Error modelling in Bayesian CSEM inversion

James Gunning
Michael Glinsky

CSIRO

General error structure in inverse problems

- Model

$$y = F(\boldsymbol{m}) + \boldsymbol{\epsilon}$$

F(m) = forward model (Maxwell etc, or approximations)

y = measured data (E fields etc)

m = gridblock resistivities, anisotropy

object-like: locations, surfaces, etc

- Error

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

Inversion approach

- Bayes

$$P(\boldsymbol{m}|y) \quad \sim \quad L(y|\boldsymbol{m})P(\boldsymbol{m})$$

- Typically;

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

- If C_d "known"…

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

- Objections to Bayes
  - Often focused on P(m)
  - Noise structure in L(y|m) often more questionable/disputable

CSIRO

Some terminology

Marginal distributions

$$P(m_i|y) = \int P(\boldsymbol{m}|y) dm_{j \neq i}$$

Model probabilities: 'evidence' or 'marginal model likelihood'

$$P(\mathcal{M}) = \int P(\boldsymbol{m}|y, \mathcal{M}) d\boldsymbol{m}$$

Typical approximations for evidence (Laplace)

$$P(\mathcal{M}) \sim |H|^{-1/2} \exp(-\chi^2(\hat{\boldsymbol{m}})/2)$$

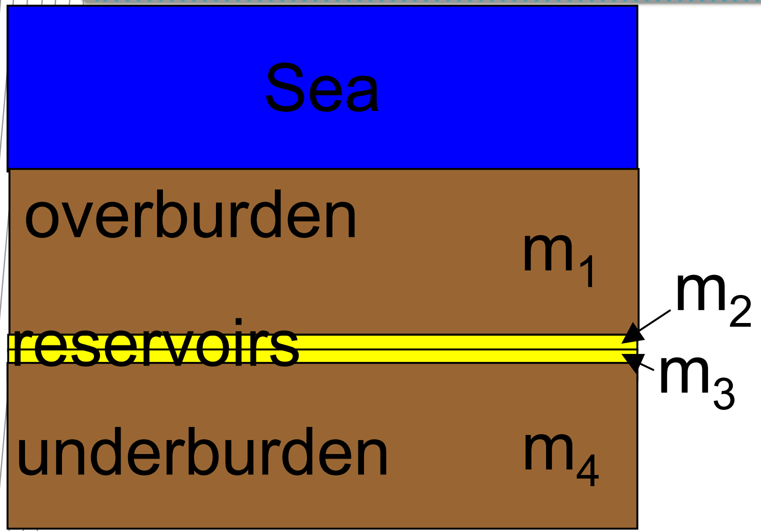
Posterior Uncertainties

- Linearisation is useful only if the geometry is *extremely* coarse. For finer models, F(m) has a huge “null-space”. But decisions needed on finer models.
- Approximate posterior distributions from Hessian usually very poor.
- Alternative sampling methods required
  - Tailored MCMC methods

    Exhaustive mode enumeration, followed by mixture of:

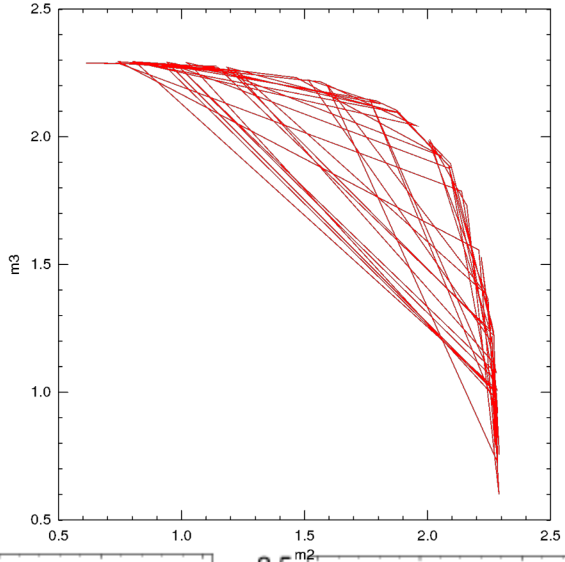
    1) Reversible jump MCMC (diffusion)
    2) Mode jumps
    3) Big jumps along constant RTP
  - Bayesianized Parametric Bootstrap

