Some Project Issues

Antanas Verikas
antanas.verikas@hh.se

IDE, Halmstad University

2013

The task

- Construct a regression model

$$\widehat{y}(\mathbf{x}) = f(\mathbf{x}, \mathbf{w}) \tag{1}$$

- Construct a classifier

$$\widehat{y}(\mathbf{x}) = \widehat{p}(c_i|\mathbf{x}) \tag{2}$$

Given

- A training set

$$\mathbf{Z}_{\mathtt{Train}} = \{\mathbf{x}(n), \mathbf{y}(n)\}_{n=1,\ldots,N_{\mathtt{Train}}} \tag{3}$$

- An outline for the project and report
- MATLAB
- Answers to some questions

Not given

- A test set

$$\mathbf{Z}_{\mathtt{Test}} = \{\mathbf{x}(n), \mathbf{y}(n)\}_{n=1,\ldots,N_{\mathtt{Test}}} \tag{4}$$

- The answers to the problems.
- Infinite time.

Meat fat content

Estimate the fat content of a meat sample on the basis of its near infrared (NIR) absorbance spectrum.

- Regression problem, $\widehat{y}(\mathbf{x}) = f(\mathbf{x}, \mathbf{w})$
- $\mathbf{x}(n) =$ a 100 channel absorbance spectrum ($D = 100$).
- $\mathbf{y}(n) =$ measured contents of moisture, fat and protein.
- $N_{\mathtt{Train}} = 172, N_{\mathtt{Test}} = 43.$

Aspects: variable transforms (PCA),
variable selection,
linear and nonlinear models,
model selection,
estimating generalization performance...

Thyroid diagnosis

Tell if a person is normal, hypothyroid, or hyperthyroid, on the basis of medical test results.

- Classification problem, $\widehat{y}(\mathbf{x}) = \widehat{p}(c_i|\mathbf{x})$
- $\mathbf{x}(n) = 21$ variables (results from different tests), some are continuous, others are discrete ($D = 21$).
- $\mathbf{y}(n) =$ class name (normal, hypothyroid, hyperthyroid).
- $N_{\mathtt{Train}} = 5000$, $N_{\mathtt{Test}} = 2200$.

Aspects: variable transforms,
variable selection,
linear and nonlinear models,
model selection,
committees of models...

Data exploration

- Scatter plots.
- Correlation coefficients.
- Histograms.
- Fisher index.
- Outlier detection.
- Data transformations:
  - Standardization,
  - PCA,
  - Box-Cox transforms.

The 1st and the 10th components

Reveals if there are any simple relationships (But misses interactions)

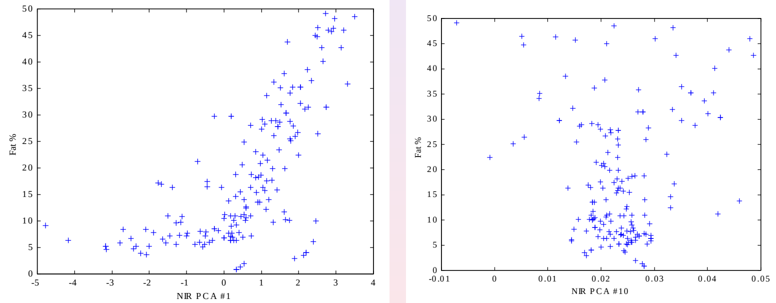


Figure: The 1st NIR principal component vs. the fat % in the meat (*left*). The 10th NIR principal component shows no similar relationship.

PCA #1-9 vs. fat %

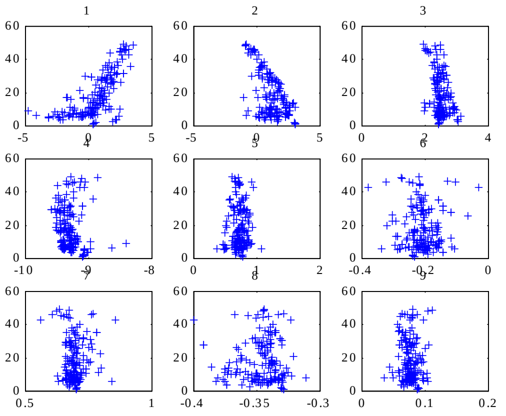


Figure: Scatter plots for PCA #1-9 vs. fat %

The 1st and the 10th components

Pearson's: $r_{xy} = \sigma_{xy} / \sqrt{\sigma_{xx}\sigma_{yy}} \in [-1, 1]$. Spearman's: (s_{xy}) uses ranks instead of values.

