Principal Component Analysis

2019 Winter Camp at PHBS

Xianhua Peng
Associate Professor
PHBS, Peking University

January 2019

Outline

1. Principal Component Analysis (PCA)
2. Empirical Principal Component Analysis
3. Applications of Principal Component Analysis

Outline

1. Principal Component Analysis (PCA)
2. Empirical Principal Component Analysis
3. Applications of Principal Component Analysis

Motivation

Let $X = (X_1, X_2, \ldots, X_n)^T$ be a n-dimensional random vector

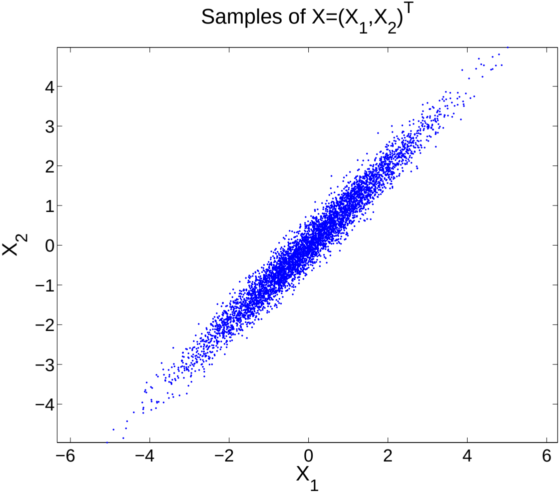
- X_i denotes the change of i-th risk factor over a specified time horizon
- These risk factors could represent
  - security prices
  - security returns
  - zero rates at different maturities
- $X_1, X_2, \ldots, X_n$ are generally correlated: $\text{Cov}(X_i, X_j) \neq 0$

Question: can we find a matrix $A \in \mathbb{R}^{n\times n}$, such that the components of random vector $Y = AX$ are uncorrelated with each other?

$$\text{Cov}(Y_i, Y_j) = 0, \ \forall i \neq j$$

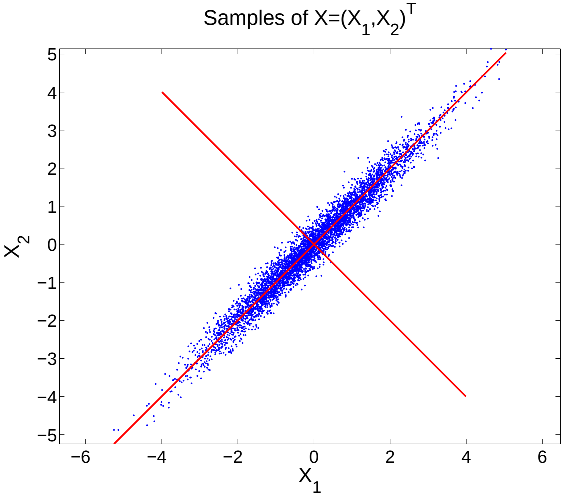
Motivation (continued)

Example of a random vector $X = (X_1, X_2)^T$ with $\mathrm{Cov}(X_1, X_2) > 0$



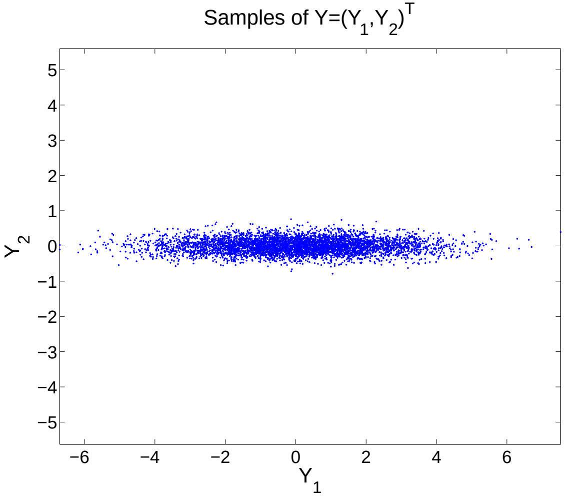
Motivation (continued)

