Value of Information Analysis in Spatial Models

Jo Eidsvik

Jo.eidsvik@ntnu.no

Plan for course

Time	**Topic**
Monday	Introduction and motivating examples
	Elementary decision analysis and the value of information
Tuesday	Multivariate statistical modeling, dependence, graphs
	Value of information analysis for dependent models
Wednesday	Spatial statistics, spatial design of experiments
	Value of information analysis in spatial decision situations
Thursday	Examples of value of information analysis in Earth sciences
	Computational aspects
Friday	Sequential decisions and sequential information gathering
	Examples from mining and oceanography

Every day: Small exercise half-way, and computer project at the end.

Other applications

- Robotics - where should drone (UAV) or submarine (AUV) go to collect valuable data?
- Automatic maintenance – how to allocation sensors to ‘best’ monitor state of system?
- Adaptive testing – which questions to ask for optimal recommendation system?
- Internet of things – which sensors should be active now?

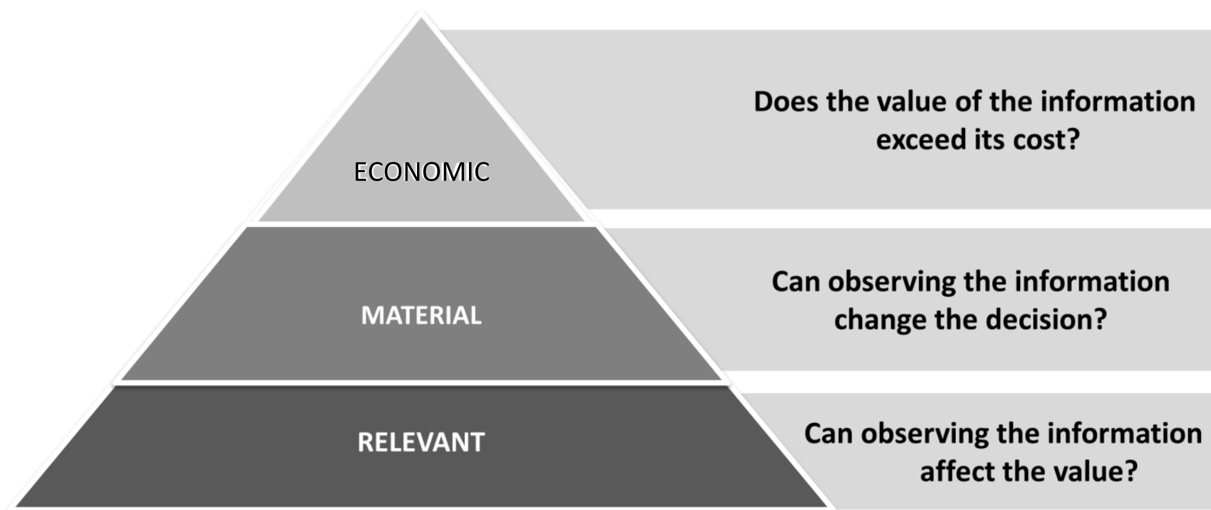
Which data are valuable?

Five Vs of big data:
- Volume
- Variety
- Velocity
- Veracity
- **Value**

We must acquire and process the data that has value!
There is often a clear question that one aims to answer, and data should help us.

Value of information (VOI)

In many applications we consider purchasing more data before making difficult decisions under uncertainty.
VOI is useful for quantifying the value of the data, before it is acquired and processed.



This pyramid of conditions - VOI is different from other information criteria (entropy, variance, prediction error, etc.)

Formula for VOI

$$PV = \max_{\boldsymbol{a} \in \boldsymbol{A}} \left\{ E\left(v(\boldsymbol{x}, \boldsymbol{a}) \right) \right\} = \max_{\boldsymbol{a} \in \boldsymbol{A}} \left\{ \int_{\boldsymbol{x}} v(\boldsymbol{x}, \boldsymbol{a}) p\left(\boldsymbol{x} \right) d\boldsymbol{x} \right\}$$

$$PoV\left(\boldsymbol{y} \right) = \int \max_{\boldsymbol{a} \in \boldsymbol{A}} \left\{ E\left(v\left(\boldsymbol{x}, \boldsymbol{a} \right) \mid \boldsymbol{y} \right) \right\} p\left(\boldsymbol{y} \right) d\boldsymbol{y}$$

$$VOI\left(\boldsymbol{y} \right) = PoV\left(\boldsymbol{y} \right) - PV.$$

The analysis is usually done for **static decisions** and **static tests**:
- We make the one-time decisions here and now.
- We can only collect the data here and now.

Sequential decisions or **sequential tests** can give benefits.

Sequential decision (prior value)

$$PV_{seq} = \max_{i_1} \left\{ E\left(v\left(x_{i_1}, a_{i_1} = 1\right)\right) + \int_{x_{i_1}} \max_{i_2 \neq i_1} \left\{0, \mathrm{ContVal}_{i_2}\left(x_{i_1}\right)\right\} p\left(x_{i_1}\right) dx_{i_1}, 0 \right\}$$

$$\mathrm{ContVal}_{i_2}\left(x_{i_1}\right) = E\left(v\left(x_{i_2}, a = 1\right) \mid x_{i_1}\right) + \int_{x_{i_2}} \max_{i_3 \neq i_1, i_2} \left\{0, \mathrm{ContVal}_{i_3}\left(x_{i_1}, x_{i_2}\right)\right\} p\left(x_{i_2} \mid x_{i_1}\right) dx_{i_2}$$

$$\mathrm{ContVal}_{i_n}\left(x_{i_1}, \ldots, x_{i_{n-1}}\right) = \max\left\{0, E\left(v\left(x_{i_n}, a_{i_n} = 1\right) \mid x_{i_1}, \ldots, x_{i_{n-1}}\right)\right\}$$

Solution is a discrete optimization method called dynamic programming:

- Go through all possible decision paths, moving forward.
- Find the optimal values winding backwards in the tree of paths .

This is costly for large systems, and several approximations exists.
Suboptimal strategies, using heuristics, are often used in practice, with success.

Sequential decisions (VOI)

$$PV_{seq} = \max_{i_1}\left\{E\left(v\left(x_{i_1}, a_{i_1} = 1\right)\right) + \int_{x_{i_1}} \max_{i_2 \neq i_1}\left\{0, \mathrm{ContVal}_{i_2}\left(x_{i_1}\right)\right\} p\left(x_{i_1}\right) dx_{i_1}, 0\right\}$$

$$PoV_{seq}\left(\boldsymbol{y}\right) = \int_{\boldsymbol{y}} \max_{i_1}\left\{\begin{matrix} 0, E\left(v\left(x_{i_1}, a = 1\right) \mid \boldsymbol{y}\right) + \\ \int_{x_{i_1}} \max_{i_2 \neq i_1}\left\{0, \mathrm{ContVal}_{i_2}\left(x_{i_1}, \boldsymbol{y}\right)\right\} p\left(x_{i_1} \mid \boldsymbol{y}\right) dx_{i_1} \end{matrix}\right\} p\left(\boldsymbol{y}\right) d\boldsymbol{y}$$

Must now use dynamic programming for different data outcomes.

VOI is difference between posterior and prior value.

Information gathering

	Perfect	Imperfect
Total	Exact observations are gathered for all variables. This is rare, occurring when there is extensive coverage and highly accurate data gathering. $\boldsymbol{y} = \boldsymbol{x}$	Noisy observations are gathered for all variables. This is common in situations with remote sensors with extensive coverage, e.g. seismic, radar, satellite data. $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\varepsilon}$
Partial	Exact observations are gathered at some variables. This might occur, for instance, when there is careful analysis of some samples, or sensors at some locations. $\boldsymbol{y}_{\mathbb{K}} = \boldsymbol{x}_{\mathbb{K}}, \quad \mathbb{K}\ \text{subset}$	Noisy observations are gathered at some variables. Examples include noisy sensors for local monitoring, testing along a survey line, etc. $\boldsymbol{y}_{\mathbb{K}} = \boldsymbol{x}_{\mathbb{K}} + \boldsymbol{\varepsilon}_{\mathbb{K}}, \quad \mathbb{K}\ \text{subset}$

Could also have **sequential** (adaptive) information gathering.

