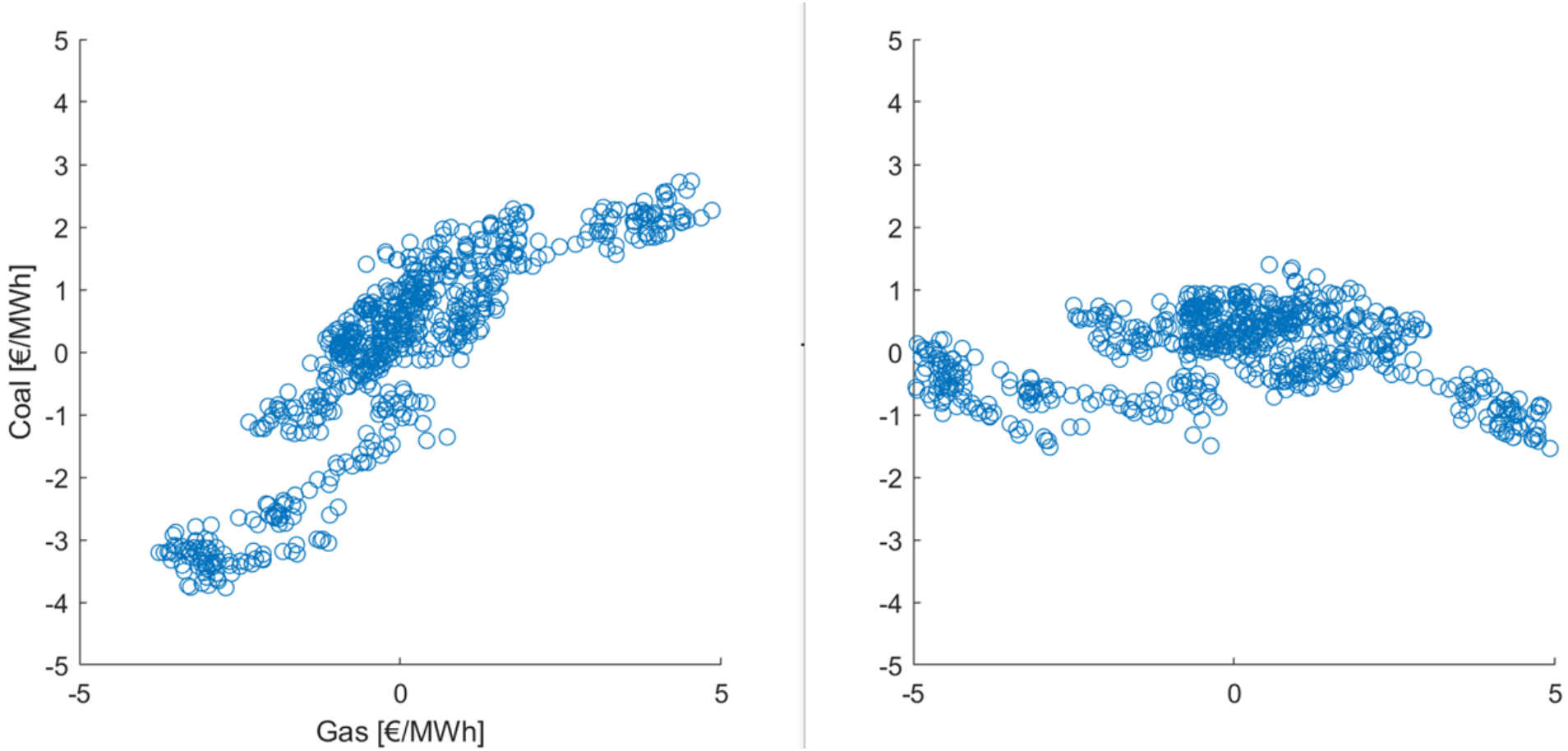
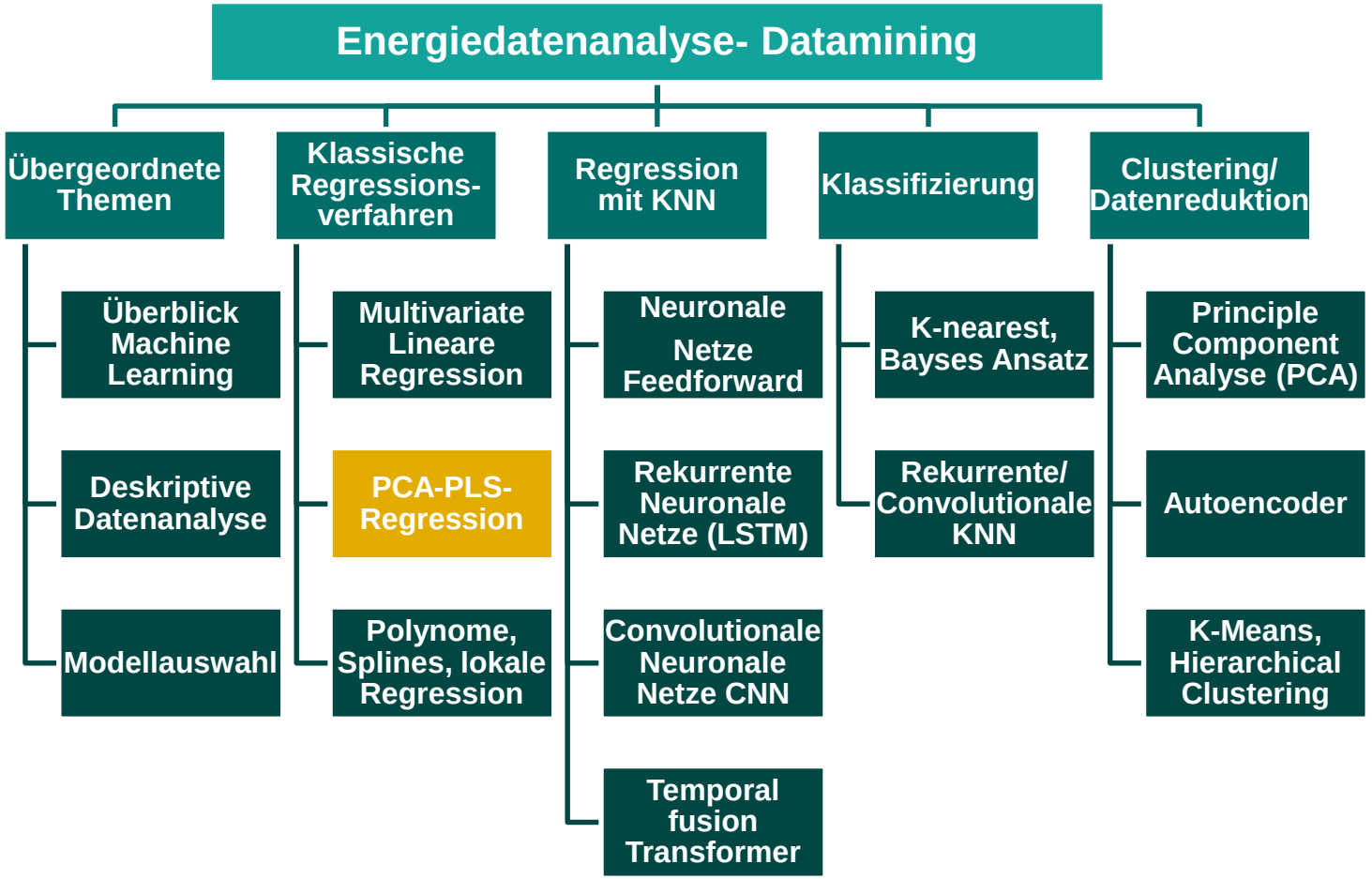



# PCA, PLS- Clustering und Regression

**Die Themengebiete der Veranstaltung verknüpfen Modelle des „Machine Learning“ mit energiewirtschaftlichen Fragestellungen**



## Zielsetzung der heutigen Vorlesung:

| | |
|---|---|
| **Thema** | Überblick über Verfahren: Principle Component Analysis, Partial Least Square Regression |

| | **Aufbau der heutigen Vorlesung:** | **Lernziel:** |
|---|---|---|
| 1 | Einführung in Funktionsweise und Methodik | Nachvollziehen der mathematischen Grundlagen |
| 2 | Energiewirtschaftliche Anwendung | Commodity-Preise und Lastanalyse |

## Agenda

1. Ridge and Lasso Regression
2. Principle Component Analysis (PCA)
3. Partial Least Square Regression (PLS)

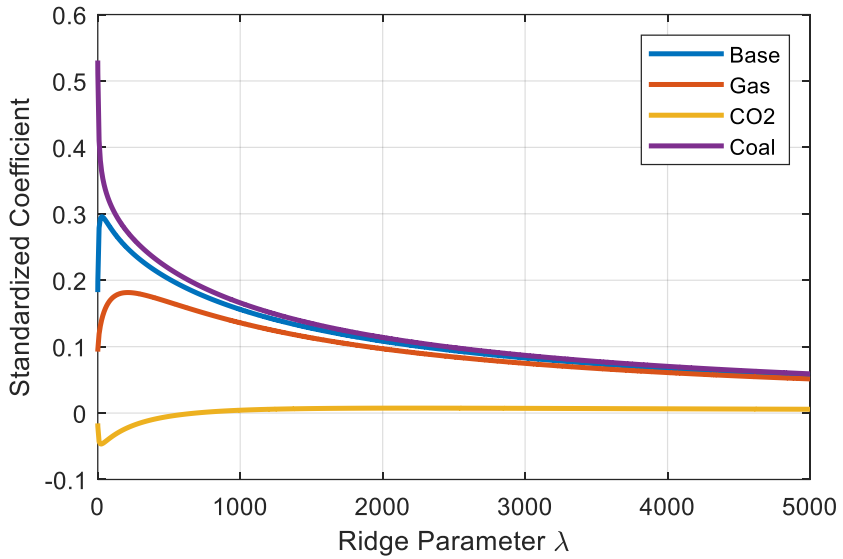
Literaturempfehlung und Quellennachweise der Abbildungen: James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics), Springer New York. Kindle-Version.

**Ausgangsproblematik: in der letzten Vorlesung haben wir gezeigt, dass auf Basis von p Variablen selbst mit effizienten Suchalgorithmen eine Vielzahl unterschiedlicher Modellhypothesen untersucht werden muss**

| | Modellhypothesen* | Modelle im CV | |
|---|---|---|---|
| **Klassische Modellauswahl** | | $p \cdot \frac{p+1}{2}$ | Grundsätzliche Schwierigkeit bei allen Ansätzen:<br><br>Es müssen alle in Frage kommenden Modellhypothesen separat geschätzt werden |
| **Best subset Selection** | $maximal$<br>$p \cdot \frac{p+1}{2}$ | $p+1$ | Bildung Modellhypothese (**Vorauswahl** relevanter Parameter) |
| **Forward Stepwise Selection** | | $maximal$<br>$p$ | Risiko:<br><br>Modelle mit potentiell niedrigen E[Test-MSE] bleiben unberücksichtigt |

**Ridge Regression (Ridge) ist eine Alternative Modellkalibrierung, die grundsätzlich alle potentiellen Regressoren in ein Modell miteinbindet**

## Auswirkung von $\lambda$ auf die Parametrierung




Ridge sollte auf standardisierte Regressoren erfolgen!

