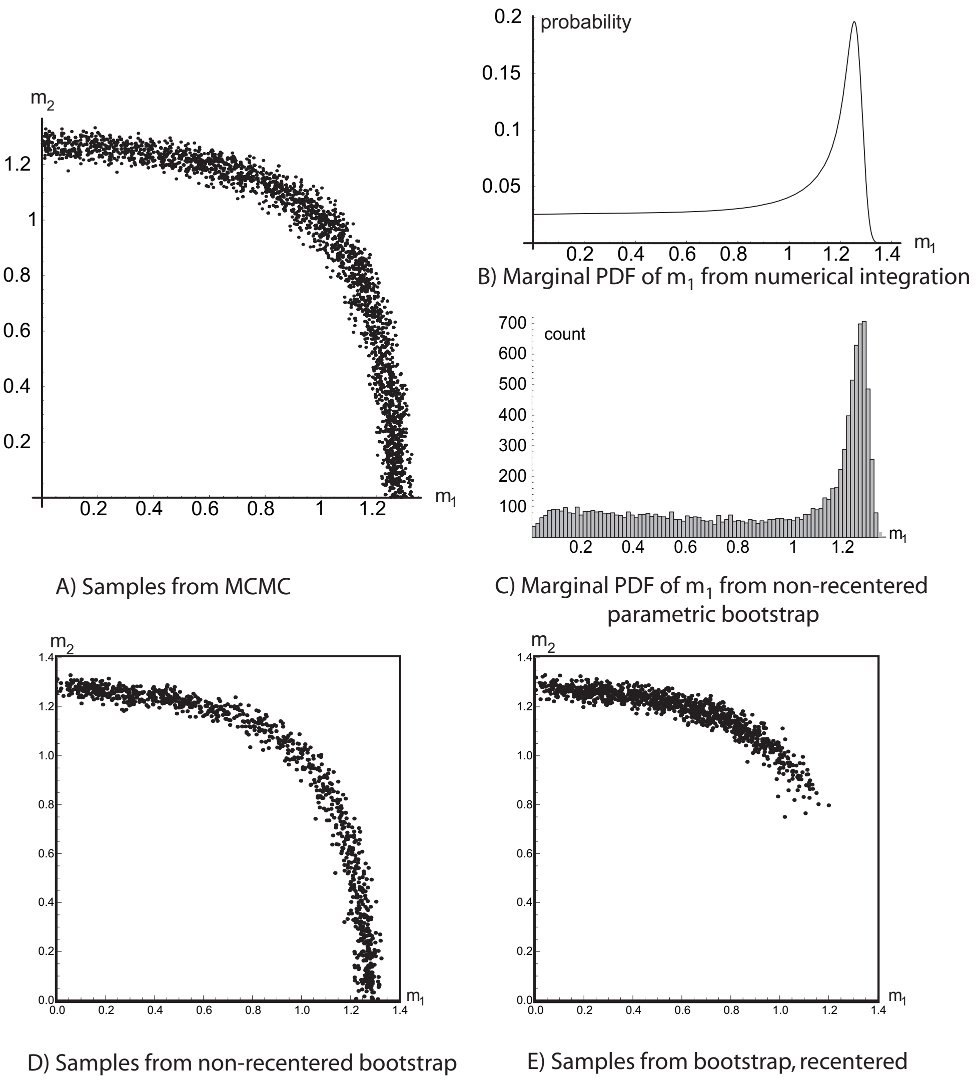

Figure 17: A) Joint samples of m_1, m_2 rigorously from MCMC. Inset B) shows the exact marginal of m_1 from numerical integration, and C) the approximate marginal of m_1 from non–recentered parametric bootstrapping. D) Samples generated by the non–recentered bootstrap, and E) Samples from the re–centered bootstrap, where the MAP model on the original "data" is at about $(0.08, 1.27)$. The re–centered 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 re–centered prior $N((0.08, 1.27), I)$, and clearly such an ensemble undersamples the region with large m_1 values compared to MCMC.

APPENDIX C

SIMPLE LINEAR PROBLEM ILLUSTRATING BAYESIAN SMOOTHING

True EM models are usually significantly nonlinear in the forward response. Here we study a more manageable toy problem which exhibits the same general character of resolution loss, and show how the use of a correlated Bayesian prior with additional smoothing length interacts with the noise level in the posterior inference. We construct a linear inversion problem where the forward model $\mathbf{F}(\mathbf{m}) = X\mathbf{m}$, with the matrix X behaving like a smoothing filter. For simplicity, let the model $\mathbf{m}$ sit on the integer lattice, so all correlation lengths are relative to this lattice also. Further, let the prior mean $\mathbf{m}_p = 0$ for simplicity, with the prior covariance $C_p(\lambda_p)$. Let also the prior $P_{\text{meta}}(\lambda_p)$ be flat in the region of interest. For a pure maximum likelihood approach in the joint space $\{\mathbf{m}, \lambda_p\}$, the optimization is then for

$$2\chi^2_{\text{Bayes}}(\mathbf{m}, \lambda_p) = (\mathbf{d} - X\mathbf{m})^T C_d^{-1}(\mathbf{d} - X\mathbf{m}) + \mathbf{m}^T C_p^{-1}(\lambda_p)\mathbf{m} + \log|C_p(\lambda_p)|, \qquad \text{(C-1)}$$

which, for fixed λ_p, has optimum at

$$\hat{\mathbf{m}}(\lambda_p) = (X^T C_d^{-1} X + C_p^{-1})^{-1} X^T C_d^{-1}\mathbf{d}. \qquad \text{(C-2)}$$

and thus reduces to a one parameter problem

$$\begin{aligned} 2\chi^2_{\text{Bayes}}(\lambda_p) &= (\mathbf{d} - X\hat{\mathbf{m}}(\lambda_p))^T C_d^{-1}(\mathbf{d} - X\hat{\mathbf{m}}(\lambda_p)) + \hat{\mathbf{m}}(\lambda_p)^T C_p^{-1}(\lambda_p)\hat{\mathbf{m}}(\lambda_p) \\ &\quad + \log|C_p(\lambda_p)| \\ &= \mathbf{d}^T (X C_p(\lambda_p) X^T + C_d)^{-1}\mathbf{d} + \log|C_p(\lambda_p)|, \end{aligned} \qquad \text{(C-3)}$$

after some algebra. Estimation of λ_p by finding the minimum of this function is a procedure we might call "MAP" estimation for meta–parameters.

It is interesting to plot this quantity $2\chi^2_{\text{Bayes}}(\lambda_p)$ for various toy model choices. The toy construction used for illustrating the recovery of a "localised" or "anomaly–like" underlying model, is:

- The lattice has $n_p = 50$ model parameters and $n_d = n_p$ model observations.
- X is a normalised, centered "box–car" smoothing filter with nonzero elements $X_{ij} = a_i$, $|i - j| \leq 5$, with the row constant a_i defined from $\sum_j X_{ij} = 1$.
- A truth–case model is set as a spike in the middle of the domain: $m_{\text{true}} = 3\delta(i = n_p/2)$.
- The truth–case "data" are obtained from $\mathbf{d} = X.m_{\text{true}} + \epsilon$, where the synthetic noise is $\epsilon_i \sim N(0, \sigma^2_{n,gen})$, with typically $\sigma_{n,gen} = 0.05$.
- The prior variance is modestly loose, with $\sigma_p = 0.5$ in the prior $C_{p,ij} = \sigma_p^2 \exp(-|i - j|/\lambda_p)$
- The observational error process is taken as independent normally distributed, so $C_d = \text{diag}(\sigma_d^2)$. Typically we have $\sigma_d = 0.03$.