Sequential information gathering

Decision maker has the opportunity of dynamic testing, where one can stop testing, or continue testing, depending on the currently available data. The sequential order of tests and the number of tests also depend on the data.

Continue testing.

$$PoV_{seqtest}\left(\boldsymbol{y}_1\right) = \int \max \left\{ \begin{array}{l} \max_{j \neq 1}\left\{CV\left(j \mid 1\right)\right\}, \\ \sum_{i=1}^{n} \max_{a_i}\left\{E\left(v\left(x_i, a_i\right) \mid \boldsymbol{y}_1\right)\right\} \end{array} \right\} p(\boldsymbol{y}_1) d\boldsymbol{y}_1$$

Stop testing.

$$CV\left(j \mid 1\right) = Cont\left(j \mid 1\right) - P_j$$

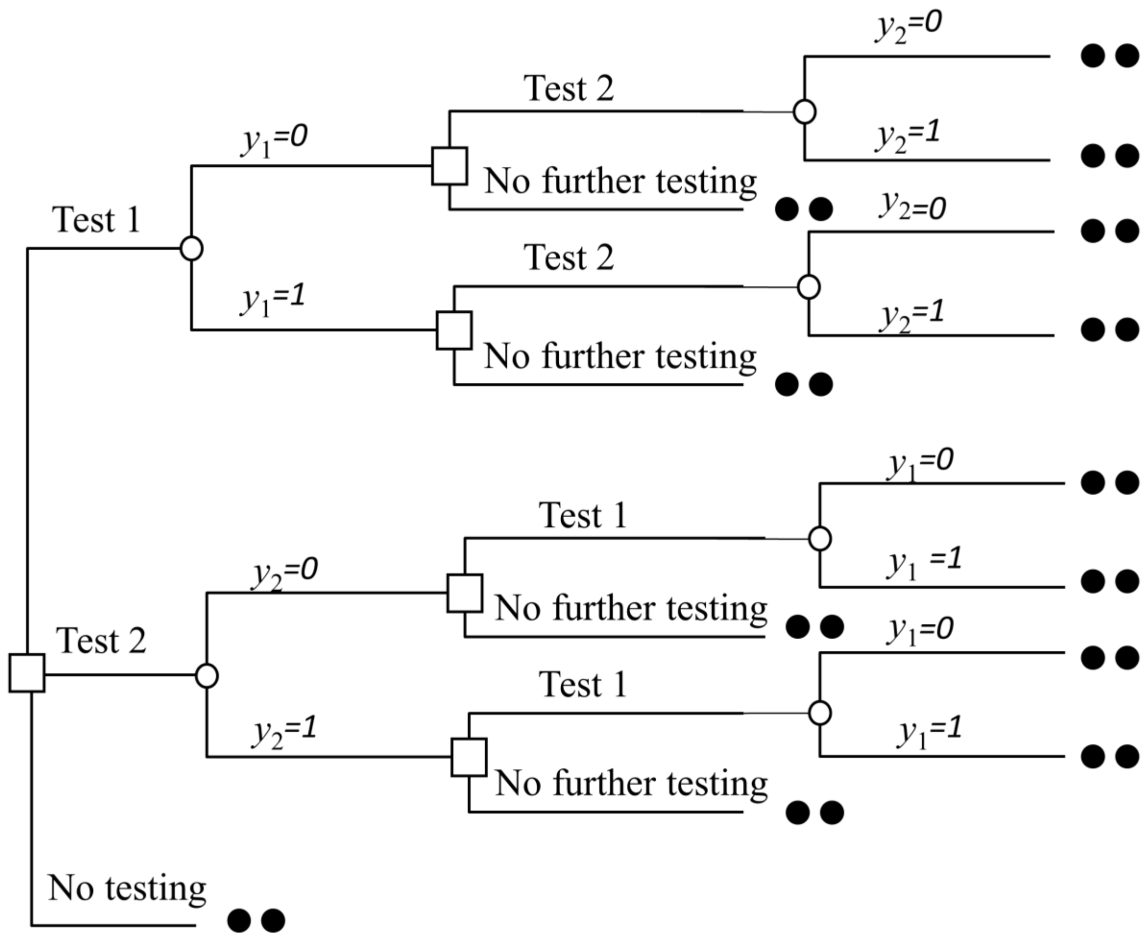
Continue testing.

$$Cont\left(j \mid 1\right) = \int \max \left\{ \begin{array}{l} \max_{k \neq 1, j}\left\{CV\left(k \mid j, 1\right)\right\}, \\ \sum_{i=1}^{n} \max_{a_i}\left\{E\left(v\left(x_i, a_i\right) \mid \boldsymbol{y}_1, \boldsymbol{y}_j\right)\right\} \end{array} \right\} p(\boldsymbol{y}_j \mid \boldsymbol{y}_1) d\boldsymbol{y}_j$$

Solution is again dynamic programming.

Stop testing.

Sequential testing– bivariate illustration



Sequential information (bivariate data)

Value with no more testing (after first test):

$$PoV(\boldsymbol{y}_1) = \int \sum_{i=1}^{n} \max_{a_i \in A_i} \left\{ E\left(v\left(x_i, a_i \right) \mid \boldsymbol{y}_1 \right) \right\} p(\boldsymbol{y}_1) d\boldsymbol{y}_1$$

Criterion for continued testing:

$$\int_{\boldsymbol{y}_2} \sum_{i=1}^{n} \max_{a_i \in A_i} \left\{ E\left(v\left(x_i, a_i \right) \mid \boldsymbol{y}_1, \boldsymbol{y}_2 \right) \right\} p\left(\boldsymbol{y}_2 \mid \boldsymbol{y}_1 \right) d\boldsymbol{y}_2 - P_2 > \sum_{i=1}^{n} \max_{a_i \in A_i} \left\{ E\left(v\left(x_i, a_i \right) \mid \boldsymbol{y}_1 \right) \right\}$$

$$PoV_{seqtest}(\boldsymbol{y}_1) = \int \max \left\{ \begin{array}{l} \int_{\boldsymbol{y}_2} \sum_{i=1}^{n} \max_{a_i} \left\{ E\left(v\left(x_i, a_i \right) \mid \boldsymbol{y}_1, \boldsymbol{y}_2 \right) \right\} p\left(\boldsymbol{y}_2 \mid \boldsymbol{y}_1 \right) d\boldsymbol{y}_2 - P_2, \\ \sum_{i=1}^{n} \max_{a_i} \left\{ E(v\left(x_i, a_i \right) \mid \boldsymbol{y}_1) \right\} \end{array} \right\} p(\boldsymbol{y}_1) d\boldsymbol{y}_1$$