Find two projection directions



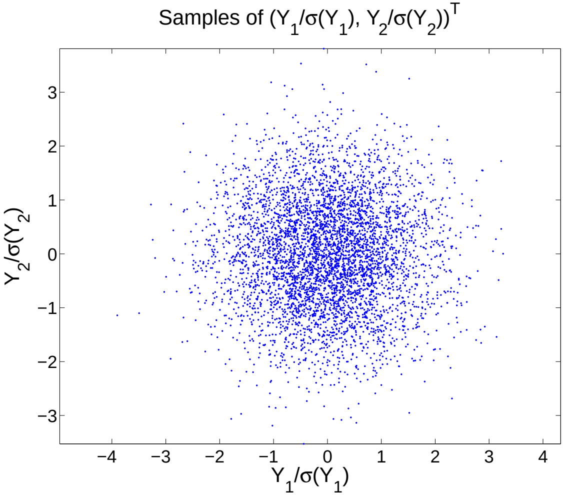
Motivation (continued)

The projected random vector $Y = (Y_1, Y_2)^T$ has uncorrelated components: $\mathrm{Cov}(Y_1, Y_2) = 0$



Motivation (continued)

The normalized random vector $\left(\frac{Y_1}{\sigma(Y_1)}, \frac{Y_2}{\sigma(Y_2)}\right)^T$



Normalization

- We first normalize each component of $X = (X_1, X_2, \ldots, X_n)^T$:
  - Remove the mean of X_i: $X_i \leftarrow X_i - E[X_i]$
  - Normalize the variance of X_i: $X_i \leftarrow \frac{X_i}{\sqrt{\mathrm{Var}(X_i)}}$
  - The normalized random vector is a linear transform of X
- After normalization, $\mu = E[X] = 0$ and $\mathrm{Var}(X_i) = 1$ for all $i = 1, 2, \ldots, n$.
- Normalizing the variance is not necessary but is a common practice.

Formulate the Problem

Let $X = (X_1, X_2, \ldots, X_n)^T$ be a random vector with $E[X] = 0$ (Var(X_i) is not necessarily equal to 1)

- Let $\Sigma := E[XX^T]$ be the variance-covariance matrix of X
- Let $Y = AX$, where $A \in \mathbb{R}^{n\times n}$ is the matrix to be determined:

$$\mathrm{Cov}(Y) = A\Sigma A^T$$

- Can we find A such that $A\Sigma A^T$ is a diagonal matrix?

The key method is the **spectral decomposition** of symmetric matrix.

Spectral Decomposition

Spectral decomposition: any real-valued symmetric matrix $\Sigma \in \mathbb{R}^{n\times n}$ can be decomposed into

$$\Sigma = \Gamma\Lambda\Gamma^T, \text{ where}$$

- Λ is a diagonal matrix $\text{diag}\{\lambda_1, \ldots, \lambda_n\}$, where $\lambda_1, \ldots, \lambda_n$ are the **eigen values** of Σ
  - Without loss of generality the eigen values can be ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - If Σ is positive semi-definite, then $\lambda_i \geq 0$, for all i
- Γ is an orthogonal matrix with the i-th column $\gamma_i \in \mathbb{R}^n$ being the i-th standardized **eigen vector** of Σ
  - $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$
  - γ_i is the i-th eigen vector: $\Sigma\gamma_i = \lambda_i\gamma_i$
  - γ_i is standardized: $\gamma_i^T\gamma_i = 1$
  - Γ is an orthogonal matrix: $\Gamma\Gamma^T = \Gamma^T\Gamma = I_n$

Principal Components Analysis (PCA)

Let $\Sigma = \Gamma\Lambda\Gamma^T$ be the spectral decomposition of Σ and define

$$Y = \Gamma^T X$$

- The covariance matrix of Y is

$$\mathrm{Cov}(Y) = \Gamma^T \Sigma (\Gamma^T)^T = \Gamma^T \Gamma \Lambda \Gamma^T (\Gamma^T)^T = \Lambda$$

- Components of Y are uncorrelated and Y is called the **principal components** of X
- The i-th component of Y is $Y_i = \gamma_i^T X$:

$$Y = \Gamma^T X = (\gamma_1, \gamma_2, \ldots, \gamma_n)^T X = \begin{pmatrix} \gamma_1^T \\ \gamma_2^T \\ \cdots \\ \gamma_n^T \end{pmatrix} X = \begin{pmatrix} \gamma_1^T X \\ \gamma_2^T X \\ \cdots \\ \gamma_n^T X \end{pmatrix}.$$

- $\mathrm{Var}(Y_i) = \lambda_i$
- The matrix Γ^T is called the matrix of **factor loadings**.