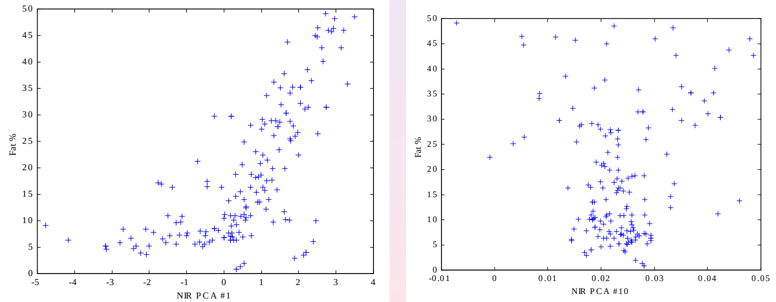


Figure: **Left**: $r_{xy} = 0.68$, $s_{xy} = 0.70$. **Right:** $r_{xy} = 0.04$, $s_{xy} = -0.01$.

PCA #1-9 vs. fat %

$$r_{xy} = \begin{pmatrix} 0.6798 & -0.5763 & -0.3944 \\ -0.1439 & 0.0586 & -0.0233 \\ 0.0062 & 0.0289 & -0.0187 \end{pmatrix}$$

$$s_{xy} = \begin{pmatrix} 0.6996 & -0.4689 & -0.3276 \\ -0.2620 & 0.0806 & -0.0363 \\ 0.0172 & 0.0079 & -0.0750 \end{pmatrix}$$

If $|r_{xy}|, |s_{xy}| > 1.96/\sqrt{N}$, the correlation is significant at the 95% level.

$N = 172 \Rightarrow 1.96/\sqrt{N} = 0.1494$

Histograms (1)

Useful in classification problems to detect class specific distribution.

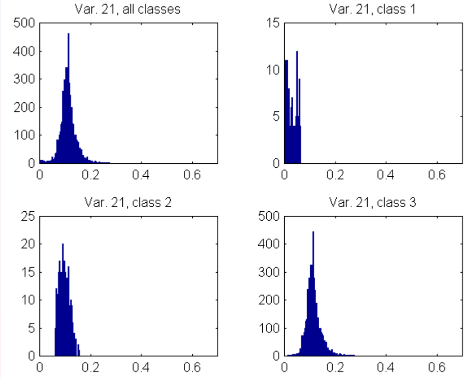


Figure: Class-specific distributions.

Histograms (2)

Detect non-normal distributions and test transformations.

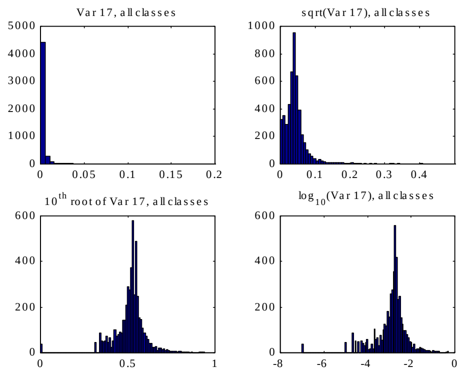


Figure: Affect of different transformations.

Standardization

Useful in classification problems to detect class specific distribution.

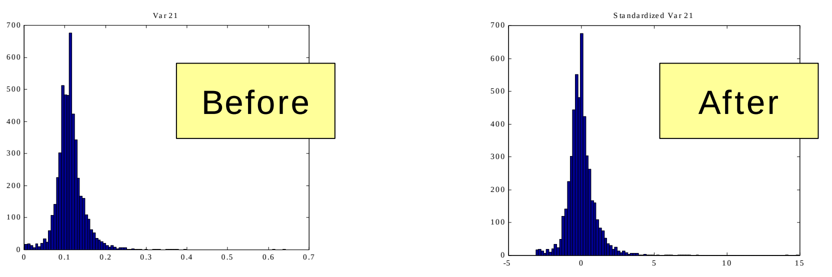


Figure: Affect of standardization.