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \bar{x_j}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i,j} - \bar{x_j}\right)^2}}$$

## Funktionsweise Ridge-Regression

- Bisheriger Ansatz im Rahmen der Regression zur Optimierung des Modellfits:

$$MSE_{Train} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{i,j}\right)^2$$

- Im Rahmen von Ridge Regression wird ein Strafterm direkt in die Zielfunktion integriert:

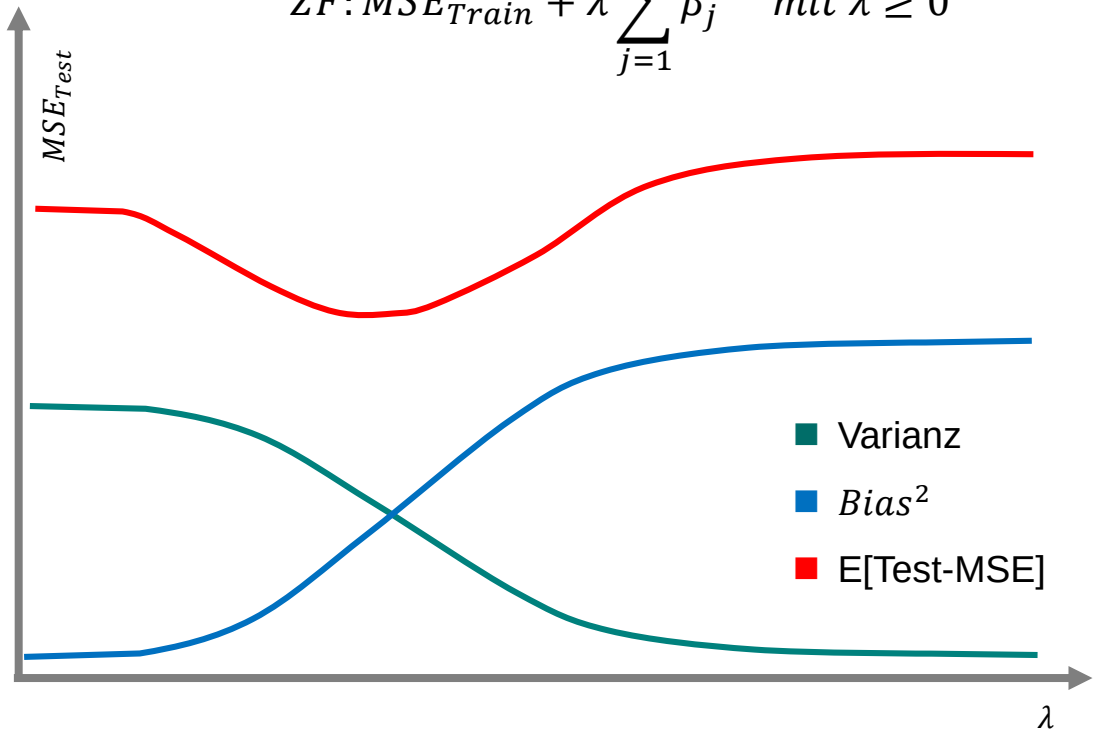
$$ZF\colon\ MSE_{Train} + \lambda\sum_{j=1}^{p}\beta_j^2 \quad mit\ \lambda \geq 0$$

- Zielfunktion besitzt Trade off zwischen Verringerung MSE und Strafterm

## Der Ridge Parameter $\lambda$ steuert die Modellkomplexität und damit das Verhältnis aus Bias und Varianz

### Auswirkung von $\lambda$ auf Varianz und Bias

$$ZF: MSE_{Train} + \lambda \sum_{j=1}^{p} \beta_j^2 \quad mit\ \lambda \geq 0$$

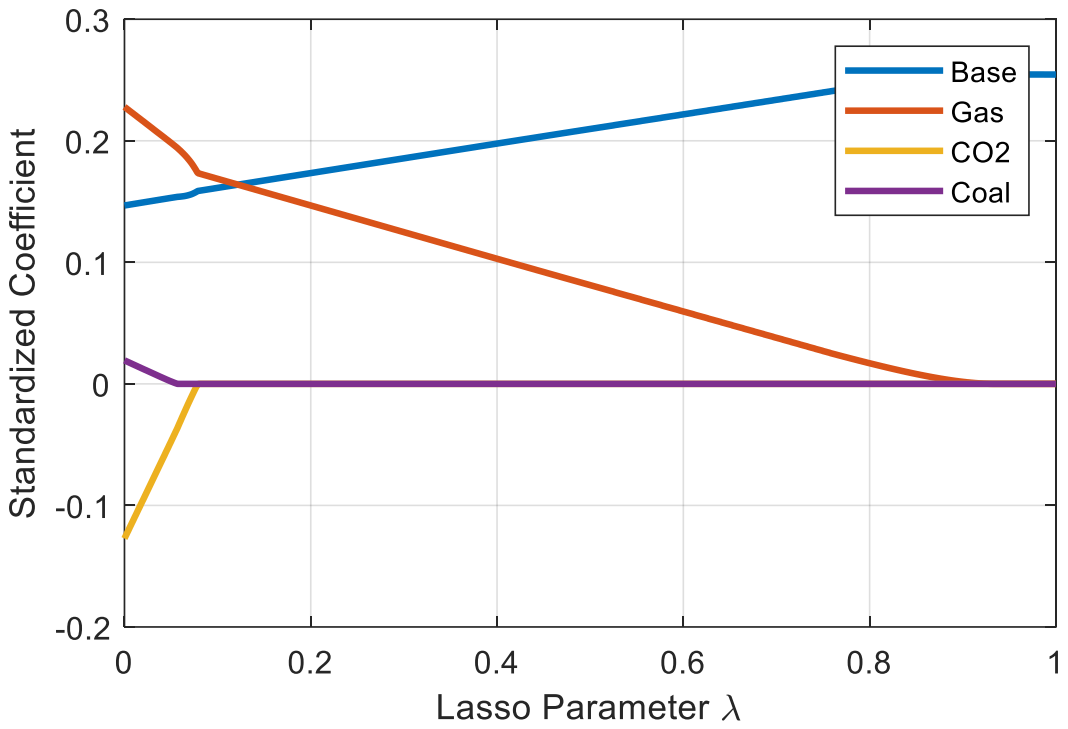



### Erläuterung

- Bei Zunahme von $\lambda$ verringern sich die Freiheitsgrade des Modells.
  - Verringerung der Varianz
  - Erhöhung des Bias
- Ridge ist in Situationen anzuwenden, wenn bei einer normalen Regression eine hohe Varianz der Parameterschätzung vorliegt
- Nachteil:
  - Trotz hohem $\lambda$ sind im verbliebenden Modell alle Regressoren vorhanden tlw. mit Koeffizienten nahe Null
  - Es kommt nicht zum Ausschluss von Variablen

# Lasso hat den Vorteil, dass die Parameter schneller gegen Null streben als bei Ridge

## Auswirkung von $\lambda$ auf die Parametrierung




## Funktionsweise Lasso-Regression

- Im Rahmen von Lasso Regression wird ein Strafterm mit Absolut-Term integriert:

$$ZF\colon MSE_{Train} + \lambda \sum_{j=1}^{p} |\beta_j| \quad mit\ \lambda \geq 0$$

- Bei hohem $\lambda$ streben einige Koeffizienten gegen Null!

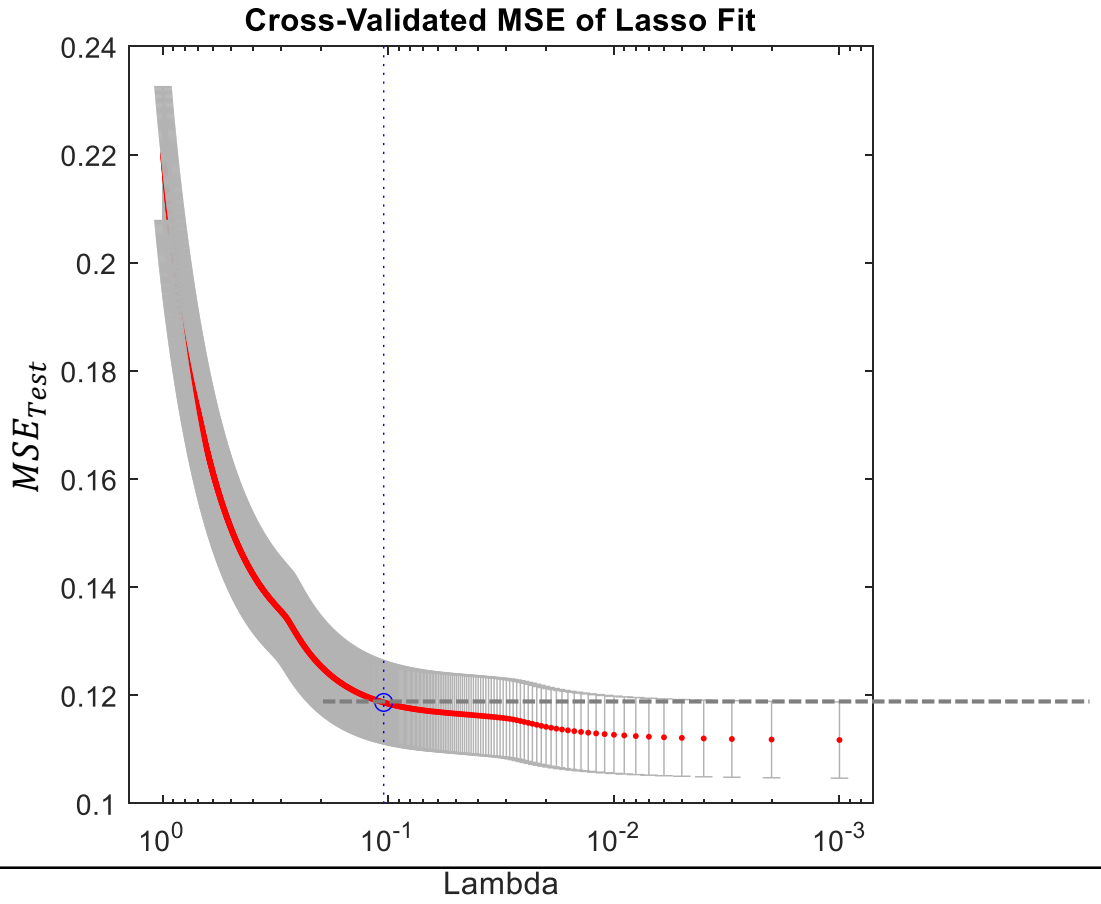
- Man kann zeigen, dass folgender Term äquivalent ist:

$$ZF\colon MSE_{Train}\ , NB\colon \sum_{j=1}^{p} |\beta_j| \leq s$$

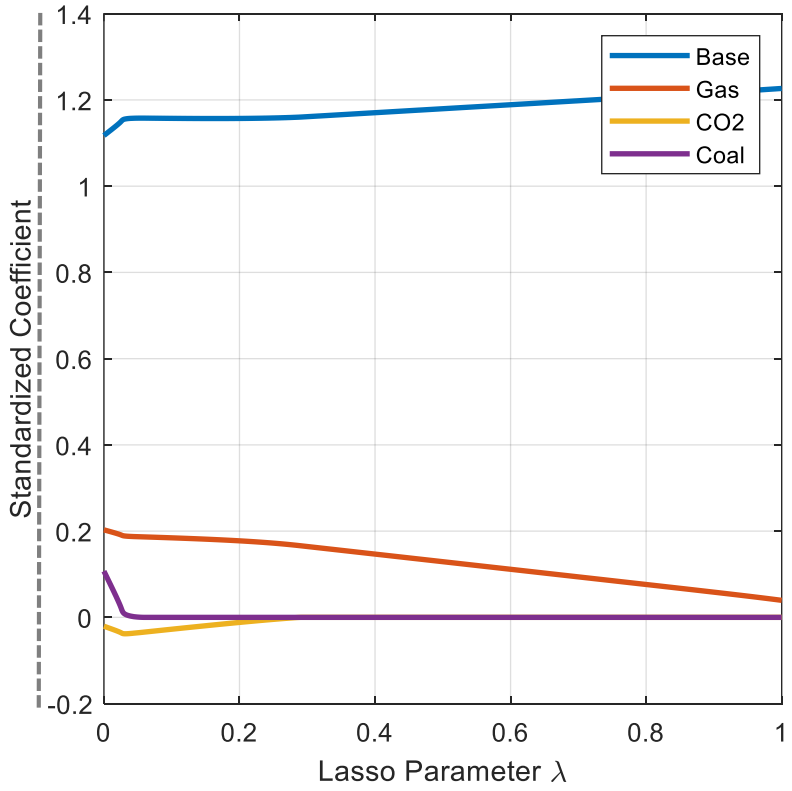
- Wie stellen wir das optimale $\lambda$ ein?

