Profile Analysis

Martin Singull

Department of Mathematics
Linköping University, Sweden

LINKÖPING
UNIVERSITY

Example 1

The same example again but now twelve subjects are asked to estimate the price of the bar.

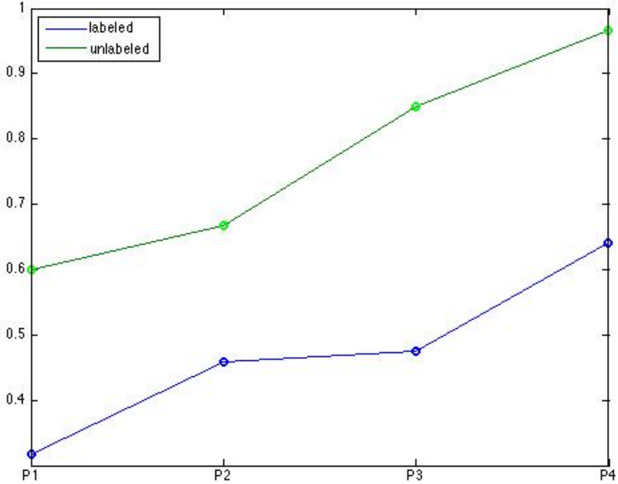
For six of the subjects, the packages

P_1: plain wrapped, unboxed,
P_2: plain wrapped, boxed,
P_3: foil wrapped, unboxed, and
P_4: foil wrapped, boxed.

have been labeled with a well-known brand name. For the remaining six subjects, no label is used.

Srivastava, M. S., & Carter, E. M. (1983). An introduction to applied multivariate statistics. North-holland.

	Subject	P_1	P_2	P_3	P_4
		Packaging			
	1	0.30	0.40	0.55	0.65
	2	0.20	0.65	0.30	0.80
labeled	3	0.30	0.50	0.50	0.70
	4	0.25	0.35	0.45	0.65
	5	0.35	0.35	0.55	0.55
	6	0.50	0.50	0.50	0.50
	mean	0.317	0.458	0.475	0.642
	1	0.40	0.40	0.60	0.60
	2	0.45	0.50	0.55	0.85
unlabeled	3	0.90	0.95	1.10	1.10
	4	0.60	0.70	0.85	0.95
	5	0.55	0.75	1.00	1.20
	6	0.70	0.70	1.00	1.10
	mean	0.600	0.667	0.850	0.967



labeled
unlabeled
1
0.9
0.8
0.7
0.6
0.5
0.4
P1
P2
P3
P4

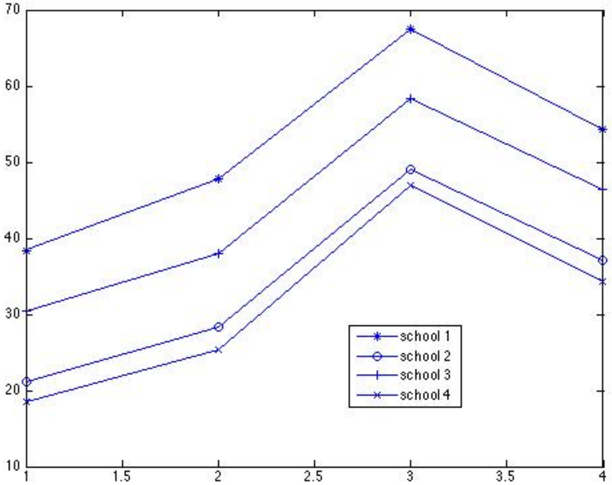
Example 2, $p = 4$ and $k = 4$ (Srivastava, 1987)

We wish to compare the performance of students from four different schools in four different subjects such as Mathematics (S_1), Science (S_2), English (S_3) and History (S_4).

Assume that we have n_i students from school $i = 1, 2, 3, 4$.

Students were required to solve problems in each subject. All the problem were planned to be of the same difficulty and the time to solve each problem was recorded. From the data (fictitious) we obtain

$$\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 38.41 & 47.81 & 67.49 & 54.30 \end{pmatrix}', n_1 = 10$$
$$\bar{\boldsymbol{x}}_2 = \begin{pmatrix} 21.06 & 28.26 & 49.10 & 37.05 \end{pmatrix}', n_2 = 15$$
$$\bar{\boldsymbol{x}}_3 = \begin{pmatrix} 30.50 & 38.05 & 58.33 & 46.41 \end{pmatrix}', n_3 = 14$$
$$\bar{\boldsymbol{x}}_4 = \begin{pmatrix} 18.53 & 25.27 & 46.99 & 34.35 \end{pmatrix}', n_4 = 12$$