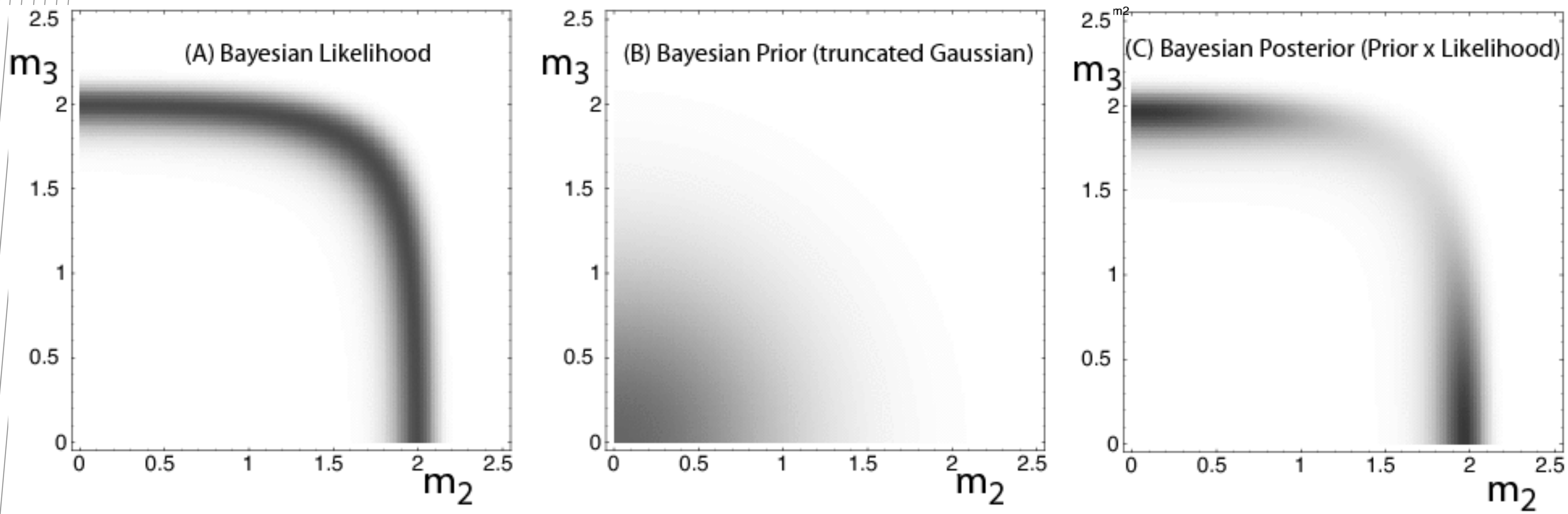
Illustrative 1D example



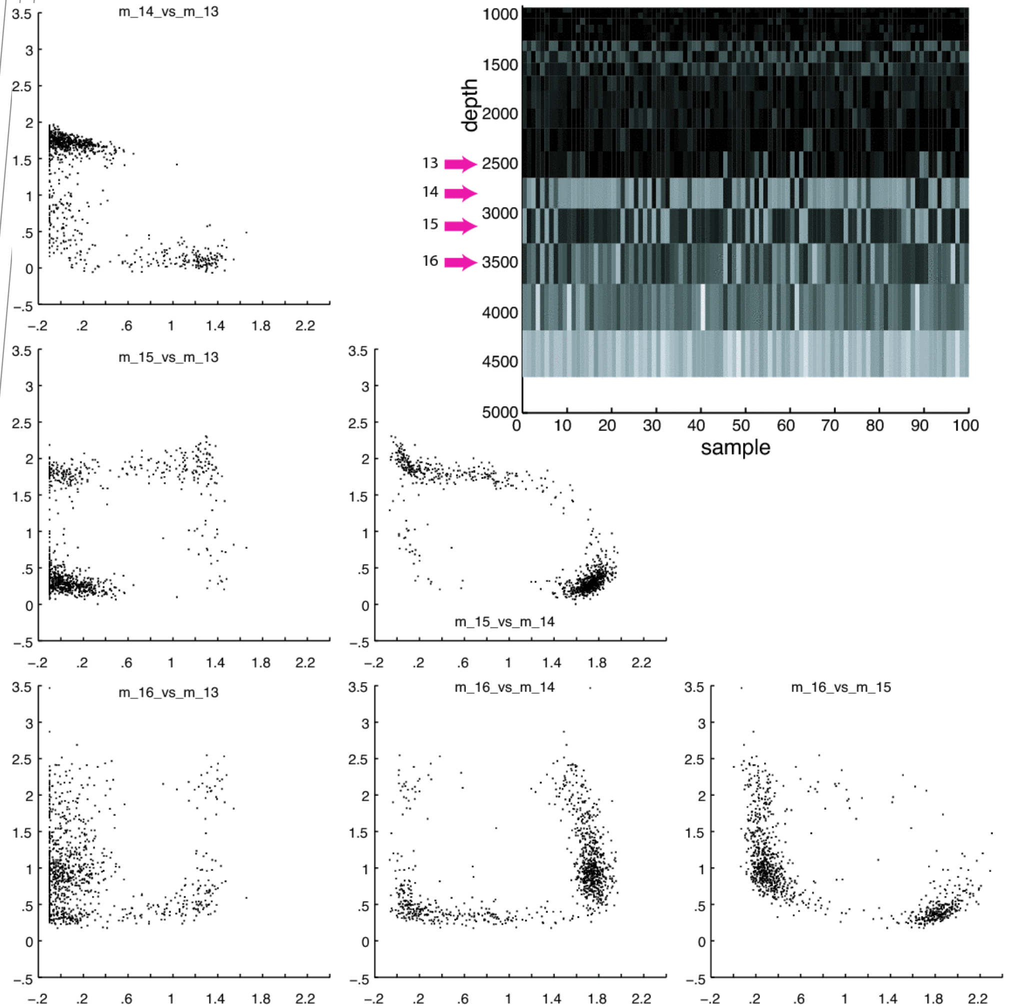
$m_i = log(\rho_i)$

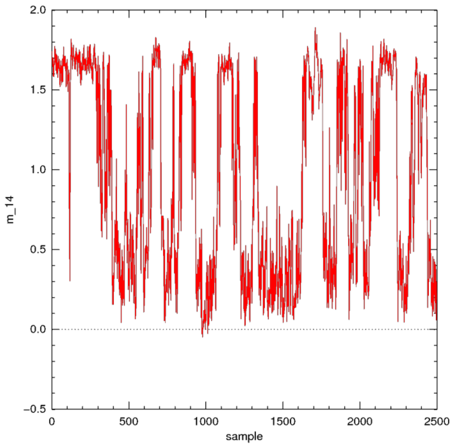
Tailored MCMC:

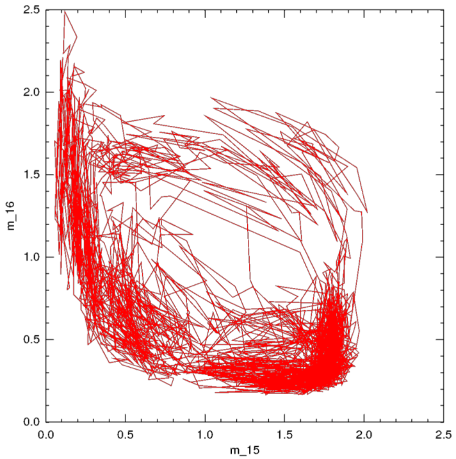




More complex MCMC example



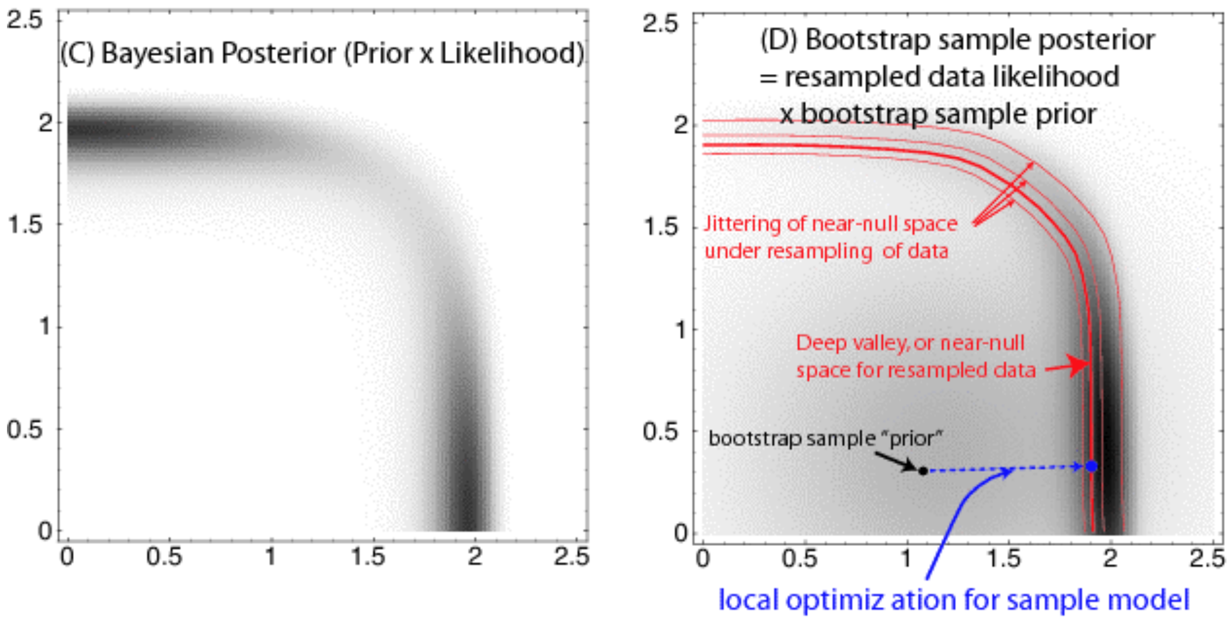




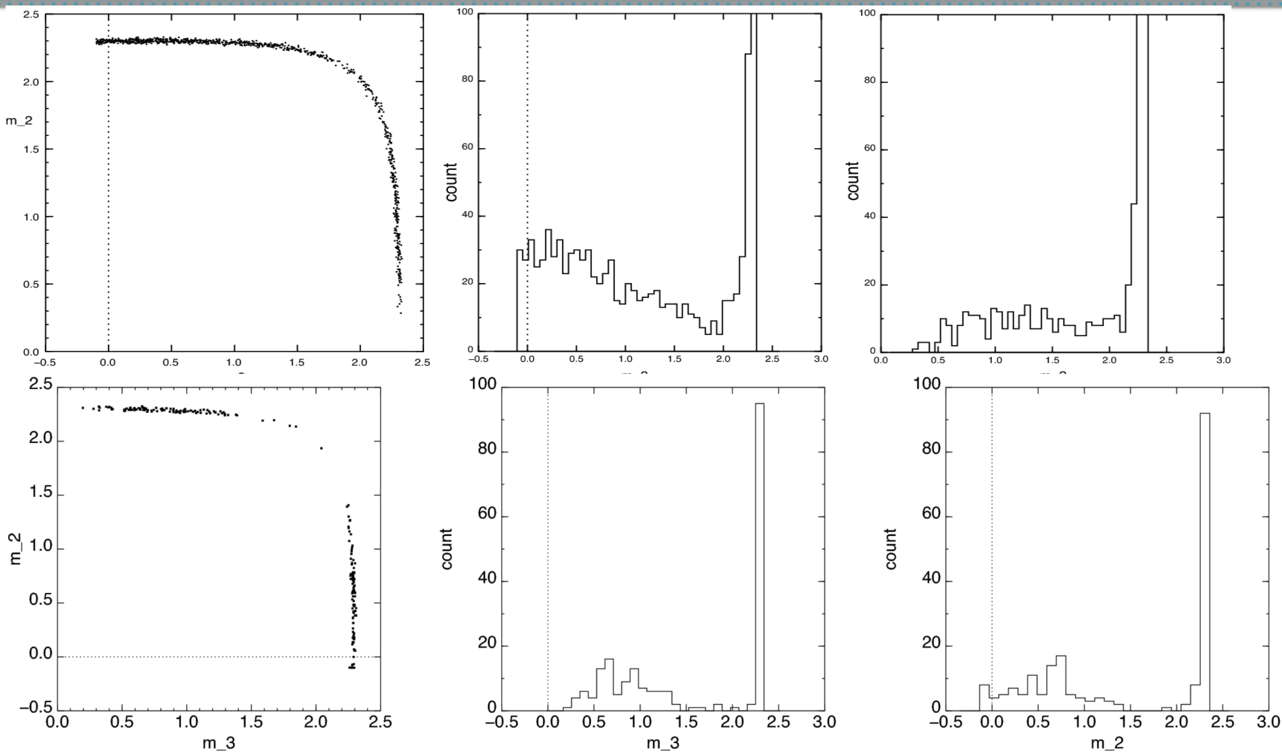
- Mixing is hard. Very many samples needed

Bayesian Parametric bootstrap

- *Randomized maximum likelihood* in some papers
- Bayesian priors treated as "extra data"
- Bootstrap data sets drawn from "best-fit" model, and MAP inversions found for each
- **Many** inversions needed. But avoids MCMC slow diffusion
- Jumps over parameter space very well



Uncertainties upshot



MCMC

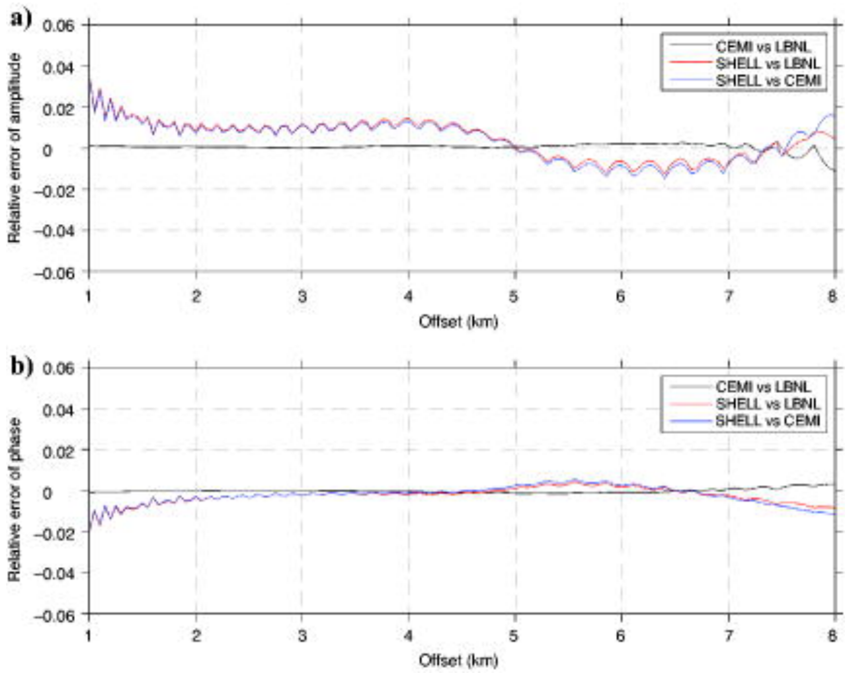
Bootstrap

- MCMC - expensive. Many forward runs
- Bootstrap - approx. but cheaper. Many inversions
- Either way: *very* fast forward models needed
  - But approximate forward models will increase the error, usually in a correlated way

Example modelling errors (1)

3D FD/FE modelling:

Darnett et al , Geophysics 2007

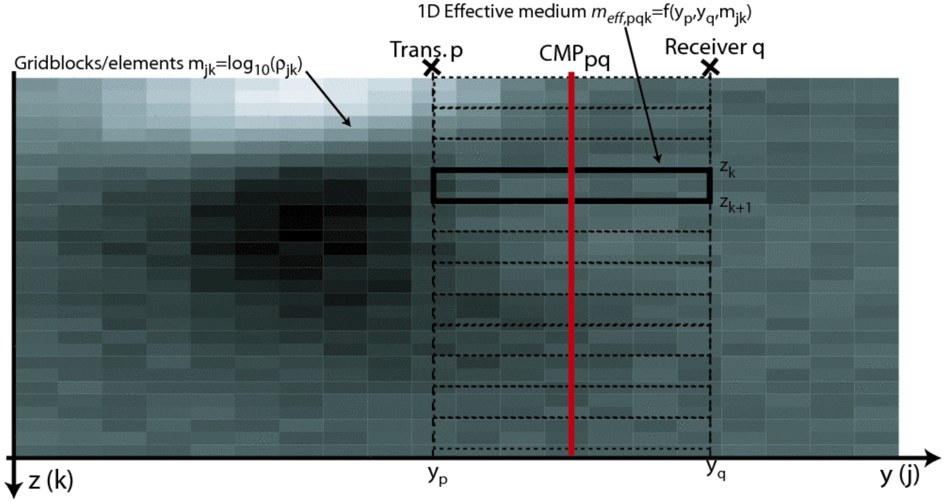


Note long correlations or trends in errors...