Whitening

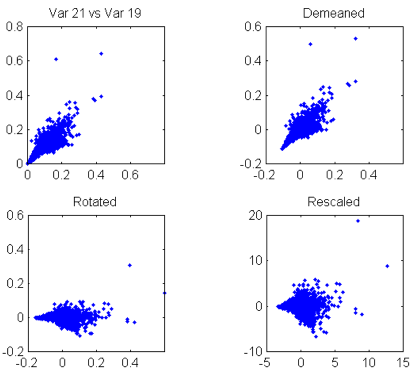
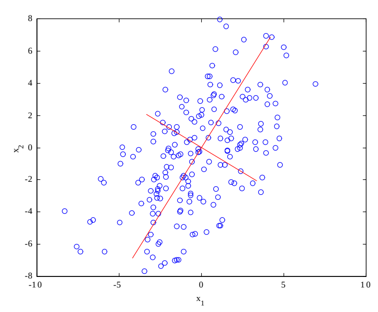


Figure: Top-left: $\mathbf{X}$, Top-right: $\mathbf{X} - \mathbf{1}\mu^T$, Bottom-left: $(\mathbf{X} - \mathbf{1}\mu^T)\mathbf{Q}$, Bottom-right: $(\mathbf{X} - \mathbf{1}\mu^T)\mathbf{Q}\mathbf{\Lambda}^{-1/2}$.

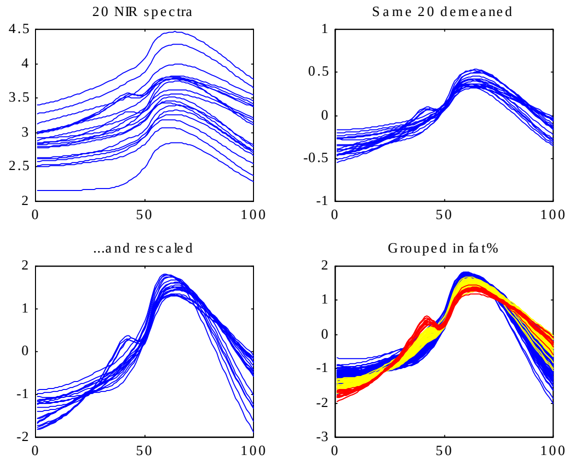
Principal components

Express the data in the new basis $\mathbf{Q}$, with eigenvectors of the data covariance matrix $\mathbf{\Sigma}$ as basis vectors.

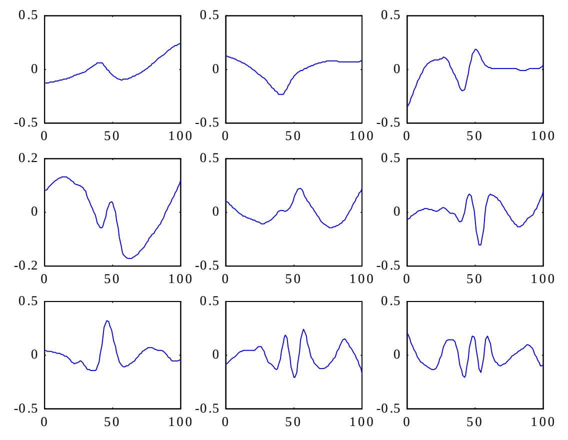
$$\mathbf{\Sigma}\mathbf{q}_i = \lambda_i\mathbf{q}_i \tag{5}$$



PCA example: NIR spectra

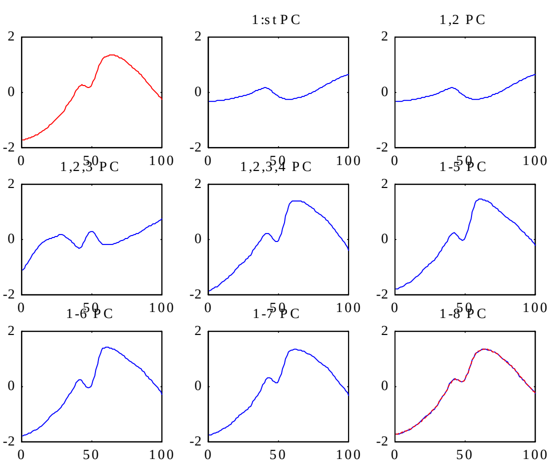


NIR: The 9 leading eigenvectors



λ_i
2.4308
0.9372
0.0489
0.0256
0.0108
0.0023
0.0014
0.0002
0.0001

NIR reconstruction with PCA



Comparing regression models

- **Standard** $t-$**test**, if E or $\log(E)$ are (approximately) normally distributed.
- Non-parametric **Wilcoxon** test, if E or $\log(E)$ are not normally distributed.
- **Paired** $t-$**test**, if we test models A and B on the same data, and residuals $e(n)$ are normally distributed.