Principal Components Analysis (continued)

Let $Y = \Gamma^T X$ be the principal components of X. Then

$$X = \Gamma Y = (\gamma_1, \gamma_2, \ldots, \gamma_n) \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix} = Y_1\gamma_1 + Y_2\gamma_2 + \ldots + Y_n\gamma_n$$

- X is a random linear combination of $\gamma_1, \gamma_2, \ldots, \gamma_n$; the combination coefficients are random and are uncorrelated

Explain Variability of X by Its Principal Components

Let $Y = \Gamma^T X$ be the principal components of X

- Total variance of principal components:

$$\sum_{i=1}^{n} \mathrm{Var}(Y_i) = \sum_{i=1}^{n} \lambda_i = \mathrm{trace}(\Lambda) = \mathrm{trace}(\Gamma^T \Sigma \Gamma)$$

$$= \mathrm{trace}(\Gamma\Gamma^T \Sigma) = \mathrm{trace}(\Sigma) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

- The proportion of variability explained by the i-th principal component

$$\frac{\mathrm{Var}(Y_i)}{\sum_{i=1}^{n} \mathrm{Var}(X_i)} = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

- The proportion of variability explained by the first k principal components

$$\frac{\sum_{i=1}^{k} \mathrm{Var}(Y_i)}{\sum_{i=1}^{n} \mathrm{Var}(X_i)} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

Explain Variability of X by Its Principal Components

Let $Y = \Gamma^T X$ be the principal components of X

- The first component $Y_1 = \gamma_1^T X$ satisfies

$$\mathrm{Var}(\gamma_1^T X) = \max\{\mathrm{Var}(a^T X) : a^T a = 1\}$$

- γ_1 is the projection direction along which the variability of X is the largest
- The second component $Y_2 = \gamma_2^T X$ satisfies

$$\mathrm{Var}(\gamma_2^T X) = \max\{\mathrm{Var}(a^T X) : a^T a = 1, a^T \gamma_1 = 0\}$$

where $a^T \gamma_1 = 0$ means that a is orthogonal to γ_1, or equivalently, $a^T X$ is orthogonal to Y_1 (i.e., $E(a^T X \cdot Y_1) = 0$).

- The successive principal components $Y_i = \gamma_i^T X$ satisfy the optimization problem

$$\mathrm{Var}(\gamma_i^T X) = \max\{\mathrm{Var}(a^T X) : a^T a = 1, a^T \gamma_j = 0, j = 1, 2, \ldots, i-1\}.$$

Outline

1. Principal Component Analysis (PCA)
2. Empirical Principal Component Analysis
3. Applications of Principal Component Analysis

Empirical PCA

- In practice, the true variance-covariance matrix is unknown and may only be estimated from historical data.
- Suppose we observe $X^t = (X_1^t, X_2^t, \ldots, X_n^t)^T$, $t = 1, 2, \ldots, T$: X^t is the t-th observation (e.g, the sample observed at time t)
- It is important to make sure that $\{X^t, t = 1, 2, \ldots, T\}$ is (weakly) stationary
  - $E(X^1) = E(X^2) = \cdots = E(X^T)$
  - $\mathrm{Cov}(X^1, X^1) = \mathrm{Cov}(X^2, X^2) = \cdots = \mathrm{Cov}(X^T, X^T)$
- If X denotes asset returns or yields, $\{X^t, t = 1, 2, \ldots, T\}$ is usually assumed to be stationary
- If X denotes asset prices, $\{X^t, t = 1, 2, \ldots, T\}$ usually have a trend and may not be assumed to be stationary

Empirical PCA (continued)

Demean the sample

- Let $\hat{\mu}_i := \frac{1}{T}\sum_{t=1}^{T} X_i^t$ be the sample mean of X_i, $i = 1, 2, \ldots, n$
- Let $Z_i^t := X_i^t - \hat{\mu}_i,\ i = 1, 2, \ldots, n, t = 1, \ldots, T$

$$Z^t = \begin{pmatrix} Z_1^t \\ Z_2^t \\ \cdots \\ Z_n^t \end{pmatrix} = X^t - \hat{\mu},\ t = 1, \ldots, T$$

Empirical PCA (continued)

- Compute the sample covariance matrix

$$\Sigma = \frac{1}{T}\sum_{t=1}^{T} Z^t (Z^t)^T$$

- Spectral decomposition of $\Sigma = \Gamma\Lambda\Gamma^T$
- Compute principal components

$$Y^t = \Gamma^T Z^t = \Gamma^T \cdot (X^t - \hat{\mu})$$

The original data X^t can be recovered from principal components by

$$X^t = \Gamma Y^t + \hat{\mu}.$$

Outline

1 Principal Component Analysis (PCA)

2 Empirical Principal Component Analysis

3 Applications of Principal Component Analysis

Zero-coupon Bond Yields

- Continuously compounded zero-coupon bond yield y_t:

$$D(0,t) = \frac{1}{e^{y_t \cdot t}} = e^{-y_t \cdot t}$$

- zero-coupon bond yield y_t is also called **zero rate** or **spot rate**.
- y_t is a function of t: it is called the **zero-coupon bond yield curve**.
- The zero-coupon bond yield curve is also called the **term structure of interest rates**.

