Master in Financial Engineering (EPFL)
Financial Econometrics

Part II: Machine learning and Asset Pricing
Lecture 2: Selection of variables—Dimension reduction methods

Florian Pelgrin

EDHEC Business School

February - June 2019

1. Dimension reduction methods
   - Overview
   - Principal components regression
   - Independent components regression
   - Partial least squares regression
   - Projection pursuit regression

4. Dimension reduction methods

4.1. Overview

- So far...methods that control variance in two different ways:
  - ✓ Subset of the original variables
  - ✓ Skrinkage of coeficients toward zero
- Common feature of these methods: use the original features/inputs/predictors.
- Variance reduction methods (generally) *transform* variables.

General principles of dimension reduction methods

- ✓ **Step 1**: Preprocessing data
- ✓ **Step 2**: Finding the components
  - A set of original predictors $x_1, x_2, \cdots, x_p$.
  - Let $z_1, \cdots, z_m$ denote $m < p$ linear combinations of original features:

$$z_k = \sum_{j=1}^{p} \phi_{j,k} x_j$$

    for some constants $\phi_{1,k}, \cdots, \phi_{p,k}$ for $k = 1, \cdots, m$.
  - Which statistical criteria? Latent *versus* observed variables?

- **Step 3**: Linear regression model

  - Fit the linear model:

$$y_i = \theta_0 + \sum_{k=1}^{m} \theta_k z_{i,k} + u_i, \quad i = 1, \cdots, n.$$

  First principle: $p+1$ coefficients $\beta_0, \cdots, \beta_p \longrightarrow m+1$ coefficients $\theta_0, \cdots, \theta_m$ (with $m < p$)!

  - Note that

$$\begin{aligned} \sum_{k=1}^{m} \theta_k z_{i,k} &= \sum_{k=1}^{m} \theta_k \underbrace{\sum_{j=1}^{p} \phi_{j,k} x_{i,j}}_{\text{reduction method}} \\ &= \sum_{j=1}^{p} \sum_{k=1}^{m} \theta_k \phi_{j,k} x_{i,j} \\ &= \sum_{j=1}^{p} \beta_j x_{i,j} \end{aligned}$$

  where $\beta_j = \sum_{k=1}^{m} \theta_k \phi_{j,k}$.

  Second principle: Dimension reduction constrains β_j estimates....

Methods

- Principal components regressions
- Independent components regressions
- Partial least squares regressions
- Projection pursuit regressions

4.2. Principal components regression

- ✓ Intuition
- ✓ Solving a two-dimensional case
- ✓ Interpretation of PCA
- ✓ Issues
- ✓ PCA-based regression

