Term Structure Models

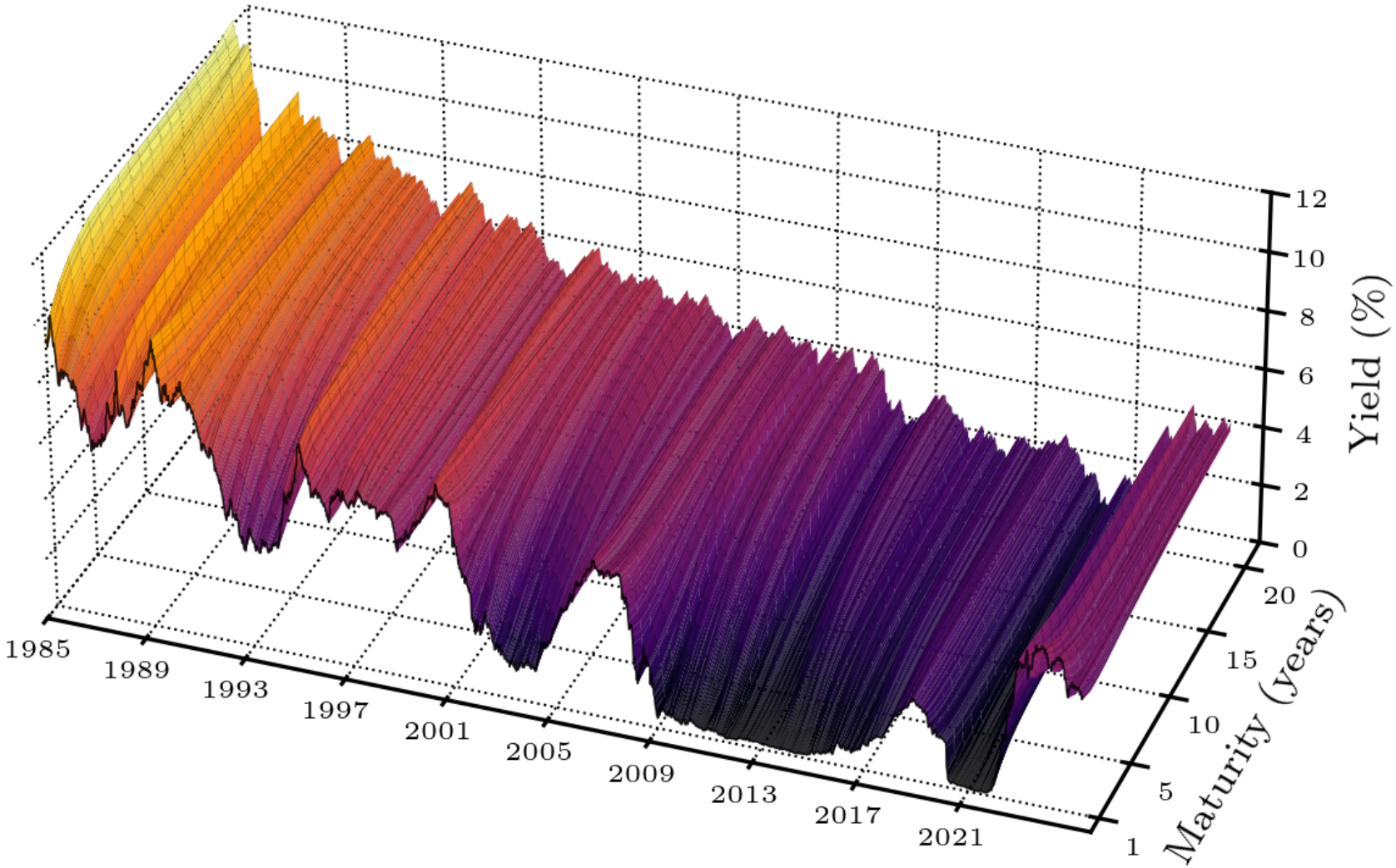
Week 4

University of Copenhagen

Last updated: May 15, 2025

Outline

US Treasury yield curve (1985 to today)



Data from Gürkaynak et al. (2007)

PART 1:
YIELD CURVE BASICS

Three curves, one story

- The main object of investigation for these lectures is the *discount curve*

$$\{P(t,T)\}_{T\geq t},$$

  where $P(t,T)$ is the time-t price of a bond paying one unit of currency at a later time $T \geq t$.
- The discount curve can also be represented in terms of the yields, where

$$P(t,T) = e^{-y(t,T)(T-t)} \;\Leftrightarrow\; y(t,T) = -\frac{\log P(t,T)}{T-t}.$$

  An alternative starting point would then be to consider the *yield curve*

$$\{y(t,T)\}_{T\geq t}.$$

- A third option would be to use forward rates, defined by $P(t,T) = e^{-\int_t^T f(t,u)\mathrm{d}u}$, resulting in the *forward rate curve*

$$\{f(t,T)\}_{T\geq t}.$$

These three curves contain the same information, so we can choose the one most convenient for us. We will mostly focus on the yield curve.