In Figure 18, we show the true model (an impulse in the middle of the domain), box-car filtered to make data, to which Gaussian noise of varying levels is added. Again, we try and recover the model by using the hierarchical Bayesian model with exponential covariance $C_{p,ij} = \sigma_p^2 \exp(-|i-j|/\lambda_p)$, estimating lambda by the minimum in the curve $2\chi^2_{\mathrm{Bayes}}(\lambda_p)$. Figure 18 shows the synthetic data sets generated using this procedure, and the estimated models at the MAP point on the right hand sides, using varying noise levels σ_n. Clearly, as the noise diminishes, the estimated correlation length λ decreases, and the inverted model is a more aggressively deconvolved estimate. This behaviour is generic to inverse problems treated in this way. A similar finding (not for a CSEM problem) – is presented, Bayesian style, in Mitsuhata (2004) and, more classically, in Ory and Pratt (1995) and O'Sullivan (1986).

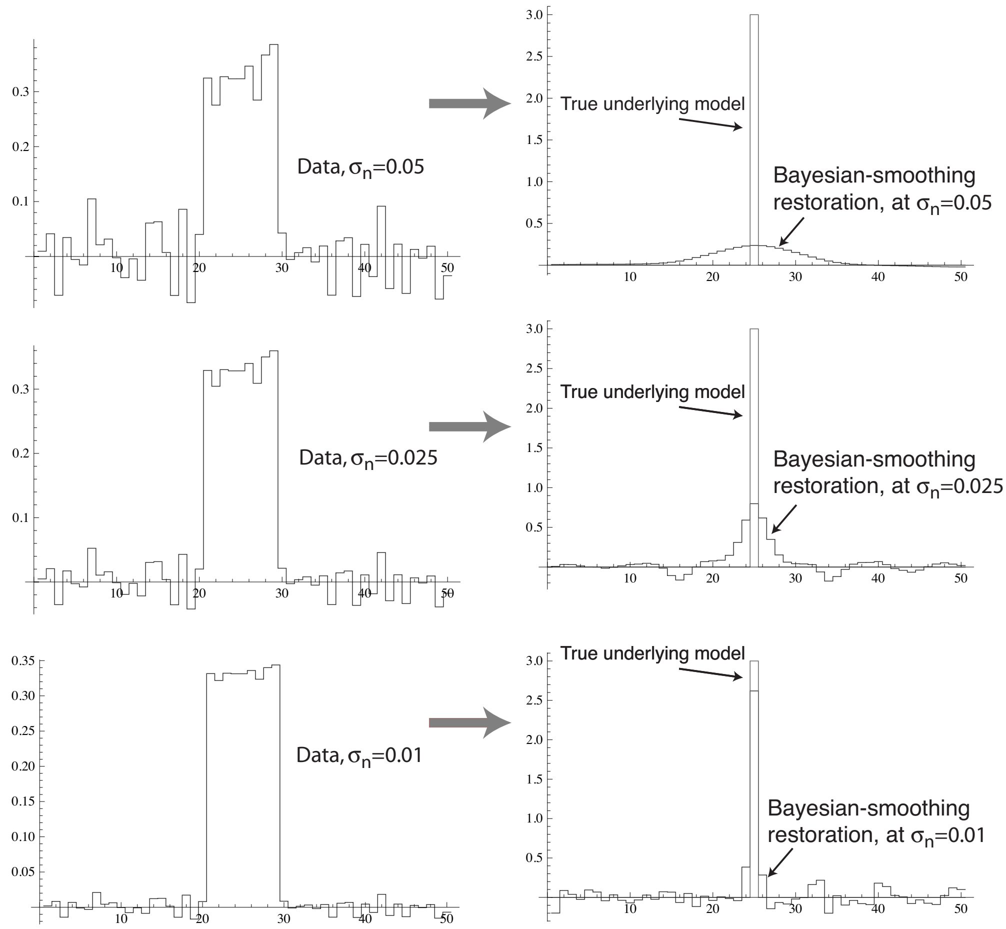

Figure 18: Bayesian–smoothing model estimation using varying noise levels and correlation parameter λ estimated from MAP values. The impulse-like true model is box-car filtered to make "data", with added noise, on the left. The MAP model recovered at the optimal (MAP–estimated) λ is overlaid on the underlying true model on the right insets.

The empirical Bayes alternative for meta-parameter estimation is to use the mode of the *marginal distribution* for λ, rather than the mode in the joint parameter space $\{\mathbf{m}, \lambda\}$.

Defining the marginal distribution by $\Pi(\lambda) = \int \Pi(\mathbf{m}, \lambda) d\mathbf{m}$, we see from equation (C-1) that the integral over $\mathbf{m}$ will be trivial, and yields

$$-2\log(\Pi(\lambda)) \sim \mathbf{d}^T (X C_p(\lambda_p) X^T + C_d)^{-1} \mathbf{d} + \log |C_p(\lambda_p)| + \log |X^T C_d^{-1} X + C_p^{-1}(\lambda_p)|, \quad \text{(C-4)}$$

which is identical to the MAP solution (eqn (C-3)), except the extra term from the determinant of the Hessian ($\log |X^T C_d^{-1} X + C_p^{-1}(\lambda_p)|$). Naturally, it is remarkably like the marginal model likelihood expression in the splitting methods, eqn (10).

Figure 19 shows plots of the two possible curves used to estimate the meta–parameter. Clearly the marginal method is more stable when the noise levels get high, as it is bounded away from infinity. For smaller noise levels, the alternatives appear to converge on to the same solution.

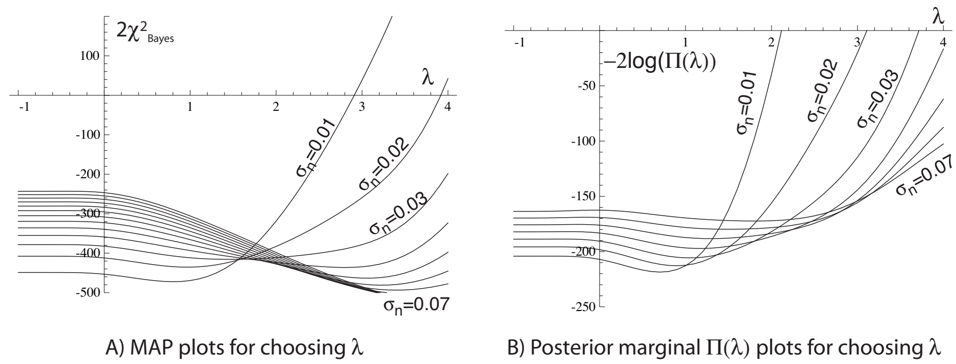

Figure 19: Alternative methods for meta–parameter λ estimation: (A), minima of the MAP point in joint parameter space $\{\mathbf{m}, \lambda\}$, with $\mathbf{m}$ analytically removed. (B) the mode of the marginal distribution of λ, $\Pi(\lambda)$. For modest and small noise levels, the alternatives are similar.