Intuition

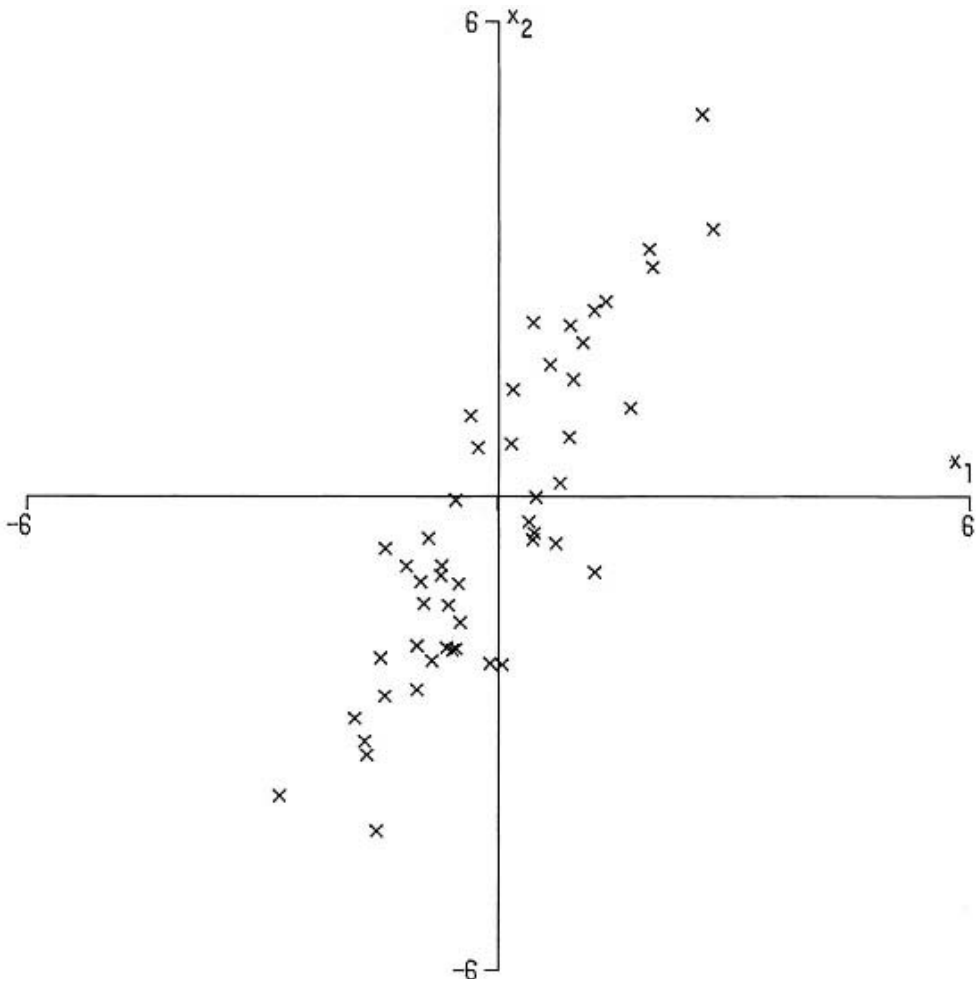

Figure: Two correlated random variables x_1 and x_2

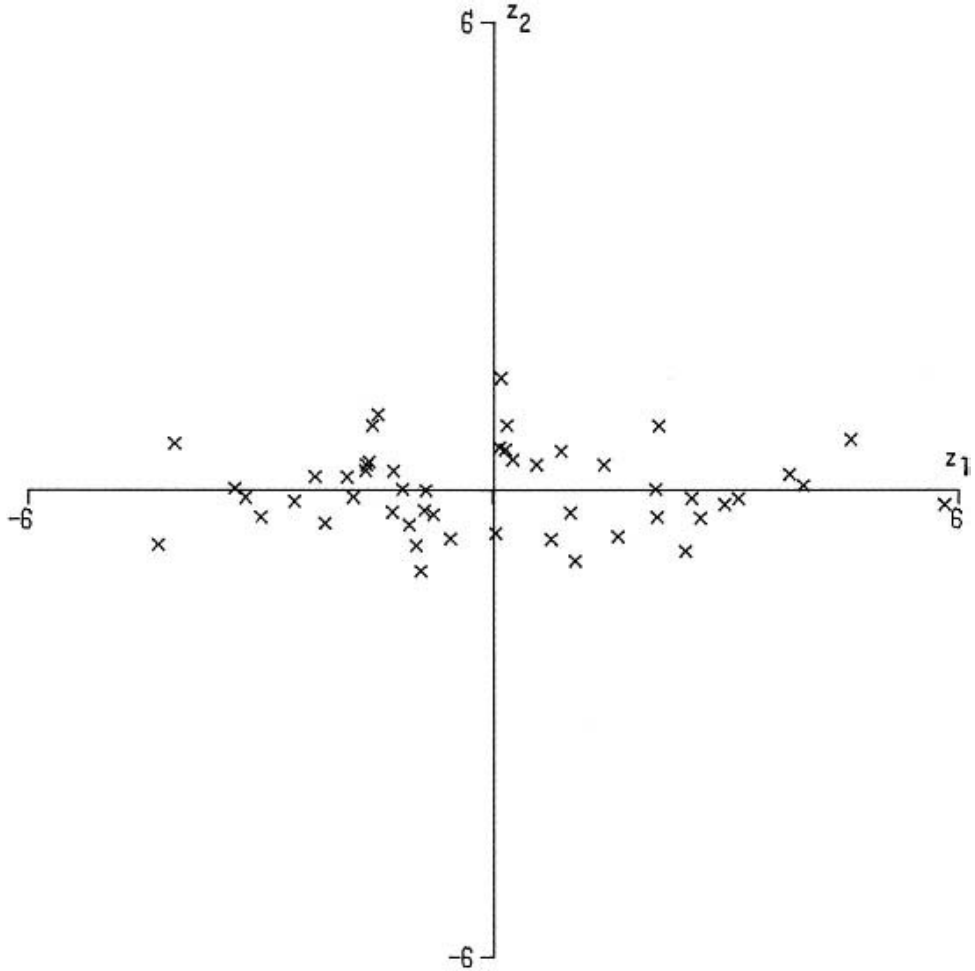

Figure: Still the same variation?

- ✓ Figure 1: The two random variables are (highly) correlated and both variables display variation!

- ✓ Figure 2: There is greater (resp., little) variation in the direction of z_1 (resp., z_2) than in either of the original variables

Main goal: Explain Figure 1 $\longrightarrow$ Figure 2, and especially **Principal component analysis**

Solving a 2-dimensional case

- ✓ Let $x = (x_1, x_2)$ denote a vector two random variables with variance-covariance (or correlation) matrix Σ (and with mean zero);

- ✓ z_1 is determined so that that a linear function of x

$$z_1 \equiv \alpha_1^\top x = a_{11}x_1 + a_{12}x_2$$

has maximum variance:

$$\max_{a_1} \mathbb{V}\left[a_1^\top x\right] \Leftrightarrow \max_{a_1} \quad a_1^\top \Sigma a_1$$

where $a_1 = A_{1.}^\top$ is the transpose of the first row of the A matrix.

However the maximum is not achieved for finite a_1!

✓ a_1 solves the constrained maximization problem Solver

$$\max_{a_1} \mathbb{V}\left[a_1^\top x\right]$$
$$\text{s.t.} \quad a_1^\top a_1 = 1$$

or

$$\Leftrightarrow \max_{a_1} \quad a_1^\top \Sigma a_1$$
$$\text{s.t.} \quad a_1^\top a_1 = 1$$

Note: This is one possible normalization constraint...

✓ z_2 is determined so that that a linear function of x

$$z_2 \equiv a_2^\top x = a_{21}x_1 + a_{22}x_2$$

has maximum variance subject to **being uncorrelated** with z_1 (and with a normalization constraint) Solver

$$\max_{a_2} \mathbb{V}\left[a_2^\top x\right]$$
$$\text{s.t.} a_2^\top a_2 = 1 \quad \text{and} \quad a_1^\top a_2 = 0.$$

More generally,

- Suppose that there are p (interrelated) variables of interest
- **Aim**: Identify a small number of **uncorrelated linear combinations** that explain most of the variability of the original data.
- `PCA` is a **data-reduction technique** and is based on the **spectral decomposition** of the covariance or correlation matrix.

Interpretation of principal components

Using the case $p = 2$, how can one interpret the principal components?

- The first-order condition writes down

$$\Sigma a_1 = \lambda a_1$$

  Therefore, in the case of the first principal component, the optimal objective function is given by

$$a_1^\top \Sigma a_1 = a_1^\top \lambda a_1 = \lambda$$

  since λ is some scalar and $a_1^\top a_1 = 1$.

- λ must be as large as possible $\Rightarrow$ a_1 is the eigenvector corresponding to the largest eigenvalue of Σ and

$$\mathbb{V}\left[a_1^\top x\right] = d_1^2.$$

- More generally, the kth principal component corresponds to the kth largest eigenvalue of Σ.

- This results from the **spectral decomposition** sd or the **singular value decomposition** svd of Σ

- The singular values d_j correspond to direction in the column space of X with more or less variance direction:

  - The first principal component has maximum variance;

  - Subsequent principal components have maximum variance subject to being orthogonal to the earlier ones (decreasing contributions of "directions");

  - The last principal component has minimum variance.

Issues

- Principal components depend on the scaling of inputs example: it is preferable to standardize them and thus make use of the correlation matrix (except if all data has the same unit);

- *Prior* selection of number of components:

  - **Cumulative percentage of total variation** : Select a cumulative percentage of total variation that the selected principal components contributed (say 80% or 90%) Notably, the ratio (for $k \leq p$)

$$\frac{\sum_{i=1}^{k} d_i}{\sum_{i=1}^{m} d_i}$$

    represents the proportion of the total variability explained by the first k principal components.

  - Note: For large p, SVD or spectral decomposition might involve issues related to the curse of dimensionality, and especially the determination of (tiny) eigenvalues.