Zero-coupon Bond Yields (continued)

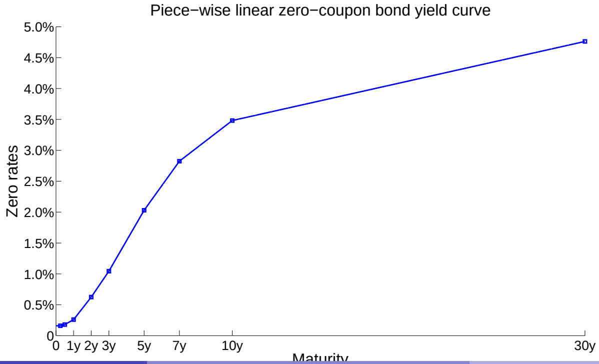
- The price of a coupon bond with face value F, coupon payments C_i at time T_i, $i = 1, 2, \ldots, m$ and maturity $T = T_m$:

$$\begin{aligned} P &= \sum_{i=1}^{m-1} D(0, T_i) \cdot C_i + D(0, T_m) \cdot (C_m + F) \\ &= \sum_{i=1}^{m-1} e^{-y_{T_i} \cdot T_i} \cdot C_i + e^{-y_{T_m} \cdot T_m} \cdot (C_m + F) \end{aligned} \tag{1}$$

Piece-wise Linear Zero-coupon Bond Yield Curve

- Assumptions (models) of the zero-coupon bond yield curve
  - Piece-wise linear assumption: y_t is constant on $[0, 0.25]$ and y_t is linear on each interval between the time points 0.25, 0.5, 1, 2, 3, 5, 7, 10, and 30

$$y_t = y_{T_{i-1}} + \frac{y_{T_i} - y_{T_{i-1}}}{T_i - T_{i-1}}(t - T_{i-1}), \forall t \in (T_{i-1}, T_i], i = 2, \ldots, 8$$



PCA of Change of Zero-coupon Curve

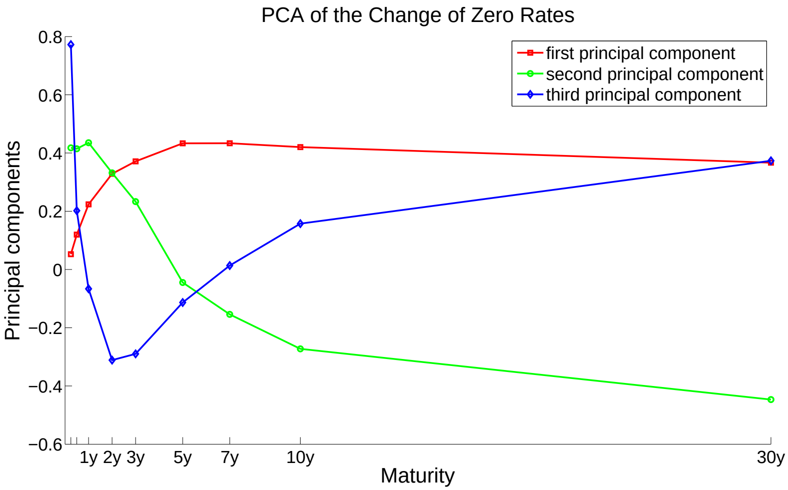
- Historical data of discount factors: “s0023_disc_factors_hist.xls"
- Compute the zero rates from the discount factors:

$$Z_i^t := -\frac{\log D(t, t+T_i)}{T_i}, i = 1, 2, \ldots, 9, t = 1, 2, \ldots, S$$

- $Z^t = (Z_1^t, Z_2^t, \ldots, Z_9^t)^T$ is the zero-coupon yield curve observed on day t, $t = 1, 2, \ldots, S$
- Compute change of zero rates: $X^t := Z^t - Z^{t-1}$
- PCA in Python: sklearn.decomposition.PCA in the package scikit-learn

