Short Course on Statistics and Uncertainty Part III

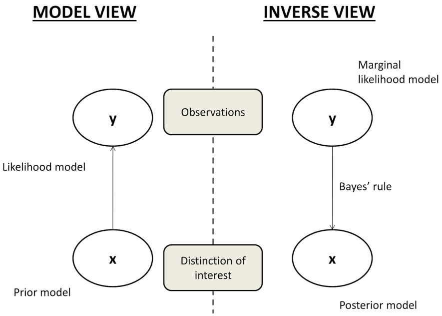
Jo Eidsvik

Department of Mathematical Sciences, NTNU, Norway

Topics

- Bayesian inversion (prior, likelihood and posterior)
- Gaussian process regression and Kriging
- Linear Bayesian inversion

Model and inversion



Model for $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$$

For the analysis, the main interest is in the conditional $p(\boldsymbol{x}|\boldsymbol{y})$.

Bayesian inversion

Model for $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$$

$p(\boldsymbol{x})$ from a priori knowledge, $p(\boldsymbol{y}|\boldsymbol{x})$ from data acquisition. Bayes' rule gives the posterior:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y})}$$

Inverse problems are often considered difficult, and requires computational approximations, except for small dimensions or linear Gaussian models

Bayesian inversion and parameters

Model for $\boldsymbol{\theta}$, $\boldsymbol{x}$ and $\boldsymbol{y}$ is

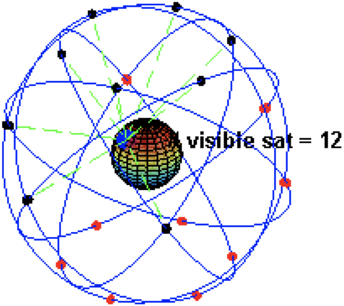
$$p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{x})$$

(Assuming conditional independence.)
Bayes' rule:

$$p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y})}$$

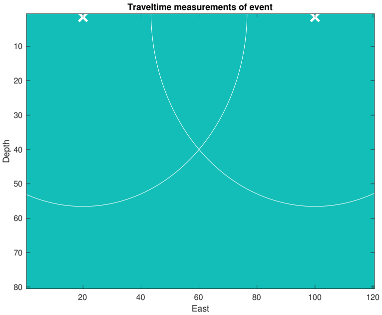
.
Parameter $\boldsymbol{\theta}$ can be specified from auxiliary data sources.

Example of this setting: Positioning from traveltime data



Localization problem from traveltime data.

2D earth - traveltime from source to receiver



Based on the traveltime data : 'Where is the source?'
There are many similar settings in Earth sciences:
Seismic data (Vertical Seismic Profiling), Sonar / acoustic data (range only data).
Earthquakes, hazards, explosions, etc. similar, with a reference (time difference).

Bayesian approach to source localization

- Prior model for source location in the subsurface (or sea).
- Likelihood model for the traveltime model, with noise characteristics.
- Bayes' rule combines these to give the posterior model.

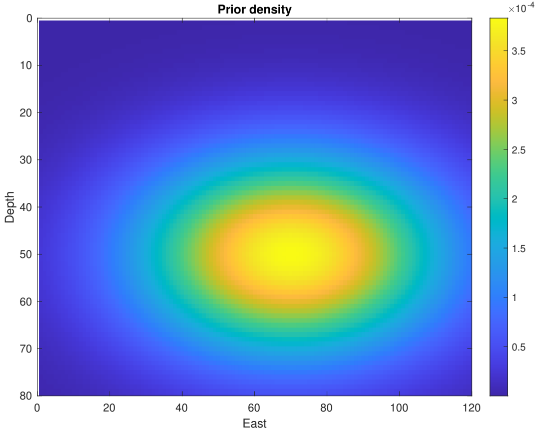
Notation

- Source location $\boldsymbol{x} = (x_1, x_2)$ (east, depth).
  Prior probability density function $p(\boldsymbol{x})$.
- Traveltime data $\boldsymbol{y} = (y_1, \ldots, y_m)$. ($m$ receivers)
  Likelihood model is defined via a conditional probability density function $p(\boldsymbol{y}|\boldsymbol{x})$.
- The solution to the inverse problem is the posterior probability density function:

Bayes' rule:

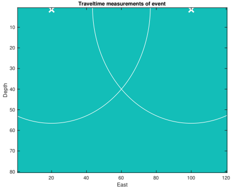
$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y})} \propto p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$$

Prior model



Gaussian distribution for $\boldsymbol{x} = (x_1, x_2)$.