- **The scree graph and the log-eigenvalue diagram**:
  - Scree graph: Plot the eigenvalues against the number of components (say, k) and detect whether there is a "break", i.e. (after joining the different points) decide at which k the slopes of the lines joining the plotted points are "steep" (resp., not "steep") to the left (resp., right) of k.
  - Log-eigenvalue diagram: Plot log-eigenvalues against the number of components.
- **Testing procedures**
- Example: example

PCA-based regression

- Estimate a linear regression models with principal components

$$y_i = \theta_0 + \sum_{k=1}^{m} \theta_k z_{i,k} + u_i, \quad i = 1, \cdots, n.$$

- Proceed with (*ex-post*) variable (principal components) selection using cross-validation techniques/standard techniques of variable selection.

- Note that the principal components corresponds to a variance reduction of the original p variables and are not determined ex-ante using their explanatory power for the dependent variables. Cross-validation techniques (or variable selection techniques) might lead to a different selection of the principal components (w.r.t. *ex-ante* selection).

- PCR can be viewed as an approximate factor model!

Issues

- PCR is not a feature selection method: each of the m principal components used in the second step is a linear combination of all original p features
- Principal components are determined without taking into account the correlation with the dependent/output variable—see partial regression models.
- PCR can mitigate overfitting especially when $m << p$.
- PCR can have poor performances when data were generated in a such a way that many principal components are required to capture adequately the response (i.e., the relative contribution of each principal component is rather weak). In contrast, PCR will tend to perform well when only the first few principal components are sufficient to capture the variation of the predictors (and the relationship with the variables).
- Forecasting principal components might be a difficult task!

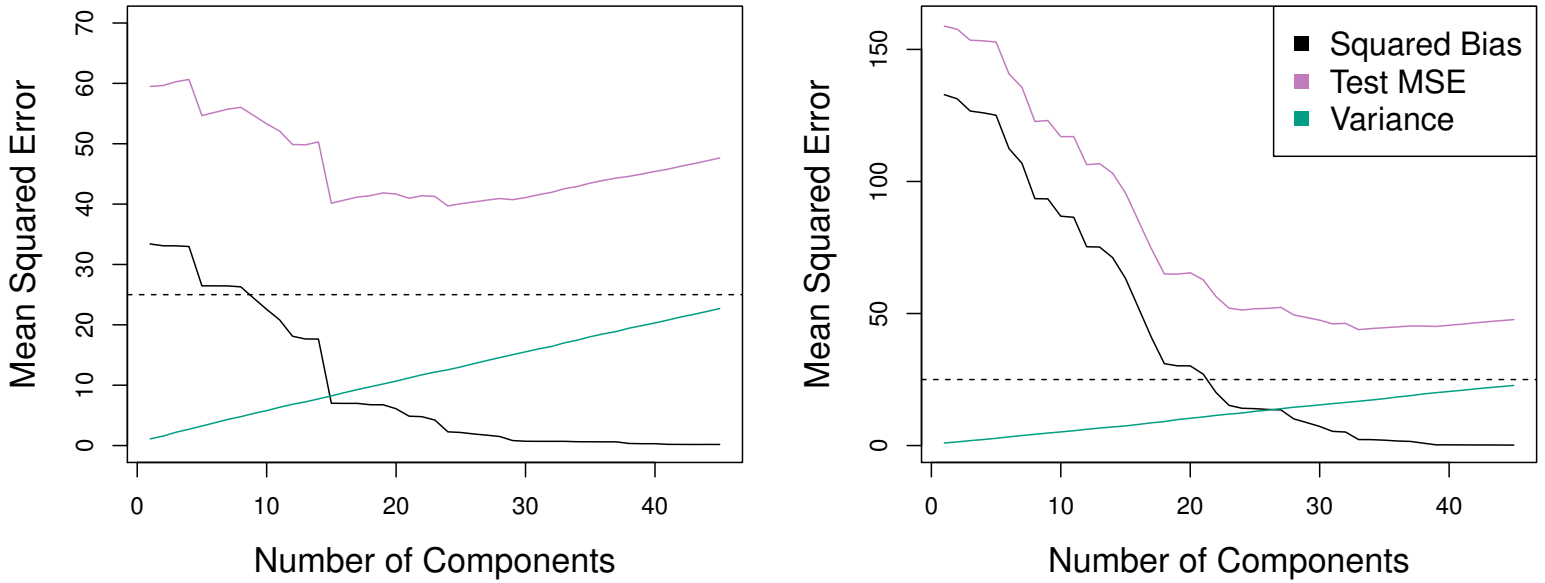

Figure: Selection of principal components

Note: Application of PCA for two simulated data. Left panel: data generated using $n = 100$ and $p = 45$. Right panel: only two of the 45 predictors are used to generate data.