APPENDIX D

BOUND-CONSTRAINED OPTIMIZATION

Here we present methods for simple box–constrained optimization of the model update **m**, equation (7).

Take each component of m to have lower and upper bounds $m_{L,i}, m_{U,i}$. Define the component-wise "box–projection" operator P by

$$P(m_i) = \begin{cases} m_i, & m_{L,i} < m_i < m_{U,i} \\ m_{L,i}, & m_i \leq m_{L,i} \\ m_{U,i}, & m_i \geq m_{U,i} \end{cases}$$

Recall also that the unconstrained full-Newton step can always be expressed in the form

$$\Delta m \equiv m' - m = -H^{-1} \nabla f,$$

where H is the current full Hessian, and ∇f the effective gradient of the function being optimized.

Given a fixed, small constant ϵ, at the kth iteration of the algorithm, we compute ‡

$$\epsilon_k = \min(\epsilon, |m - P(m - \nabla f)|),$$

and define an *active* set of constraints (lower, upper, combined) by the set $\mathcal{A}_k$, with

$$\mathcal{A}_k^L = \{i| \; m_i - m_{L,i} < \epsilon_k, \frac{\partial f}{\partial m_i} > 0\} \quad \text{(D-1)}$$

$$\mathcal{A}_k^U = \{i| \; m_{U,i} - m_i < \epsilon_k, \frac{\partial f}{\partial m_i} < 0\} \quad \text{(D-2)}$$

$$\mathcal{A}_k = \mathcal{A}_k^L \cup \mathcal{A}_k^U \quad \text{(D-3)}$$

(roughly, the components both near the boundary and driven towards it by the gradient). Given the active set $\mathcal{A}_k$, a modified Hessian H_M is constructed as

$$H_{M,ij} = \begin{cases} H_{ii}\delta_{ij}, & i,j \in \mathcal{A}_k \text{ and} \\ H_{ij} & \text{otherwise} \end{cases}$$

with corresponding modified gradient

$$\nabla f_{M,i} = \begin{cases} H_{ii}(m_i - m_{L,i}), & i \in \mathcal{A}_k^L \\ H_{ii}(m_i - m_{U,i}), & i \in \mathcal{A}_k^U \\ \nabla f_i & i \notin \mathcal{A}_k \end{cases}$$

‡An alternative is

$$\epsilon_k = \min(\epsilon, \{|m_i - P(m - \nabla f)_i|\}).$$

Line-search version

Here, we solve for the full projected Newton update

$$\Delta m = -H_M^{-1}.\nabla f_M.$$

Stability is then imposed by a a Arjimo–like backtracking search in α along the direction

$$m' = P(m + \alpha \Delta m),$$

which eventually amounts to a stable Newton–like step in the subspace of non–active constraints, with active constraints quadratically converging to the correct active bound in concert. In practice, due to the poor scaling, it is very common that the projection operator P rotates the search direction substantially from the Newton direction. See Figure 20 for an example. Some subtlety in implementation is required here. Thus, although we can guarantee that the modified Newton direction Δm is a descent direction, i.e. $\nabla f.\Delta m < 0$, since H_M is positive definite, we may well have $\nabla f.(m' - m) > 0$, since the projection operator can exert a further rotation.

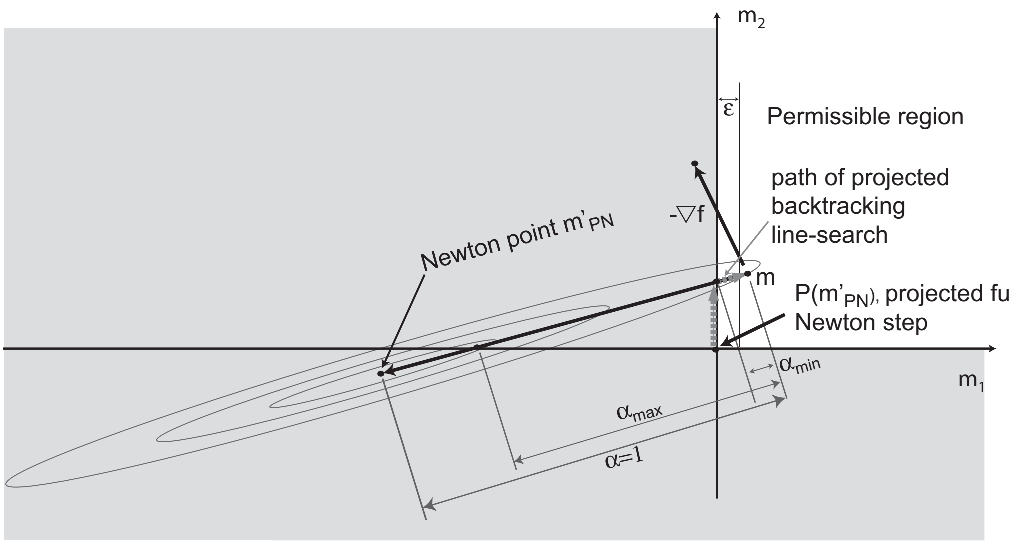

Figure 20: Typical Projected Newton step for a badly scaled problem, for a point just before the onset of active constraints. Note the rotation of the full Newton step by the projection operator. The projected direction can be nearly opposite the downhill gradient.

A few more details of the backtracking are important. We define

- $\alpha_{\max}$ as the maximum fractional reduction in the modified Newton step Δm that leaves $P(m + \alpha\Delta m)$ unchanged. Backtracking values $\alpha > \alpha_{\max}$ are pointless, as they map to the same point, so the backtracking must start no further away than $\alpha_{\max}$.
- $\alpha_{\min} \equiv \min\{|P(m'_{PN})_i - m_i|/|m'_{PN,i} - m_i|\}$, for "free" components i, i.e. those components where $\alpha|m'_{PN,i} - m_i| < \epsilon_{\text{FP}}$, where ϵ_{FP} is about floating–point machine

epsilon. This is roughly the "shortest backtrack" available over all components which are not sitting on a constraint boundary: see Figure 20.

By default we start the backtrack at $\alpha = \alpha_{\min}$, since the Newton–quadratic convergence is not expected to set in until the active set has stabilised, where the active components are settled on the boundary, and then $\alpha_{\min}$ is not determined by these active components. A final step–limiting (trust–region) restriction on α is to reduce it by factors of 2 until $\max\{|P(\alpha\Delta m + m)_i - m_i|\} < \Delta m_{\max} \approx 0.3$. This occurs typically only a few times early on in the outer loop.