Likelihood model

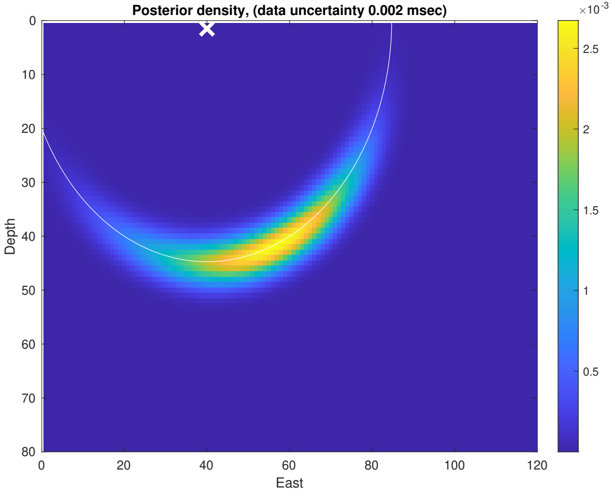


Traveltime measurements $j = 1, \ldots, m$ are defined by

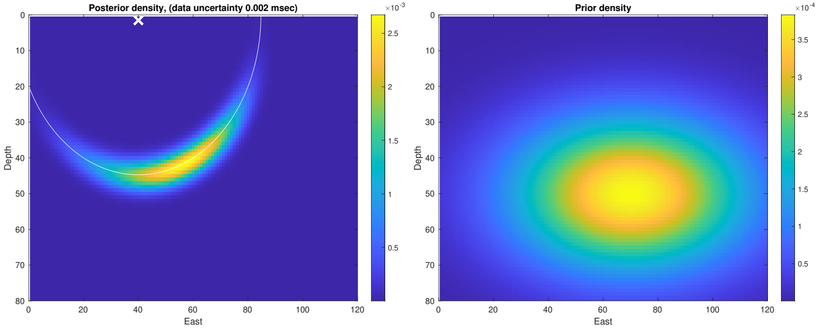
$$y_j = \sqrt{(s_{1,j} - x_1)^2 + (s_{2,j} - x_2)^2}/v + \epsilon_j, \quad \epsilon_j \sim N(0, r^2)$$

Receiver(s) at surface $(s_{1,j}, s_{2,j})$, $j = 1, \ldots, m$.
Assume conditionally independent errors between sensors.

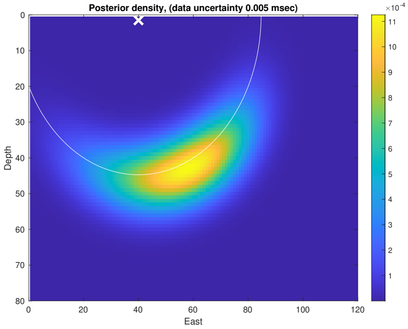
Posterior model (1 sensor, accurate measurement)

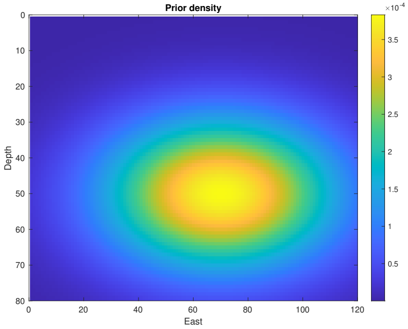


Posterior model (1 sensor, accurate measurement)