70
60
50
40
30
20
10
1
1.5
2
2.5
3
3.5
4
school 1
school 2
school 3
school 4

Profile analysis of several groups

Considered the following three hypotheses:

1. $H_1 : \boldsymbol{\mu}_i - \boldsymbol{\mu}_k = \gamma_i \mathbf{1}_p,\ i = 1, \ldots, k-1$ vs. $A_1 \neq H_1$
   (parallelism – no interaction)
2. $H_2|H_1 : \gamma_i = 0,\ i = 1, \ldots k-1,$ vs. $A_2 \neq H_2|H_1$
   (same level)
3. $H_3|H_1 : \boldsymbol{\mu}_\bullet = \gamma_k \mathbf{1}_p$ vs. $A_3 \neq H_3|H_1$
   (flatness – no row effect)

Here $\boldsymbol{\mu}_\bullet = N^{-1} \sum_{i=1}^{k} n_i \boldsymbol{\mu}_i$ and the scalars γ_i are unknown.

Srivastava, M. S. (1987). Profile analysis of several groups. *Communications in Statistics - Theory and Methods*, 16(3):909–926.

Model

Let $\boldsymbol{x}_{ij}$ be p-dimensional random vectors independent distributed as $\boldsymbol{x}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})'$, $\boldsymbol{\Sigma} > 0$, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$ and $N = n_1 + \cdots + n_k$.

This model can be written as (observe that it is transposed to the usual observation matrix)

$$\boldsymbol{X} \sim N_{N,p}\left(\boldsymbol{AM}, \boldsymbol{I}_N, \boldsymbol{\Sigma}\right),$$

where

$$\begin{aligned} \boldsymbol{X} &= \left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k\right)', \\ \boldsymbol{X}_i &= \left(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}\right), \\ \boldsymbol{M} &= \left(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\right)' \end{aligned}$$

and

$$\boldsymbol{A} = \mathrm{diag}\left(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_k}\right).$$

The likelihood function

The likelihood function is now given by

$$c|\boldsymbol{\Sigma}|^{-\frac{N}{2}} etr\left\{ -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{V} + (\boldsymbol{Y} - \boldsymbol{\eta})\boldsymbol{\Xi}^{-1}()' + N(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_\bullet)()'\right]\right\},$$

where c is a constant,

$$\begin{aligned}
\boldsymbol{V} &= \boldsymbol{X}'\left(\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\right)\boldsymbol{X} : p \times p \\
&\qquad\qquad (\boldsymbol{V} \text{ is the within sum of squares}), \\
\boldsymbol{Y} &= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_k, \ldots, \bar{\boldsymbol{x}}_{k-1} - \bar{\boldsymbol{x}}_k) : p \times (k-1), \\
\boldsymbol{\eta} &= \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_k, \ldots, \boldsymbol{\mu}_{k-1} - \boldsymbol{\mu}_k\right) : p \times (k-1), \\
\bar{\boldsymbol{x}}_i &= \frac{1}{n_i}\boldsymbol{X}_i\boldsymbol{1}_{n_i} : p \times 1, \\
\bar{\boldsymbol{x}} &= \frac{1}{N}\boldsymbol{X}'\boldsymbol{1}_N : p \times 1 \qquad \text{and}
\end{aligned}$$

the matrix

$$\boldsymbol{\Xi} = \mathrm{diag}\left(\frac{1}{n_1}, \dots, \frac{1}{n_{k-1}}\right) + \frac{1}{n_k}\mathbf{1}_{k-1}\mathbf{1}'_{k-1}$$

with

$$\boldsymbol{\Xi}^{-1} = \mathrm{diag}\left(n_1, \dots, n_{k-1}\right) - \frac{1}{n}\boldsymbol{n}_{k-1}\boldsymbol{n}'_{k-1},$$

where $\boldsymbol{n}_{k-1} = (n_1, \dots, n_{k-1})'$. This matrix can be used for the between sum of squares

$$\boldsymbol{H} = \boldsymbol{Y}\boldsymbol{\Xi}^{-1}\boldsymbol{Y}' = \boldsymbol{Z}\boldsymbol{Z}',$$

where $\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{\Xi}^{-1/2}$.