In principle, the projection operator should perform no further rotations of the Newton direction if we restrict the backtrack to $0 < \alpha < \alpha_{\min}$. Full steps of $\alpha_{\min}$ are expected to be taken early as as the trajectory "crashes" into the wall. For severe cases, we may find that $\nabla f.\alpha_{\min}\Delta m \approx \delta > 0$, which appears like a failure to find a descent direction. The usual reason is that large values of the gradient ∇f are pointing into active constraints, the Newton direction is very nearly perpendicular to the gradient, and the dot–product is only just positive from floating point error. Another possible cause, just prior to convergence, is $|m'_{PN} - m|$ becomes very small, with the same floating–point error consequence.

Typically, the poor scaling and the magic constant ϵ interact such that some significant slowing of the convergence occurs until the set of active constraints stabilises, and Newton–like behaviour emerges in the relevant subspace.

Trust–region version

It is possible to hybridize the active set and projections described above with a trust-region method like the Levenberg-Marquardt algorithm, where the Newton update is modified to a form like $(H_M + \lambda_{\text{LM}} I)\Delta m = -\nabla f_M$. Here λ_{LM} is the Marquardt parameter, and is adjusted at each step to compromise best between steepest–descent and full Newton. The general idea is to use the active set information such that, once the active set settles in, the conventional Levenberg-Marquardt machinery will operate in the subspace of inactive constraints, with active components carried along harmlessly.

Trust-region methods have a reputation for better navigation and greater robustness in the remote parts of the fitting surface, where residuals will be large. The algorithm used closely follows that of Madsen et al. (2004), except we hybridize it with the active set detection thus:

- The active set is detected as before, and off–diagonal elements zeroed to form H_M
- If the active set changes, the Marquardt parameter λ_{LM} is reset to an initial guess, based on Gershgorin bounds of the maximum eigenvalue of the "inactive" submatrix of H.
- No preliminary step–reduction to $\alpha_{\min}$ is applied. The contracting trust region controls the step length.
- Solution of the modified problem $(H + \lambda_{\text{LM}} I)\Delta m = -\nabla f_M$ are first projected back into the feasible region before the usual algorithms for adjusting the Marquardt scaling parameter are applied.

On average the trust region method appears faster than the line–search in a majority of cases, but this is probably dominated by the snaking character of the optimization surface with modestly large residuals, rather than theoretical behaviours close to the final optimum.

APPENDIX E

SUGGESTIONS FOR DEALING WITH MORE COMPLEX NOISE

In real data, the noise process is a complex mixture of instrumental, processing, environmental/cultural noise, and modelling noise, which we might write symbolically as

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_{\text{instr/proc}} + \boldsymbol{\epsilon}_{\text{env}} + \boldsymbol{\epsilon}_{\text{model}}.$$

In the deep marine environment, external EM noise sources $\boldsymbol{\epsilon}_{\text{env}}$ are not as serious as on land, and one might reasonably hope that service companies' instruments are calibrated within say, 5%. So some parts of the external and instrumental noise at least are uncorrelated, and hopefully, close to zero mean. Indeed, the independence assumption in modelling noise is critical to the success of stacking techniques in reducing external noise, and this is heavily used in the processing steps for CSEM data.

The most difficult piece of the noise is often the modelling error $\boldsymbol{\epsilon}_{\text{model}}$, because many effects caused by inadequacy of the forward model will appear as systematic structure in the residuals. In the context of this paper, serious systematic effects might occur from e.g., anisotropy, bathymetry approximations, and other 3D heterogeneity effects in the subsurface. Moreover, the size and nature of these systematic effects may not ever be able to be estimated from the data itself, so the accomodation of them must depend on some subjective judgement of the user. An example of such a systematic error is shown in fig. 21. It is often possible (even in parsimonious models) to find model fits that remove the systematic structure, and too rigorously penalise misfits such as those shown in fig. 21 under the uncorrelated Gaussian model. The severity of such exclusions naturally increases with the tightness of the error bars, but also significantly with the number of data.

There are at least 3 possible approaches to allow for less exclusive likelihood responses in the presence of systematic errors.

- Reduce the *effective* number of data. Since correlated residuals contain partially redundant information, the number of "true" *effective independent* measurements may be very much less than the raw data count. Various ways to dilute the data can be considered. See "Data reduction" below.
- Explicitly try to model the noise as a mixture of systematic components and random components. Particular structural forms for the systematic components must be chosen. See "Explicit systematics modelling" below.
- Try to absorb correlated residuals into a general correlation–matrix structure which will more flexibly accomodate residual trends. The central ideas involve a Bayesian spin on some statistical apparatus called *shrinkage*, which strikes a compromise between a prior noise covariance and the rank–deficient sample covariance. More broadly, the full noise covariance becomes part of the inference problem, and the extended theory will aim to produce predictive distributions for the important model parameters (including model selection aspects) which marginalise over this uncertain noise covariance.

Of these 3 ideas, the first is simplest to implement, and is closely related to frequentist approaches to data analysis. The second idea is not too hard for systematic components

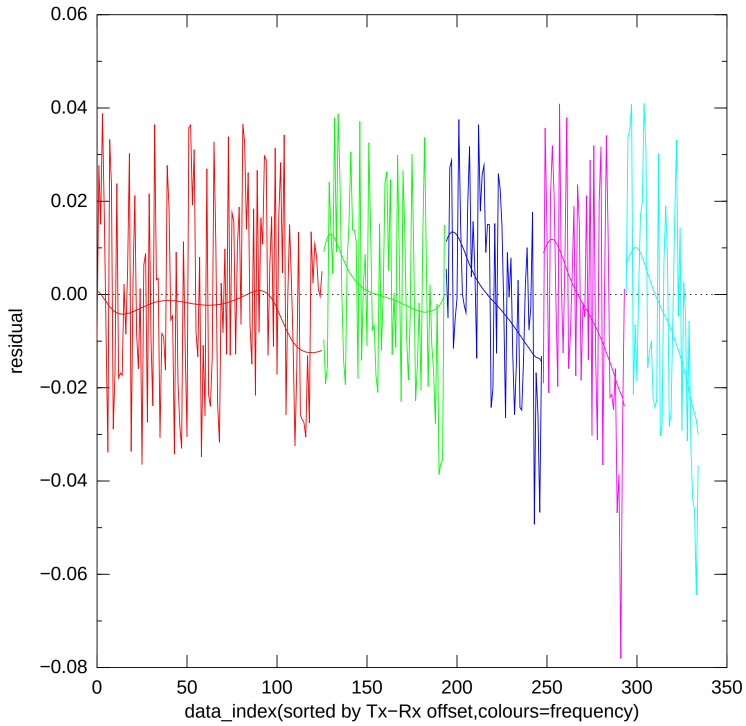

Figure 21: Residual errors in $|E|_{\text{inline}}$, for a model fit where it is desired to treat the residual trend as probabilistically 'acceptable'. This data set has 100m offset spacings.

that have a relatively simple form, but is difficult for measurements that may oscillate rapidly with offset, e.g. phase. The third idea is the most technically difficult, since putting a diffuse prior on the covariance structure in even simple Gaussian linear models makes for analytical and computational challenges (Gelman et al., 1995). We will illustrate the first two of these ideas in turn, but defer the detail for the third approach to a forthcoming paper.