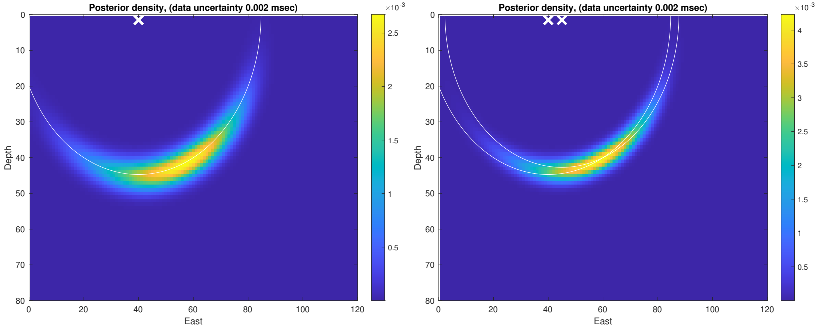


Posterior model (1 sensor, inaccurate measurement)

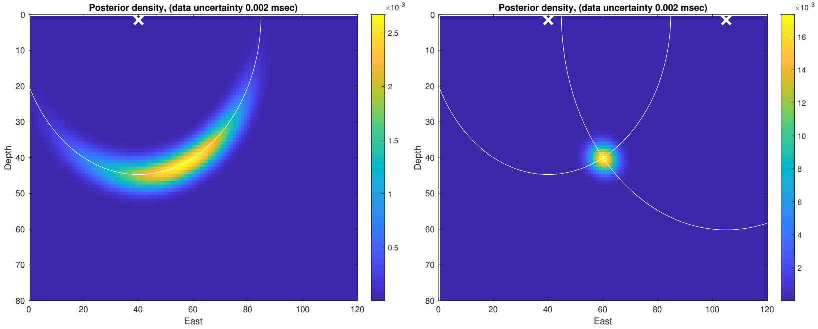




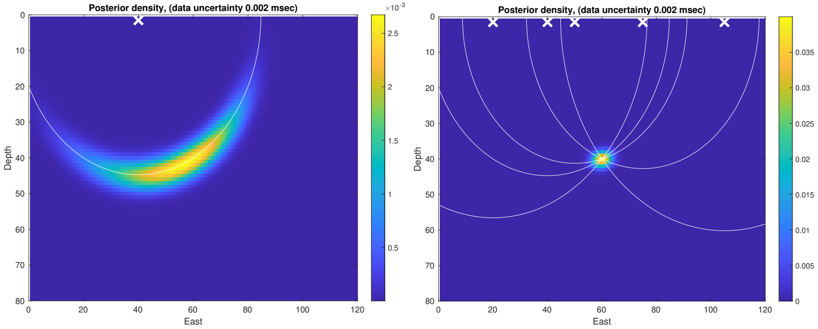
Posterior model (2 sensors, poor design)



Posterior model (2 sensors, good design)



Posterior model (5 sensors)



Bayesian inversion

The source location problem is a classic example of an inverse problem. The forward model (time) is easy to calculate. But the inverse is difficult in real-world problem which are often of much higher dimensions than 2 as we have here with position in (east, depth), and more non-linear.
A gerenal approach to Bayesian inversion:

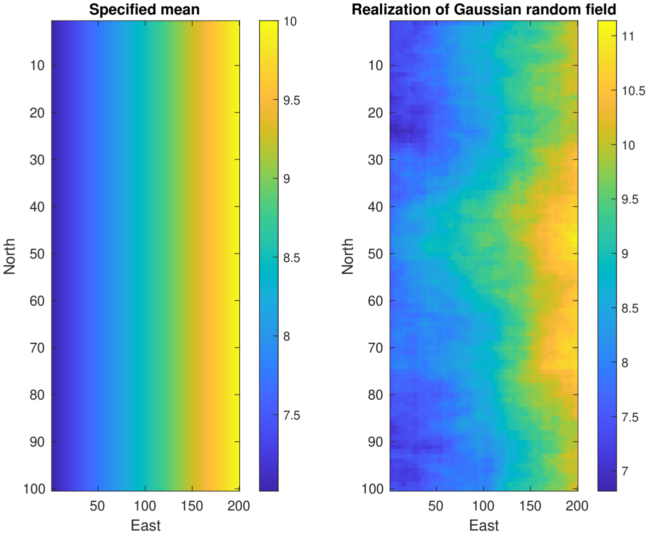
- Prior model for (spatial) variables of interest.
  Usually a *Gaussian model*, bringing in smoothness and regularization.
- Likelihood model for the link to data and the acquisition assumption.
  Focus of today will be a *linear Gaussian* likelihood model.
- Bayes' rule gives the posterior model.
  With the assumption of linearity and Gaussian densities, this posterior will also be *Gaussian*.