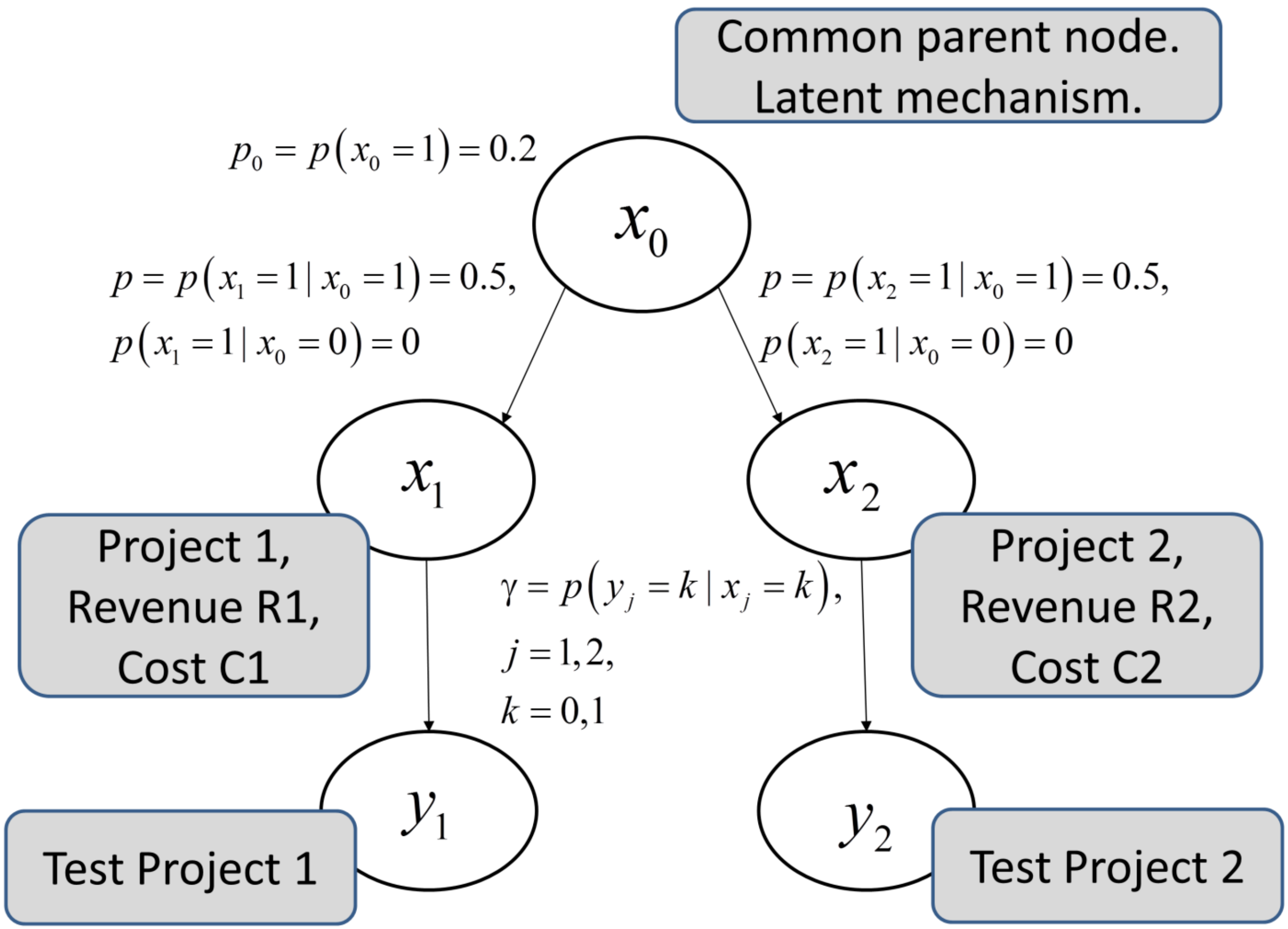
Continue testing when the additional expected value of more testing excceds the price.

Sequential decisions and testing

Of course, one can have both sequential decisions and sequential data gathering!
Even more complex to evaluate and interpret, but it also gives opportunities.

Not considered here.

Bivariate projects example



Bivariate projects example

Need to frame the **decision situation**:

- Can one freely select (profitable) projects, or must both be selected? Free selection.
- Does value decouple? No coupling.
- Can one do sequential selection? Non-sequential.

Need to study **information gathering** options:

- Imperfect test, or perfect test? Study both.
- Can one test both prospects, or only one (total or partial)? Study both.
- Can one perform sequential testing? **Compare non-sequential / sequential testing.**

Focus on information gathering.

Bivariate projects example - static

$$\mathrm{R}_1, \mathrm{R}_2, \qquad C_1 = C_2 = \mathrm{Cost}$$

$$PV = \sum_{i \in \{1,2\}} \max\left\{0, \mathrm{R}_i \cdot p\left(x_i = 1\right) - \mathrm{Cost}\right\}$$

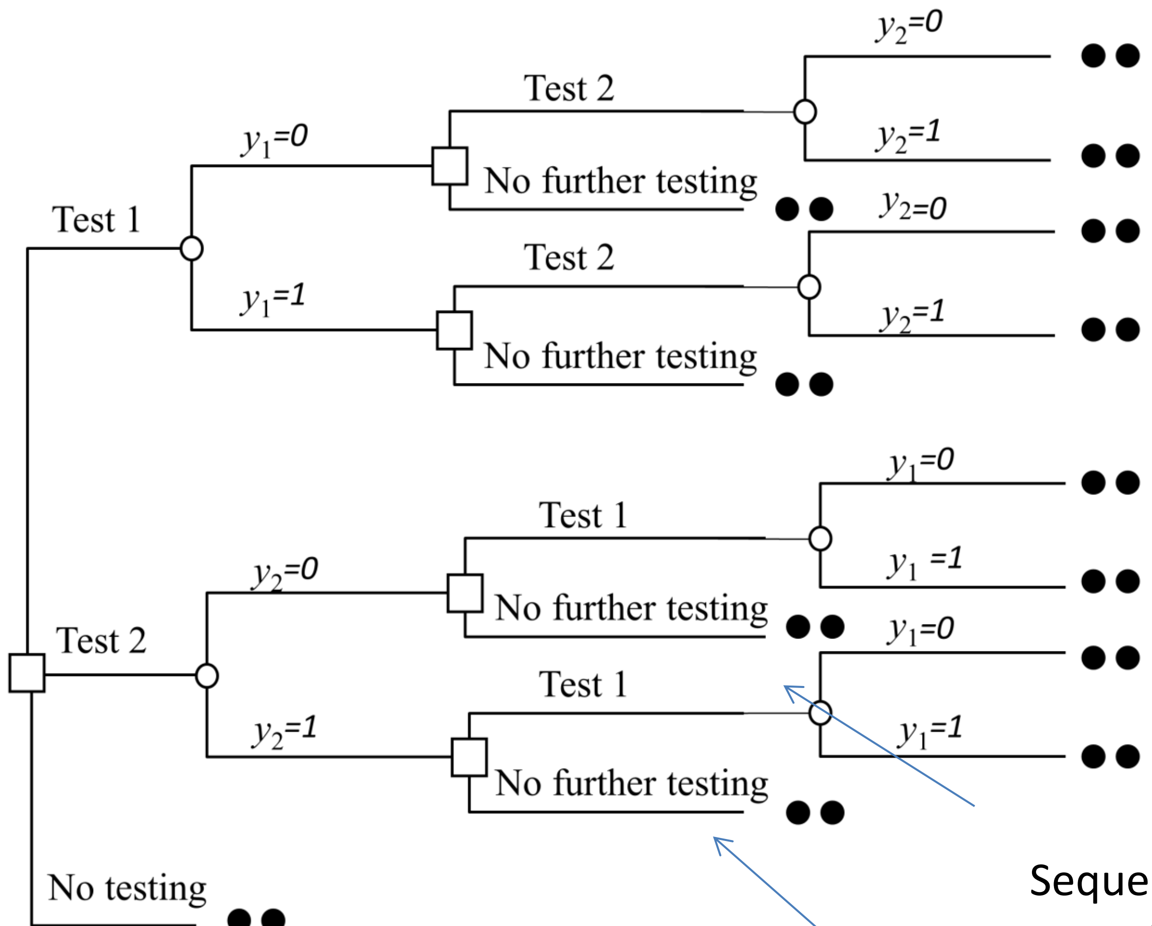
Total imperfect information, **static**

$$PoV\left(\boldsymbol{y}\right) = \sum_{i \in \{1,2\}} \max\left\{0, \mathrm{R}_i \, p\left(x_i = 1 | \boldsymbol{y}\right) - \mathrm{Cost}\right\} p\left(\boldsymbol{y}\right)$$

$$VOI\left(\boldsymbol{y}\right) = PoV\left(\boldsymbol{y}\right) - PV$$

With sequential testing, we do one test, and can continue with second, if it is worthwhile to do so.