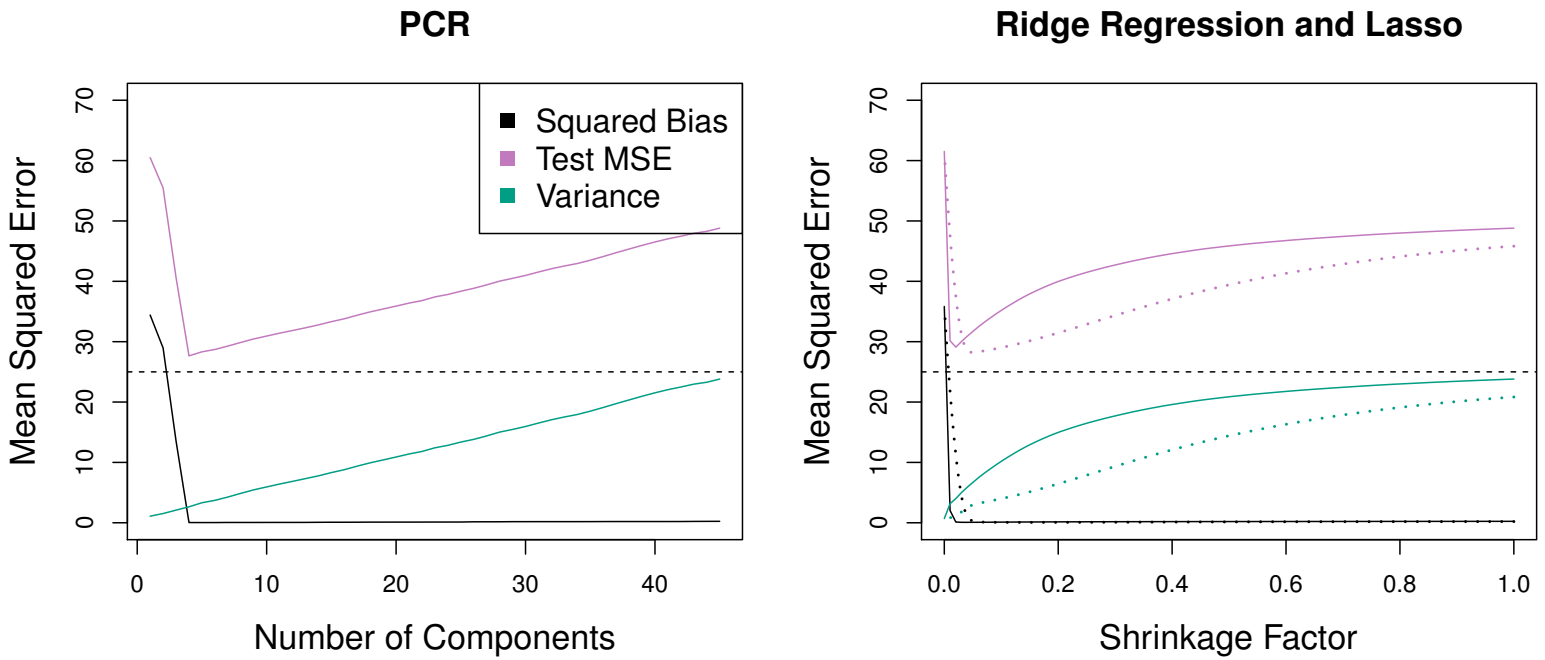

Figure: Comparison PCA, Lasso and Ridge

Note: PCR, Ridge regression and Lasso regression are subsequently applied to a simulated data set in which the first five principal components contain all relevant information. The dotted line corresponds to the irreducible error.

3.3. Independent components regression

Example:

- Suppose that one observes x_1 and x_2 and that these two observations are explained by two unobserved variables (signals) s_1 and s_2 as follows:

$$\begin{aligned} x_1 &= a_{11}s_1 + a_{12}s_2 \\ x_1 &= a_{21}s_1 + a_{22}s_2 \end{aligned}$$

  i.e.,

$$x = As$$

  where A is an unknown matrix and s is unobserved.

- Questions:
  - Can one identify s?
  - Can one use some information (statistical properties) of s to provide an estimate of A?

- ✓ Especially, two key ingredients:
  - Assume that s_1 and s_2 are **statistically independent** (e.g., at each time instant t);
  - At least, **one component of s is not Gaussian**.
- ✓ Then Independent Components Analysis (ICA) provides an approach to estimate A and to identify s.
- ✓ Note: ICA is also known as a method of *blind source separation* or *blind signal separation*—one observes a source and wish to identity original signals (i.e., independent components).

- More generally, consider

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n, \text{for all } i$$

i.e.,

$$x = As = \sum_{i=1}^{n} a_i s_i$$

where s_i is a column vector.

- This is a **generative model**: observed data is generated by a process of mixing the components s_i.

- Without loss of generality, there is no noise, A is assumed to be a square (invertible) matrix and all variables (x and s) have mean zero and unit variance. After estimating A, one can compute:

$$s = Wx$$

where $W = A^{-1}$.

How does it work?

- **Go beyond the normality assumption**: Assume that the mixing matrix, A, is orthogonal and, s_1 and s_2 are gaussian. Then x_1 and x_2 are gaussian, uncorrelated, and of unit variance. Their joint density is given by

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left(x_1^2 + x_2^2\right)\right)$$

The graphical representation bivariate normal shows that the density is completely symmetric.

What is the implication for ICA? It does not contain any information on the directions of the columns of the mixing matrix A—A cannot be estimated.

More rigorously,

- The distribution of any orthogonal transformation of the gaussian vector (x_1, x_2) has exactly the same distribution as (x_1, x_2), and that x_1 and x_2 are independent.
- In the case of gaussian variables, one can only estimate the ICA model **up to an orthogonal transformation**.
- Hence the matrix A is not identifiable for gaussian independent components.
- Actually, if just one of the independent components is gaussian, the ICA model can still be estimated!

Departure for normality is critical: Statistical criterion?

Why independence and nongaussianity?

- ✓ The intuition relies on the **central limit theorem**:
  - In general, the sum of two (n) independent (identically distributed) random variables is more "normal" ("Gaussian") than any of the random variable.
  - Combine this "result" with the need to go beyond normality: Find some linear combinations such that the nongaussianity is maximized!

Consider the estimation of one independent component.

- Let y denote a linear combination of x

$$y = w^\top x = \sum_{i=1}^{n} w_i x_i$$

where w is an unknown vector. If w is one row of the inverse of A, then y is obviously one of the independent components

- Problem is equivalent to find w...and especially as being one row of the inverse of A.

- Here comes the intuition of "CLT"! Let z denote $z = A^\top w$, one has

$$\begin{aligned} y &= w^\top \underbrace{x}_{=As} = w^\top As \\ &= z^\top s \end{aligned}$$

y is a linear combination of s_i with weights given by z_i and $z^\top s$ is "more" Gaussian than any of the s_i.

- ✓ **Optimization problem**: Find w such that to maximize the nongaussianity of $w^\top x$
- ✓ Measure of nongaussianity?
  - Kurtosis;
  - Negentropy;
  - Minimization of mutual information;
  - ML-based estimation;
  - Etc.
- ✓ Note: There are $2n$ local maxima in a n-dimensional space: two for each independent component s_i and $-s_i$ (defined up to multiplicative sign).

1. Kurtosis-based estimation

- (Excess) Kurtosis

$$\texttt{Kurt}(y) = \mathbb{E}\left[y^4\right] - 3\left\{\mathbb{E}\left[y^2\right]\right\}^2$$

  or (with unit variance)

$$\texttt{Kurt}(y) = \mathbb{E}\left[y^4\right] - 3$$

- Negative (resp., positive) kurtosis corresponds to a sub-Gaussian (resp. super-Gaussian) vector;
- Use the absolute value transformation (since identified up to a multiplicative sign restriction);
- Two interesting properties:

$$\texttt{Kurt}(x_1 + x_2) = \texttt{Kurt}(x_1) + \texttt{Kurt}(x_2) \quad x_1 \text{ and } x_2 \text{ are independent}$$
$$\texttt{Kurt}(\alpha x_1) = \alpha^4 \texttt{Kurt}(x_1).$$

Example: Two-dimensional case

- Let y denote

$$\begin{aligned} y &= w^\top x = w^\top A s \\ &= \underbrace{z^\top}_{w^\top A} s = z_1 s_1 + z_2 s_2. \end{aligned}$$

- The maximization program writes

$$\begin{aligned} &\max_{z_1, z_2} \left| \mathtt{Kurt}(y) \right| \\ &\text{s.t.} \mathbb{V}(y) = 1 \end{aligned}$$

where

$$\begin{aligned} &\mathtt{Kurt}(y) = z_1^4 \mathtt{Kurt}(s_1) + z_2^4 \mathtt{Kurt}(s_2) \\ &\mathbb{V}(y) = 1 \Leftrightarrow z_1^2 + z_2^2 = 1 \end{aligned}$$

with s_1 and s_2 have unit variance.