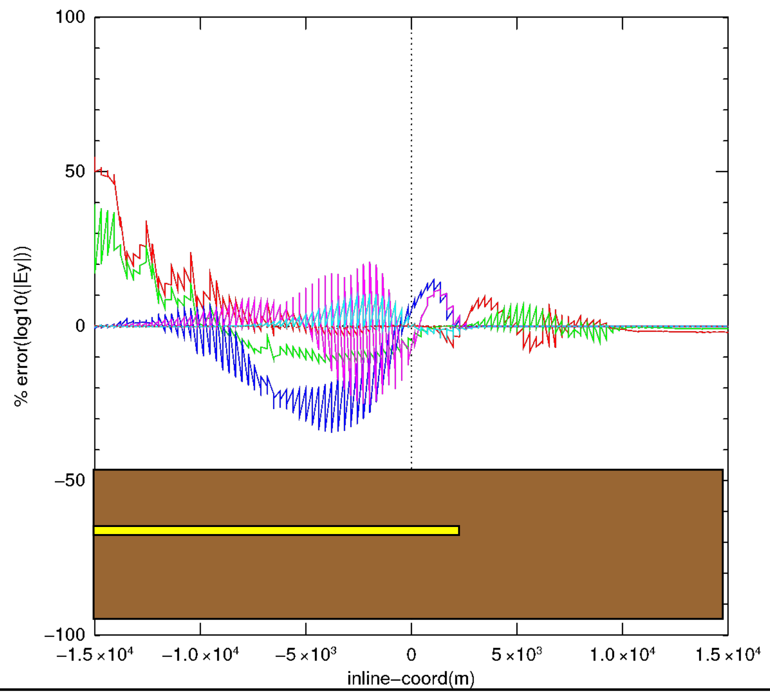
Errors probably below instrument errors ?

Example modelling errors (2)

1D effective media CMP models



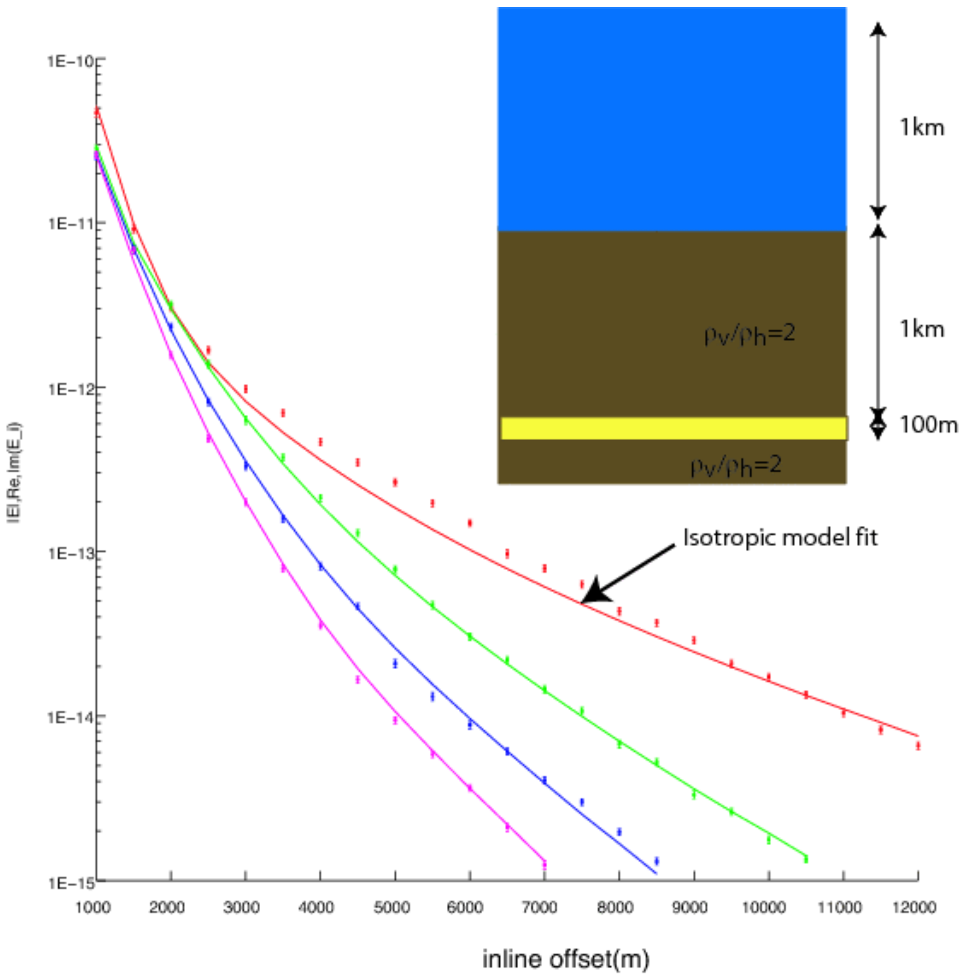
Effective-media framework



Errors (modestly favourable case)

Example modelling errors (3)

- Layering known, Anisotropy neglected



Some things we know about model inference from linear theory

- OLS estimates (C_d "known")

$$\begin{aligned} \hat{\boldsymbol{m}} &= (X^T C_d^{-1} X)^{-1} X^T C_d^{-1} y \\ \mathrm{cov}(\hat{\boldsymbol{m}}) &= (X^T C_d^{-1} X)^{-1} \end{aligned}$$