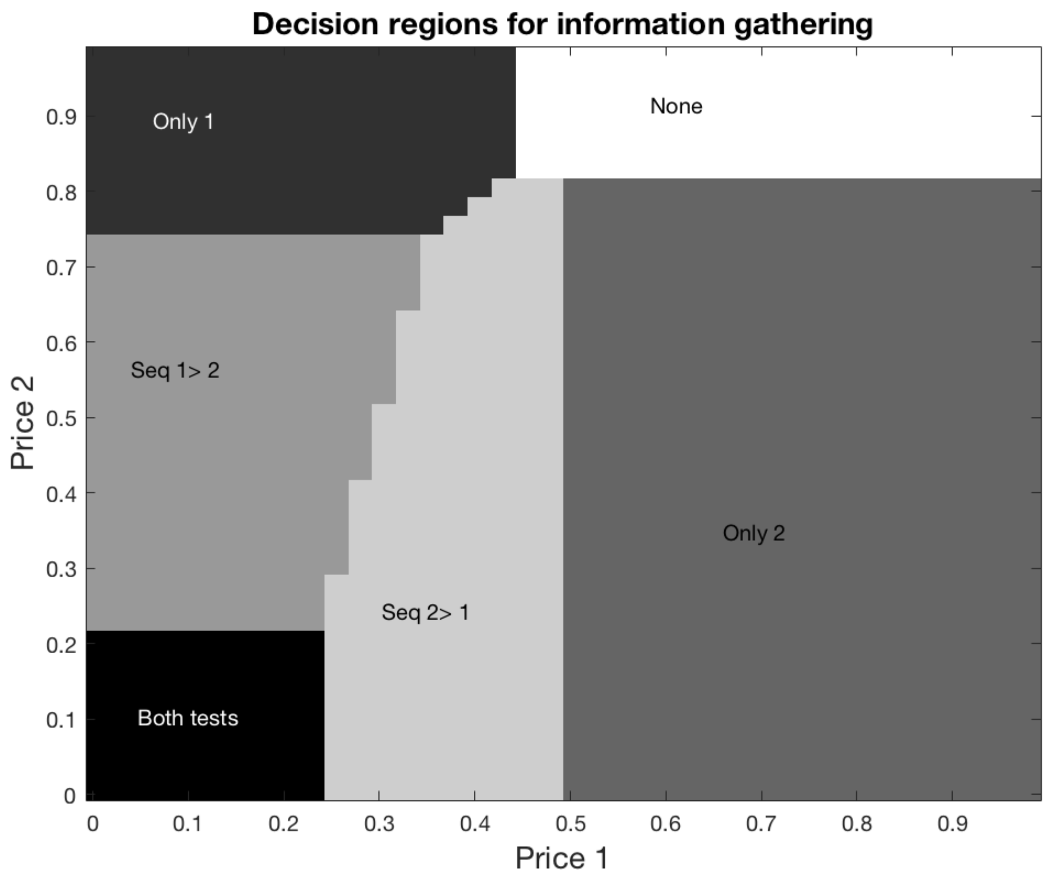
Sequential testing– Bivariate projects



Sequential has the opportunity to test two.

Static, stops after first.

Decision regions for project testing



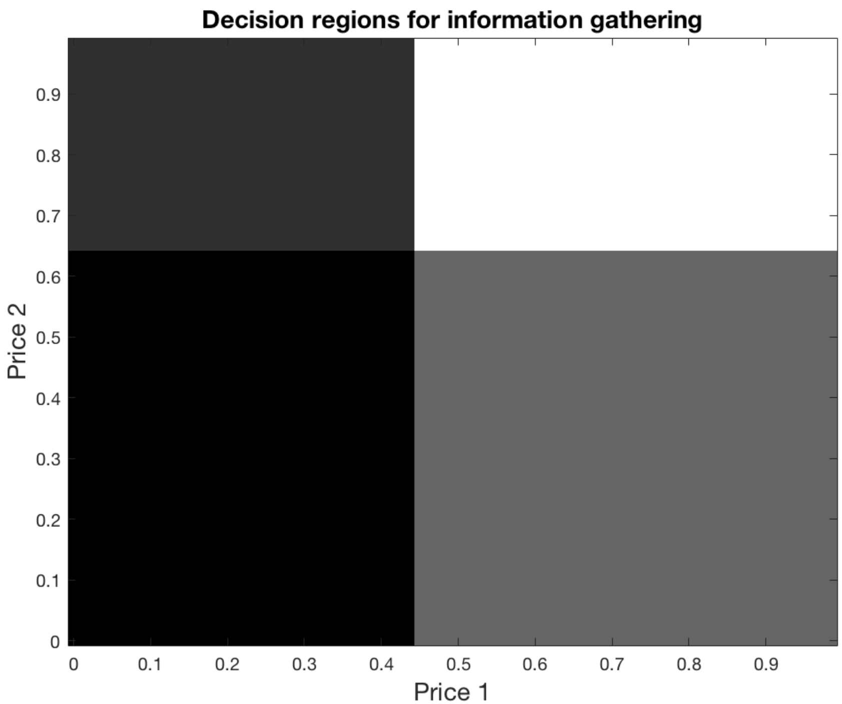
Sequential testing is better than static (only 1, or 2 tests), for a large set of price levels.

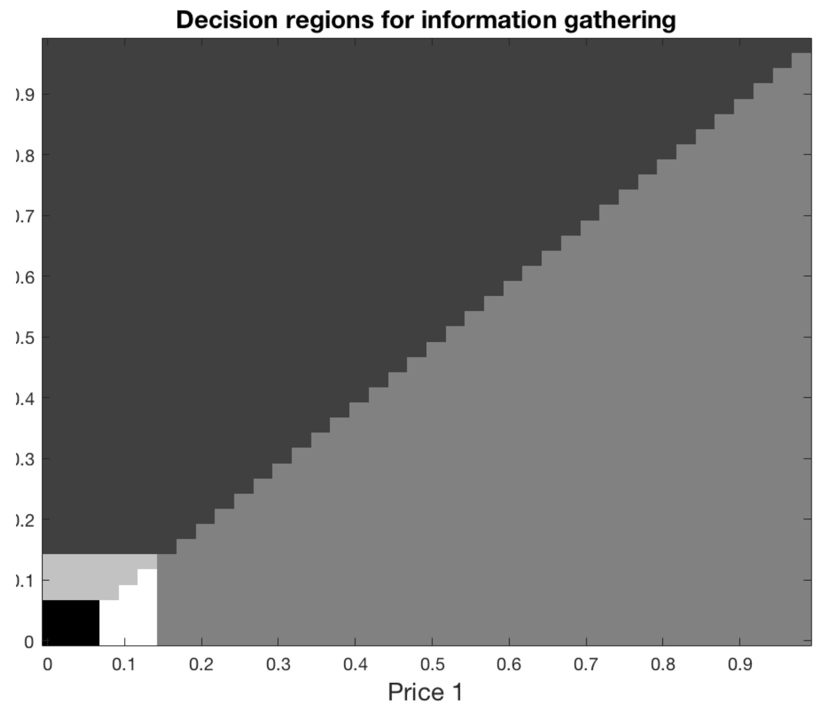
Decision regions for other correlations

Exercise:

1. **Discuss what the decision regions would look like for almost no correlation (but same marginal probabilities).**
2. **Discuss what the decision regions would look like for very high correlation (but still same marginal probabilities).**

Decision regions for low/high corr





Low dependence: VOI is additive, no value in sequential testing.

High dependence: One only is usually preferred, because partial test extremely informative. Sequential not needed.

Myopic strategy for information

Myopic (near-sighted) is a common strategy for sequential problems.
It considers only one-stage at a time, not looking into the 'future'.

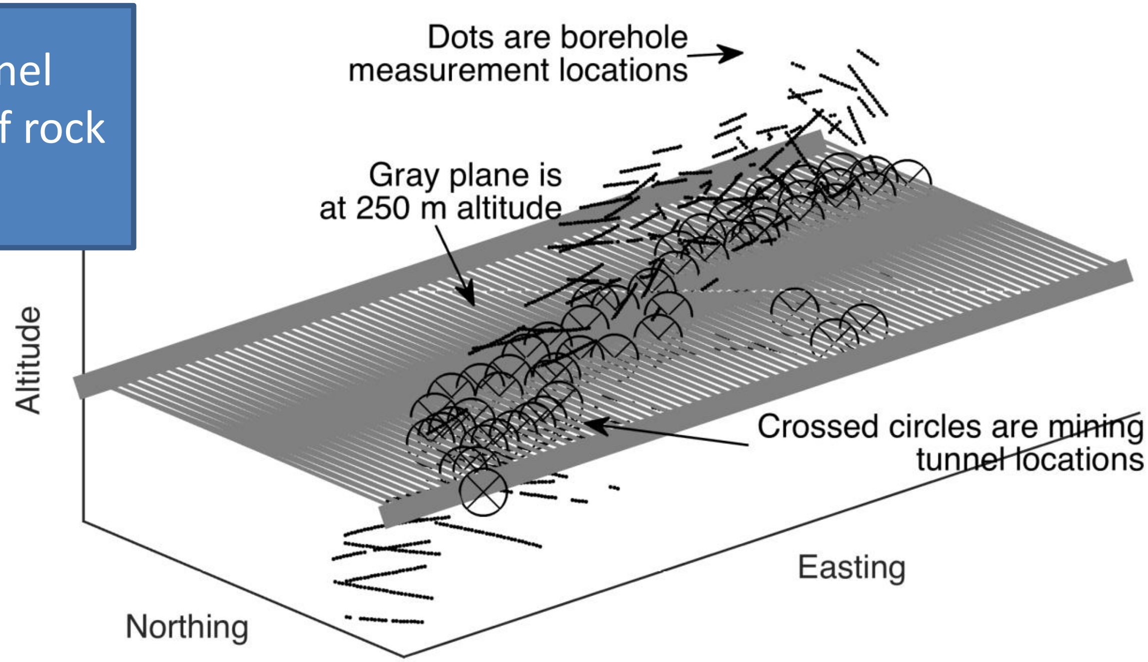
1. Find best first data design, using one-stage, if any give positive VOI. 1 level
2. Collect data (by simulation) using best design.
3. Update probability distributions, conditional on the data.
4. Find second best data, using one-stage, in new model, if any give positive VOI.
5. Collect second data (by simulation from new model) using best design. 2 level
6. Update probability distributions, conditional on the data.
7. Find third best data, using one-stage, in new model, if any give positive VOI.