- (Up to a sign restriction), two corner solutions $(z_1 = 1, z_2 = 0)$ and $(z_1 = 0, z_2 = 1)$!

2. Negentropy-based estimation

- Entropy: Degree of information that the observation of the variable provides. The more "random" (i.e., unpredictable and unstructured) the variable is, the larger its entropy

$$H(y) = \begin{cases} -\sum \mathbb{P}(Y = y_i) \log\left(\mathbb{P}(Y = y_i)\right) & \text{discrete r.v.} \\ -\int f(y) \log\left(f(y)\right) & \text{continuous r.v.} \end{cases}$$

- Fundamental result: A Gaussian variable has the largest entropy among all random variables of equal variance! The Gaussian distribution is the "most random" or the "least structured" of all distributions.

- Negentropy or differential entropy can be defined as

$$J(y) = H\left(y_{\text{Gaussian}}\right) - H(y)$$

Generally, this requires a nonparametric estimate of the pdf, which might turn to be difficult in practise. Some approximations have been provided in the statistical literature.

3. Mutual information-based estimation

- Using the concept of differential entropy, the mutual information, denoted I, between p (scalar) random variables is defined as follows:

$$I(y_1, \cdots, y_p) = \sum_{i=1}^{p} H(y_i) - H(y).$$

  This is equivalent to the well-known Kullback-Leibler divergence between the joint density and the product of its marginal densities/

- This measure is always non-negative, and zero if and only if the variables of interest are statistically independent.

- Key advantage: Take into account the whole dependence structure and not only the variance-covariance structure (as in PCA).

- Mutual information is a natural information-theoretic measure to capture the independence of random variables.

- Therefore the ICA problem is equivalent to find w so that the mutual information of the (transformed) components s_i is minimized.

- Minimizing mutual information is (roughly) equivalent to find directions in which the negentropy is maximized. For instance, if the y_i's ($i = 1, \cdots, p$) are uncorrelated of unit variance, it can be shown that the fundamental relationship between negentropy and mutual information is (with $n = p$):

$$I(y_1, \cdots, y_n) = C - \sum_i J(y_i)$$

  for some constant C.

- Minimizing mutual information amounts of finding maximally nongaussian directions.

3. Maximum likelihood estimation

- In the (noisy-free) ICA model, the likelihood function can be directly formulated as:

$$L = \sum_{t=1}^{T} \sum_{i=1}^{n} \log \left\{ f_i \left(w_i^\top x_t \right) \right\} + T \log \left\{ |\det(W)| \right\}$$

  where $n = p$, $W = (w_1, \cdots, w_n)^\top$ denotes the matrix A^{-1}, the f_i's are the density functions of the s_i (here assumed to be known), and x_t, $t = 1, \cdots, T$ are the realizations of x.

- This is essentially equivalent to the minimization of mutual information.

- Another related contrast function (derived from the neural network literature) is the output entropy (or information flow). The so-called principle of network entropy (maximization) or infomax is equivalent to maximum likelihood estimation.

3.4. Partial least squares regression

Motivation

- PCR amounts of identifying linear combinations or directions that best represent the features/inputs/predictors. These directions are identified in an *unsupervised* way, i.e. the output/dependent variable/response does not supervise the identifications of the principal components.

- Directions that best explain the predictors are not necessarily the best ones to use for predicting the response.

- Partial Least Squares Regression (PLSR) is a dimension reduction technique that identifies linear combinations of the original variables in a *supervised* way, i.e. PLSR aims at finding directions that explain both the predictors (through the variance-covariance matrix) and the response (through the correlation between the dependent variable and the directions).

How does it work?

- Consider p standardized predictors. Let $z_1, \cdots, z_m$ denote $m < p$ linear combinations of original features:

$$z_k = \sum_{j=1}^{p} \phi_{j,k} x_j$$

for some constants $\phi_{1,k}, \cdots, \phi_{p,k}$ for $k = 1, \cdots, m$.

- The *first direction* z_1 is computed by setting each $\phi_{j,1}$ equal to the coefficient from the simple linear regression of y onto x_j:

$$y_i = x_{j,i}\phi_{j,1} + u_{j,i}.$$

By definition, the OLS estimate of $\phi_{j,1}$ is proportional to the correlation between y and x_j.

- By computing $z_1 = \sum_{j=1}^{p} \phi_{j,1} x_j$, the highest weight is placed on variables that are most strongly related to the response.

- To identify the *second direction*, z_2, one must whiten the response from the first direction, i.e. the second direction might capture the remaining information that has not been explained by the first PLS direction. This can be done in two steps:
  - The first step is thus to regress each variable on z_1 and then to take the corresponding residuals:

    $$x_j = z_1 \beta_j + u_j.$$

    The corresponding residual $\widehat{u}_j$ corresponds to the information unexplained by the first direction z_1.
  - The second step consists of using these *orthogonalized* variables in the same way as z_1 is computed.
- This iterative procedure can be repeated m times to identify the directions $z_1, \cdots, z_m$.
- In a final step, one regress y onto these m directions!

Algorithm: Partial Least squares regression

1. Standardize each predictor and set $\widehat{y}^{(0)} = \bar{y}e_n$ and $x_j^{(0)} = x_j$ for $j = 1, \cdots, p$.

2. For $m = 1, \cdots, p$:

   - ✓ $z_m = \sum_{j=1}^{p} \widehat{\phi}_{j,m} x_j^{(m-1)}$ where $\widehat{\phi}_{j,m} = \langle x_j^{(m-1)}, y \rangle = \left(x_j^{(m-1)} \right)^\top y$ is the OLS estimate of the simple linear regression of y onto $x_j^{(m-1)}$;
   - ✓ $\widehat{\theta}_m = \langle z_m, y \rangle / \langle z_m, z_m \rangle$ is the OLS estimate of the simple linear regression of y onto z_m;
   - ✓ $\widehat{y}^{(m)} = \widehat{y}^{(m-1)} + \widehat{\theta}_m z_m$;
   - ✓ Orthogonalize each $x_j^{(m-1)}$ w.r.t. z_m: $x_j^{(m)} = x_j^{(m-1)} - \left[\langle z_m, x_j^{(m-1)} \rangle / \langle z_m, z_m \rangle \right] z_m$ for $j = 1, \cdots, p$.