**Die Auswahl des Koeffizienten $\lambda$ (Hyperparameter) erfolgt mit Hilfe von Cross Validation Techniken!**

## Verlauf des 10-Fold CV




## Verlauf der Koeffizienten

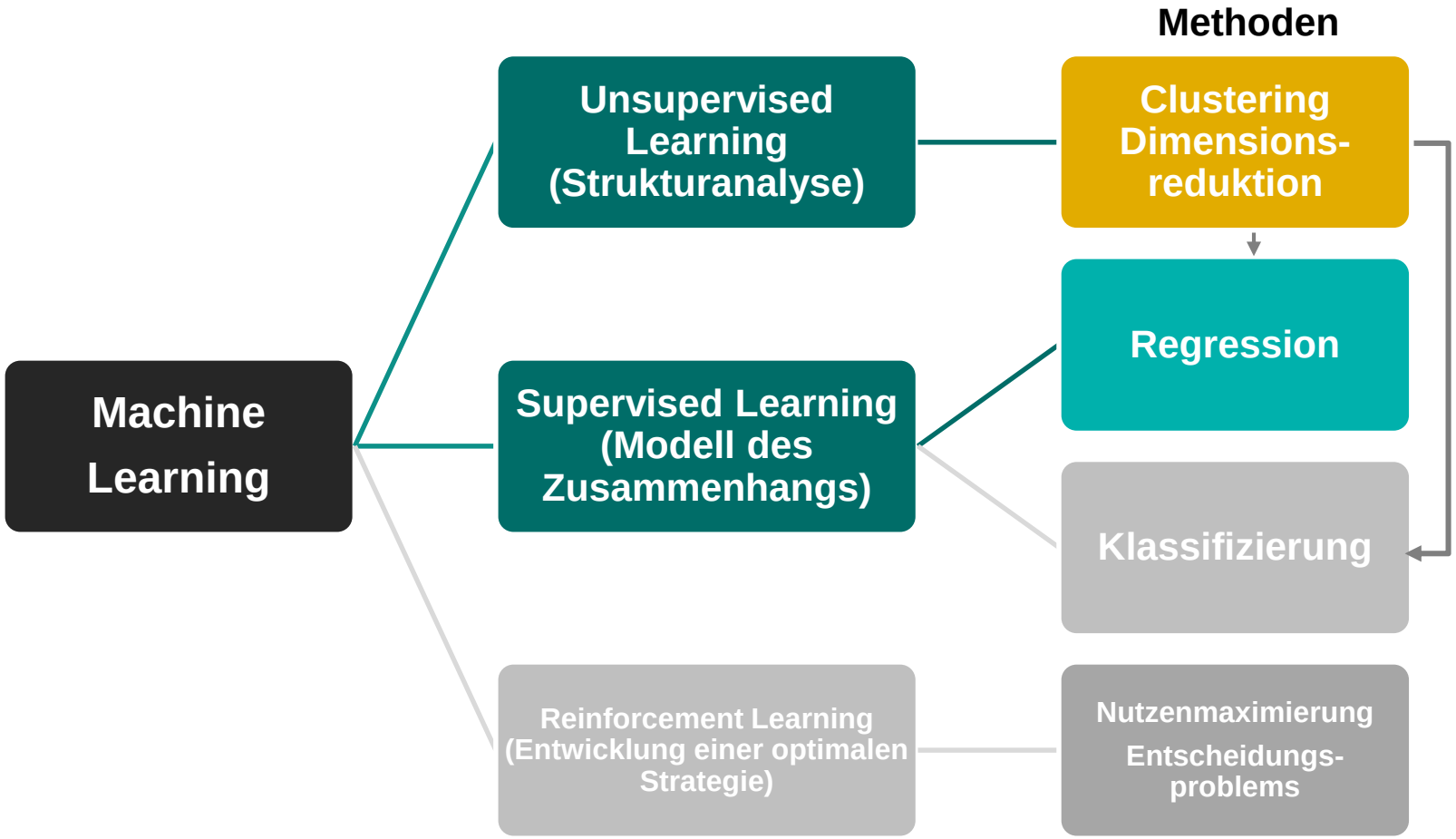


# Agenda

1. Ridge and Lasso Regression
2. Principle Component Analysis (PCA)
3. Partial Least Square Regression (PLS)

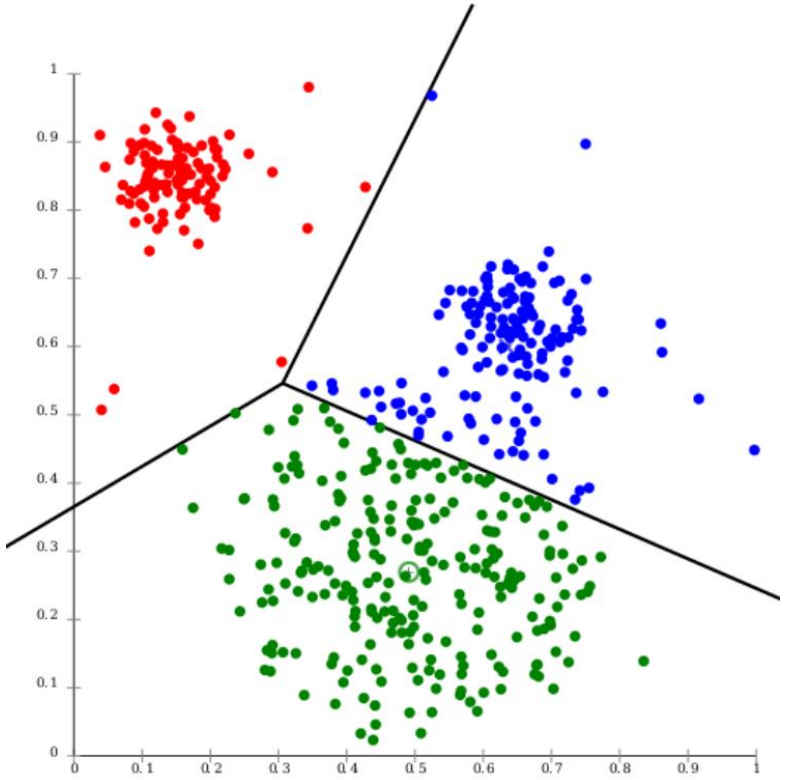
Literaturempfehlung und Quellennachweise der Abbildungen: James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics), Springer New York. Kindle-Version.

**Die betrachteten quantitative Methoden fallen unter die Rubrik Machine Learning**
**Ziel: Aufzeigen Beziehungen in Datenstrukturen mit Hilfe von Algorithmen**



## Unsupervised Learning versucht durch Gruppierung Datenstrukturen und Muster aufzuzeigen

### Clustering




- Unsupervised Learning versucht versteckte Muster und Strukturen in Daten aufzudecken.
- Es wird explizit nicht der Zusammenhang zwischen einer Input und einer Outputgröße betrachtet.
- Clustering ist dabei die gängigste Methodik.
- Principle Component Analysis ist ein Verfahren um Strukturen zu analysieren.

- Wichtige Arten von Unsupervised Learning:

  1. **Dimensionsreduktion**

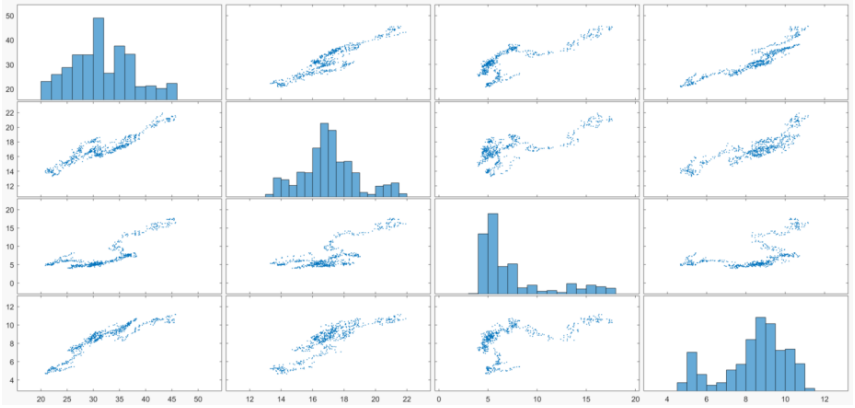
  2. Clusterverfahren

Quelle: Wikipedia

**Ausgangssituation bei Betrachtung vieler Variablen:**
**Das Vorliegen von Multikollinearität kann die Anwendung einer Regression erschweren**

## Ausgangssituation

| | Strom (Peak) | Gas | CO2 | Kohle |
|---|---|---|---|---|
| **Strom (Peak)** | 100% | 92% | 80% | 92% |
| **Gas** | 92% | 100% | 75% | 87% |
| **CO2** | 80% | 75% | 100% | 53% |
| **Kohle** | 92% | 87% | 53% | 100% |



## Herausforderung

- Multikollinearität liegt vor, wenn zwei oder mehrere Variablen eine **hohe Korrelation** aufweisen.
- Def. Multikollinearität :
  **Vektoren** sind kollinear, wenn sie **linear abhängig** sind.
- Eine perfekte Kollinearität macht die rechnerische Durchführung der linearen Regressionsanalyse unmöglich.
- Resultierende Probleme:

**Instabile Parameterschätzung:**

**Erhöhung der Standardabweichung der Parameter-schätzung $\hat{\beta}$**

**Hypothesentest führt zu keinen stabilen Ergebnissen**

## Mit Hilfe des Variance Inflation Factor (VIF) kann das Ausmaß der Kollinearität gemessen werden

### Betrachtung der Varianz eines Parameters bei isolierter Betrachtung eines Regressors zur Erklärung von y

1. Regression isoliert auf den zu betrachtenden Regressor

$$y = \beta_0 + \beta_k x_k + \epsilon$$

2. Berechnung der Varianz des Schätzers

$$Var_{Min}(\hat{\beta}_k) = \frac{\hat{\sigma}_{ee}}{\sum_{i=1}^{n}(x_{ki} - \bar{x}_k)^2}$$

3. Veränderung der Varianz von $\hat{\beta}_k$ bei Hinzunahme weiterer Regressoren zur Erklärung von y

$$y = \beta_0 + \beta_1 x_1 \ldots + \beta_k x_k + \cdots + ß_n x_n + \epsilon$$

$$Var(\hat{\beta}_k) = Var_{Min}(\hat{\beta}_k) \cdot [\frac{1}{1 - R_k^2}] \leftarrow VIF_k$$

### Herleitung von $R_k^2$ und VIF

- Berechung von $R_k^2$ : Modell zur Erklärung von $x_k$ durch die weiteren Variablen

$$x_k = \beta_0^* + \sum_{i \neq k} \beta_i x_i + \epsilon^* \rightarrow \mathrm{R_k^2}$$

- Das resultierende Bestimmtheitsmaß $\mathrm{R_k^2}$ beeinflusst die Varianz des Schätzer bei Hinzunahme weiterer Regressoren

$$VIF_k = \frac{Var(\hat{\beta}_k)}{Var_{Min}(\hat{\beta}_k)} = [\frac{1}{1 - R_k^2}]$$

- Hinweis zur alternativen Berechnung: VIFs sind auch die Diagonalelemente der Inversen der Korrelationsmatrix

Die Annahme bei dieser Betrachtung ist, dass sich $\hat{\sigma}_{ee}$ nicht verändert bei den zwei Regressionen von y.