PCA of Change of Zero-coupon Curve (continued)

The graph of γ_1, γ_2, and γ_3



PCA of Change of Zero-coupon Curve (continued)

- The first principal component γ_1 can usually be interpreted as (approximate) parallel shift in the zero curve
- The second component represents a flattening or steepening of the zero curve
- The third component represents the curvature of the curve
- The first three principal components explain 95% of the total variability:

$$\frac{\lambda_1}{\sum_{i=1}^{9}\lambda_i} = 0.76, \frac{\lambda_1+\lambda_2}{\sum_{i=1}^{9}\lambda_i} = 0.89, \frac{\sum_{i=1}^{3}\lambda_i}{\sum_{i=1}^{9}\lambda_i} = 0.95$$

Using PCA to Building Factor Models

- Recall that $X = \Gamma Y + \mu$
- Partition the matrix Γ into $\Gamma = [\Gamma_1, \Gamma_2]$, where

$$\Gamma_1 = (\gamma_1, \gamma_2, \ldots, \gamma_k) \in \mathbb{R}^{n \times k}, \Gamma_2 = (\gamma_{k+1}, \ldots, \gamma_n) \in \mathbb{R}^{n \times (n-k)}$$

- $X = (\gamma_1 Y_1 + \cdots + \gamma_k Y_k) + (\gamma_{k+1} Y_{k+1} + \cdots + \gamma_n Y_n) + \mu$
- If first k principal components explain sufficiently large amount of total variability, then $\varepsilon := \gamma_{k+1} Y_{k+1} + \cdots + \gamma_n Y_n$ is small in magnitude and can be viewed as noise, and

$$X = \gamma_1 Y_1 + \cdots + \gamma_k Y_k + \epsilon + \mu$$

This is a k-factor model for X.

Using PCA to Building Factor Models (continued)

ε can be ignored to obtain an exact k-factor model

$$X = \gamma_1 Y_1 + \cdots + \gamma_k Y_k + \mu$$

- The joint distribution of the first k principal components can be estimated by using data $Y_i^1, Y_i^2, \ldots, Y_i^S$, $i = 1, 2, \ldots, k$
  - $(Y_1, Y_2, \ldots, Y_k)^T$ can be assumed to have a $N_k(0, \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_k\})$ distribution
  - $(Y_1, Y_2, \ldots, Y_k)^T$ can be assumed to have a multivariate-t distribution
- This reduces the number of risk factors from n to $k \ll n$.

Portfolio Immunization

- A fund manager at a pension fund or a life insurance company needs to pay out a stream of liability ℓ_i at time T_i, $i = 1, 2, \ldots, n$, to retirement policy holders.
- The current zero-coupon yield curve is $\{y_t, t \geq 0\}$.
- The present value $L(y)$ of the liability stream is $L(y) = \sum_{i=1}^{n} e^{-y_{T_i} \cdot T_i} \cdot \ell_i$.
- The fund manager is facing interest rate risk: when interest rates change from y to $y + X$, $L(y)$ changes to

$$L(y + X) = \sum_{i=1}^{n} e^{-(y_{T_i} + X_{T_i}) \cdot T_i} \cdot \ell_i$$

Portfolio Immunization (continued)

- The fund manager is managing an investment portfolio which will generate cash to pay off the liability
- The portfolio is composed of x_i units of bond i, $i = 1, 2, \ldots, n$.
- The currnt value of the portfolio: $P(y) = \sum_{i=1}^{n} x_i \cdot P_i(y)$
- When interest rates change from y to $y + X$, the value of the portfolio changes to $P(y + X) = \sum_{i=1}^{n} x_i \cdot P_i(y + X)$

Portfolio Immunization (continued)

We want to construct a portfolio with x_i units of bond P_i such that the portfolio value matches the liability value no matter how the zero curve changes:

$$L(y+X) \approx P(y+X), \text{ for small random change } X$$

Portfolio Immunization

Portfolio of bonds can be immunized under the k-factor model of the change of zero-coupon yield curve