3. Recover the slope coefficients $\widehat{\beta}_j = \sum_{k=1}^{m} \widehat{\theta}_k \widehat{\phi}_{j,k}$ or regress y onto the m directions $z_1, \cdots, z_m$.

Maximization problem

- The mth PLS direction $\widehat{\phi}_m = (\widehat{\phi}_{m,1}, \cdots, \widehat{\phi}_{m,p})^\top$ solves:

$$\begin{aligned} \max_{\alpha} \quad & \texttt{Corr}^2(y, x\alpha)\, \mathbb{V}(x\alpha) \\ \text{s.t.} \quad & \|\alpha\| = 1 \quad \text{and} \quad \alpha^\top S\widehat{\phi}_\ell = 0, \text{for } \ell = 1, \cdots, m-1 \end{aligned}$$

  where S is the sample covariance (correlation) matrix.

- In contrast, the mth PCA direction v_m solves:

$$\begin{aligned} \max_{\alpha} \quad & \mathbb{V}(x\alpha) \\ \text{s.t.} \quad & \|\alpha\| = 1 \quad \text{and} \quad \alpha^\top S v_\ell = 0, \text{for } \ell = 1, \cdots, m-1 \end{aligned}$$

  where the conditions $\alpha^\top S v_\ell = 0$ ensure that the mth principal component is uncorrelated with previous ones.

- There is a trade-off between the correlation component and the variance component in the PLS regression.
- However, the variance term tends generally to dominate such that PLS regression behaves "much like" PC regressions (and Ridge regression).
- While the supervised dimension reduction of PLS generally reduces the bias at the expense of an increasing variance (especially w.r.t. PC regression).

Projection pursuit regression: a first pass

PC and PLS regressions

- Final step corresponds to the estimation of

$$y_i = \theta_0 + \sum_{k=1}^{m} \theta_k z_{i,k} + u_i, \quad i = 1, \cdots, n.$$

where

$$\begin{aligned} \sum_{k=1}^{m} \theta_k z_{i,k} &= \sum_{k=1}^{m} \theta_k \sum_{j=1}^{p} \phi_{j,k} x_{i,j} \\ &= \sum_{j=1}^{p} \sum_{k=1}^{m} \theta_k \phi_{j,k} x_{i,j} \\ &= \sum_{j=1}^{p} \beta_j x_{i,j} \end{aligned}$$

with $\beta_j = \sum_{k=1}^{m} \theta_k \phi_{j,k}$.

Projection pursuit regression considers the following additive model:

$$\begin{aligned} y_i &= \theta_0 + \sum_{j=1}^{m} f_i\left(x_i^\top \beta_j\right) + u_i, \quad i = 1, \cdots, n \\ &= \theta_0 + \sum_{j=1}^{m} f_i\left(z_j\right) + u_i, \quad i = 1, \cdots, n \end{aligned}$$

where f_j is a sequence of m initially unknown (single valued) smooth functions (Ridge functions).

Instead of modeling each response as a linear combination of the explanatory variables or directions, PPR models each response as a sum of functions of linear combination of the predictors. For a given set of data $\{(y_i, x_i)\}_{i=1}^n$, the minimization problem writes:

$$\min_{f_j, \beta_j} \quad \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{m} f_j(x_i^\top \beta_j) \right]^2.$$

- PPR is expected to perform better in the presence of significant nonlinearities, especially if nonlinearites are well approximated by Ridge function (i.e., functions that only vary in one direction in $\mathbb{R}^p$).

- PPR approximations are dense in the sense that any function of p variables can be arbitrarily closely approximated by Ridge function approximations for large enough m.

Technical appendix

Appendix 1: Principal component analysis with $p = 2$

First principal component...

- ✓ The Lagrangian function writes down:

$$\mathcal{L}(\lambda; a_1) = a_1^\top \Sigma a_1 - \lambda(a_1^\top a_1 - 1)$$

where λ is the Lagrange multiplier.

- ✓ F.O.C.

$$(\Sigma - \lambda I_p)\, a_1 = \mathbf{0}$$

where I_p is the identity matrix of order p.

- ✓ Therefore λ is an eigenvalue of Σ and a_1 is the corresponding eigenvector.

return

Second principal component...

- ✓ The Lagrangian function writes down:

$$\mathcal{L}(\lambda, \phi; a_1) = a_2^\top \Sigma a_2 - \lambda(a_2^\top a_2 - 1) - \phi a_2^\top a_1$$

where λ is the Lagrange multiplier.

- ✓ F.O.C.

$$(\Sigma - \lambda I_p)\, a_2 - \phi a_1 = \mathbf{0}$$

Therefore

$$a_1^\top (\Sigma - \lambda I_p)\, a_2 - \phi a_1^\top a_1 = \mathbf{0}$$

and thus $\phi = 0$ and

$$(\Sigma - \lambda I_p)\, a_2 = \mathbf{0}.$$

λ is an eigenvalue of Σ and a_2 is the corresponding eigenvector...□

return

Appendix 2: Singular Value Decomposition (SVD)

Definition

Let X denote an $n \times p$ matrix. The singular value decomposition (SVD) takes the form:

$$X = UDV^\top$$

where

- ✓ U is a $n \times p$ *orthogonal* matrix whose columns, u_j $j = 1, \cdots, p$, span the column space of X;
- ✓ V is a $p \times p$ *orthogonal* matrix whose columns, u_j $j = 1, \cdots, p$, span the row space of X;
- ✓ D is a $p \times p$ *diagonal* matrix with diagonal entries or singular values $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$.

return

Definition (Spectral decomposition)

Any symmetric matrix $\Sigma \in \mathcal{M}_{p\times p}$ can be written as

$$\Sigma = VDV^\top$$

where $D = \texttt{Diag}(d_1, \cdots, d_p)$ denotes the diagonal matrix of eigenvalues of Σ (descending order $d_1 > d_2 > ... > d_p$) and V is an orthogonal matrix, $VV^\top = V^\top V = I_p$, whose columns are the eigenvectors of length 1 (see SVD).

return

- Using the SVD of the centered matrix X, the (sample) covariance matrix has the following decomposition

$$\begin{aligned} S = \frac{X^\top X}{n} &= \frac{1}{n}\left(UDV^\top\right)^\top\left(UDV^\top\right) \\ &= \frac{1}{n}VDU^\top UDV^\top \\ &= \frac{1}{n}VD^2V^\top \end{aligned}$$

  which is the eigenvalue decomposition (**spectral decomposition theorem**) of S. Especially, the columns of V—so-called eigenvectors—are the principal components or Karhunen-Loeve) directions of X (up to a scaling factor).

return

Appendix 3: Data Preprocessing

Example: Let x_1 and x_2 denote two random variables. The variance-covariance matrix of (x_1, x_2) (respectively, $(10x_1, x_2)$) is given by Σ_1 (resp., Σ_2):

$$\Sigma_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$$

- The first principal component of Σ_1 is $0.707x_1 + 0.707x_2$;
- The first principal component of Σ_2 is $0.998x_1 + 0.055x_2$!

return

Appendix 4: Example PCA

- Data: 50 key UK government yield curves;
- Period: 2005-2007;
- Objectives: Extract some principal components and provide some "approximation" of each individual series.