# Mit Hilfe des Variance Inflation Factor (VIF) kann das Ausmaß der Kollinearität gemessen werden

## Modelle zur Erklärung des Strompreises

Reines Gasmodell

| | Estimate | SE | tStat | pValue |
|---|---|---|---|---|
| x1 | 2.3369 | 0.0078023 | 299.51 | 0 |

Gas-, Kohle und CO2-Modell

Estimated Coefficients:

| | Estimate | SE | tStat | pValue |
|---|---|---|---|---|
| x1 | 0.779 | 0.020885 | 37.299 | 4.8628e-166 |
| x2 | 0.65311 | 0.013204 | 49.464 | 1.7315e-226 |
| x3 | 2.5504 | 0.041299 | 61.754 | 1.1273e-279 |

## Variance Inflation Factor (VIF) = $\frac{Var(\hat{\beta}_k)}{Var_{Min}(\hat{\beta}_k)}$

VIF mit Base-Preis

| VIF | | | |
|---|---|---|---|
| Base | Gas | CO2 | Coal |
| 51,82 | 7,72 | 12,89 | 30,06 |

VIF nur Brennstoff und CO2-Preise

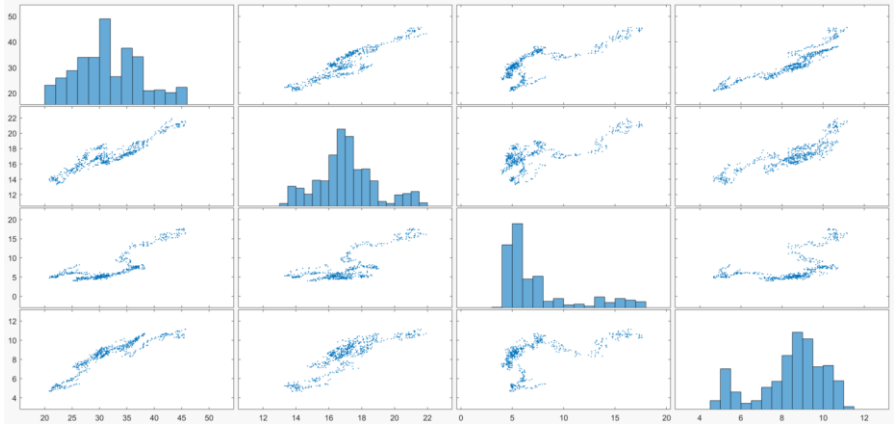
| VIF | | |
|---|---|---|
| Gas | CO2 | Coal |
| 7,70 | 2,65 | 4,69 |

Hinweis: ab einen Wert von 10 ist sicher mit Kollinearität zu rechnen!

**Die Principle Component Analysis (PCA) ist ein Instrument zur Datenreduktion, um Multikollinearität zu vermeiden**

## Ausgangssituation

| | Strom (Peak) | Gas | CO2 | Kohle |
|---|---|---|---|---|
| **Strom (Peak)** | 100% | 92% | 80% | 92% |
| **Gas** | 92% | 100% | 75% | 87% |
| **CO2** | 80% | 75% | 100% | 53% |
| **Kohle** | 92% | 87% | 53% | 100% |



## Herausforderung

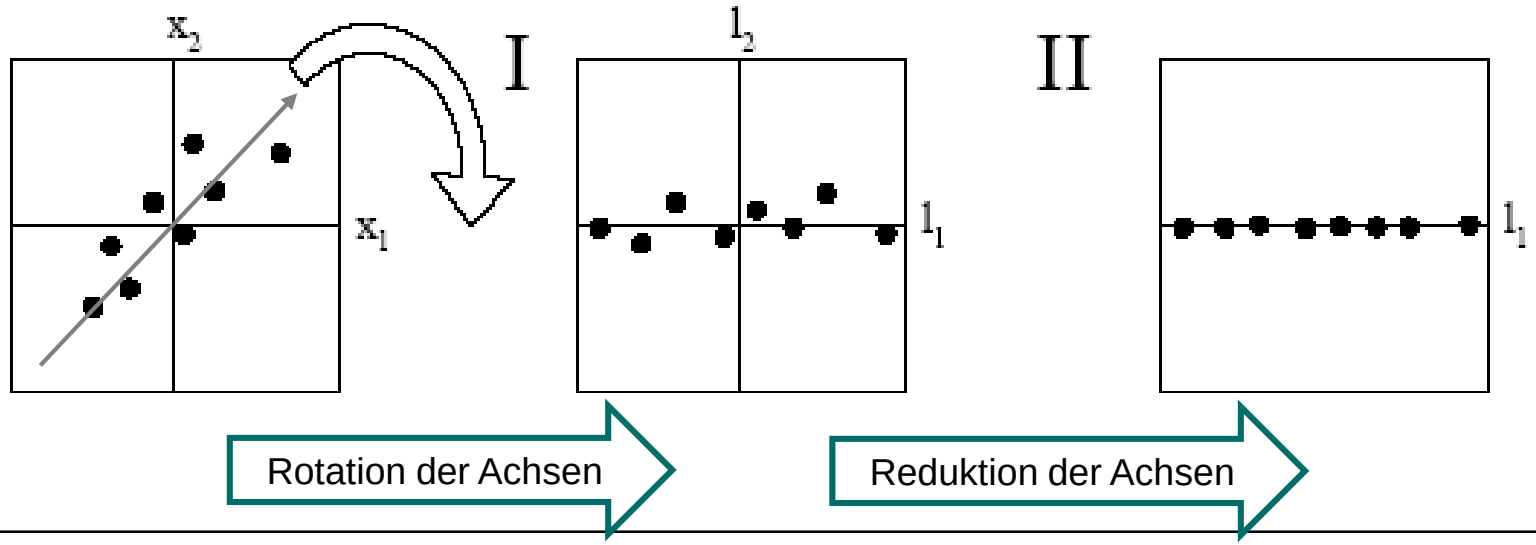
- Häufig sind Datenstrukturen hochdimensional und hochkorreliert
- Zugrunde liegende Gemeinsamkeiten und Unterschiede sind in den Datenstrukturen nicht direkt erkennbar
- Zielsetzung PCA:
  - Vermeidung von Multikollinearität (wenn zwei oder mehr erklärende Variablen eine sehr starke Korrelation besitzen)

    Reduzierung der Regressorbasis -> bessere Generalisierung des Modells

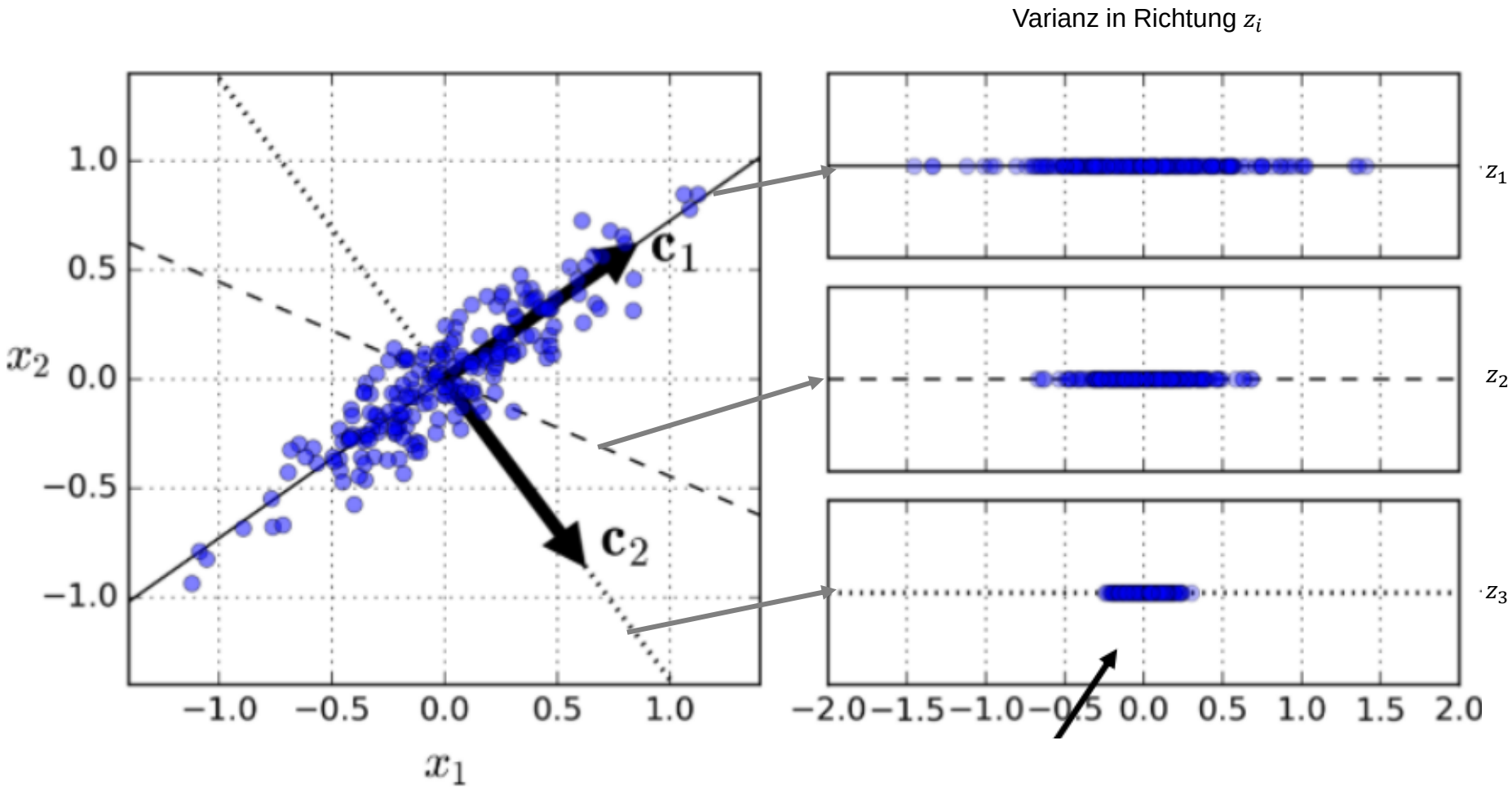
**Zielstellung: Transformation des Datensatzes und Reduktion der Anzahl der Dimensionen ohne deutlichen Verlust der Variabilität der Ursprungsdaten**

Methodik:

- Transformation der Ursprungsdaten in ein Set orthogonal zueinander stehender Variablen (Principle Components).
- Hierzu werden die Dimensionen (Achsen) des Ursprungssystem rotiert.
- Im neuen „Koordinatensystem“ besitzt die Principle Component den höchsten Erklärungsgehalt zur Variabilität der Ursprungsdaten
- Achsen, bezüglich derer die Daten kaum Varianz anweisen, können weggelassen werden.