Comparing regression models: Example (1)

- Estimate mean and variance for the generalization error (or log error) using $K-$fold cross validation. $\log(E)$ is often more normal than E.
- Compare means using $t-$test (or Wilcoxon test):

$$|\mu_A - \mu_B| > 1.96 \frac{\sigma_{AB}}{\sqrt{K}}$$

$E_{\texttt{Test,A}}$	$E_{\texttt{Test,b}}$
0.6488	3.3016
0.1891	0.6465
1.1335	5.0245
1.3333	12.473
0.3178	1.2315
3.2901	5.8008
3.2843	8.6193
0.9631	0.4997
1.3872	0.5822
1.1908	4.3542

Comparing regression models: Example (2)

$$\mu_A = \frac{1}{10}\sum_{k=1}^{10} \log(E_{k,A}) = 0.0013$$

$$\mu_B = \frac{1}{10}\sum_{k=1}^{10} \log(E_{k,B}) = 0.9253$$

$$\sigma_A = \sqrt{\frac{1}{9}\sum_{k=1}^{10} [\log(E_{k,A}) - \mu_A]^2} = 0.9034$$

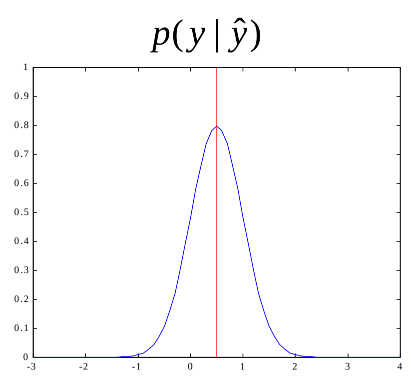
$$\sigma_B = \sqrt{\frac{1}{9}\sum_{k=1}^{10} [\log(E_{k,B}) - \mu_B]^2} = 1.1898$$

Comparing regression models: Example (3)

$$\begin{aligned}
\sigma_{AB} &= \sqrt{\sigma_A^2 + \sigma_B^2} = 1.4940 \\
|\mu_A - \mu_B| &= 0.9240 < 1.96 \frac{\sigma_{AB}}{\sqrt{10}} = 0.9260 \\
&\Rightarrow \textbf{Not significant difference!}
\end{aligned}$$

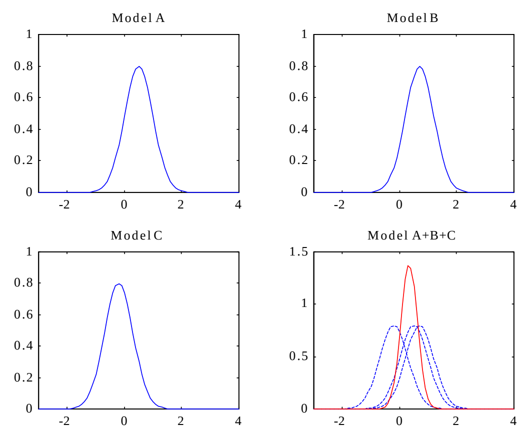
Interpreting the output: Regression

$$\begin{aligned} \widehat{y}(\mathbf{x}) &= \langle y(\mathbf{x}) \rangle_\varepsilon \\ \sigma_\varepsilon^2 &\approx \mathtt{MSE}_{\mathtt{Test}} \\ y(\mathbf{x}) &= \widehat{y}(\mathbf{x}) \pm 1.96\sqrt{\mathtt{MSE}_{\mathtt{Test}}} \end{aligned}$$



Combining models

$$\hat{y}_{\mathrm{com}} = \frac{\sum_k \hat{y}_k / \mathrm{MSE}_k}{\sum_k 1/\mathrm{MSE}_k}$$



Comparing classification models

- Use K-fold cross-validation method.
- Use McNemar's test for paired testing

$$z = \frac{|n_A - n_B| - 1}{\sqrt{n_A + n_b}}$$

$n_A =$ mistakes made by A but not B and vice versa.

If $z > 1.96$ then we have a significant difference at the 95% confidence level.

Error bars

Error bars on the classification error:

$$R \quad = \quad \widehat{R} \pm 1.96\sqrt{\frac{\widehat{R}(1-\widehat{R})}{N}}$$

$\widehat{R} =$ estimate of the classification error on out-of-sample test data.