- Let $y = (y_1, y_2, \ldots, y_n)^T$ be the current zero-coupon yield curve, where y_i is the zero rate of maturity T_i, $i = 1, 2, \ldots, n$.
- Bond portfolio value as a function of zero-coupon yield curve

$$P(y) = P(y_1, y_2, \ldots, y_n) = \sum_{j=1}^{m} e^{-y_{t_j} \cdot t_j} C_j$$

- Note that in general y_{t_j} is a function of y.
- The k-factor model of the change of zero curve obtained by PCA

$$X = \gamma_1 Y_1 + \cdots + \gamma_k Y_k + \mu$$

- Change of portfolio value

$$P(y + X) - P(y) = P(\gamma_1 Y_1 + \gamma_2 Y_2 + \cdots + \gamma_k Y_k + \mu + y) - P(y)$$

- Risk factors are the principal components $Y_1, Y_2, \ldots, Y_k$

Portfolio Immunization (continued)

Change of portfolio value

$$
\begin{aligned}
P(y+X)-P(y) &= P(\gamma_1 Y_1 + \gamma_2 Y_2 + \cdots + \gamma_k Y_k + \mu + y) - P(y) \\
&\approx \left(\frac{\partial P(y)}{\partial y_1}, \frac{\partial P(y)}{\partial y_2}, \ldots, \frac{\partial P(y)}{\partial y_n}\right)(\gamma_1 Y_1 + \cdots + \gamma_k Y_k + \mu) \\
&= \nabla P(y)^T (\gamma_1 Y_1 + \gamma_2 Y_2 + \cdots + \gamma_k Y_k + \mu) \\
&= \nabla P(y)^T \gamma_1 Y_1 + \nabla P(y)^T \gamma_2 Y_2 + \cdots + \nabla P(y)^T \gamma_k Y_k + \nabla P(y)^T \mu
\end{aligned}
$$

Percentage change of portfolio value

$$
\frac{P(y+X)-P(y)}{P(y)} = \frac{\nabla P(y)^T \gamma_1}{P(y)} Y_1 + \cdots + \frac{\nabla P(y)^T \gamma_k}{P(y)} Y_k + \frac{\nabla P(y)^T \mu}{P(y)}
$$

- $\frac{\nabla P(y)^T \gamma_i}{P(y)}$ is the sensitivity of the percentage change of portfolio w.r.t. the i-th factor

Portfolio Immunization (continued)

Consider a bond portfolio with x_i units of bond i with price $P_i(y)$, $i = 1, 2, \ldots, N$

- Total portfolio value: $P(y) = \sum_{i=1}^{N} x_i P_i(y)$
- Change in the value of the bond portfolio:

$$P(y+X) - P(y) = \sum_{i=1}^{N} x_i (P_i(y+X) - P_i(y))$$

$$\approx \sum_{i=1}^{N} x_i \left(\nabla P_i(y)^T \gamma_1 Y_1 + \nabla P_i(y)^T \gamma_2 Y_2 + \cdots + \nabla P_i(y)^T \gamma_k Y_k \right)$$

$$+ \sum_{i=1}^{N} x_i \nabla P_i(y)^T \mu$$

Portfolio Immunization (continued)

- Percentage change of the portfolio value

$$
\begin{aligned}
&\frac{P(y+X)-P(y)}{P(y)} \\
&= \sum_{i=1}^{N} \frac{x_i P_i(y)}{P(y)} \frac{\nabla P_i(y)^T \gamma_1}{P_i(y)} Y_1 + \cdots + \sum_{i=1}^{N} \frac{x_i P_i(y)}{P(y)} \frac{\nabla P_i(y)^T \gamma_k}{P_i(y)} Y_k \\
&+ \sum_{i=1}^{N} x_i \frac{\nabla P_i(y)^T \mu}{P(y)}
\end{aligned}
$$

Portfolio Immunization (continued)

Let $L(y)$ be the present value of the future stream of liabilities. We want to construct a portfolio with x_i units of bond $P_i(y)$ such that the portfolio is immunized against the move of the zero curve

$$L(y+X) \approx P(y+X), \text{ for small random change } X$$

- Matching current present value:

$$L(y) = \sum_{i=1}^{N} x_i P_i(y)$$

- Matching sensitivities w.r.t. principal components

$$\frac{\nabla L(y)^T \gamma_j}{L(y)} = \sum_{i=1}^{N} \frac{x_i P_i(y)}{P(y)} \frac{\nabla P_i(y)^T \gamma_j}{P_i(y)}, j = 1, \ldots, k$$

$$\frac{\nabla L(y)^T \mu}{L(y)} = \sum_{i=1}^{N} \frac{x_i P_i(y)}{P(y)} \frac{\nabla P_i(y)^T \mu}{P_i(y)}$$