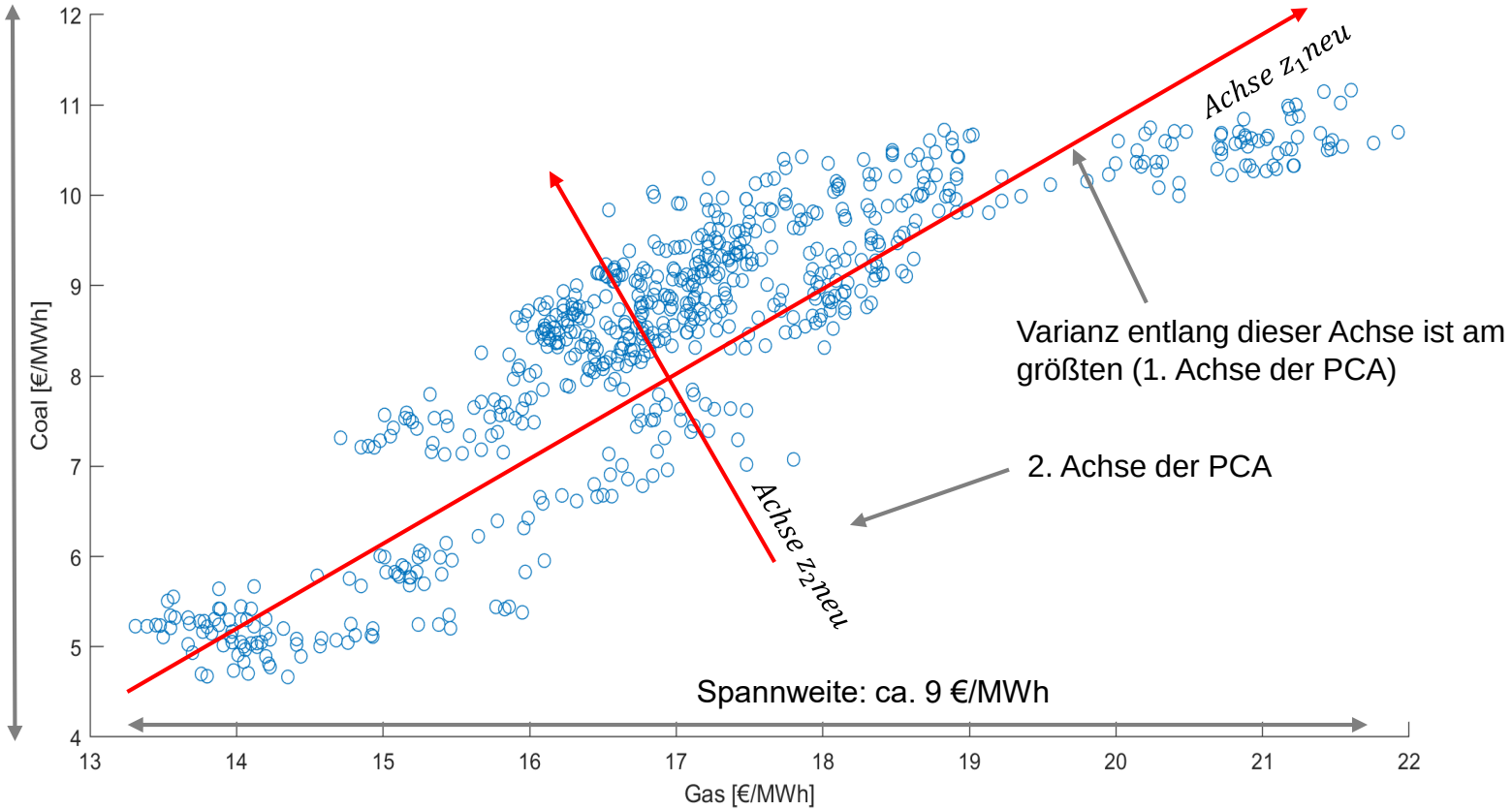


**Die PCA findet neue Achsen $z_i$ (Komponenten) für die Daten, so dass die Daten bezüglich dieser Achsen eine möglichst große Varianz aufweisen.**




Hinweis: in der Darstellung links ist die dritte Dimension zu $x_3$ nicht dargestellt

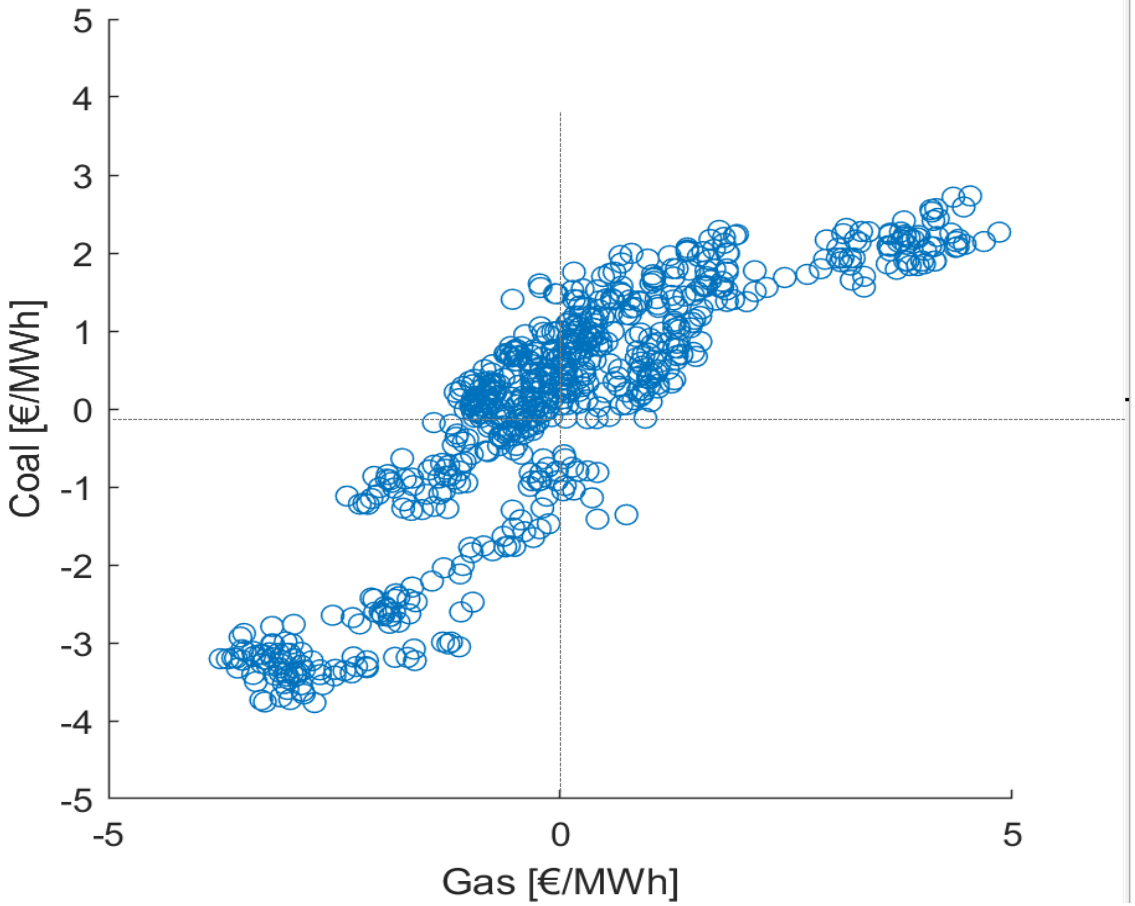
**Die PCA bestimmt neue Achsen $z_i$(Komponenten), so dass die Daten bezüglich dieser Achsen eine möglichst große Varianz aufweisen.**




Die Principle Components werden orthogonal in Richtung größter Variabilität gebildet.

# In Vorbereitung der PCA sollten die Merkmale standardisiert werden

## Standardisierung



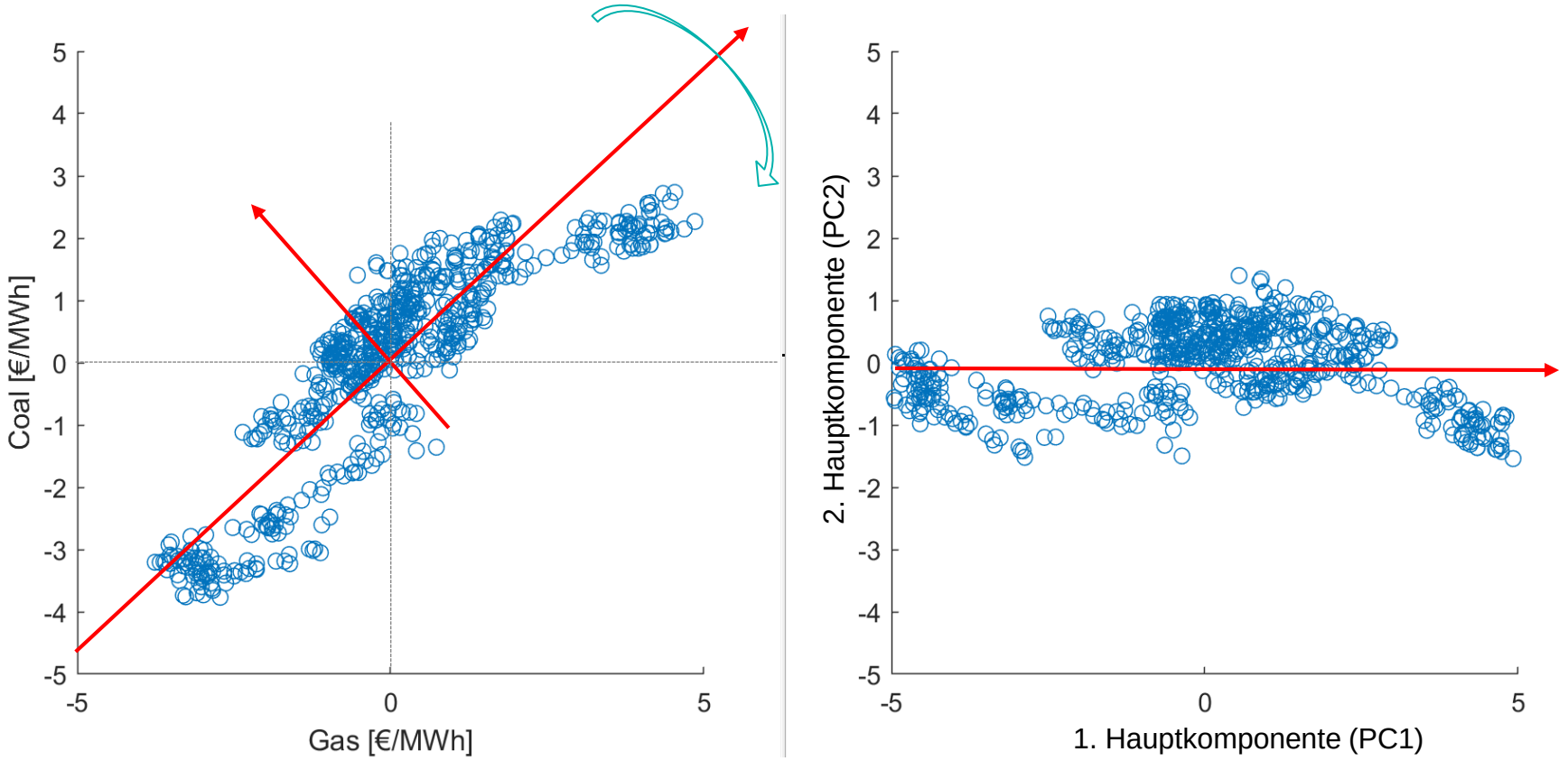

## Erläuterung

- Im Rahmen der Standardisierung wird jedes Merkmal um seinen Mittelwert und Standardabweichung bereinigt

$$Z = \frac{X - \bar{X}}{\sigma}$$

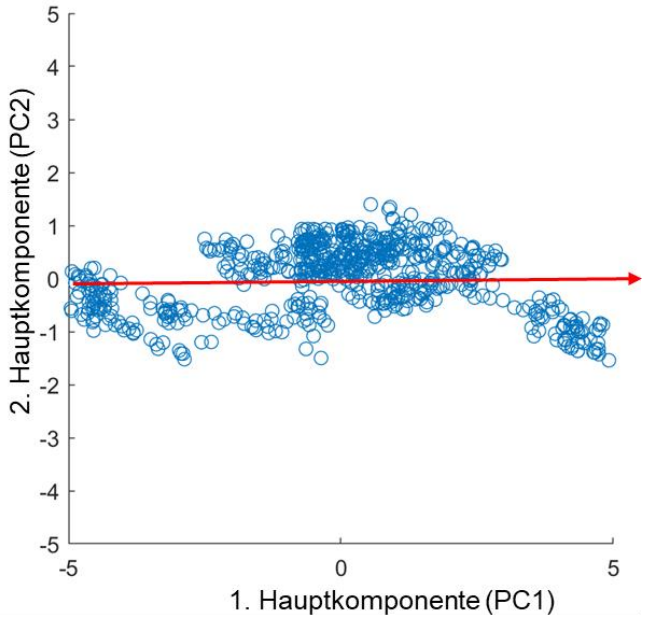
- Mit Hilfe der Standardisierung können auch Daten sehr unterschiedlicher Dimensionierung im Rahmen der PCA bearbeitet werden.