3 level

.....

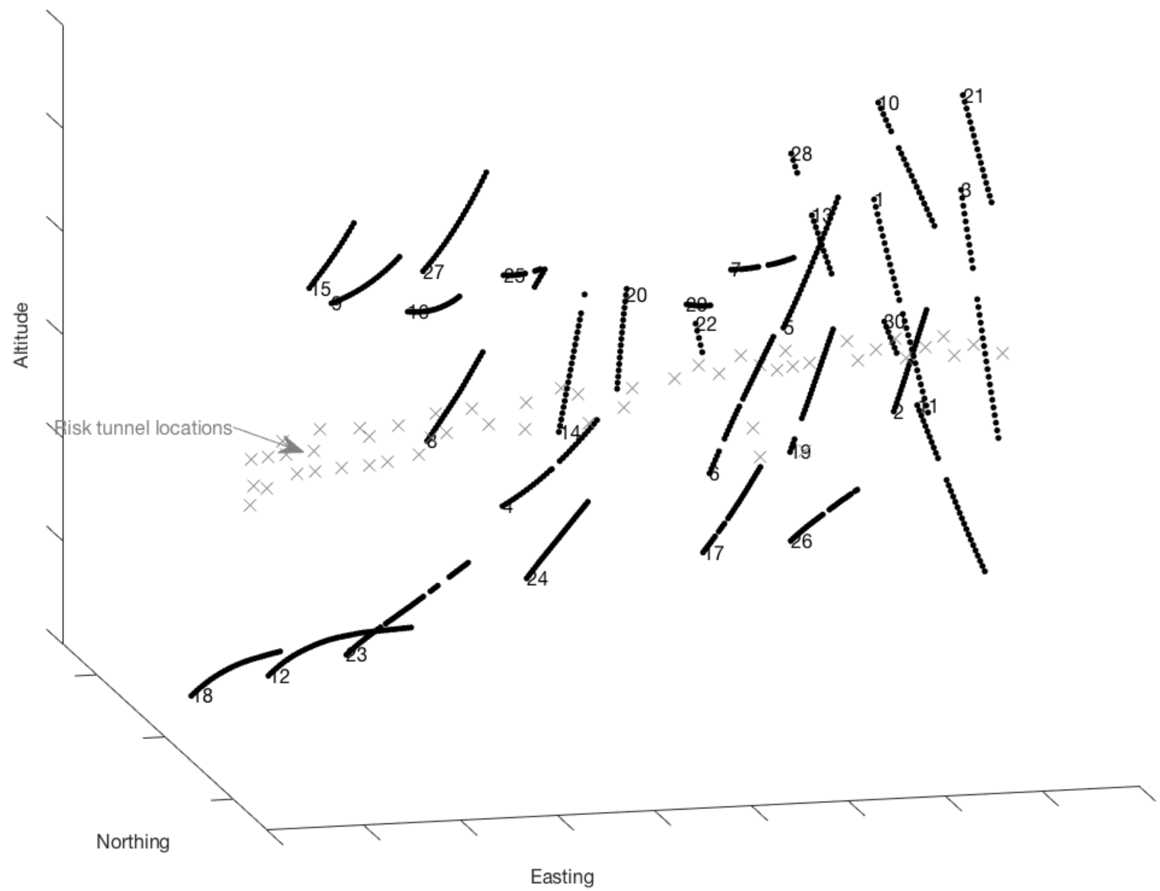
We will rock you – mining hazard example

Add support at tunnel locations where risk of rock fall is too high.



What is the value of additional borehole information about spatial joint counts?

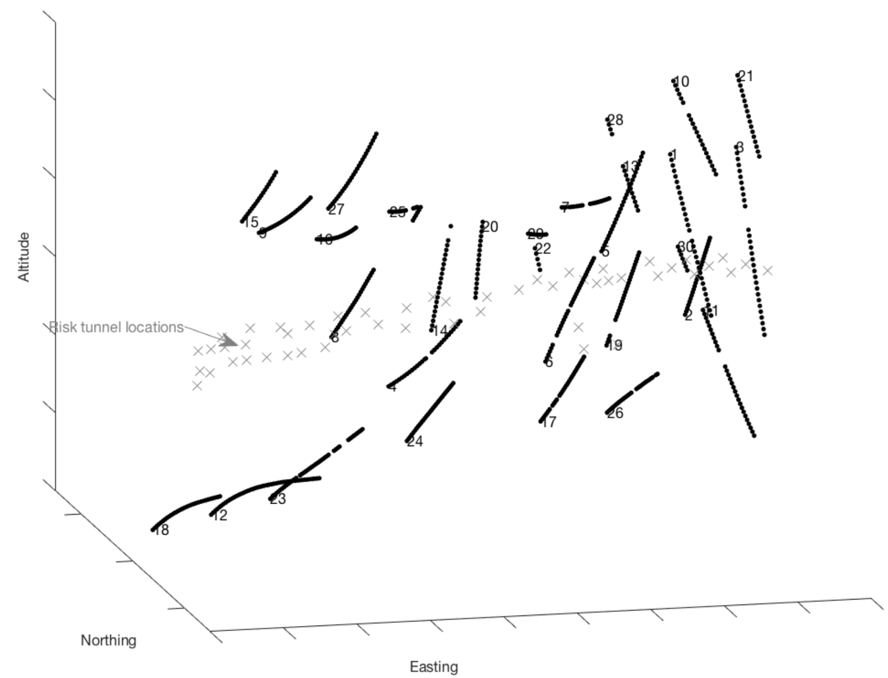
Mining risk locations and boreholes



Is the borehole data valuable for rock-hazard decisions?

Mining risk locations and boreholes

Sequential data gathering opportunities:
 - drill one borehole at the time, process data, see model result, then consider if more data should be gathered, and where.



Models for mining risk and data

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

Prior for joints.

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x} + N(\boldsymbol{0}, \tau^2 \boldsymbol{I})$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{F}\boldsymbol{x}, \tau^2 \boldsymbol{I})$$

Likelihood,
design matrix, picks
borehole locations.

The Poisson counts are rather large, Gaussian model is a reasonable approximation.