Gaussian random field model

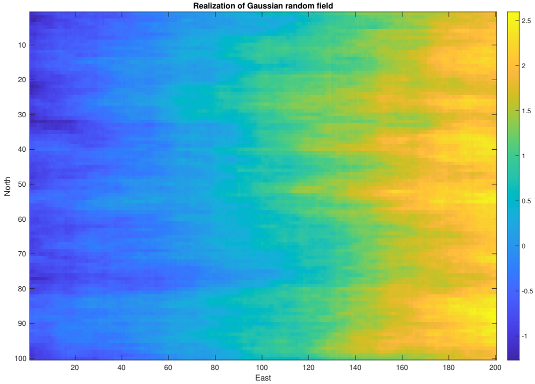
$x(\boldsymbol{s}) = \mu(\boldsymbol{s}) + z(\boldsymbol{s})$, $\boldsymbol{s} \in \mathcal{D} \in \mathcal{R}^d$ (Today $\mathcal{R}^2$ or $\mathcal{R}^3$.)

- $\mu(\boldsymbol{s})$ defines the spatial trend. Often depends on covariates, in a regression model: $\mu(\boldsymbol{s}) = \boldsymbol{h}(\boldsymbol{s})\boldsymbol{\beta}$.
- $z(\boldsymbol{s})$ is a zero-mean structured (spatially correlated) Gaussian process.
- Close sites are very correlated. Sites far away are less correlated.

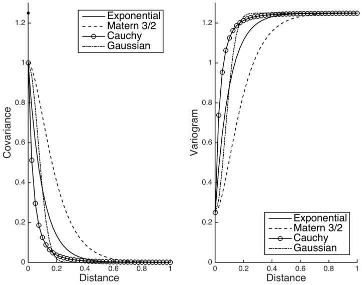
Trend and realization



Lower trend

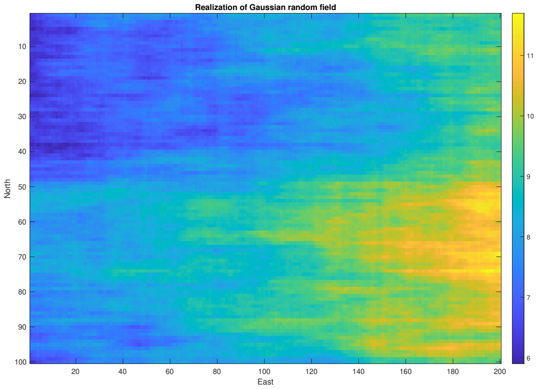

Realization of Gaussian random field
North
10
20
30
40
50
60
70
80
90
100
East
20
40
60
80
100
120
140
160
180
200
2.5
2
1.5
1
0.5
0
-0.5
-1

Covariance functions and variograms

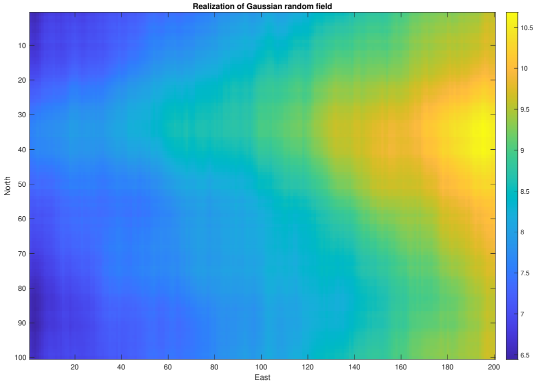


These valid choices of models give a positive definite covariance matrix for any discretization of the domain.