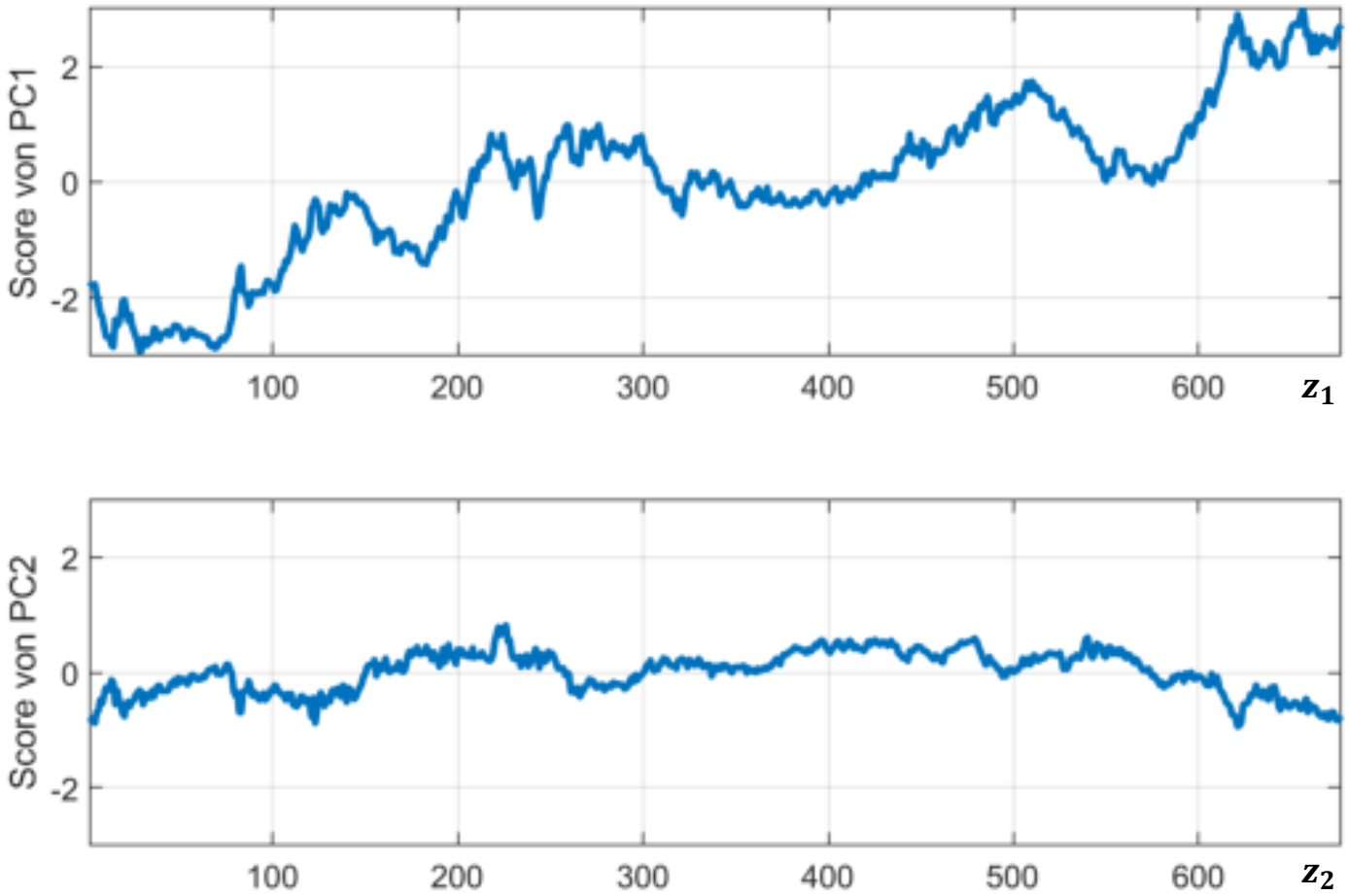
**Durch die PCA erfolgt eine Rotation der Achsen mit resultierenden neuen „Variablen“ (Scores oder nachfolgend Z genannt)**




Achsen, bezüglich derer die Daten kaum (oder keine) Varianz anweisen, können später weggelassen werden (→ Dimensionsreduktion).

**Verlauf der „Scores“ (Z) (latente Faktoren) kann in zeitlicher Dimension dargestellt werden. Die Zeitreihen Z sind so nicht direkt beobachtbar, sondern ergeben sich aus der PCA(X)**



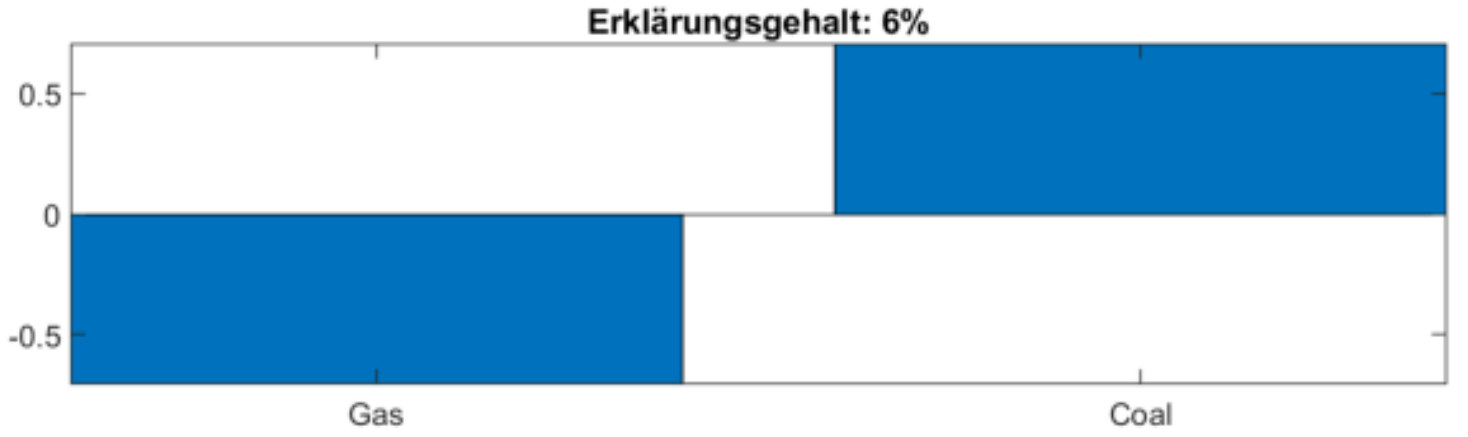




*nicht unmittelbar sichtbar oder zu erfassen

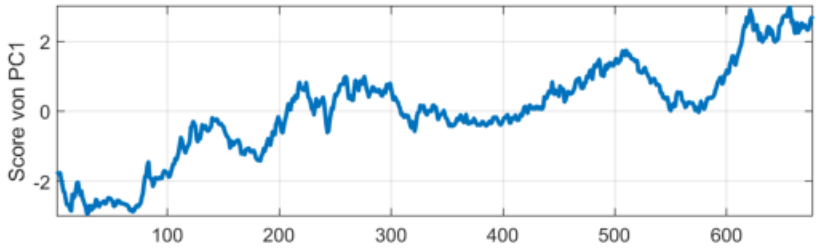
**Im Rahmen der PCA werden auch Gewichte ermittelt ($PC_{\text{coeff}}$) mit Hilfe derer die ursprünglichen Daten rekonstruiert werden können.**

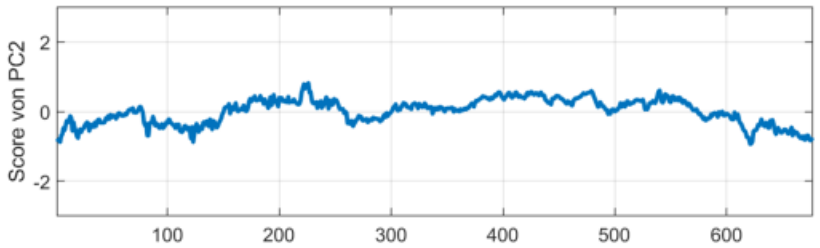
PC1

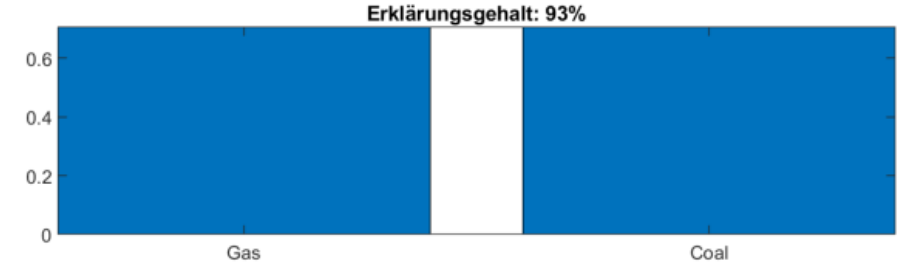

PC2

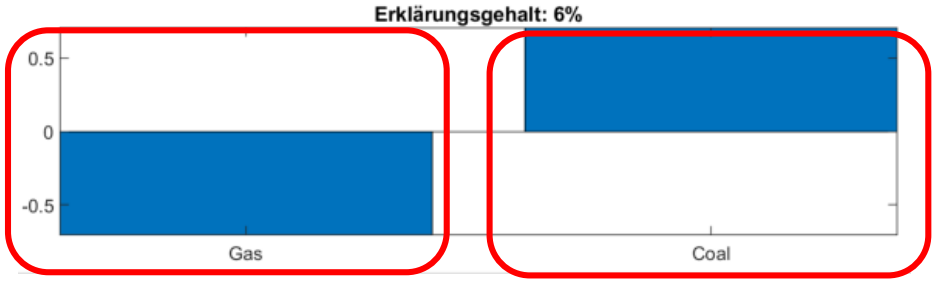


**Die Rekonstruktion der ursprünglichen Daten X erfolgt mit einer Linearkombination der Faktorladungen Z und den Koeffizienten $PC_{\text{coeff}}$**