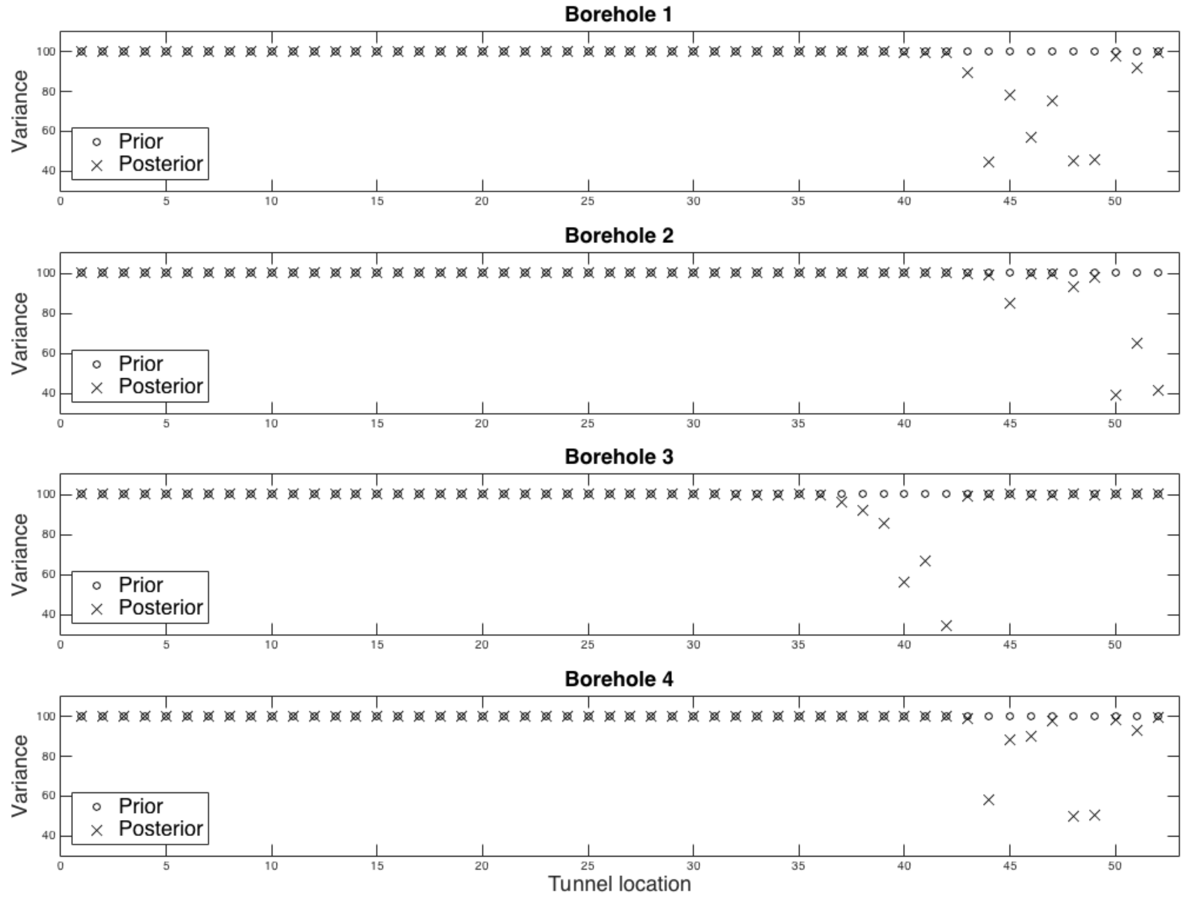
Conditioning – Gaussian models

$$E\left(\boldsymbol{x} \mid \boldsymbol{y}\right) = \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{F}^{t}\left(\tau^{2}\boldsymbol{I} + \boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^{t}\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\mu}\right)$$

$$Var\left(\boldsymbol{x} \mid \boldsymbol{y}\right) = \boldsymbol{\Sigma} - \boldsymbol{R}, \qquad \boldsymbol{R} = \boldsymbol{\Sigma}\boldsymbol{F}^{t}\left(\tau^{2}\boldsymbol{I} + \boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^{t}\right)^{-1}\boldsymbol{F}\boldsymbol{\Sigma}$$

$$r_i = \sqrt{R_{ii}}$$

Mining risk – variance reduction



Information gathering in single boreholes.

VOI - Gaussian models

High flexibility:
Can select individual units.

$$PV = \sum_{i=1}^{n} \max\left\{0, E\left(x_i\right)\right\} = \sum_{i=1}^{n} \max\left\{0, \mu_i\right\}$$

Value decouples to sum.

$$\boldsymbol{R} = \boldsymbol{\Sigma F}^t \left(\tau^2 \boldsymbol{I} + \boldsymbol{F \Sigma F}^t\right)^{-1} \boldsymbol{F \Sigma} \qquad r_i = \sqrt{R_{ii}}$$

$$PoV\left(\boldsymbol{y}\right) = \sum_{i=1}^{n} \int \max\left\{0, E\left(x_i \mid \boldsymbol{y}\right)\right\} p\left(\boldsymbol{y}\right) d\boldsymbol{y} = \sum_{i=1}^{n} \left(\mu_i \Phi\left(\mu_i \big/ r_i \right) + r_i \phi\left(\mu_i \big/ r_i \right) \right)$$

$\phi\left(z\right), \Phi\left(z\right)$ standard Gaussian density and cumulative function

Mining risk – myopic scheme

1. Find best single borehole posterior values, if any.

$$PoV\left(\boldsymbol{y}_j\right)=\sum_{i=1}^{n}\left(\mu_i\Phi\left(\mu_i/r_{i,j}\right)+r_{i,j}\phi\left(\mu_i/r_{i,j}\right)\right)-P_j$$

2. Collect data in this borehole. $\boldsymbol{y}_j$

3. Update the model

$$\boldsymbol{\mu}=\boldsymbol{\mu}+\boldsymbol{\Sigma}\boldsymbol{F}_j^t\left(\tau^2\boldsymbol{I}+\boldsymbol{F}_j\boldsymbol{\Sigma}\boldsymbol{F}_j^t\right)^{-1}\left(\boldsymbol{y}_j-\boldsymbol{F}_j\boldsymbol{\mu}\right)$$

$$\boldsymbol{\Sigma}=\boldsymbol{\Sigma}-\boldsymbol{R}_j,\qquad \boldsymbol{R}_j=\boldsymbol{\Sigma}\boldsymbol{F}_j^t\left(\tau^2\boldsymbol{I}+\boldsymbol{F}_j\boldsymbol{\Sigma}\boldsymbol{F}_j^t\right)^{-1}\boldsymbol{F}_j\boldsymbol{\Sigma}$$

$$r_{i,j}=\sqrt{R_{ii,j}}$$

4. Stop testing or continue testing.

$$Stop=\sum_{i=1}^{n}\max\left\{0,\mu_i\right\}$$

Find largest among all k. $\longrightarrow$

$$Cont\left(\boldsymbol{y}_k\right)=\sum_{i=1}^{n}\left(\mu_i\Phi\left(\mu_i/r_{i,k}\right)+r_{i,k}\phi\left(\mu_i/r_{i,k}\right)\right)-P_k$$

Etc….