MLEs under A_1 and H_1

The MLEs under A_1, i.e., no mean structure, are given by

$$\widehat{\boldsymbol{\mu}}_\bullet = \bar{\boldsymbol{x}}, \quad \widehat{\boldsymbol{\eta}} = \boldsymbol{Y} \quad \text{and} \quad N\widehat{\boldsymbol{\Sigma}} = \boldsymbol{V}.$$

The first hypothesis is given by

$$H_1 : \boldsymbol{\mu}_i - \boldsymbol{\mu}_k = \gamma_i \mathbf{1}_p,\ i = 1, \ldots, k-1 \quad \Leftrightarrow \quad H_1 : \boldsymbol{\eta} = \mathbf{1}_p \boldsymbol{\gamma}',$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{k-1})'$. The MLEs under H_1 are

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_\bullet = \bar{\boldsymbol{x}}, \qquad \widehat{\boldsymbol{\gamma}}' &= \left(\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\boldsymbol{V}^{-1}\boldsymbol{Y} \quad \text{and} \\ N\widehat{\boldsymbol{\Sigma}} &= \boldsymbol{V} + \left(\boldsymbol{Y} - \mathbf{1}\widehat{\boldsymbol{\gamma}}'\right)\boldsymbol{\Xi}^{-1}\left(\right)' = \ldots = \\ &= \boldsymbol{V} + (\boldsymbol{I} - (\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}'\boldsymbol{V}^{-1})\boldsymbol{H}()'.\end{aligned}$$

LRT

The LRT, for the parallel hypothesis $H_1 : \boldsymbol{\eta} = \mathbf{1}_p \gamma'$ is given by

$$\Lambda_{H_1} = \frac{|N\widehat{\boldsymbol{\Sigma}}_{A_1}|}{|N\widehat{\boldsymbol{\Sigma}}_{H_1}|} = \ldots =$$
$$= \left| \boldsymbol{I} + \boldsymbol{Z}' \left(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\mathbf{1}\left(\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\boldsymbol{V}^{-1} \right) \boldsymbol{Z} \right|^{-1}$$

and we reject H_1 for small values of Λ_{H_1}.

Lemma

Let $\boldsymbol{C}$ be a $(p-1)\times p$ matrix of rank $p-1$ such that $\boldsymbol{C}\mathbf{1}=\mathbf{0}$. Let $\boldsymbol{V}$ be a $p\times p$ positive definite matrix. Then

$$\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}')^{-1}=\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\mathbf{1}(\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{V}^{-1}.$$

Using the lemma, the LRT can be rewritten as

$$\Lambda_{H_1}=|\boldsymbol{I}_{p-1}+(\boldsymbol{C}\boldsymbol{H}\boldsymbol{C}')(\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}')^{-1}|^{-1}=\frac{|\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}'|}{|\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}'+\boldsymbol{C}\boldsymbol{H}\boldsymbol{C}'|}$$

Canonical reduction

One can use a canonical reduction to find the distribution of the LRT. Let $\boldsymbol{Q} : p \times p$ be an orthogonal matrix such that

$$\boldsymbol{Q} = \begin{pmatrix} p^{-1/2}\mathbf{1}_p & \boldsymbol{Q}_1 \end{pmatrix},$$

Consider the transformation

$$\boldsymbol{Z}^* = \boldsymbol{Q}'\boldsymbol{Z} = \begin{array}{c} 1 \\ p-1 \end{array} \begin{pmatrix} \boldsymbol{z}_1^{*\prime} \\ \boldsymbol{Z}_2^* \end{pmatrix}$$

and

$$\boldsymbol{V}^* = \boldsymbol{Q}'\boldsymbol{V}\boldsymbol{Q} = \begin{array}{c} 1 \\ p-1 \end{array} \begin{pmatrix} v_{11}^* & \boldsymbol{v}_{12}^{*\prime} \\ \boldsymbol{v}_{12}^* & \boldsymbol{V}_{22}^* \end{pmatrix}.$$

Parallelism: $H_1 : \boldsymbol{\eta} = \mathbf{1}_p \gamma'$

Theorem

The LRT Λ_{H_1} can be written as

$$\Lambda_{H_1} = \frac{|\boldsymbol{V}_{22}^*|}{|\boldsymbol{V}_{22}^* + \boldsymbol{Z}_2^* \boldsymbol{Z}_2^{*\prime}|},$$

Under H_1, $\boldsymbol{Z}_2^*$ and $\boldsymbol{V}_{22}^*$ are independently distributed as

$$\boldsymbol{Z}_2^* \sim N_{p-1,k-1}\left(\mathbf{0}, \boldsymbol{\Sigma}_{22}^*, \boldsymbol{I}_{k-1}\right)$$

and

$$\boldsymbol{V}_{22}^* \sim W_{p-1}\left(\boldsymbol{\Sigma}_{22}^*, N-k\right).$$

The null distribution

Theorem

The distribution of Λ_{H_1} is the same as the distribution of the product of $p-1$ independent beta random variables with parameters $\frac{1}{2}(N-k+1-i)$ and $\frac{1}{2}(k-1)$, where $i=1,\dots,p-1$.