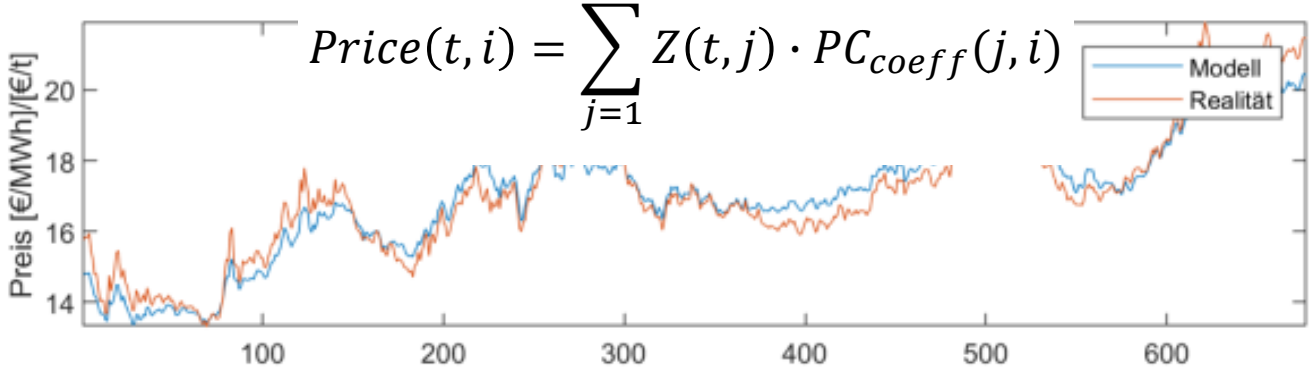



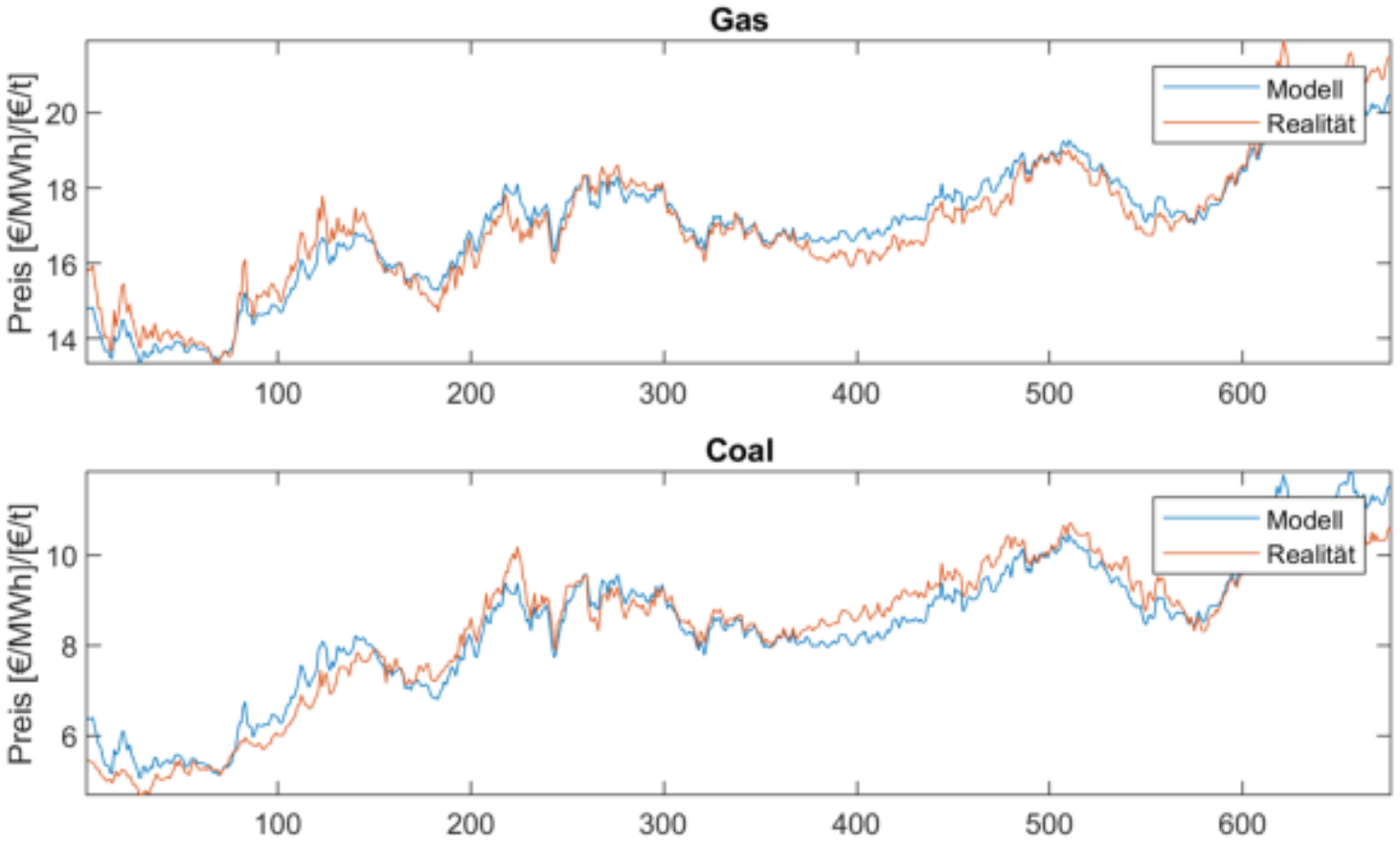

Rekonstruktion

$$Price(t,i) = \sum_{j=1}^{n} Z(t,j) \cdot PC_{coeff}(j,i)$$



**Bei einer Rekonstruktion der ursprünglichen Daten mit reduzierter Anzahl der latenten Faktoren gehen Informationen der Ursprungsdaten verloren.**

Rekonstruktion der ursprünglichen Daten mit der ersten Principle Component.




Erkenntnis:

Bereits ein latenter Faktor beschreibt sehr gut die gemeinsame Variabilität von Gas und CO2-Preisen

Die Gas und CO2- Preise können im Verlauf zu einem großen Anteil mit $z_1$ rekonstruiert werden

## Die mathematische Ableitung der Principle Components erfolgt mit Hilfe einer Zielfunktion

### Ableitung der Principle Components

Transformierte Variable:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Zielfunktion: Maximiere Varianz von $Z_1$:*

$$MAX\ ZF = \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{j=1}^{p}\phi_{j,1}X_{i,j}\right)^2 = \frac{1}{n}\sum_{t=1}^{n}Z_{t,1}^2$$

Restriktion:

$$\sum_{j=1}^{p}\phi_{j,1}^2 = 1$$

Varianz des neuen Datensatzes:

$$Var(Z_i) = \frac{1}{n}\sum_{t=1}Z_{t,1}^2$$

### Bestimmung der Principle Components

- Mittelwertbereinigung der Ursprungsdaten X → resultierende Principle Components mittelwertbereinigt
- Die erste Hauptkomponente eines Satzes von Variablen $X_1, X_2, \ldots, X_p$ ist die normierte lineare Kombination dieser.
- $\phi$ sind Freiheitsgrade
- $Z_i$(t) entspricht dem Score der i-ten Principle Component.
- Ziel ist es die Varianz von $Z_i$ zu maximieren unter der Nebenbedingung $\sum_{j=1}^{p}\phi_{j1}^2 = 1$
- Das Optimierungsproblem wird mit Hilfe einer Eigenwertzerlegung auf Basis der Kovarianzmatrix ermittelt.

* $E[Z_1] = 0, da\ die\ daten\ vorab\ standardisiert\ wurden!$

## Die latenten Faktoren z können wiederum direkt Rahmen einer multiplen linearen Regression zur Erklärung von y verwendet werden

- $z_1, z_2, \dots, z_M$ repräsentieren M **p**-Lineare Kombinationen der ursprünglichen p Regressoren x.

$$z_{m,t} = \sum_{i=1}^{p} \phi_{i,m} x_{i,t}$$

- Aus dem Regressionsmodell mit p Regressoren kann hierdurch ein Modell mit **M** Regressoren für die zu erklärende Variable $\boldsymbol{y_t}$ abgeleitet werden.

$$y_t = \theta_0 + \sum_{m=1}^{M} \theta_m z_{m,t} + \epsilon_t$$

Regressionsgleichung mit M < p Variablen

M < p bedeutet, dass nicht alle resultierenden latenten Faktoren für die Regression verwendet werden.

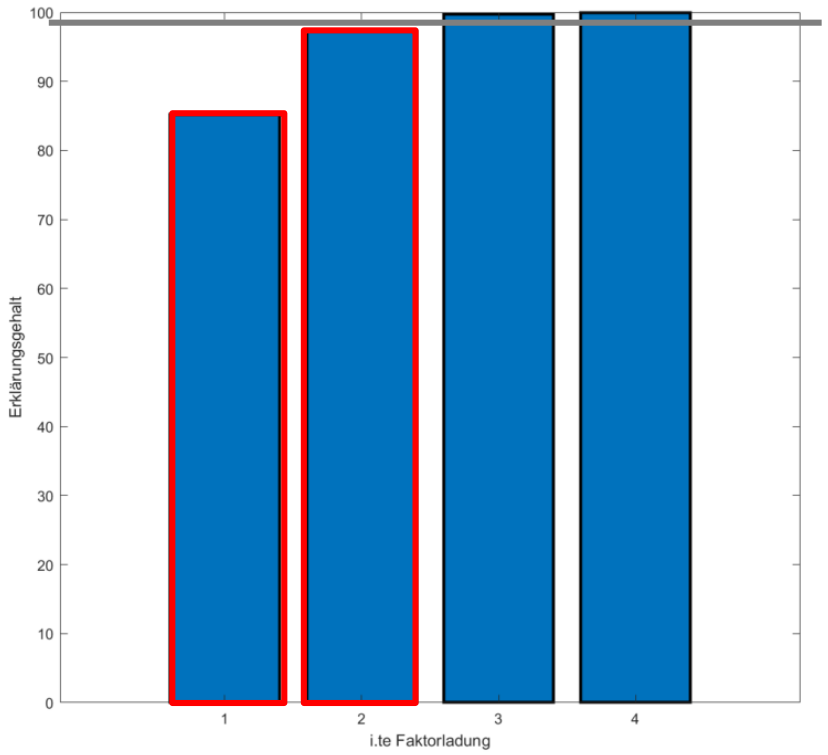
- Konsequenzen:
  - das Modell zur Beschreibung von y kann parameterärmer ausgestaltet werden,
  - durch die Orthogonalität von z (Unabhängigkeit) kann Multikollinearität vermieden werden.