Uncertainty depends on C_d at *leading* order

Robust components roughly independent of C_d

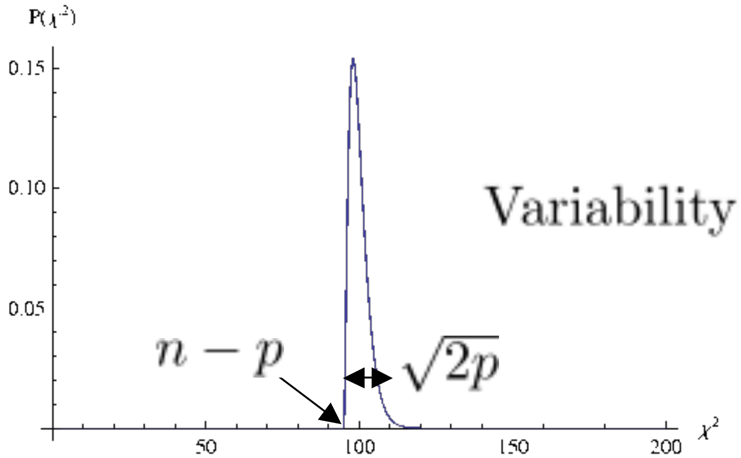
- Assumed covariance $C_d \rightarrow C_{eff}$

$$\begin{aligned} \hat{\boldsymbol{m}}' &= (X^T C_{\mathrm{eff}}^{-1} X)^{-1} X^T C_{\mathrm{eff}}^{-1} y \\ \mathrm{cov}(\hat{\boldsymbol{m}}') &= (X^T C_{\mathrm{eff}}^{-1} X)^{-1} \\ C_{\mathrm{eff}} &= \mathrm{diag}\{\sigma_i^2\} \\ \text{so} \qquad X^T C_{\mathrm{eff}}^{-1} X &\sim O(n) \end{aligned}$$

- $\hat{\boldsymbol{m}}' \approx \hat{\boldsymbol{m}}'$ for 'robust' components (dominant eigenvalues of C_d)
- Bias mainly a problem if $\mathrm{cov}(\hat{\boldsymbol{m}}')$ gets too small

Variability in χ^2

$\chi^2 = (y - Xm)^T C_D^{-1}(y - Xm) \sim \chi^2_p$, offset by $n - p$ if noise 'correct'



Variability in $\chi^2_{\mathrm{MS}} = \chi^2/n \sim \sqrt{p}/n$

What if we do 'C_D known up to a scalar σ^2...'

$$L(y|m) \sim \frac{\exp(-\frac{1}{2}(y - Xm)^T(\sigma^2 C_D)^{-1}(y - Xm))}{|\sigma^2 C_D|^{1/2}} \cdot \underbrace{\frac{1}{\sigma^2}}_{\text{Jeffreys}}$$

Then std.dev$(\sigma^2) \sim 1/\sqrt{2\nu}$, *but still*

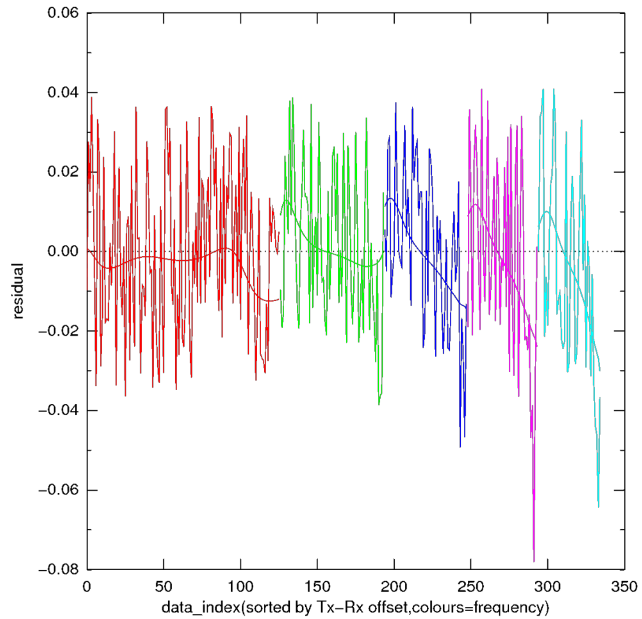
$$\chi^2_{\mathrm{MS}} = (y - Xm)^T C_D^{-1}(y - Xm)) \sim F(p, n - p)$$

with asymptotic standard deviation

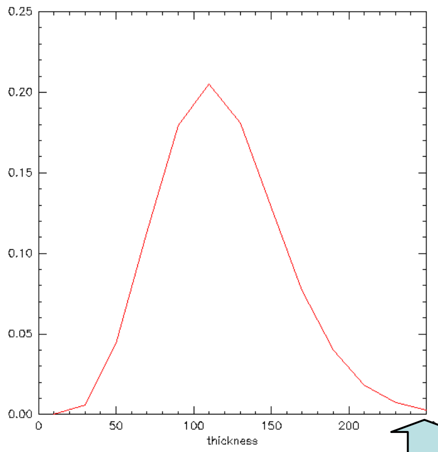
$$\mathrm{std.dev}(\chi^2_{\mathrm{MS}}) \sim \sqrt{2p}$$

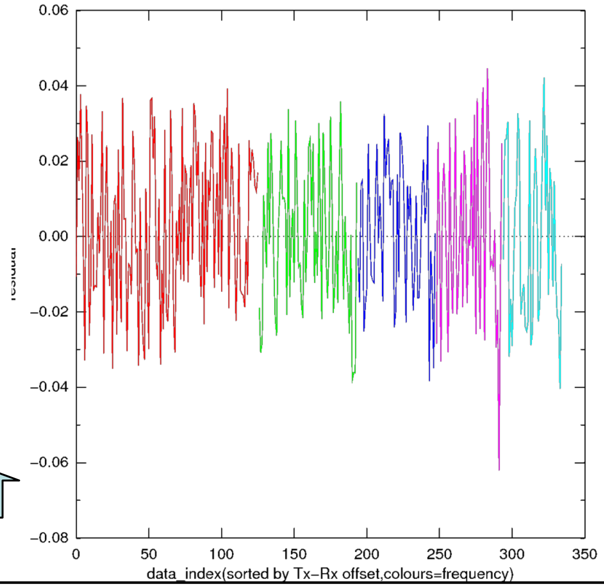
Example problem and 3 approaches

- Test data set with residuals modified to give the effect of trend with offset
    - 250m thick resistor at 850m



- Inversion using zero-mean noise, posterior of thickness from marginalisation

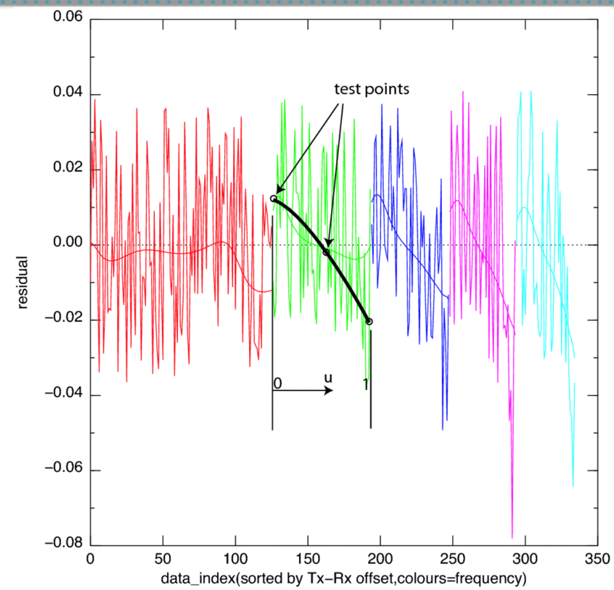




Inversion gives thickness pdf with mean 110m, and 250m above upper 5% quantile.