To illustrate these ideas, we use an example data set corresponding to a single 250m resistor buried 850m below 1km of seawater. The residuals have had systematic error added of the form illustrated in fig. 21. It is desired to find strategies that render model fits with residual trend structure *more* likely, relative to models that can flatten the residuals, and which usually overwhelm the former in terms of posterior odds when the error model is strictly i.i.d Gaussian.

Data reduction

In traditional linear regression theory, posterior covariances for the model parameters take the form

$$\tilde{C} = (X^T C_d^{-1} X)^{-1}$$

where X is the sensitivity/Jacobian/design matrix, and C_d the matrix of data errors. Typically, with independent data, the internal operations count in this expression scales as $\sim n_d$, so the posterior variances scale as $1/n_d$ (standard deviations as $n_d^{-1/2}$). This is the central limiting behaviour that explains the success of stacking methods in reducing noise in geophysical processing. If the choice of C_d is correct, these methods yield what are called *consistent* estimators, i.e. the parameter confidence intervals contract with n_d around the true value. However, it is known that if the errors are heteroscedastic or correlated, ordinary least–squares is an inconsistent estimator, i.e. the confidence intervals contract around the wrong value.

Thus, when the data are correlated, C_d should become a fuller matrix, the condition number of C_d deteriorates (some of the eigenvalues get appreciably smaller), and the posterior variance increases. The "correct" form of C_d thus automatically compensates for the fact that the data are, in effect, being "double" counted. Since we do not know what form C_d should actually take, and may be limited to using a code where is is purely diagonal (and that is the extent of our control of it), we may simulate the effect of correlation by inflating the elements $C_{d,ii}$ by n_d/n_{eff}, where $n_{\text{eff}} \leq n_d$ is an estimated *effective* number of independent measurements. In other words, wind up the error bars by $\sqrt{n_d/n_{\text{eff}}}$. This will make $\tilde{C}$ scale as $\sim 1/n_{\text{eff}}$.

An example involving straight–line curve fitting to data with correlated errors may be helpful here. Consider data y measured at x–points $i = 1, 2 \ldots n$, with the parameters $m = \{\text{intercept,slope}\}$. X has n rows $\{1, i\}$. Suppose there is a systematic measurement error, so C_d has the form

$$C_d = \begin{pmatrix} 1 & \rho^2 & \rho^2 & \ldots \\ \rho^2 & 1 & \rho^2 & \ldots \\ \rho^2 & \rho^2 & 1 & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \rho^2 & \rho^2 & \ldots & 1 \end{pmatrix} \tag{E-1}$$

with $-1 < \rho < 1$. This sort of mimics a badly calibrated "ruler". One can figure out $\tilde{C} = (X^T C_d^{-1} X)^{-1}$ algebraically, so, for example,

$$stdv(\text{intercept}) = \sqrt{cov(\hat{m})_{11}} \sim \sqrt{4/n + (1 - 4/n)\rho^2} \qquad \text{large } n \tag{E-2}$$

which shows that the *intercept* uncertainty, if $\rho \neq 1$, ought never to central limit for large data (i.e. it's never more certain than ρ). This is a good warning bell for what will happen when we use C_{eff}: when the data covariance is diagonal, all the posterior uncertainties will scale like $1/\sqrt{n}$, so obviously some error–bar stretching with increasing n will be required to keep $\sqrt{cov(\hat{m})_{11}}$ from shrinking too much.

Equation (E-2) also invites a speculative generalisation. Covariances C_d of the form (E-1) arise from measurements with a mixture of uncorrelated ($\epsilon_{u,i}$) and systematic noise (ϵ_s), for example,

$$\epsilon_i = (1 - \rho^2)^{1/2} \epsilon_{u,i} + \rho \epsilon_s$$

with $\epsilon_{u,i}, \epsilon_s$ both $N(0, 1)$. It is clear that ρ is something like the ratio of systematic noise RMS power to total noise RMS power. The uncertainty in the slope in (E-2) stops improving when $n \approx O(1/\rho^2)$, so we might speculative that the number of "effective measurements" to use is

$$n_{\text{eff}} \approx \min\left(\frac{\text{total noise power}}{\text{systematic noise power}}, n\right) \tag{E-3}$$

This can be a *very* modest number when the relative power of the systematic effects is strong.

The code has a very simple implementation of the "dilution" rule described above. Use the flag `--dilute 10`, where the integer is the number n_{eff}. If systematic effects are overwhelming, use $n_{\text{eff}} = 1$: this has the effect of diluting the statistical power of the whole data set to that of a single measurement, and consequently no inference will be more precise than the mean error–bar supplied. This is probably the simplest option for the average user to wield.

An alternative, more frequentist idea, is simply to subselect data (data hiding). The stability of inferences from different data subsets can illumine problems with poorly known parameters very nicely. As an example, figure 22 shows the variety of posterior inferences for resistor thickness (obtained using model–study sweeps) for the 4–frequency data whose truth–case residuals are shown in figure 21. The volatility of the subsetted data inferences seems a good diagnostic of the sensitivity of this particular inference (a comparable plot of, say the RTP, would be much more stable).

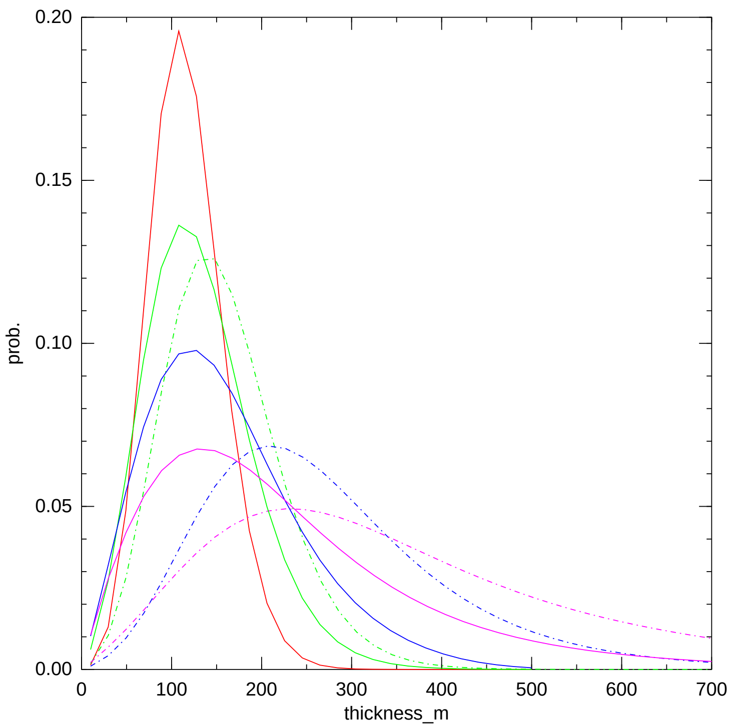

Figure 22: Marginal distributions for thickness for the test data with systematic infection. The solid curves are (i) the full data set, red, $n_d = 670$: ii) green, subjected to n_{eff} dilution, with $n_{\text{eff}} = 335$ iii) blue, $n_{\text{eff}} = 168$, iv) magenta, $n_{\text{eff}} = 84$, i.e. factors of 2 reduction. The *dashed* curves are marginals computed for data sets of matching size, but formed by balanced decimation of the data. Any of these forms of decimation pulls the "truth–case" 250m thickness in and away from the upper tail of the full data set distribution.

