Information approach to channels selection: stellar occultation measurements

V.Sofieva and E.Kyröla

Introduction

Overdetermined inverse problem

Detector: charge-coupled device (CCD);

Wavelength range 250-675 nm

GOMOS	COALA
1500 spectral channels	900 spectral channels
70 samples at different altitudes	70 samples at different altitudes
over 100 000 measurements	over 60 000 measurements

Retrieval:

ozone, NO_2, NO_3 , air, aerosol vertical profiles

Question: how to use the data efficiently?

Subset selection

Goals

—To reduce dimension of the problem (to speed up data processing)

—To detect the most informative channels (optimization of instrument design)

Selection criteria

? Accuracy of retrieval	*Function of altitude and retrieved quantity*
? Spatial resolution	
Information content	*single parameter*

Selection procedure

Direct solution is impossible

Application of information theory

Entropy of continuous pdf *P(x)*

$$H(P) = -\int P(x)\log(P(x)\,|\,M)\,dx$$

Information content

$$I = H(P_{before}) - H(P_{after})$$

For *Gaussian distribution*

$$H(P) = \tfrac{1}{2}\ln|C| + const$$

Linear model

$$y = Ax + \varepsilon$$

x and ε - Gaussian random variables

Statistical inversion

$$C^{-1} = A^T C_\varepsilon^{-1} A + C_a^{-1}$$

Information content of measurement

$$I = \tfrac{1}{2}\ln\left|E + A^T C_\varepsilon^{-1} A C_a\right|$$

Removal of channels ⇒ loss of the information content

Optimization problems

- *OP1:*
  - Choose the minimal set of measurements providing the information content $I_{0.}$
- *OP2:*
  - Choose the subset of *m* channels from *M* so that the information content is maximal.
- *Solutions can be not unique*
- *OP1 or OP2? Practical considerations:*
  - Only informative measurements subset is needed
  - Instrumental design: we can choose only spectral band,

not individual pixels

Selection procedures

- **Sequential selection (SS) (Rodgers, 1996)**
  - The change in the information content on introducing channel k

$$\delta I_k = \frac{1}{2}\ln\left(1+\frac{1}{\sigma_k^2}a_k C_{k-1} a_k^T\right)$$

  - $C_o = C_a$
  - Algorithm - dependent definition of the information content of individual channel
  - Relatively expensive computationally
- **Sequential deselection (SD)**

– The change in the information content caused by removal k-th channel

$$\delta I_k = -\frac{1}{2}\ln\left(1-\frac{1}{\sigma_k^2}a_k C_{k-1} a_k^T\right)$$

  - Algorithm - dependent definition of the information content of individual channel
  - expensive computationally
- **Fast algorithm of channel selection (IIC)**
  - Selection according to initial information content
  - Algorithm - independent definition of the information content of individual channel; convenient for comparison
  - Very fast

Optimality of selecting procedures.

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 4 \\ 4 & 3.1 \\ 3 & 1 \end{bmatrix}$$

$$C_a = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

$$C_\varepsilon = I$$

- Problem : to choose two most informative channels
- The most informative – {2,4}

  $I_{\{2,4\}}$=4.52
- Selected channels:

  - sequential selection: {1,3}

  - sequential de-selection: {1,3}

  - fast algorithm: {1,4}

  $I_{\{1,3\}}$=4.51 $I_{\{1,4\}}$=4.39
- all the procedures may lead to solutions not optimal in sense of OP2

Optimality of selecting procedures: Monte-Carlo test.

- Randomly generated matrix A (10 x 3) of forward model A_i ~U [0,10]
- 1000 different cases

Case	% of correct solutions			Mean loss of information, % $I = \frac{1}{n}\sum_{i=1}^{n}\frac{I^{opt}-I^{sel}}{I^{opt}}\times 100\%$		
	SS	SD	ICC	SS	SD	ICC
Selected channels=3						
C_a=100 E	37.6	84.2	43.2	1.03	0.07	3.05
C_a=10 E	37.6	84.0	43.2	1.38	0.09	3.89
Selected channels=6						
C_a=100 E	83.5	98.4	75.0	0.026	0.0006	0.044
C_a=10 E	83.8	98.4	74.8	0.034	0.0008	0.059

- All the procedures find combinations having high information content
- Sequential deselecting procedure showed the best result
- Numerical efficiency : ICC > SS > SD