For large N, the asymptotic null distribution of Λ_{H_1} is given by

$$-\left(N-\frac{1}{2}(k+p+1)\right)\ln\Lambda_{H_1}\sim\chi^2_{(p-1)(k-1)}.$$

Level hypothesis: $H_2|H_1 : \gamma = \mathbf{0}$

The estimator for the covariance matrix under the level hypothesis: $H_2|H_1 : \gamma = \mathbf{0}$ is given by

$$N\widehat{\boldsymbol{\Sigma}}_{H_2|H_1} = \boldsymbol{V} + \boldsymbol{Y}\boldsymbol{\Xi}^{-1}\boldsymbol{Y}' = \boldsymbol{V} + \boldsymbol{H}.$$

Hence, the LRT is given by

$$\Lambda_{H_2|H_1} = \frac{|N\widehat{\boldsymbol{\Sigma}}_{H_1}|}{|N\widehat{\boldsymbol{\Sigma}}_{H_2|H_1}|} = \frac{|\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}' + \boldsymbol{C}\boldsymbol{H}\boldsymbol{C}'|}{|\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}'|}\frac{|\boldsymbol{V}|}{|\boldsymbol{V} + \boldsymbol{H}|}$$

Using the canonical reduction, the LRT for the second hypothesis $H_2|H_1 : \gamma = \mathbf{0}$ is

$$\Lambda_{H_2|H_1} = \frac{|\boldsymbol{V}^*|\left|\boldsymbol{I} + \boldsymbol{Z}^{*\prime}\left(\boldsymbol{V}^{*-1} - \boldsymbol{V}^{*-1}\boldsymbol{e}\left(\boldsymbol{e}'\boldsymbol{V}^{*-1}\boldsymbol{e}\right)^{-1}\boldsymbol{e}'\boldsymbol{V}^{*-1}\right)\boldsymbol{Z}^*\right|}{|\boldsymbol{V}^* + \boldsymbol{Z}^*\boldsymbol{Z}^{*\prime}|}$$

$$= \cdots = \frac{v_{1.2}^*}{v_{1.2}^* + \boldsymbol{y}_2^{*\prime}\boldsymbol{y}_2^*},$$

where $\boldsymbol{e} = \begin{pmatrix}1 & 0 & \dots & 0\end{pmatrix}' : p \times 1$,

$$v_{1.2}^* = v_{11}^* - \boldsymbol{v}_{12}^{*\prime}\boldsymbol{V}_{22}^{*-1}\boldsymbol{v}_{12}^*,$$
$$\boldsymbol{y}_2^{*\prime} = \left(\boldsymbol{I} - \boldsymbol{Z}_2^{*\prime}\boldsymbol{V}_{22}^{*-1}\boldsymbol{Z}_2^*\right)^{-1/2}\left(\boldsymbol{z}_1^* - \boldsymbol{Z}_2^*\boldsymbol{V}_{22}^{*-1}\boldsymbol{v}_{12}^*\right).$$

$\boldsymbol{y}_2^*$ and $v_{1.2}^*$ are independently distributed as

$$\boldsymbol{y}_2^* \sim N_{k-1}\left(\mathbf{0}, \sigma_{1.2}^*\boldsymbol{I}_{k-1}\right) \quad \text{and} \quad \frac{v_{1.2}^*}{\sigma_{1.2}^*} \sim \chi^2\left(N - k - p + 1\right).$$

The null distribution

Theorem

Rejecting the hypothesis $H_2|H_1$ for small values of $\Lambda_{H_2|H_1}$ is equal to reject the hypothesis for large values of

$$F = \frac{1-\Lambda_{H_2|H_1}}{\Lambda_{H_2|H_1}} = \frac{\boldsymbol{y}_2^{*\prime}\boldsymbol{y}_2^*}{v_{1.2}^*},$$

where

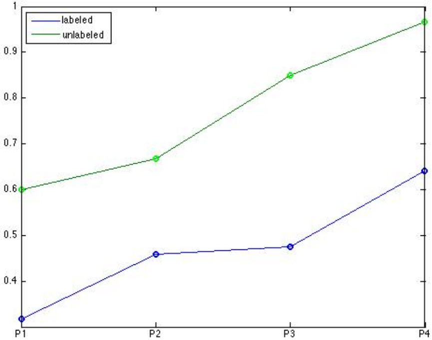
$$\begin{aligned} v_{1.2}^* &= v_{11}^* - \boldsymbol{v}_{12}^{*\prime}\boldsymbol{V}_{22}^{*-1}\boldsymbol{v}_{12}^*, \\ \boldsymbol{y}_2^{*\prime} &= \left(\boldsymbol{z}_1^{*\prime} - \boldsymbol{v}_{12}^{*\prime}\boldsymbol{V}_{22}^{*-1}\boldsymbol{Z}_2^{*\prime}\right)\left(\boldsymbol{I} - \boldsymbol{Z}_2^{*\prime}\boldsymbol{V}_{22}^{*-1}\boldsymbol{Z}_2^*\right)^{-1/2}. \end{aligned}$$