Explicit systematics modelling

Closer inspection of figure 21 shows that, in this instance at least, the systematic component seems to be almost linear in offset. For measurements like $|E|$ and $|B|$, such behaviour will probably automatically follow if the model error is expanded as a perturbation series in likely model deficiencies, e.g. anisotropy, or weak 3D effects etc. It thus seems reasonable to try and model this component explicitly, so we will use the form

$$\mathbf{d} = \mathbf{F}(\mathbf{m}) + \underbrace{C_d^{1/2} X}_{\equiv X_N} .\mathbf{m}_n + \boldsymbol{\epsilon}$$

where $\mathbf{F}(\mathbf{m})$ is the CSEM model, and $C_d^{1/2}$ sets the error scale for a general regression $X\mathbf{m}_n$ of the systematic parts, using a *systematic* part of the model $\mathbf{m}_n$. The leftover noise $\boldsymbol{\epsilon}$ will be modelled as $N(0, C_d)$. Typically, e.g. we will use $\mathbf{m}_n$ to be the list of regression coefficients required to fit a polynomial trend in offset to each set of residuals for a given frequency, for measurements like $|E|, |B|$. To set reasonable bounds on how much systematic is "tolerable", a suitable prior for $\mathbf{m}_n$ must be supplied, and for this reason, writing the regression in dimensionless form is helpful. See figure 23 for a picture of how the regression might look.

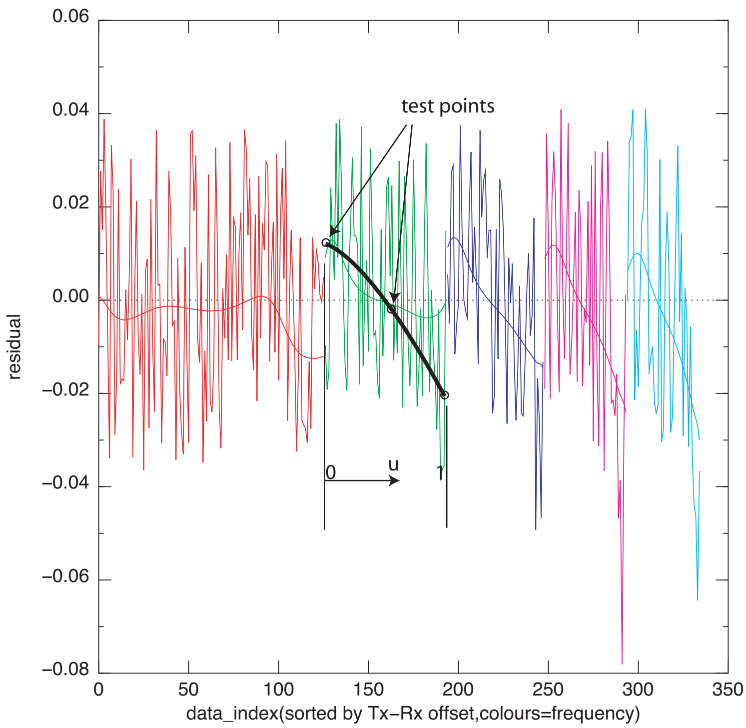

Figure 23: Trend fitting of a frequency group in $|E|_{\text{inline}}$, showing test points used in assembling the prior.

Specifically, for each frequency f, let's use the dimensionless offset

$$u = \frac{\text{offset} - \text{offset}_{\text{min}}}{\text{offset}_{\text{max}} - \text{offset}_{\text{min}}}$$

so $0 \leq u \leq 1$. For order N_p polynomial regression in u, X has blocks of form $X_{ij} = u_i^{j-1}$ (e.g. rows $(1, u_i, u_i^2 \ldots, u_i^{N_p})$), and there will be as many diagonal blocks in X of such form as there are frequencies. The equations below use the dimensioned form $X_N = C_d^{1/2} X$, which implicitly trusts the chosen scaling in C_d to remove the overall exponential decays caused by the absorption etc (i.e C_d should be well chosen!). One can almost think of the extra regression as a systematics–stripping process working on the logarithmic data scale.

This form fits within the general model we have already outlined, but now the complete model vector is $\mathbf{M} = \{\mathbf{m}, \mathbf{m}_n, \ldots\}$. It is not immediately clear how to directly specify a reasonable prior on the polynomial coefficients $\mathbf{m}_n$, but we can indirectly induce this by asking what the effect of the prior distribution is at a set of "test" points evenly distributed along the offset range. Roughly speaking, we do not wish the systematic prediction at each of these test points to dominate the random component by more than some prescribed $O(1)$ ratio, say σ_{sys}. Specifically, at test points $U_k = 0, 1/N_p, \ldots, 1$, the systematic prediction component is $X_{\text{test}}.\mathbf{m}_n$, where the matrix X_{test} is again of block form, with rows $(1, U_k, U_k^2 \ldots, U_k^{Np})$. A Gaussian form of the test-point prior is then

$$\exp(-||X_{\text{test}}.\mathbf{m}_n||^2/2\sigma_{\text{sys}}^2) = \exp(-\mathbf{m}_n^T.(X_{\text{test}}^T.X_{\text{test}}/\sigma_{\text{sys}}^2).\mathbf{m}_n/2)$$

Thus the prior on the systematic regression terms is of form $\mathbf{m}_n \sim N(0, C_{sys})$, where $C_{sys}^{-1} = X_{\text{test}}^T.X_{\text{test}}/\sigma_{\text{sys}}^2$, so sensibly has a preference for zero mean (no systematic at all), but a tolerance settable by σ_{sys}. Comparable systematic and random components are permitted when $\sigma_{\text{sys}} = 1$, but quite large systematic terms may be admissable when, e.g. $\sigma_{\text{sys}} = 5$. Thus Np and σ_{sys} are user–specified parameters for the systematic noise, dependent again on subjective judgements.