Nachweis: Auf Basis der Principle Components kann ein entsprechendes Regressionsmodell für die zu erklärende Variable $\boldsymbol{y_i}$ abgeleitet werden.

$$\sum_{m=1}^{M} \theta_m z_{m,t} = \sum_{m=1}^{M} \theta_m \sum_{i=1}^{p} \phi_{i,m} x_{i,t} = \sum_{i=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{i,m} x_{i,t} = \sum_{i=1}^{p} \beta_i x_{i,t} \text{ mit } \beta_i = \theta_m \sum_{m=1}^{M} \theta_m \phi_{i,m}$$

## Die notwendige Anzahl der latenten Faktoren als Regressorbasis von y lässt sich mit Hilfe der Eigenwerte der PCA ableiten

### Grenzwert für Anzahl der zu betrachtenden Faktorladungen




### Erläuterung

- Mit Hilfe der Eigenwerte (EV) kann der Erklärungsgehalt zur Gesamtvarianz der i.ten Faktorladung bestimmt werden.

$$Erklärungsgehalt(i) = \frac{EV(i)}{\sum EV}$$

- Der Erklärungsgehalt der einzelnen Faktorladungen ist hierbei immer abnehmend.
- Aus diesem Grunde kann die kumulierte Darstellung der Erklärungsgehalte verwendet werden und ab Erreichung eines definierten Zielwertes die Auswahl getroffen werden.
- Es gibt keine statistische Berechnungsmethodik; vielmehr versucht man die Faktorladungen abzuleiten in Abhängigkeit:
  - der Anzahl der Faktorladungen,
  - der Höhe des gewünschten Gesamterklärungsgehalt,
  - des Grenznutzen der letzten zu betrachtenden Faktorladung

# Agenda

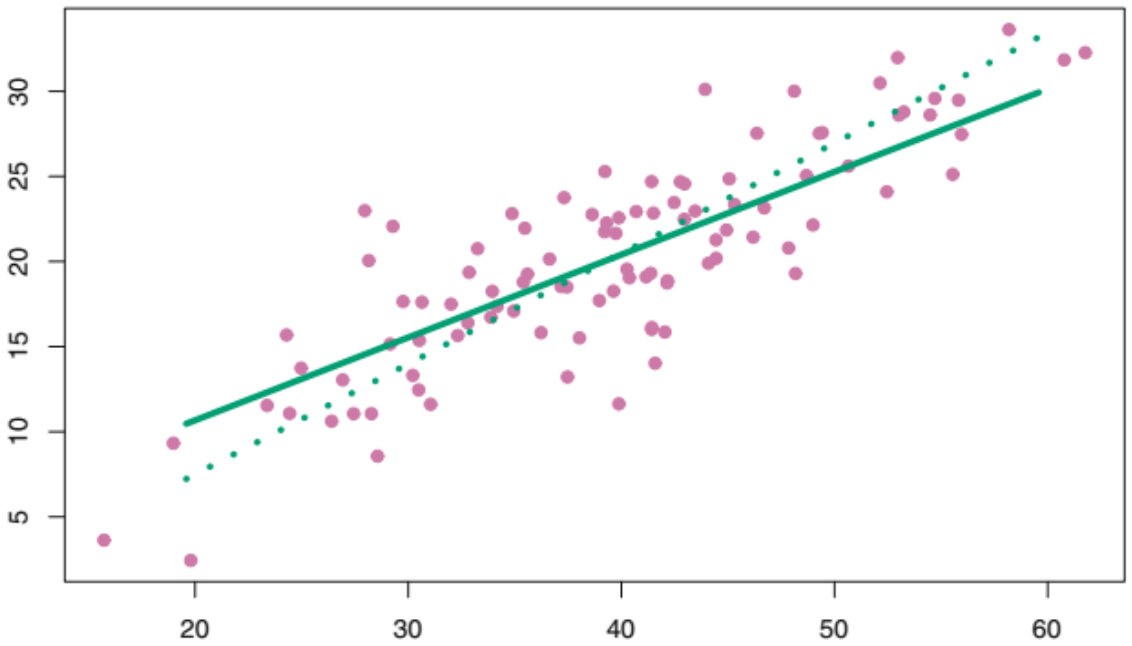
1. Ridge and Lasso Regression
2. Principle Component Analysis (PCA)
3. Partial Least Square Regression (PLS)

Literaturempfehlung und Quellennachweise der Abbildungen: James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics), Springer New York. Kindle-Version.

# Funktionsweise der Methodik Partial Least Squares

## PLS orientiert sich stärker am Zusammenhang zur Beobachtungsgröße

Gestrichelte Linie (PLS-Regression)

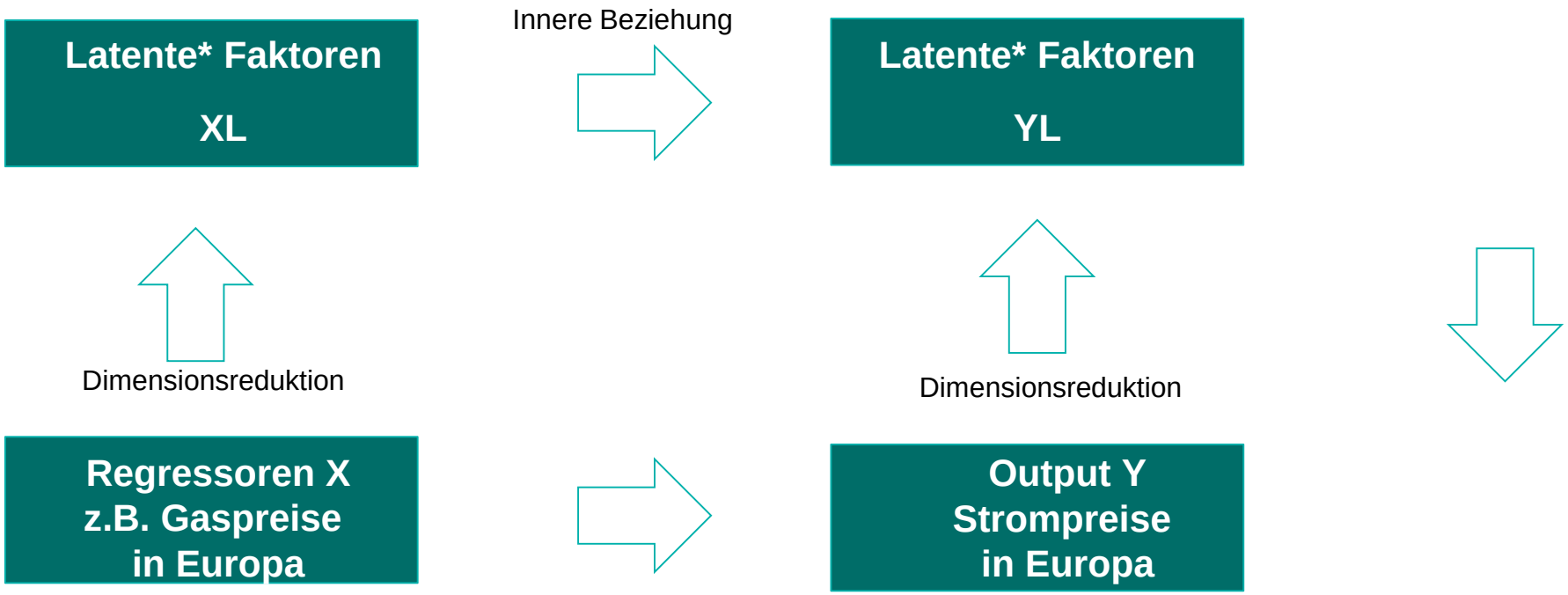



## Erläuterung

- PCA-Regression identifiziert Linearkombinationen der ursprünglichen **Regressoren**, welche bestmöglich deren Variabilität beschreiben.
- Der Identifikationsprozess wird **unabhängig** von einer Beobachtungsgröße y vorgenommen.
- Es gibt keine „Garantie“, dass die Principle Components ebenfalls bestmöglich die Beobachtungsgröße y beschreiben.
- Die PCA- Regression kann einen sehr geringen Erklärungsgehalt aufweisen, falls Y stark mit einer Variablen $X_i$ korreliert ist, welche eine geringe Varianz im Datensatz von **X** besitzt.
- Partial Least Square berücksichtigt die Beobachtungsgröße y im Rahmen der Dimensionsreduktion.

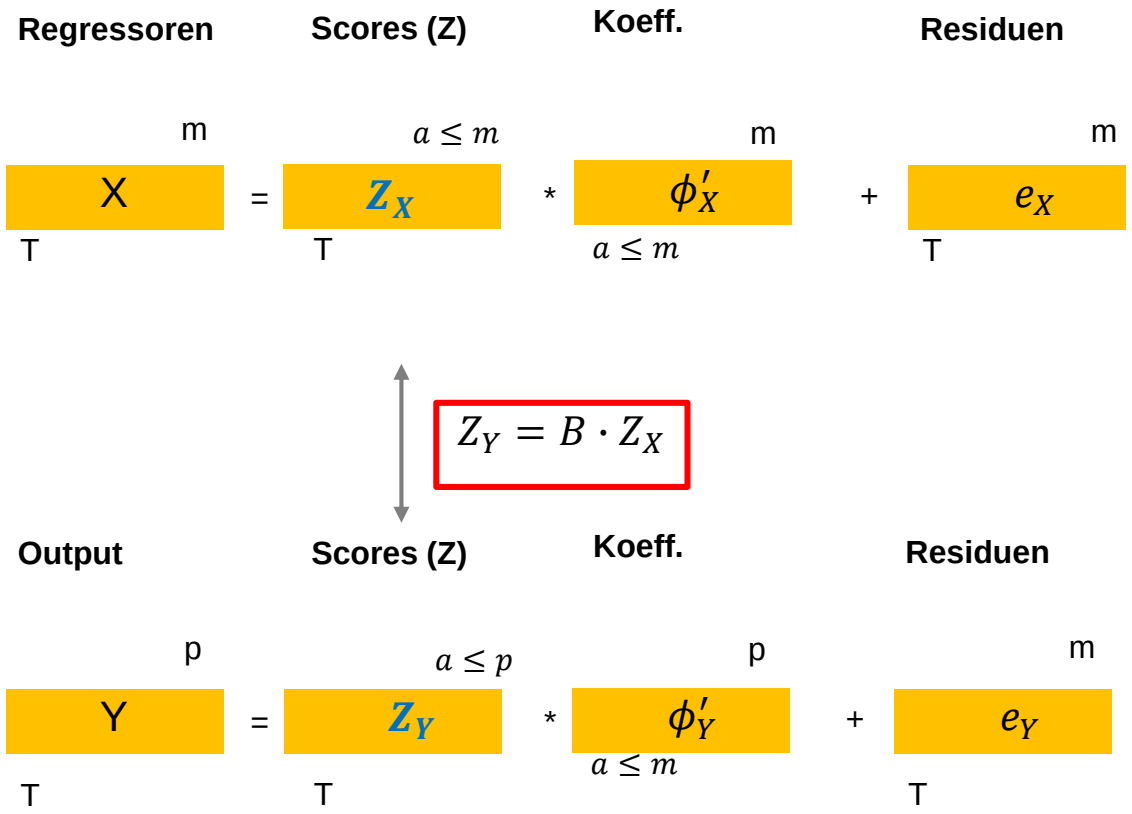
## Vorgehensweise Partial Least Square Regression

PLS-Regression führt sowohl für X als auch für Y eine Dimensionsreduktion durch.




*Latent bedeutet, dass diese Faktoren (Variablen) nicht unmittelbar beobachtbar sind

# Ziel der PLS Regression ist es die Kovarianz zwischen X und Y zu maximieren




- Das Ziel von PLS ist es, die Norm von $e_Y$ zu minimieren und gleichzeitig eine Korrelation zwischen X und Y zu erhalten.
- Hierzu werden:
  - Regressoren und Output in Komponenten zerlegt (analog zur PCA)
  - und die resultierenden Komponenten $Z_Y$ und $Z_X$ durch Regression untereinander verknüpft:

$$Z_Y = B \cdot Z_X$$

- Diese Gleichung nennt man auch die "innere Beziehung".

## Aufgabe

- Aufgabe:
  - Durchführung PCA auf Daten Strommarkt.xlsx

## Leitfragen

- Was ist der Hintergrund/ Anlass für die Anwendung von Ridge und Lasso Regression?
- Wie funktionieren Lasso- und Ridge -Regression?
- Wie wird der optimale Parameter $\lambda$ bestimmt?
- In welchen Situation kann eine Principle Component-Analyse (PCA) sinnvoll sein?
- Wie funktioniert eine Principle Component-Analyse?
- Warum bezeichnet man die resultierenden Faktoren als „latent“?
- Warum besitzen die Faktoren keine Korrelation?
- Womit kann man die benötigte Anzahl der Faktoren abschätzen, um die Ursprungsvariabilität näherungsweise zu reproduzieren?
- Wie funktioniert eine PCA-Regression?
- Unter welchen Umständen kann eine PCA-Regression „schlechte“ Resultate liefern?
- Wie funktioniert die PLS-Regression?