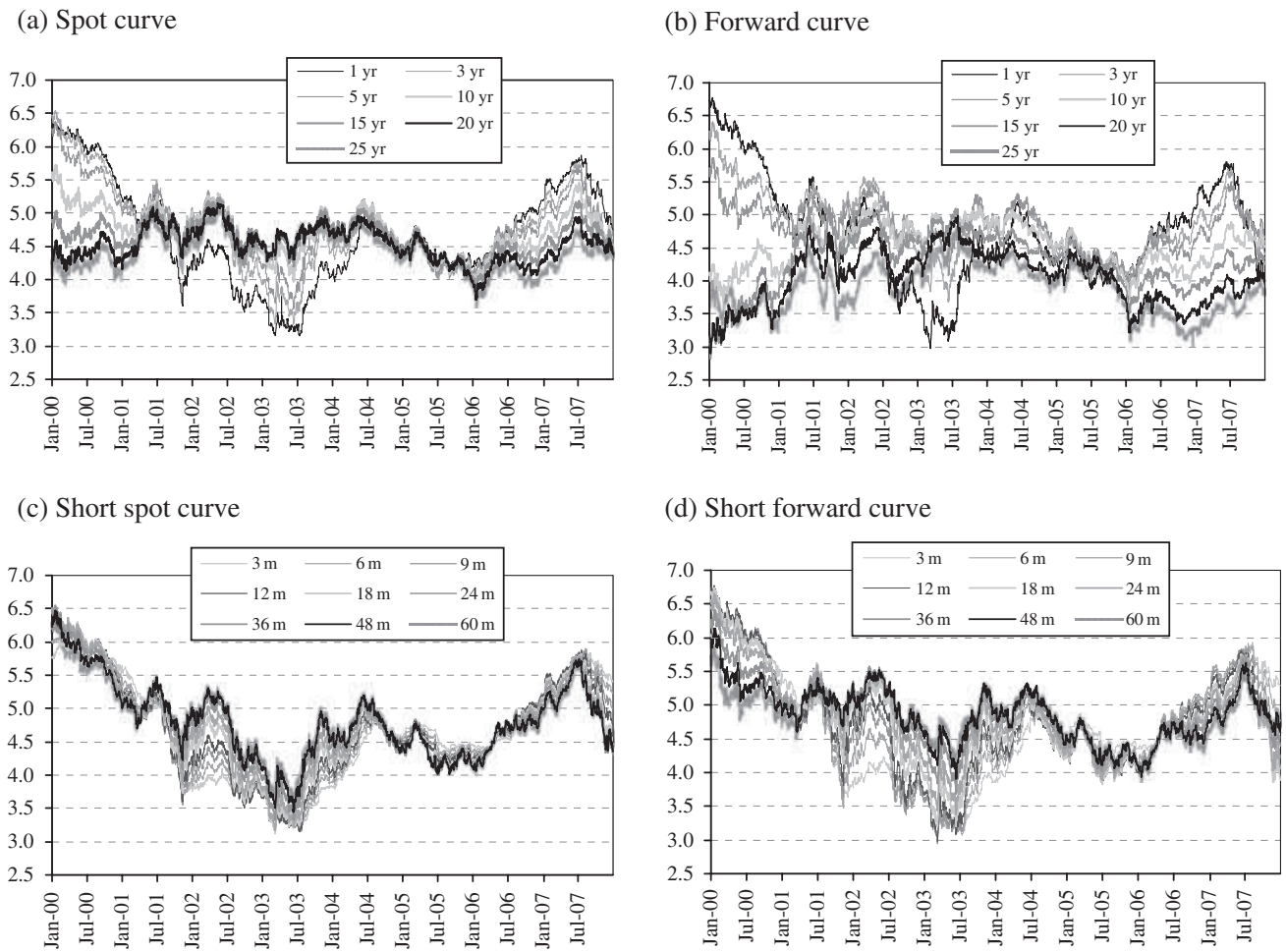

Figure: Initial series

Maturity	1 yr	2 yr	3 yr	4 yr	5 yr	7 yr	10 yr	15 yr	20 yr	25 yr
1 yr	1.000	0.925	0.877	0.843	0.809	0.744	0.674	0.615	0.558	0.501
2 yr	0.925	1.000	0.990	0.972	0.947	0.891	0.827	0.773	0.717	0.657
3 yr	0.877	0.990	1.000	0.994	0.979	0.937	0.883	0.833	0.781	0.723
4 yr	0.843	0.972	0.994	1.000	0.995	0.968	0.924	0.880	0.831	0.776
5 yr	0.809	0.947	0.979	0.995	1.000	0.987	0.955	0.917	0.871	0.819
7 yr	0.744	0.891	0.937	0.968	0.987	1.000	0.989	0.963	0.923	0.877
10 yr	0.674	0.827	0.883	0.924	0.955	0.989	1.000	0.989	0.957	0.918
15 yr	0.615	0.773	0.833	0.880	0.917	0.963	0.989	1.000	0.988	0.962
20 yr	0.558	0.717	0.781	0.831	0.871	0.923	0.957	0.988	1.000	0.992
25 yr	0.501	0.657	0.723	0.776	0.819	0.877	0.918	0.962	0.992	1.000

Figure: Correlation matrix of selected UK spot rates

Table: Eigenvalue decomposition

Component	1	2	3	4	5	6
Eigenvalue	45.524	3.424	0.664	0.300	0.062	0.019
% Variation	91.05%	6.85%	1.33%	0.60%	0.12%	0.04%
Cumulative %	91.05%	97.90%	99.22%	99.82%	99.95%	99.98%

$\Rightarrow$ The first three components together explain over 99% of the variation...