Since the added systematic is linear, all the existing machinery can be used, excepting the forward is augmented to $\mathcal{F}(\mathbf{M}) = F(\mathbf{m}) + X_N \mathbf{m}_n$. The full prior precision matrix for the model vector $\mathbf{M}$ is now of form

$$C_p^{-1} = \begin{pmatrix} \mu\partial^T\partial + \mathcal{W}_p(\mu) & 0 \\ 0 & C_{sys}^{-1} \end{pmatrix}$$

and the new mean is $\mathbf{M}_p = (\mathbf{m}_p, \mathbf{0})$. The full system Jacobian is $\mathbf{X} = (J \;\; X_N)$, so the Hessian and gradient terms equivalent to equation (8) are

$$-\nabla\chi_m^2 = \begin{pmatrix} J^T \\ X_N^T \end{pmatrix} (C_d^{-1}/\sigma_N^2)(\mathbf{d} - \mathcal{F}(\mathbf{m})) + \begin{pmatrix} (\mu\partial^T\partial + \mathcal{W}_p)(\mathbf{m}_p - \mathbf{m}) \\ -C_{sys}^{-1}\mathbf{m}_n \end{pmatrix}$$

and

$$H = \begin{pmatrix} J^T \\ X_N^T \end{pmatrix} (C_d^{-1}/\sigma_N^2) \begin{pmatrix} J & X_N \end{pmatrix} + C_p^{-1}$$

The expressions for the marginal model likelihoods (e.g. equation 10) are effectively the same, except we need extra pieces for the systematic components. Equation (10) now has the extra terms on the RHS:

$$+\tfrac{1}{2}\mathbf{m}_n^T.C_{sys}^{-1}.\mathbf{m}_n - \tfrac{1}{2}\log(|C_{sys}^{-1}|) + \tfrac{dim(\mathbf{m}_n)}{2}\log(2\pi).$$

and of course the understanding $\mathbf{F} \to \mathcal{F}$, and the Hessian H above. The optimization machinery is exactly as before, and naturally there will be no bounds on the systematic components $\mathbf{m}_n$ in the projected Newton or Marquardt methods. At present the implementation acts only on the subsets of $\mathbf{d}$ that are $|E|$ data, and the regression is against dimensionless offset. The flag `--systematics d,Np,sigma` invokes this mode of data fitting. Real and imaginary parts of E,B, or phases, show too complex a behaviour to regress in this simple way with offset.

Example
Here we compute a thickness marginal distribution for the preceeding example, using all the data, but allowing quadratic ($N_p = 2$) systematic trends with $\sigma_{sys} = 2, 1, 0.5$. With 4 frequencies, this allows another 12 parameters into the model, so it is not the most parsimonious model recommended by the MML criterion, but is does broaden the range of thicknesses and depths covered by the marginal distributions.

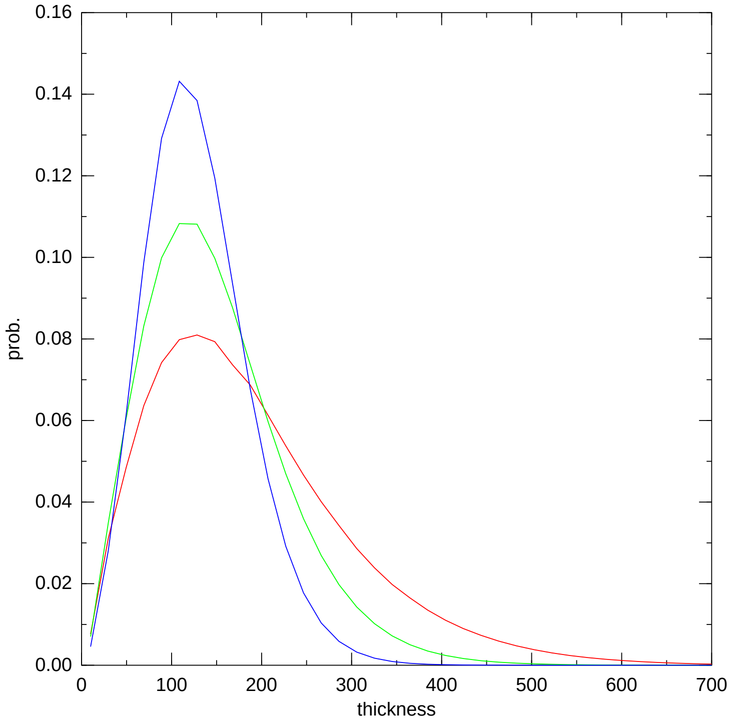

Figure 24: Growth of thickness marginals with degree of relative tolerance of systematic errors in inline $|E|$ data, using quadratic ($N_p = 2$) systematics with offset. Curves are (i) red, $\sigma_{sys} = 2$, (ii) green, $\sigma_{sys} = 1$, (iii) blue, $\sigma_{sys} = 0.5$.

A digression: a comparison with histogramming techniques for model uncertainty

One idea for assessing model uncertainty that circulates in the CSEM community is a scheme that amounts to exhaustively enumerating a set of models that satisfy some thresholding criterion on the χ^2_{RMS} misfit, and forming model estimates and uncertainty estimates from the generated set. Typically, the idea is to finely raster–scan a set of models in p dimensions, forward modelling each, and store all models whose χ^2_{RMS} misfit falls within some Δ of the globally "best model", say $\chi^2_{\min}$. Thus we have the set

$$D = \{m | \chi^2(m) < \chi^2_{\min} + n\Delta\}$$

which is approximated by, say regular, finely spaced model points in p dimensional space.

If there is some parameter q of particular interest, we might wish to find its "mean" and "variance" under this set. A natural "expectation" operator using this set can be defined as

$$\hat{q}_D = \frac{\int_D q dm}{\int_D dm}$$

and the dispersion (variance) of q as

$$Var(q)_D = \frac{\int_D (q - \hat{q}_D)^2 dm}{\int_D dm}.$$

The question then arises as to what properties this formalism has compared to traditional statistical techniques. Two examples might help.

Example 1. Suppose we wish to estimate the mean and the mean–uncertainty (variance) in a set of n random numbers y_i. Then $\chi^2 = \sum_i^n (y_i - m)^2$, and the formulae above work out to

$$\hat{m}_D = \frac{1}{n}\sum_i^n y_i \qquad \text{(E-4)}$$

$$Var(m)_D = \Delta/3 \qquad \text{(E-5)}$$

Example 2. Suppose we wish to estimate the same for a set of n random p–dimensional samples $\mathbf{y}_i$. Then $\chi^2 = \sum_i^n ||(\mathbf{y}_i - \mathbf{m})||^2$, and we get

$$\hat{\mathbf{m}}_D = \frac{1}{n}\sum_i^n \mathbf{y}_i \qquad \text{(E-6)}$$

$$Var(\mathbf{m}_i)_D = \Delta/(p+2) \qquad \text{(E-7)}$$

In both cases, the mean estimate is the same as the classical maximum–likelihood estimate, but the variance is effectively *prescribed* by Δ. In classical work, we would have $Var(\mathbf{m}_i) = 1/n$, so if the noise model is correctly specified, the mean estimate correctly central–limits to the correct value (its uncertainty contracts with increasing n), and the estimator is said to be *consistent*. By contrast, in the formalism above, the specification of Δ amounts to declaring the number of effective data informing the variance estimate:

$$n_{\text{eff}} = (p+2)/\Delta$$

Practioners typically use numbers like $\Delta = 0.1$, so the number of "effective data" in the model is not large: e.g. in 3 dimensions, this is roughly like 50 measurements.

By and large, statisticians of all flavours find this kind of construction eccentric. Some of the reasons are:

- The averaging formulae amount to having uniform (constant probability) on all models in D, and zero for all other models. For example, if D spans models having 5% to 8% misfit, it seems wierd that these should have equal weight in estimating some parameter, but that a 8.00001% misfit model has zero weight.