Stellar occultation measurements

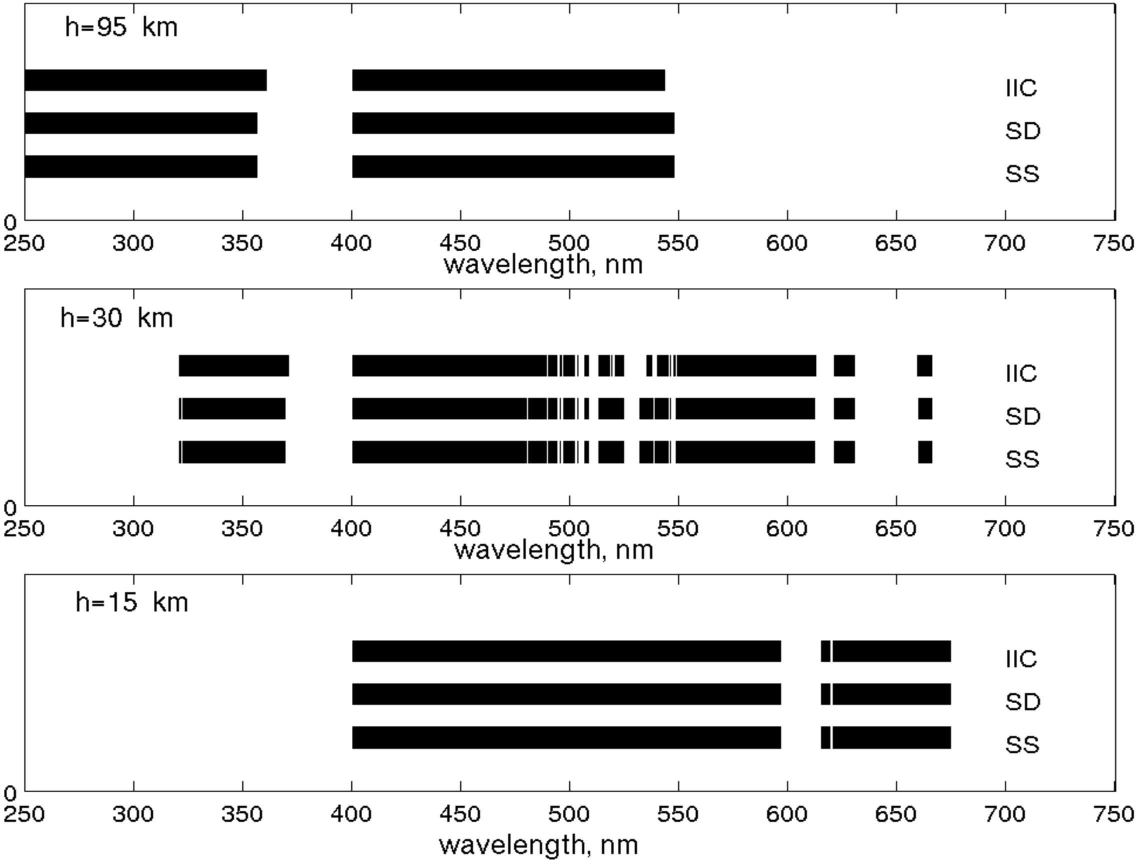
- GOMOS (~1500 spectral channels) and COALA (~ 900 channels)

in UV-VIS range

- 2 stage inversion : spectral inversion and vertical inversion
- we handle spectral inversion problem
- Linearized problem $\tau = \Sigma N + \varepsilon$
- separate inverse problem for each altitude
- Ozone, NO_2 , NO_3, air and aerosol