Problem is *not* overfitted

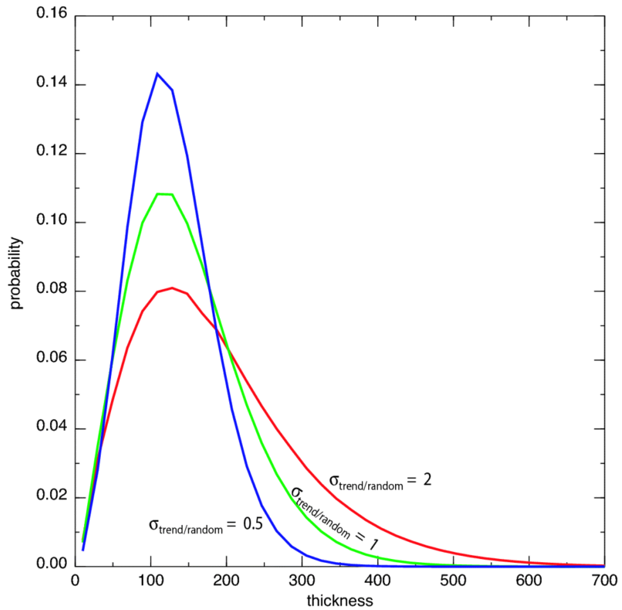
1) Explicit regression for trends/systematics



- noise="trend(u)+random"

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

- Amounts to particular parametric forms of overall noise covariance
- User specifies form of trends, plus
  trend power/random power ratio
- Gaussian prior for m_n comes from trend/power ratio
- Joint inversion for m and m_n (block-augmented Gauss-Newton framework)
- Probably too messy for effects that "thrash" (e.g. Re(E) ,Im(E))



2) Effective data reduction

Toy problem inspiration: fit straight line to data $y = Xm + \epsilon$ where

$$\epsilon_{\text{eff},i} = \rho\epsilon_0 + \sqrt{1-\rho^2}\epsilon_i = \text{systematic+random}, \qquad \rho^2 = \frac{\text{systematic power}}{\text{total noise power}}$$

So

$$C_D = \begin{pmatrix} 1 & \rho^2 & \rho^2 & \dots \\ \rho^2 & 1 & \rho^2 & \dots \\ \rho^2 & \rho^2 & 1 & \dots \\ \dots & \dots & \dots & \dots \\ \rho^2 & \rho^2 & \dots & 1 \end{pmatrix}$$

Then cov({slope,intercept}) $= (X^T C_D^{-1} X)^{-1}$, and

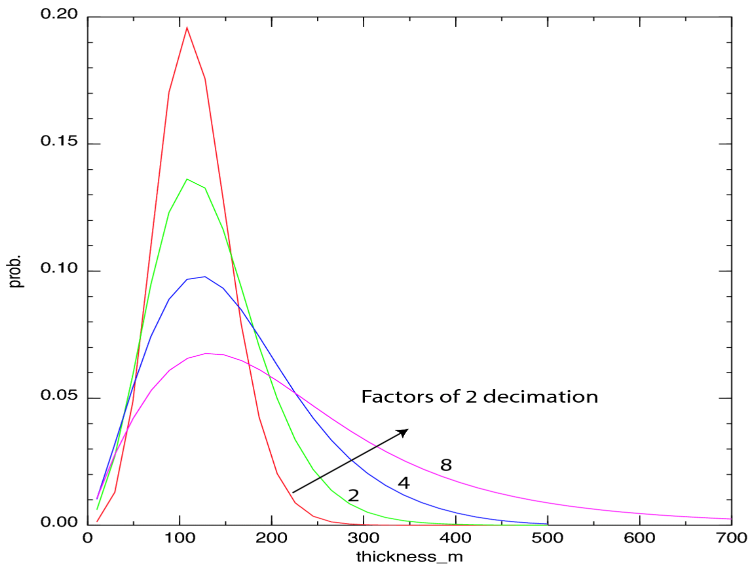
$$\text{std.dev(intercept)} \sim \sqrt{4/n + (1-4/n)\rho^2} \qquad \text{flattens when } n = O(1/\rho^2)$$

Heuristic: inflate error bars as if $n_{\text{eff}} = 1/\rho^2$, i.e. $\sigma_i \to \rho\sqrt{n}\sigma_i$, so

$$(X^T C_{\text{D,eff}} X)^{-1} \to 1/\rho^2$$

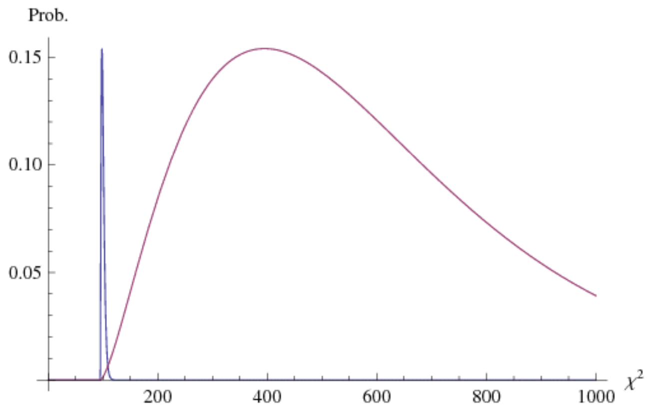
2) Effective data reduction con't

- Test problem:



Limiting case $\rho \to 1$, where $n_{\text{eff}} = 1$

$$L(y|m) \sim \exp(-\chi^2_{\text{RMS}}/2)$$



Upper bound provable from properties of the trace:
if $\boldsymbol{\epsilon}$=errors normalised by $C_{ii}^{1/2}$

$$\boldsymbol{\epsilon}^T C^{-1} \boldsymbol{\epsilon} > \frac{1}{tr(C)} ||\boldsymbol{\epsilon}||^2 = \frac{||\boldsymbol{\epsilon}||^2}{n}$$

$\therefore L$ is a 'bounding' likelihood

CSIRO

3) Averaging over the unknown noise covariance

- Use hierarchical model

$$p(\mathbf{m}|y) \sim N(y - F(\mathbf{m})|C)p(\mathbf{m})p(C)$$

- Inverse-Wishart prior

$$p(C) = \mathcal{W}^{-1}(\nu, C_0) = \frac{|\frac{1}{2}C^{-1}\nu C_0|^{\nu/2}}{Z_{\nu d}|C|^{(d+1)/2}} \exp(-\tfrac{\nu}{2}\mathrm{tr}(C^{-1}C_0))$$