Larger variance and correlation



More smoothness



Monte Carlo sampling of a Gaussian process

Simulation on a finite number of grid cells.
Prior mean $\boldsymbol{\mu}$.
Cholesky matrix $\boldsymbol{LL}' = \boldsymbol{\Sigma}$. Standard deviations: $\sigma_i = \sqrt{\Sigma(i,i)}$, for all i.

- Specify parameters (mean and covariance).
- Set a grid of n locations (discretize the spatial domain).
- Generate independent standard normal variables: *z=randn(n,1)*
- Use the Cholesky matrix to get correlated variables: *v=L*z*
- Add the mean *x=mu+v*

Data and Gaussian posterior model

Prior model:

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

We collect data according to a design.
This defines a matrix $\boldsymbol{F}$ (potentially picking observation sites) and measurement noise (covariance matrix $\boldsymbol{R}$).

$$\boldsymbol{y}|\boldsymbol{x} \sim N(\boldsymbol{F}\boldsymbol{x}, \boldsymbol{R})$$

With Gaussian assumptions.

$$\boldsymbol{x}|\boldsymbol{y} \sim N(\boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{F}'(\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}' + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\mu}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{F}'(\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}' + \boldsymbol{R})^{-1}\boldsymbol{F}\boldsymbol{\Sigma})$$

Interpretation of conditional distribution

With just univariate x and a single data y, ($\boldsymbol{F} = 1$):

$$x|y \sim N(\mu + \frac{\sigma^2}{\sigma^2 + r^2}(y - \mu), \sigma^2(1 - \frac{\sigma^2}{\sigma^2 + r^2}))$$

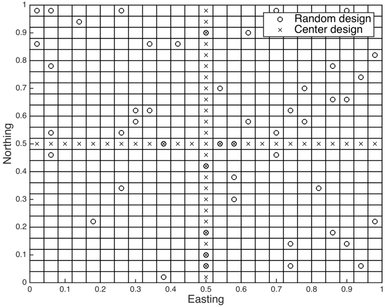
- Conditional mean is a weighting of prior mean and data: $\frac{r^2}{\sigma^2+r^2}\mu + \frac{\sigma^2}{\sigma^2+r^2}y$.
- The weights depend on the prior uncertainty σ^2 and the measurement uncertainty r^2.
- The conditional variance stays near σ^2 if r^2 is large. The conditioning has little effect.
- The conditional variance goes to 0 when r^2 is very small.

Properties of conditional distribution

$$\boldsymbol{x}|\boldsymbol{y} \sim N(\boldsymbol{\mu} + \boldsymbol{\Sigma F}'(\boldsymbol{F\Sigma F}' + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\mu}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma F}'(\boldsymbol{F\Sigma F}' + \boldsymbol{R})^{-1}\boldsymbol{F\Sigma})$$

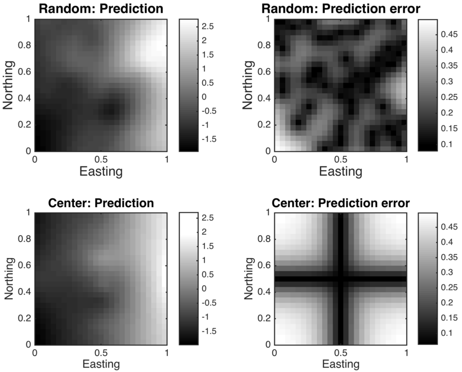
- In geostatistics and spatial interpolation this is sometimes called Kriging.
- Conditional mean is linear in the data.
- Conditioning to data changes the prediction at un-observed sites. And more so for large correlation.
- Conditional variance is reduced from the initial, the reduction depends on the design and the measurement accuracy.
- Conditional variance does not depend on the data. It can be computed before the data acquisition.
- Conditional standard deviations are square root of diagonal elements of covariance matrix.