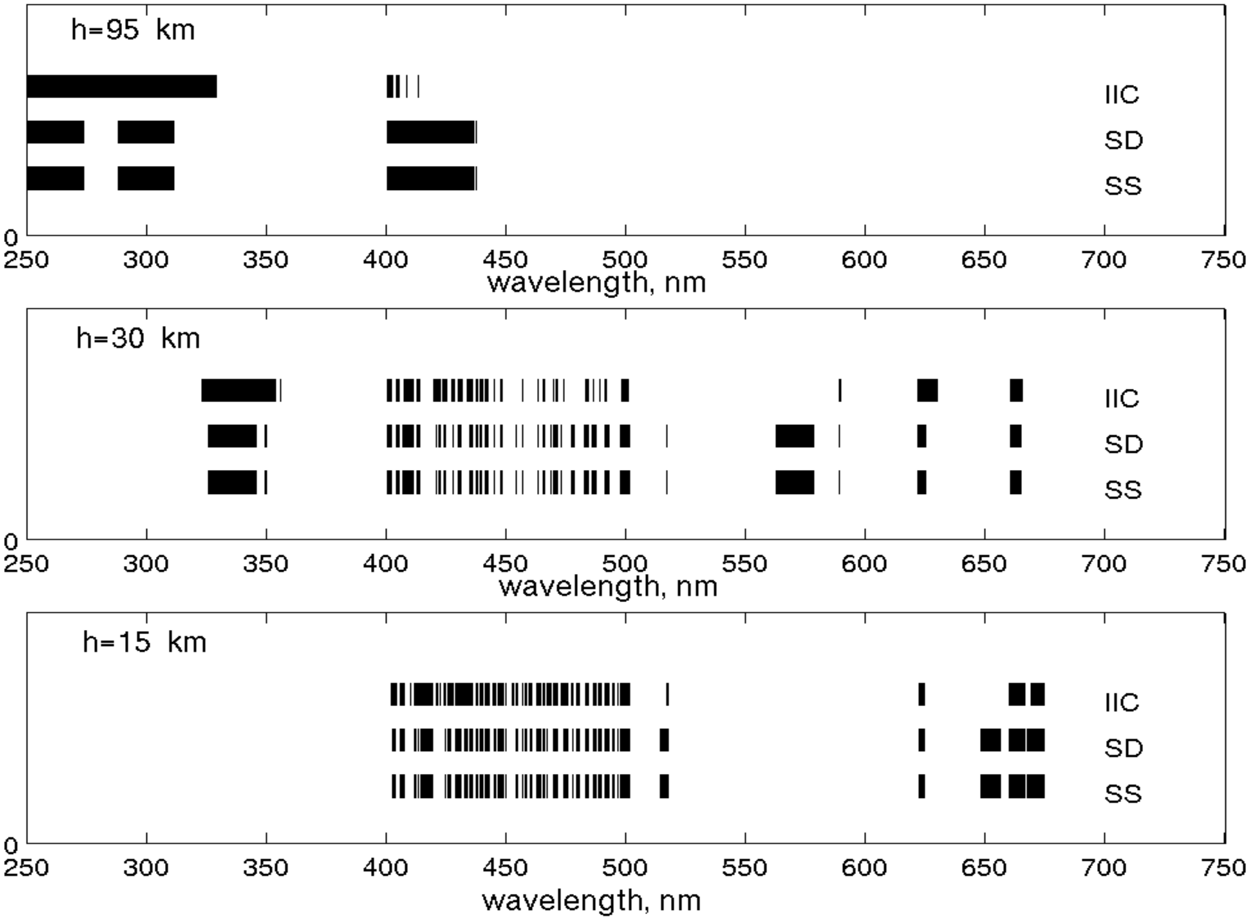
60 % of most informative channels

GOMOS ;
magnitude=2
T=10 000 K
wide prior



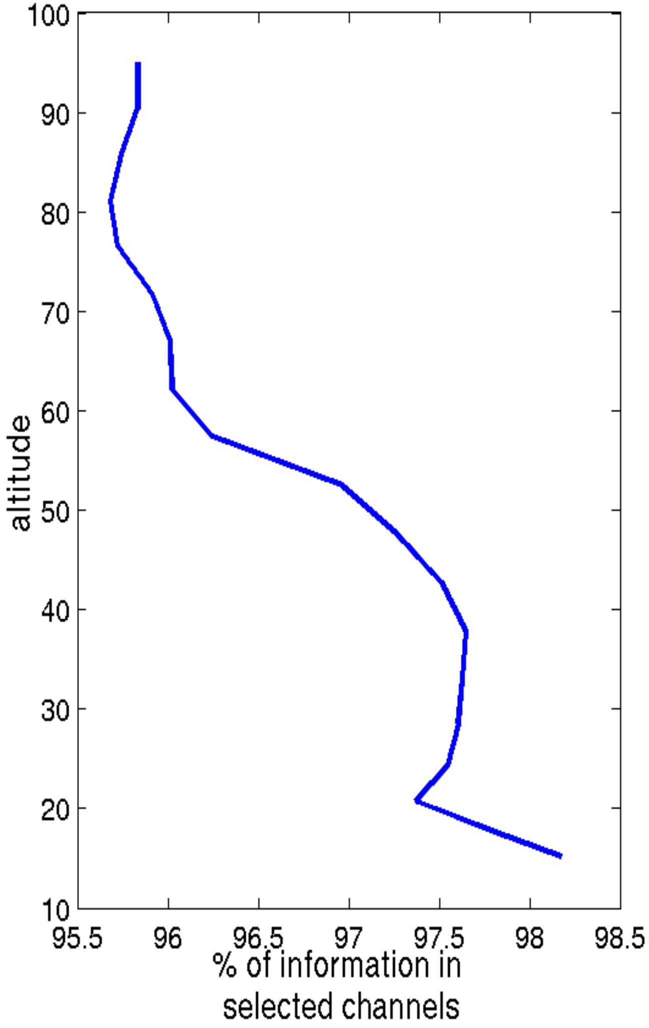
% of information	SS	SD	ICC
95 km	97.250 %	97.253 %	97.24 %
30 km	98.240 %	98.243 %	98.23 %
15 km	99.42 %	99.42 %	99.42 %
Computer time	45 sec	40 sec	< 1 sec