- mean

$$\langle C \rangle = \frac{\nu C_0}{\nu - d - 1}$$

- variance

$$\mathrm{Var}(C_{ii}) = \frac{2\nu^2 C_{0,ii}^2}{(\nu - d - 1)^2(\nu - d - 3)}$$

- ν ~ "number of prior samples" in estimating C_0

Noise-covariance updating

- Given noise δy and scatter matrix S (e.g. in Linear model):

$$\boldsymbol{\delta} y = \mathbf{y} - X.\mathbf{m} \qquad S = \boldsymbol{\delta}\mathbf{y}\boldsymbol{\delta}\mathbf{y}^T$$

- Covariance update

$$\begin{aligned} \pi(C, \boldsymbol{\delta}\mathbf{y}|\nu, C_0) &\sim N(\boldsymbol{\delta}\mathbf{y}, C)\mathcal{W}^{-1}(\nu C_0; \nu) \\ &\sim \mathcal{W}^{-1}(1+\nu, S+\nu C_0) \end{aligned}$$

- Posterior mean is then

$$\langle C \rangle|_{\boldsymbol{\delta}\mathbf{y}} = (S+\nu C_0)/(1+\nu-d-1)$$

- This is an example of a "**shrinkage**" estimator: eigenvalues of S are squeezed towards those of C_0

CSIRO

Marginalising over unknown covariance

- Effective marginal posterior…

$$
\begin{aligned}
\pi(\mathbf{m}|\mathbf{y},\nu,C_0) &\sim \int \pi(C,\boldsymbol{\delta}\mathbf{y}|\nu,C_0)dC \sim \frac{|\nu C_0|^{\nu/2}}{|S+\nu C_0|^{(1+\nu)/2}} \\
&\sim \frac{1}{|I+\nu^{-1}C_0^{-1}S|^{(1+\nu)/2}} \sim \frac{1}{(1+\chi^2(\boldsymbol{m})/\nu)^{(1+\nu)/2}}
\end{aligned}
$$

- Compare to known-noise case ($\nu \rightarrow \infty$)

$$
\pi(\mathbf{m}|\mathbf{y},C_D) \sim \exp(-\chi^2(\boldsymbol{m})/2)
$$

Marginalising for overall model evidence

- Laplace-like approximation...

$$
\begin{aligned}
\pi(\boldsymbol{y}|\nu, C_0) &\sim \int \frac{1}{(1+\frac{1}{\nu}[\chi^2(\hat{\boldsymbol{m}})+\frac{1}{2}(\mathbf{m}-\hat{\mathbf{m}})^T H(\mathbf{m}-\hat{\mathbf{m}})])^{(\nu+1)/2}} d\boldsymbol{m} \\
&\sim |H|^{-1/2}(1+\chi^2(\hat{\boldsymbol{m}})/\nu)^{-(\nu-p+1)/2}
\end{aligned}
$$

- C.f. usual expression with known noise

$$
\pi(\boldsymbol{y}|C_D) \sim |H|^{-1/2}\exp(-\chi^2(\hat{\boldsymbol{m}})/2)
$$

- How to choose C_0 and ν ? Maximise marginal over free parameters C_0 and ν. Choose C_0 from suitable subspace etc. Leads to EM algorithms (Chen '79 and followers)
- Findings:

  Limitation $\nu > d$ is annoying

  Prior structure in C_0 has strong influence and easy to mis-specify

CSIRO

Everybody loves their Shrink(age) ....

- Stein (1975)

$$\begin{aligned} S &= U\Lambda U^T \quad \text{eigenvalue decomp., } d \text{ dimensions, } N \text{ samples, with} \\ C &= U\Lambda' U^T \\ \lambda_i &\to N\lambda_j/(N - d + 1 + 2\lambda_j \sum_{i \neq j} 1/(\lambda_j - \lambda_i)) \end{aligned}$$

Needs isotonizing. Also doesn't really work for $N << d$

- Stein (1982), minimax ($N = 1$)

$$\lambda_i \quad \to \quad \frac{1}{2 + d - 2i}\lambda_i = \{\chi^2/d, 0, 0 \ldots\}$$

...not invertible. Amounts to scaling all errors by $\sqrt{d}$

- Haff (1980). Empirical Bayes. Extension to $N = 1 < d$ by Ledoit (2001)

$$C = \frac{d-4}{d}\mathrm{tr(S)}I + S/2$$

Invertible, but flattens all posterior model probabilities!

- Friedman (1989)...Leung (1998)

$$C = (1-\alpha)\frac{\mathrm{tr(S)}}{d}I + \alpha S$$

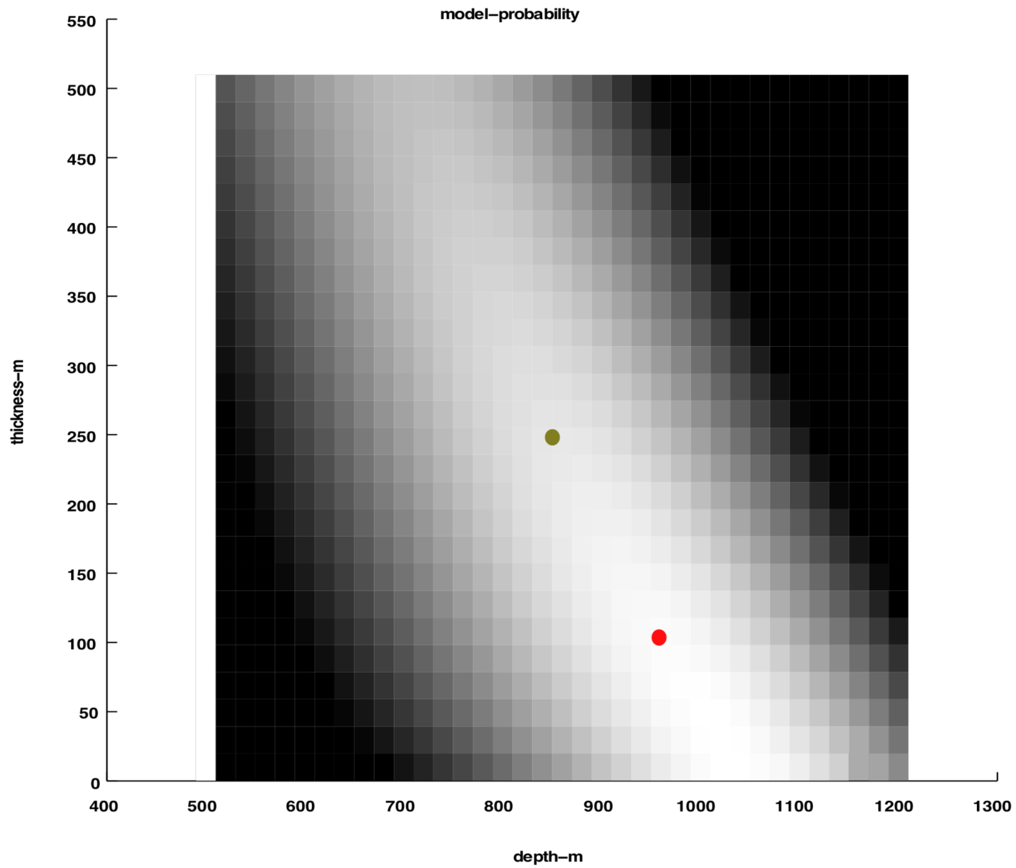
α from cross–validation (Friedman) or $\alpha = N/(N+2)$ (Leung, large N)