- If some kind of "smoother", decaying (as opposed to the boxcar function above) penalization of varying misfits is used to address the problem above, and one penalises misfits in χ^2 on a scale set by $n\Delta$, the set D then becomes the whole space, and one quickly arrives at heuristic forms like

$$\hat{q} = \frac{\int q e^{-\chi^2(m)/n\Delta} dm}{\int e^{-\chi^2(m)/n\Delta} dm}.$$

  In other words, the Δ parameter is an obvious adjustment of the error bars along the same lines as we have already described (in this case, e.g. $\sigma_{\text{eff}} \leftarrow \sigma\sqrt{n\Delta}$, if σ is a typical error bar on the diagonal of C_d). So there is nothing to be gained in this construction other than offering an alternative way to specify the number of "effective" data operating under independent Gaussian errors.

- The actual computational cost of approximating the set D is exponential in the number of parameters p. It is hard to imagine the set being accurate enough unless each dimension were gridded into at least 20 slices, so the raw cost goes at least like 20^p. Even at 10ms cost per forward model, this gets astronomical at 6 or 7 parameters. One could argue that the scan can be reduced by convexity arguments, branch and bound methods, etc, etc, but the proper implementation of these ideas amounts to conceding that the problem is best solved by extensive use of optimisation, which is what the current implementation does. A non–optimization based approach to uncertainty is unthinkable for higher dimensional CSEM problems, and too hard for even 1D problems past about 7 parameters. (6 parameters is already 100s of hours at the accuracy suggested above, compared to, say, bootstrapping, which scales as a low power of p, and can run in minutes or hours for up to 20 dimensional models).

- The construction does have the interesting merit that the mean estimates is the same as the classical maximum–likelihood formula, i.e. it can central limit (and will "stack" nicely to reduce noise, in geophysical parlance), but the fixing of Δ effectively prevents its estimated precision from ever increasing.

- The statistical community has been deeply interested for many decades in the problem of how to estimate parameters in data fitting when the properties of the noise are poorly known. There is an enormous literature on this: (i) in Bayesian treatments, typically the emphasis is on simultaneously estimating model and dispersion parameters using hierarchical models with "loose" or diffuse parameterisation of the noise. In frequentist approaches, noise covariance estimation using shrinkage and sandwich estimators are typical themes. Typically, the posterior distributions (or confidence intervals in frequentist regression) end up with fatter tails for model parameters (e.g. t-distributions), which reflect the additional uncertainty caused by ignorance of the noise. These approaches are complex and often have computational challenges, but this community clearly feels that these approaches are more satisfactory and defendable than ad hoc techniques like thresholding χ^2.

Summary

We do not yet have an established recipe for modelling noise. Virtually without exception, we recommend against interpolating data, and strongly suggest using data whose spatial Nyquist rate is the greater of the acquisition spacing and a spacing that make sense in terms of the inferable resolution (they are usually closely related). Since the noise will have in general components that can never be estimated by the data itself, a modest suggestion is to try and calibrate this by forward modelling and inversion experiments on systems that will generate such artifacts, e.g. anisotropic layer-cake models. Where there are suspected to be strong systematic effects from modelling assumptions or simplifications, reduction of the statistical power by the dilution technique is probably the simplest tool for the user.

REFERENCES

Bertsekas, D. P., 1982, Projected Newton methods for optimization problems with simple constraints: SIAM J. Control and Optimization, **20**, 221–246.

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**, WA85–WA95.

Constable, S., 2006, Marine electromagnetic methods – a new tool for offshore exploration: The Leading Edge, **25**, 438–444.

Constable, S. and L. J. Srnka, 2007, An introduction to marine controlled-source electromagnetic methods for hydrocarbon exploration: Geophysics, **72**, WA3–WA12.

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.

Denison, D. G. T. et al., 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. and P. Stark, 2002, Inverse problems as statistics: Inverse Problems, **18**, R55–R97.

Farquharson, C. and D. Oldenburg, 2004, A comparison of automatic techniques for estimating the regularization parameter in non-linear inverse problems: Geophysical Journal International, **156**, 411–425.

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 Jan 2010.

———, 2010, Resolution and uncertainty in 1D CSEM inversion: a Bayesian approach and open-source implementation.: Geophysics, **75**.

Gunning, J. and M. Glinsky, 2004, Delivery: an open-source model-based Bayesian seismic inversion program: Computers and Geosciences, **30**, 619–636.

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**, 382–401.

Kelley, C. T., 1987, 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**, F9–F20.

Kitanidis, P., 1995, Quasi–Linear Geostatistical Theory for Inversing: Water Resources Research, **31**, 2411–2419.

———, 1999, Generalized covariance functions associated with the Laplace equation and their use in interpolation and inverse problems: Water Resources Research, **35**, 1361–1367.

Koontz, R. S. J. and B. Weiss, 1982, A modular system of algorithms for unconstrained minimization: Technical report, Comp. Sci. Dept., University of Colorado at Boulder.

Liu, J. S., 2003, Monte Carlo Strategies in Scientific Computing: Springer.

Loseth, L., 2007, Modelling of controlled source electromagnetic data: PhD thesis, Department of Physics, NTNU.

Madsen, K., H. B. Nielsen, and O. Tingleff, 2004, Methods for non-linear least squares problems (2nd ed.): Informatics and Mathematical Modelling, Technical University of Denmark, DTU.

Malinverno, A. and R. L. Parker, 2006, Two ways to quantify uncertainty in geophysical inverse problems: Geophysics, **71**, W15–W27.

Mitsuhata, Y., 2004, Adjustment of regularization in ill-posed linear inverse problems by the empirical Bayes approach: Geophysical Prospecting, **52**, 213–239.

Neal, R. M., 2003, Slice sampling: Annals of Statistics, **31**, 705–741.

Nocedal, J. and S. J. Wright, 1999, Numerical optimization: Springer.

Oliver, D., N. He, and A. Reynolds, 1996, Conditioning permeability fields to pressure data: Presented at the Proceedings of the 5th European Conference on the Mathematics of Oil Recovery.

Ory, J. and R. G. Pratt, 1995, Are our parameter estimators biased - the significance of finite-difference regularization operators: Inverse Problems, **11**, 397–424.

O'Sullivan, F., 1986, A statistical perspective on ill-posed inverse problems: Statistical Science, Vol. 1, **1**.

Press, W. H. et al., 1992, Numerical Recipes The Art of Scientific Computing. 2nd ed.: 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.

Snieder, R., 1998, The role of nonlinearity in inverse problems: Inverse Problems, **14**, 387–404.

Tarantola, A., 1987, Inverse Problem Theory, Methods for Data Fitting and Model Parameter Estimation: Elsevier Science Publishers, Amsterdam.

Tompkins, M. J. and L. J. Srnka, 2007, Marine controlled-source electromagnetic methods introduction: Geophysics, **72**, WA1–WA2.

Verrill, S., 2005, Java translation of UNCMIN. `http://www1.fpl.fs.fed.us/optimization.html`, accessed 2 Feb 2010.