Table: Eigenvectors

Eigenvector	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6
6 mth	0.0675	0.3464	0.6878	0.4409	0.3618	0.2458
1 yr	0.1030	0.3536	0.3272	0.007	-0.4604	-0.4910
⋮			⋮			⋮
10 yr	0.1471	0.034	-0.0970	0.1669	-0.0390	-0.0430
⋮			⋮			⋮
25 yr	0.1400	-0.1541	0.1535	-0.1633	0.1037	-0.1979

$\Rightarrow$

$$\Delta R_{6\text{mth}} \approx \mathbf{0.0675\Gamma_1 + 0.3464\Gamma_2 + 0.6878\Gamma_3}$$

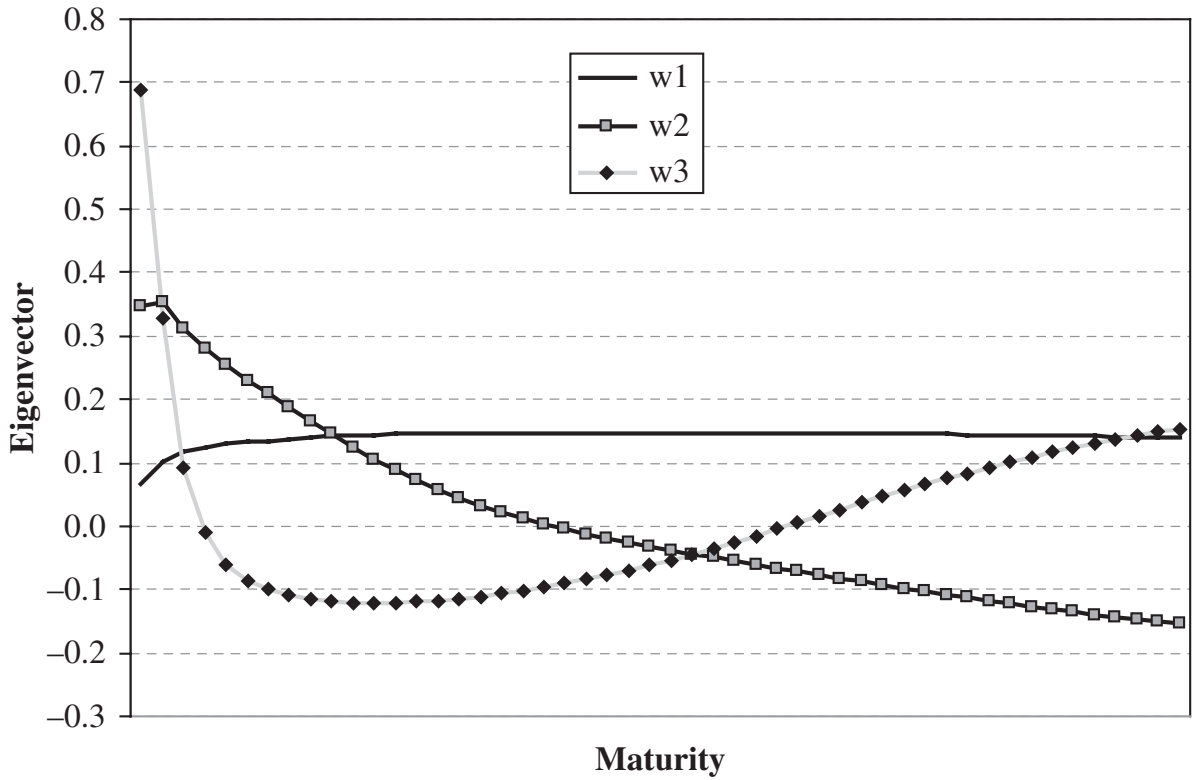

Figure: Eigenvectors of the UK daily spot rate correlation matrix

return

Appendix 5: Bivariate normal pdf

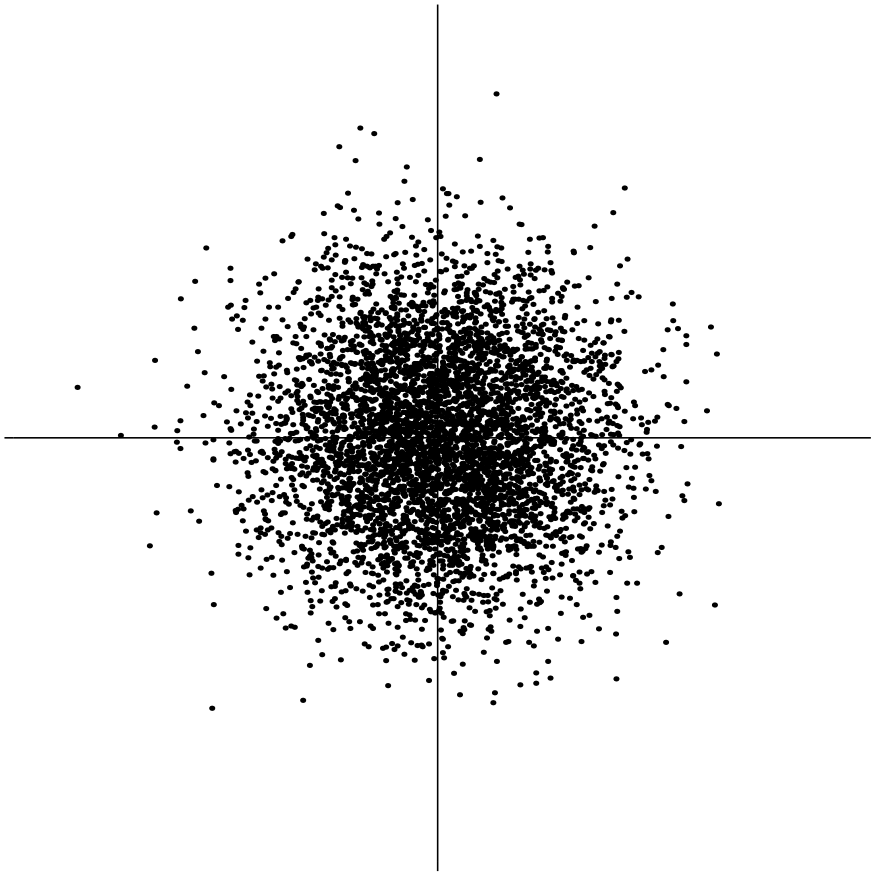

Figure: Joint distribution of two independent Gaussian variables

return