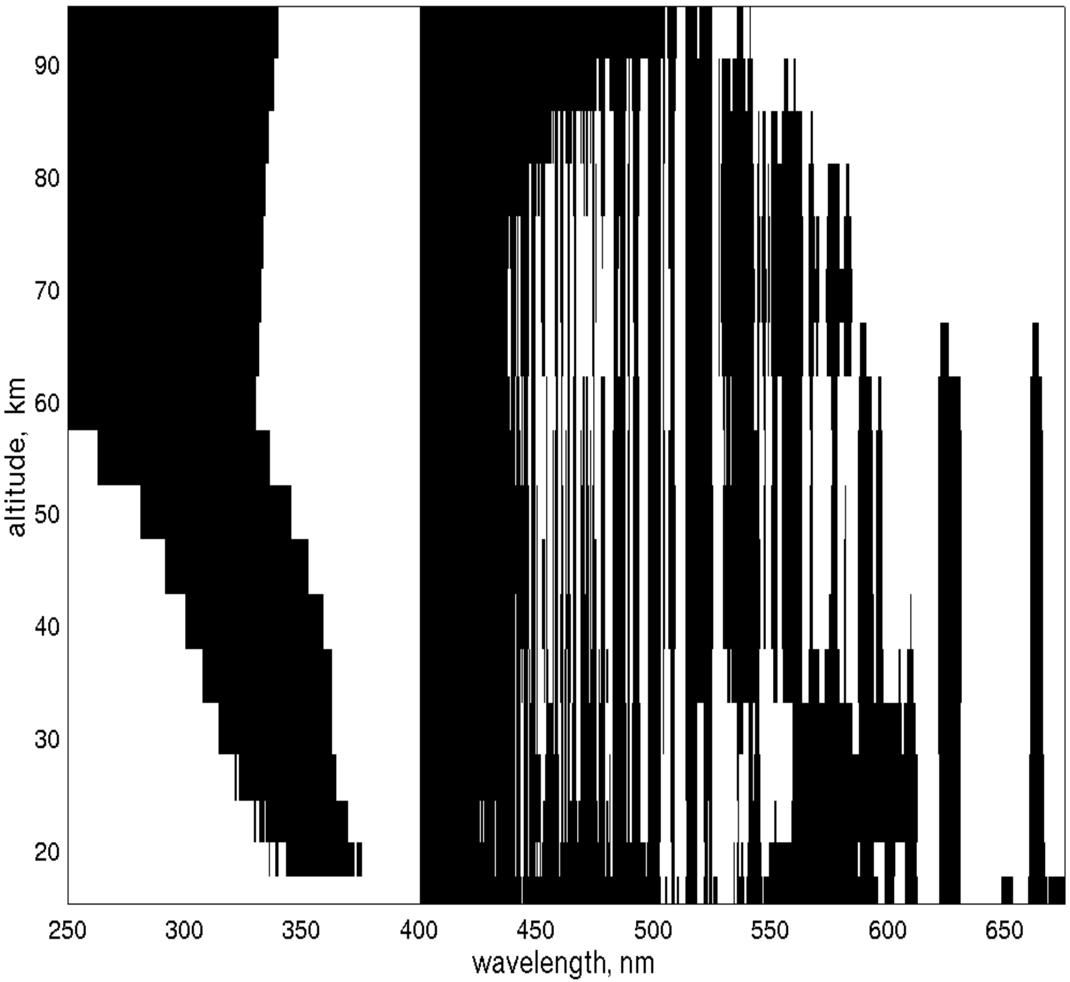
20 % of most informative channels



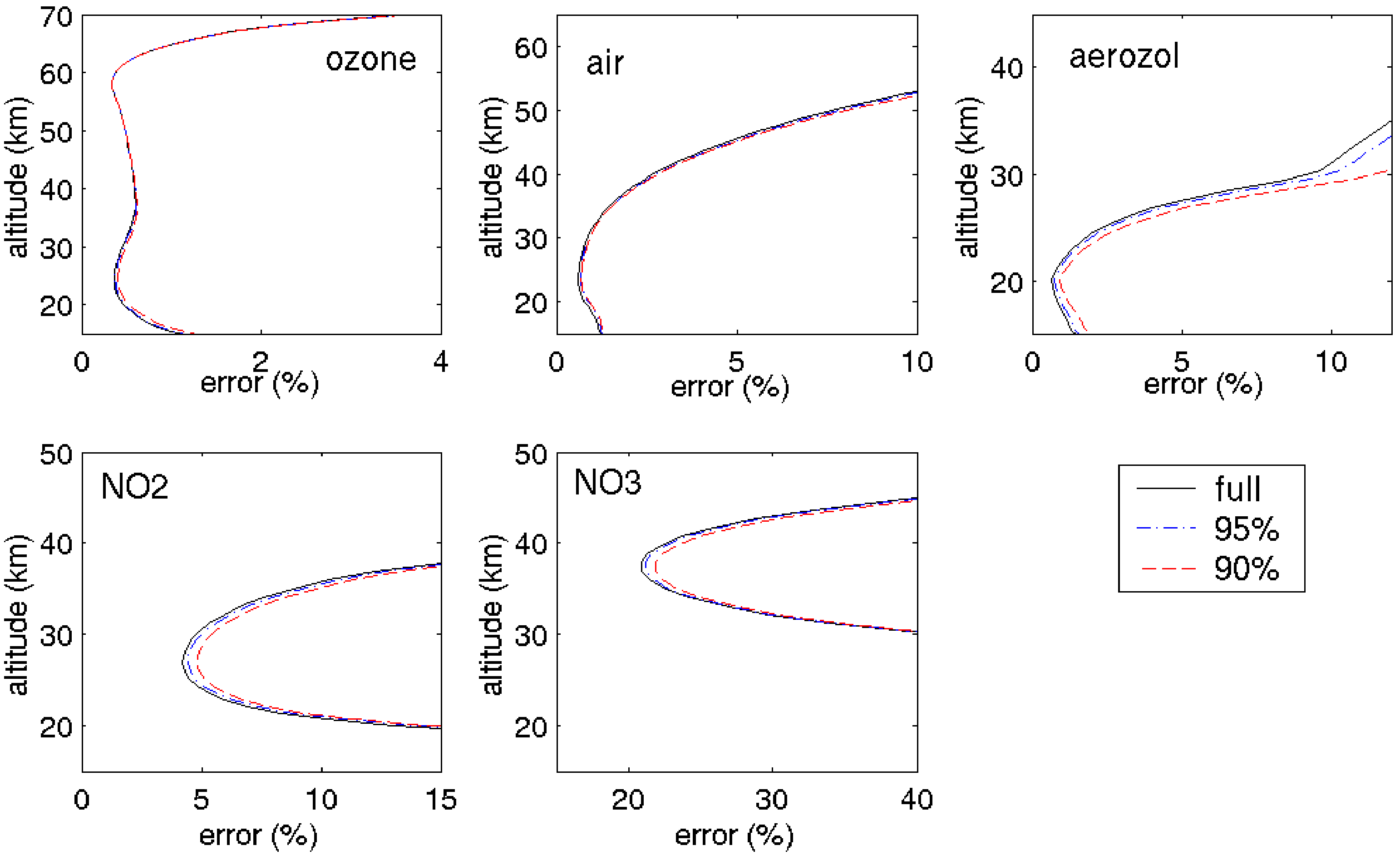
% of information	SS	SD	ICC
95 km	83.720 %	83.723 %	78.421 %
30 km	89.260 %	89.259 %	83.361 %
15 km	88.740 %	88.742 %	88.126 %
Computer time	15 sec	70 sec	< 1 sec

50 % most informative channels for all altitudes





Accuracy of reconstruction in COALA measurements using the full number of channels, and the channels containing 95% and 90 % of the information



Summary and discussion

- All of the selecting procedures are appropriate for selection the high informative subset of measurements(OP1)

- The sequential deselecting procedure gives the best approximation of the global optimum of the problem(OP2)

- The most informative channels for ozone, NO_2, $N0_3$, air and aerosol retrieval from GOMOS and COALA measurements cover the most of UV-visible wavelength range with spectral gap 360- 400 nm

- The measurement and retrieval can be provided with smaller number of spectral channels without any significant reduction of performance

- Information approach is applied to GOMOS baseline inversion. Difficulty: altitude dependence of information content and, as consequence, the selected channels. It can be avoided using one-step inversion (future work)