The null distribution of F is given by

$$\frac{N-k-p+1}{k-1}\,F \sim F_{k-1,N-k-p+1}.$$

Example 1, cont.

P_1: plain wrapped, unboxed,
P_2: plain wrapped, boxed,
P_3: foil wrapped, unboxed, and
P_4: foil wrapped, boxed.



$$H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \gamma \mathbf{1}_4, \quad \text{vs.} \quad A_1 \neq H_1$$

$$-\left(N - \frac{1}{2}(k + p + 1)\right) \ln \Lambda_{H_1} = 3.6169$$

with $N = 12, k = 2, p = 4$ and $c = \chi^2_{(p-1)(k-1),0.95} = 7.8147$

Since $k = 2$ we could use an exact F-test instead (as before).

Hence, we can't reject H_1, i.e., the profiles are similar.

$$H_2|H_1 : \gamma = 0, \quad \text{vs.} \quad A_2 \neq H_2|H_1$$

$\Lambda_{H_2|H_1} = 0.4368$ and $\frac{N-k-p+1}{k-1}\, F = 9.0243$

with $c = F_{k-1,N-k-p+1,0.95} = 5.5914$.

Hence, reject $H_2|H_1$, i.e., the profiles are not on the same level.

$H_3|H_1 : \boldsymbol{\mu}_\bullet = \gamma_k \mathbf{1}_p$

We wish to test the hypothesis

$$H_3|H_1 : \boldsymbol{\mu}_\bullet = \gamma_k \mathbf{1}_p \quad \text{vs.} \quad A_3 \neq H_3|H_1 \quad \text{(flatness – no row effect)},$$

where $\boldsymbol{\mu}_\bullet = N^{-1}\sum_{i=1}^{k} n_i \boldsymbol{\mu}_i$ and the scalars γ_i are unknown.

The MLE of $\boldsymbol{\Sigma}$ under H_1 is given above (page 10), and the MLE of $\boldsymbol{\Sigma}$ under H_3 is given by

$$N\widehat{\boldsymbol{\Sigma}}_{H_3|H_1} = \boldsymbol{V} + \left(\boldsymbol{Y} - \mathbf{1}\widehat{\gamma}'\right)\boldsymbol{\Xi}^{-1}\left(\right)' + N(\bar{\boldsymbol{x}} - \hat{\gamma}_k\mathbf{1})()',$$

where $\hat{\gamma}_k = \dfrac{\bar{\boldsymbol{x}}'\boldsymbol{V}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1}}$.

LRT

Hence, the LRT rejects the hypothesis $H_3|H_1$ for small values of

$$\Lambda_{H_3|H_1} = \frac{|\boldsymbol{V} + \left(\boldsymbol{Y} - \mathbf{1}\widehat{\gamma}'\right)\boldsymbol{\Xi}^{-1}()'|}{|\boldsymbol{V} + \left(\boldsymbol{Y} - \mathbf{1}\widehat{\gamma}'\right)\boldsymbol{\Xi}^{-1}()' + N(\bar{\boldsymbol{x}} - \hat{\gamma}_k\mathbf{1})()'|} = \ldots =$$
$$= \frac{1}{1 + N\bar{\boldsymbol{x}}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}' + \boldsymbol{C}\boldsymbol{H}\boldsymbol{C}')^{-1}\boldsymbol{C}\bar{\boldsymbol{x}}},$$

for some matrix $\boldsymbol{C}$ such that $\boldsymbol{C}\mathbf{1} = \mathbf{0}$.

Hence, the hypothesis $H_3|H_1$ is rejected if

$$N\bar{\boldsymbol{x}}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}' + \boldsymbol{C}\boldsymbol{H}\boldsymbol{C}')^{-1}\boldsymbol{C}\bar{\boldsymbol{x}} \geq \frac{p-1}{N-p+1}F_{1-\alpha}(p-1, N-p+1).$$

Linköping University - Research that makes a difference

LINKÖPING
UNIVERSITY