Example : Data designs (F)



Example: Predictions and prediction standard deviation

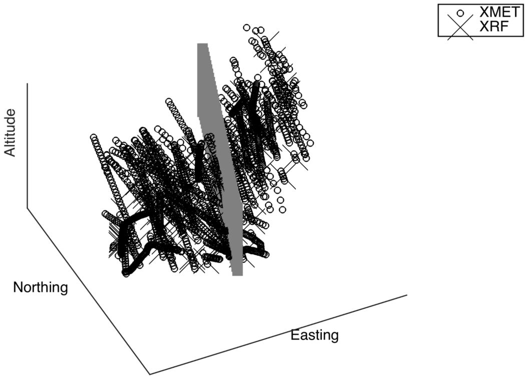
Prediction is the conditional mean. The prediction error is here extracted from the conditional covariance matrix : standard deviations are the square root of the diagonal.



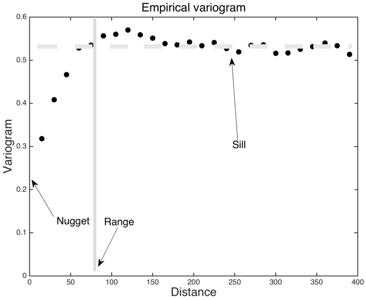
Case : mining data

- Data at 1871 locations where oxide measurements are gathered.
- Covariate from prior geological understanding $h(\boldsymbol{s}) = [1, \text{min.index}(\boldsymbol{s})]$. Regression model (trend) fit from available data.
- Spatial covariance is a Matern-type. Information will propagate from data locations.
- Two data types: some data made in the lab (very accurate, r^2 is small), others on-location with a handheld instrument (inaccurate, r^2 is large). The formulation allows them to be weighted differently and in a consistent manner.

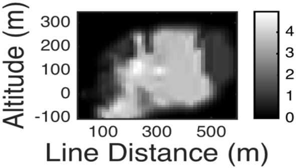
Grade prediction from boreholes

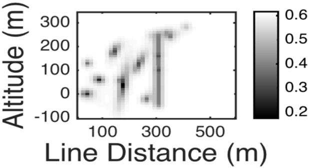


Variogram


Empirical variogram
Variogram
Distance
0
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300
350
400
Nugget
Range
Sill

Predictions and prediction standard deviations





Linear forward model

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Linear combinations of data can also be captured in the matrix $\boldsymbol{F}$. Differences, weighted averages, convolutions. This is common in e.g. seismic data or in medical tomography.

$$\boldsymbol{y}|\boldsymbol{x} \sim N(\boldsymbol{F}\boldsymbol{x}, \boldsymbol{R})$$

The posterior expression still holds:

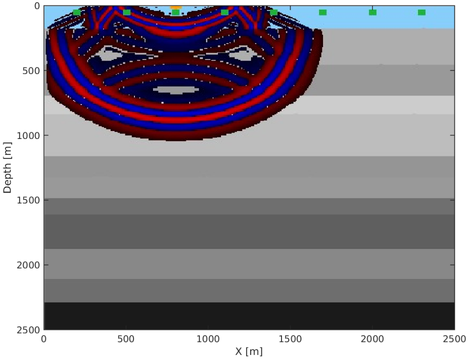
$$\boldsymbol{x}|\boldsymbol{y} \sim N(\boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{F}'(\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}' + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\mu}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{F}'(\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}' + \boldsymbol{R})^{-1}\boldsymbol{F}\boldsymbol{\Sigma})$$

Seismic Amplitude-versus-angle inversion

Processed seismic Amplitude versus angle data can be considered to be a linear operator of the elastic properties. Then: $\boldsymbol{F} = \boldsymbol{W}\boldsymbol{A}\boldsymbol{D}$.