Shrinkage on test problem

- Modified Friedman/Leung:

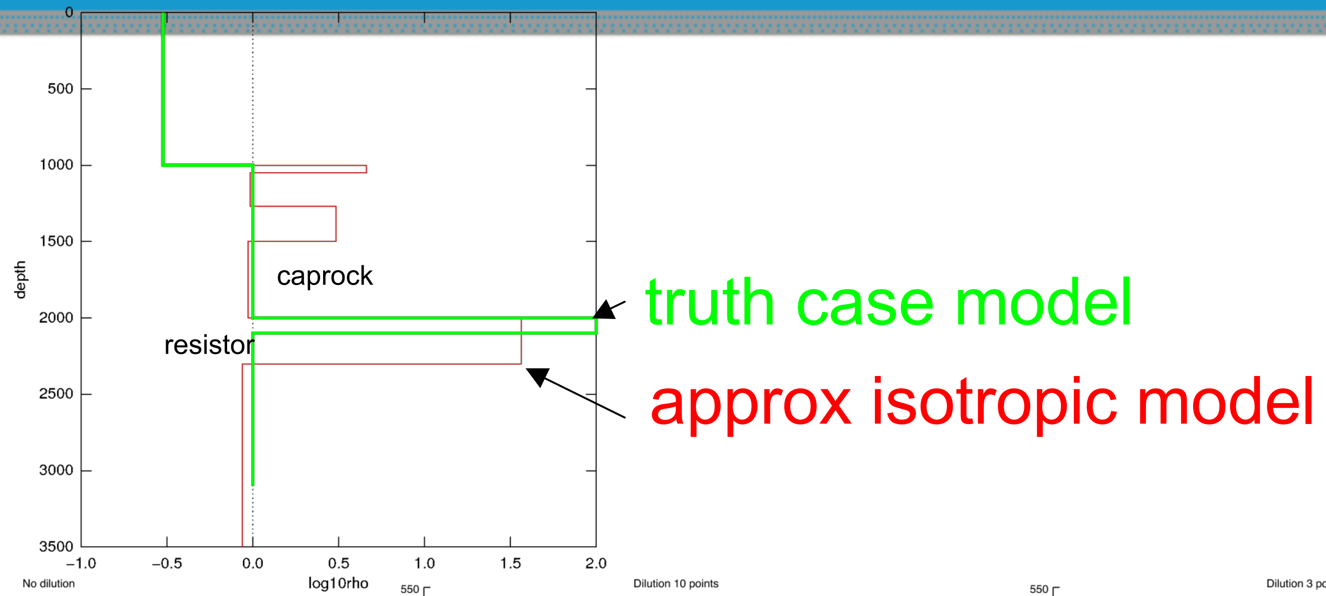
$$C = \alpha \frac{d}{\mathrm{tr}(S)} S + (1-\alpha) I$$

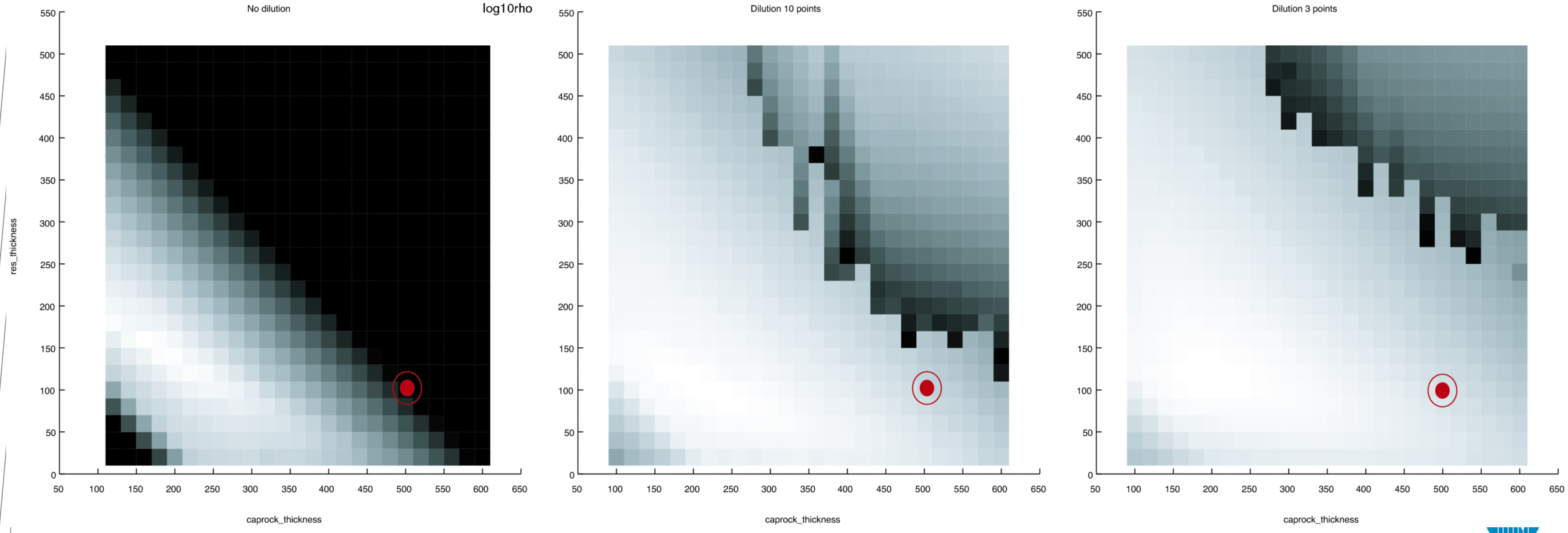
α = "fractional power non–random noise"



Calculation with α=0.1

Anisotropy example again





CSIRO

Conclusions

- Independent Gaussian noise too optimistic given likely level of modelling noise, even in heavyweight codes.
- Explicit removal or error trends possible with extra systematic-trend parameters. Probably fragile.
- Error-bar inflation based on “RMS power” works, and easy to implement. Rather ad hoc theory
- Use of shrinkage probably a better theory. More obvious absorption of bias terms into covariance structure. Shrinkage fraction probably not inferable from data.