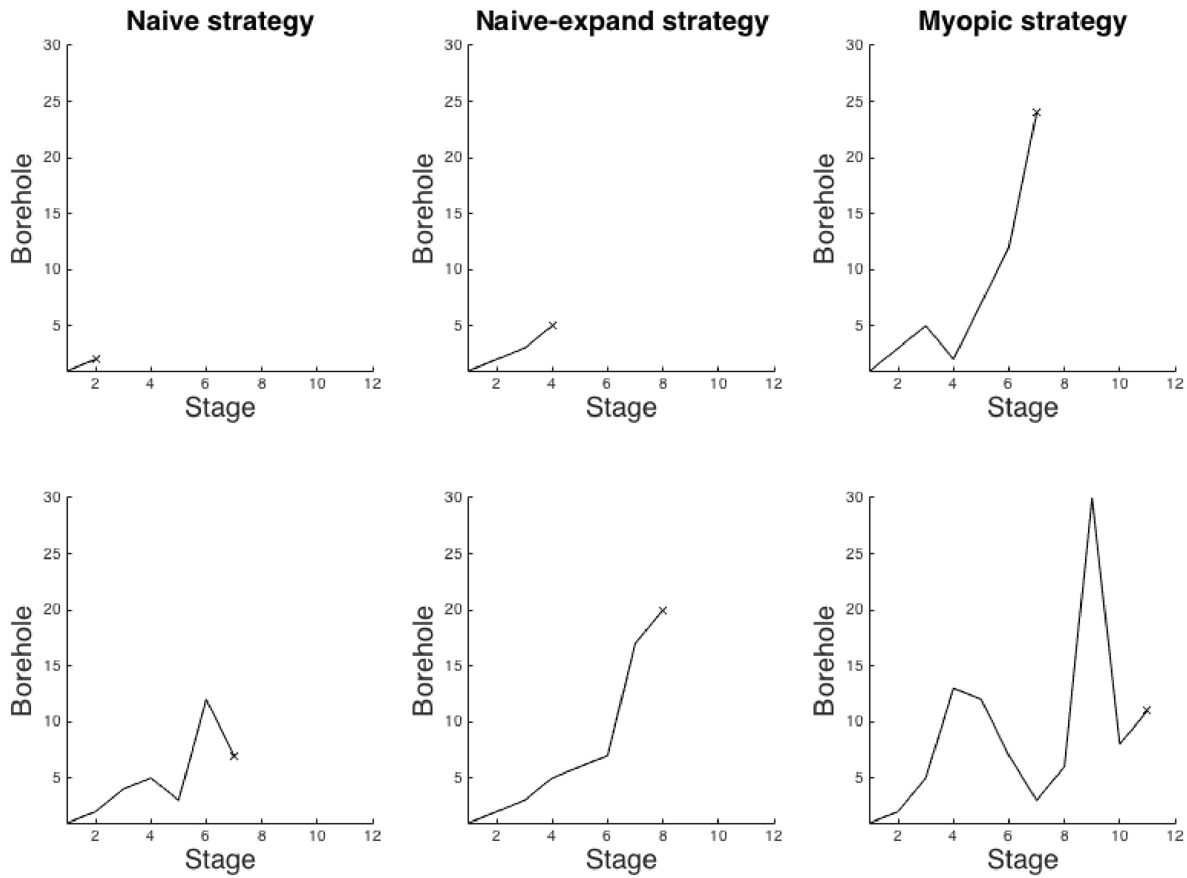
Mining risk - results

We compare myopic with other strategies.

	Static	Naive	Naive +	Myopic
VOI(1)	3.6	5.9	13.5	17.0
Ave.depth	1	2.4	6.5	8.7

There is clearly much to gain by having the opportunity to do sequential tests.

Mining risk – ‘depths’ for two data samples



Mining risk - results

We compare results for two different correlations.

Corr:225m	Static	Naive	Naive +	Myopic
VOI(1)	3.6	5.9	13.5	17.0
Ave.depth	1	2.4	6.5	8.7

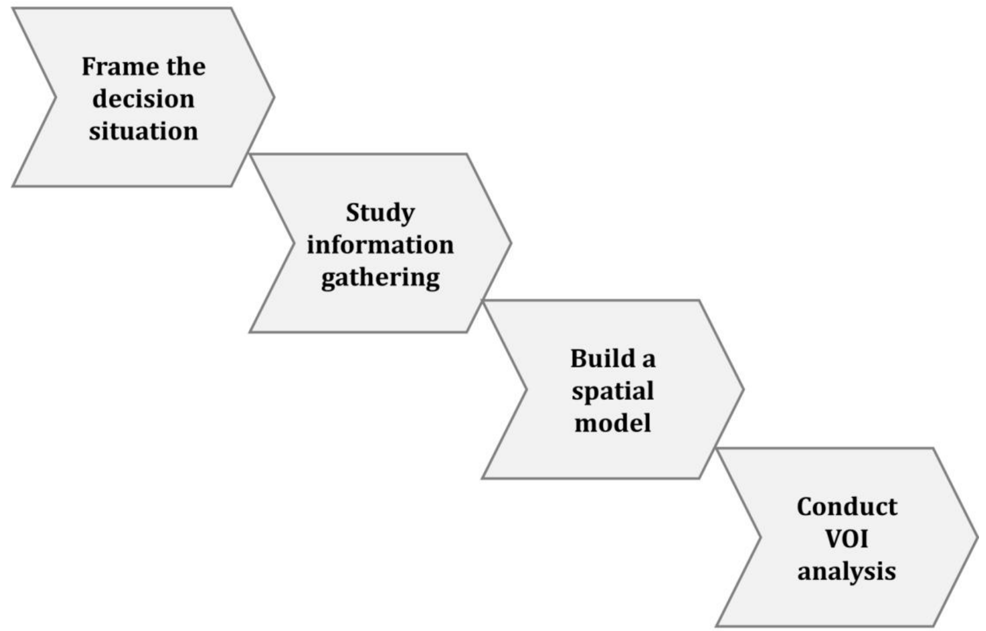
Corr:300m	Static	Naive	Naive +	Myopic
VOI(1)	5.1	8.0	18.6	25.3
Ave.depth	1	2.4	5.6	10.2

- For all strategies value increases with more correlation.
- The depth (number of stages) of myopic increases with more correlation.

- Myopic takes 100 x more computing time than static.
- How much better would more complex strategies do?? Unknown.

AUV temperature example

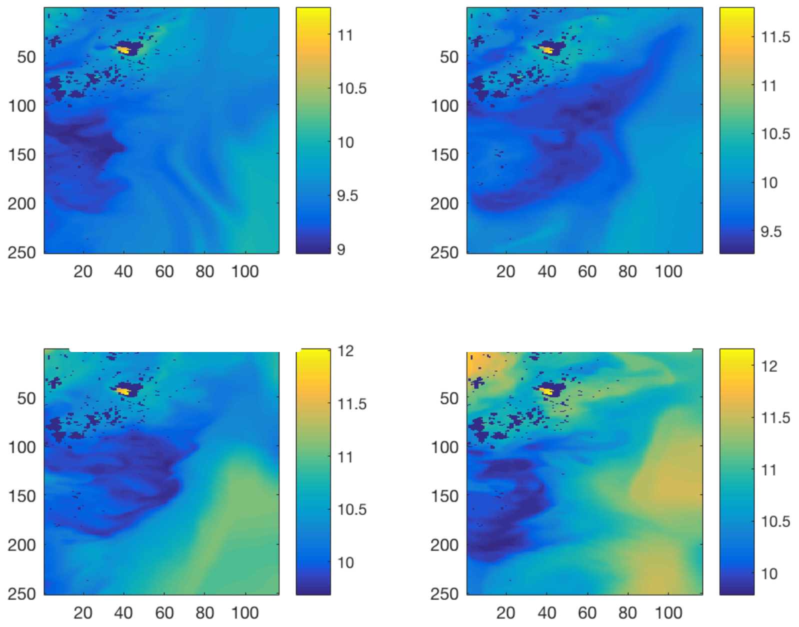
- Goal (value) is to detect large spatial gradients in ocean temperature.
- Autonomous underwater vehicle (AUV) information. Where? And in what sequence?
- Model for temperature is represented by Gaussian spatial process.
- VOI analysis uses analytical approach and myopic heuristics.



Fossum et al., 2017.

Classifying ocean temperature variability

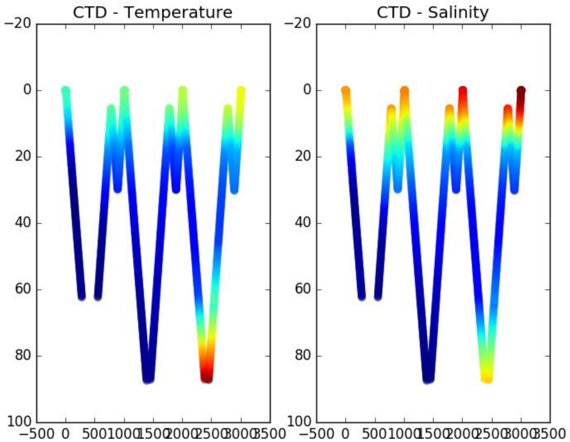
Prior realizations



Questions:

- Environmental challenges
- Fish farming
- Algae bloom

Typical AUV data



Gaussian prior and likelihood

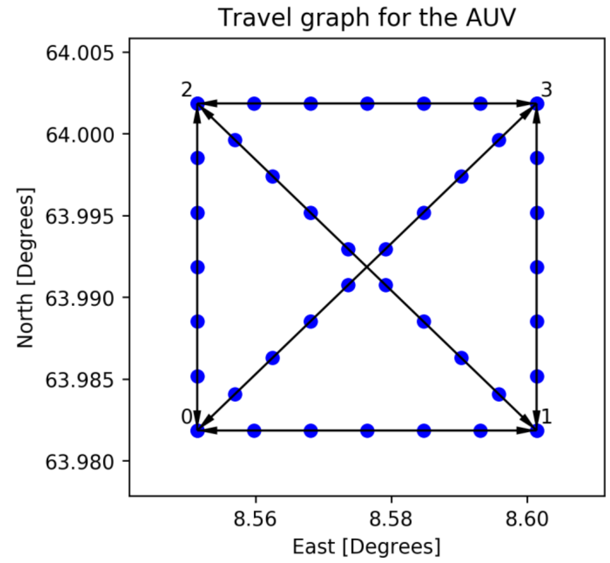
$$p(\boldsymbol{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Gaussian spatial process prior for temperatures.