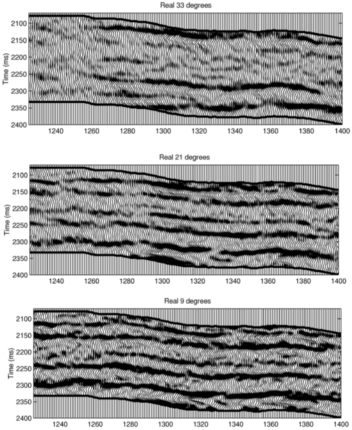
- $\boldsymbol{D}$ is a difference operator. Seismic waves are reflected when properties in the subsurface change.
- $\boldsymbol{A}$ consists of physical weights defined by Aki-Richards coefficients (depending on angles of incidence and background V_p/V_s ratio.
- $\boldsymbol{W}$ defines a wavelet convolution operator mimicking the seismic source signature.

Seismic Data



'Raw' seismic data is highly non-linear. Processing steps are often done, and in one domain the seismic amplitude versus angle data are close to linearly related to the elastic properties (V_p, V_s and ρ in the subsurface).

Seismic Amplitude versus Angle Data



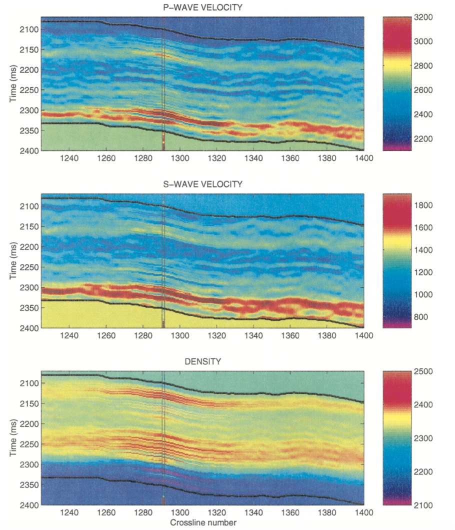
Prior

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

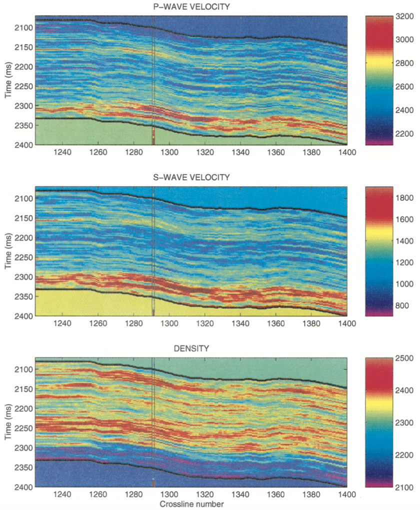
- $\boldsymbol{\mu}$ is separate for elastic properties $\log V_p$, $\log V_s$ and $\log \rho$. Often with a depth trend for each.
- $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \otimes \boldsymbol{S}$ has a 3×3 covariance matrix between the elastic properties and a $N \times N$ spatial correlation matrix between all N sites ($n = 3N$).

Buland and Omre (2003), Buland et al. (2003).

Linear Bayesian inversion : mean

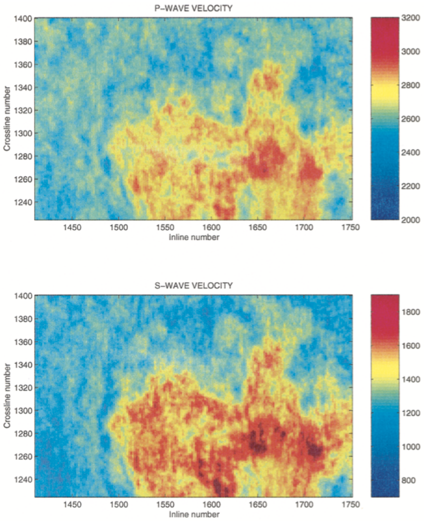


Linear Bayesian inversion : sample



Fast inversion, with uncertainty quantification.

Linear Bayesian inversion : slice mean



Non-linear approaches

- Markov chain Monte Carlo sampling.
- Ensemble Kalman filtering.