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x} + N(\boldsymbol{0}, \tau^2 \boldsymbol{I})$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{F}\boldsymbol{x}, \tau^2 \boldsymbol{I})$$

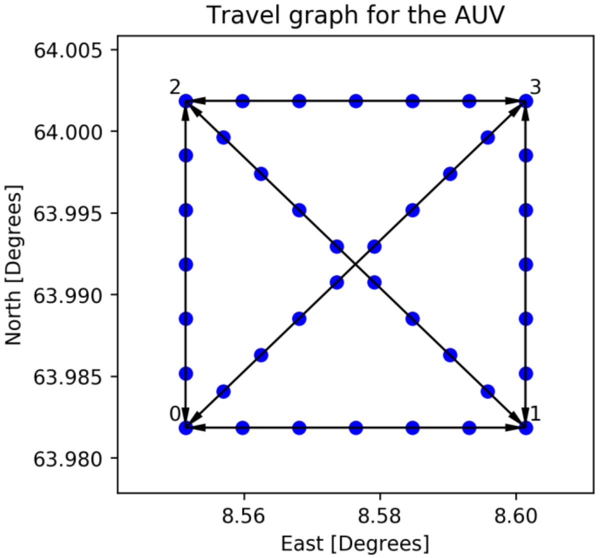
Likelihood, design matrix, picks data locations, for every time step.



Goal of surveying

The main task for the AUV is to detect large gradients in temperature which are linked to algea bloom.

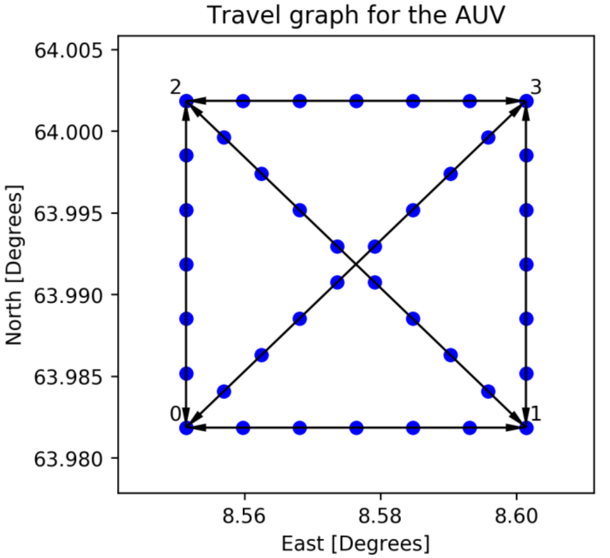
Waypoints in survey design for AUV.



Adaptive sequential algorithm

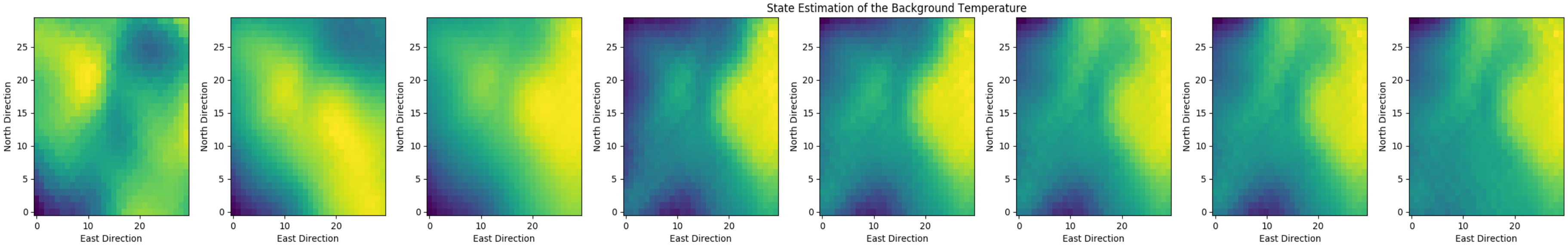
1. Find next best survey line (if any) from analytic VOI, of all possible survey lines.
2. Collect temperature data along currently best survey line.
3. Update temperature model in entire spatial domain given survey data.
4. Go to 1.

Myopic heuristic for dynamic program.

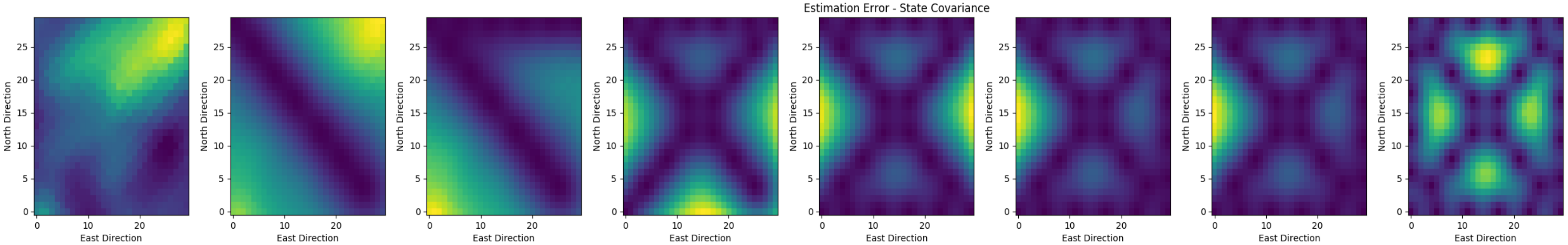


Results of adaptive algorithm

Mean of one survey realization



Variance of one survey realization



Wrap up:

- VOI is applied to a sequential search for good data designs.
- The design will depend on the data, and the results can be averaged over the data, to approximate value of different strategies.

- The process is spatio-temporal – extensions required.

Project: Sequential VOI in Gaussian models

Consider again the 25x25 grid, with a Gaussian process prior.

Assume the situation from Norwegian wood example, with low decision flexibility, harvest all or nothing.

Use the myopic strategy to find sequential data designs along the 25 North-South lines. The price of a test is P=0.1.

How many tests are done before we stop? (varies with data samples)
What tests are usually done? (varies with data samples)

Myopic scheme

1. Find best single NS line, if any.

$$\boldsymbol{R}_j = \boldsymbol{\Sigma}\boldsymbol{F}_j^t\left(\tau^2\boldsymbol{I}+\boldsymbol{F}_j\boldsymbol{\Sigma}\boldsymbol{F}_j^t\right)^{-1}\boldsymbol{F}_j\boldsymbol{\Sigma},$$

$$r_{w,j} = \sqrt{\sum\sum R_{ii',j}}, \qquad \mu_w = \sum \mu_i$$

$$PoV\left(\boldsymbol{y}_j\right) = \left(\mu_w\Phi\left(\mu_w / r_{w,j}\right) + r_{w,j}\phi\left(\mu_w / r_{w,j}\right)\right) - P_j$$

2. Collect data for this line. $\boldsymbol{y}_j$

3. Update the model

$$\boldsymbol{\mu} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{F}_j^t\left(\tau^2\boldsymbol{I}+\boldsymbol{F}_j\boldsymbol{\Sigma}\boldsymbol{F}_j^t\right)^{-1}\left(\boldsymbol{y}_j - \boldsymbol{F}_j\boldsymbol{\mu}\right)$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma} - \boldsymbol{R}_j,$$

4. Stop testing or continue testing.

$$Stop = \max\left\{0, \mu_w\right\}, \qquad \mu_w = \sum_{i=1}^{n} \mu_i$$

Find largest among all k.

$$Cont\left(\boldsymbol{y}_k\right) = \left(\mu_w\Phi\left(\mu_w / r_{w,k}\right) + r_{w,k}\phi\left(\mu_w / r_{w,k}\right)\right) - P_k$$

Use updated mean and covariances.

Etc....

Suggested solutions to projects

http://folk.ntnu.no/joeid/Monday_project.m

http://folk.ntnu.no/joeid/Tuesday_project.m

http://folk.ntnu.no/joeid/Wednesday_project.m

http://folk.ntnu.no/joeid/Thursday_project.m

http://folk.ntnu.no/joeid/Friday_project.m