Yield curve construction meet messy reality

- We do not observe any of the above-mentioned curves directly, so we must infer them from market observables.
- Consider a collection of N securities whose time-t prices can be written as a linear combination of M discount bonds

$$V_i(t) = \sum_{j=1}^{M} c_{ij} P(t, T_j),$$

  for an increasing series of maturities $t < T_1 < \cdots < T_M \leq T^*$.
- Many securities can be written in the above form, including coupon bonds, fixed-floating swaps and forward rate agreements.
- One must however be careful when selecting securities. Ideally, we want a set of securities with
  - no option-like features
  - similar liquidity
  - no market segmentation
- **The goal:** we wish to go from a finite collection of these securities to a complete yield curve $y(t, T)$ for $T \geq t$:

$$\{V_i(t)\}_{i=1,\ldots,N} \longrightarrow \{y(t, T)\}_{T \geq t}.$$

Yield curve construction methods

- One common way to solve the above is using splines. These methods range from simple piecewise linear yield curves to very complex splines. The more advanced methods are often used to ensure that the resulting curve exhibits certain desirable properties (e.g, sufficient smoothness).
- We will however consider a more parsimonious approach, where we assume some parametric functional form on the yield curve, and find the parameters θ with the best (approximate) fit to the securities,

$$\theta^* \triangleq \min_{\theta \in \Theta} \sum_{i=1}^{N} w_i \left(V_i^{\text{True}} - V_i^{\text{Model}}(\theta) \right)^2$$

  where θ is the vector of parameters, w_i are weights and $V_i^{\text{Model}}(\theta)$ is the price of security i when priced using the model-implied yield curve.
- Preferably, we want a functional form that is
  - simple and tractable,
  - ... but flexible enough to fit the many different empirically observed curves.
- We next consider one such functional form.

The Nelson-Siegel model

The arguably most well-known model of this type is the one proposed by Nelson and Siegel (1987), stated below in terms of forward rates.

Nelson and Siegel (1987) model
The time-0 forward rate at time τ is given by

$$f^{[\mathrm{NS}]}(\tau;\theta) = \beta_0 + \beta_1 e^{-\lambda\tau} + \beta_2 \lambda\tau e^{-\lambda\tau}$$

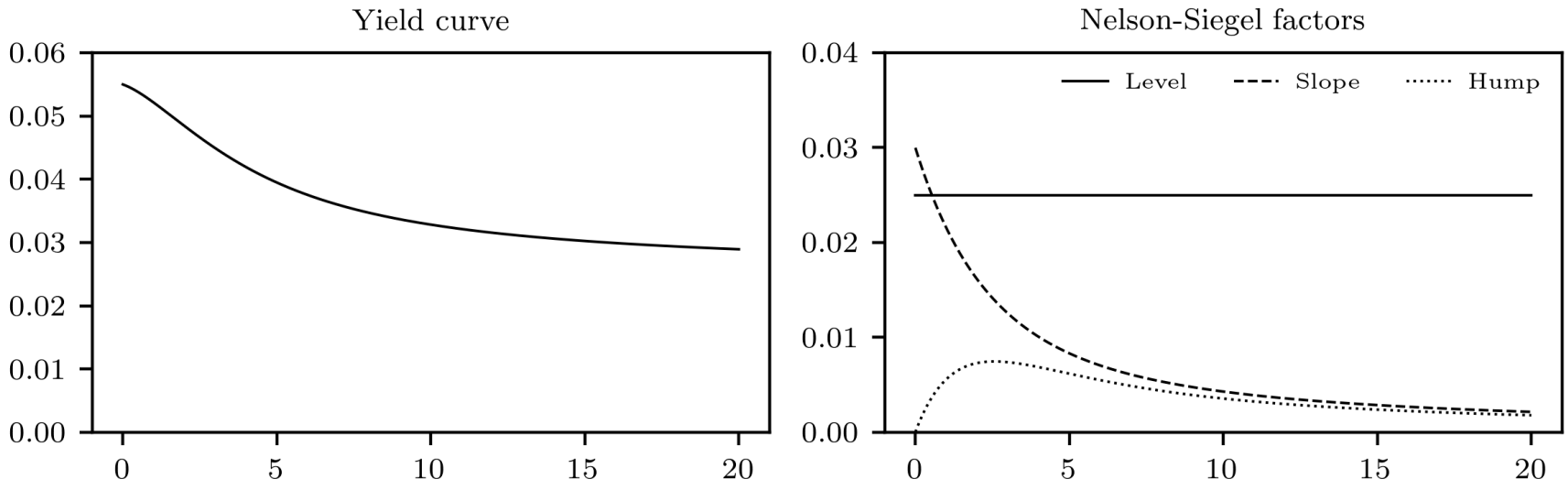
for parameters $\theta = (\beta_0, \beta_1, \beta_2, \lambda)$ with $\lambda > 0$.

Notice that

$$f^{[\mathrm{NS}]}(0) = \beta_0 + \beta_1,$$
$$\lim_{\tau\to\infty} f^{[\mathrm{NS}]}(\tau) = \beta_0.$$

Thus β_0 is the asymptote of the forward rate curve.

Decomposing the Nelson-Siegel yield curve



Example of a Nelson-Siegel yield curve with $\lambda = 0.7$

The equivalent representation in terms of yields is given by

$$y^{[\mathrm{NS}]}(\tau;\theta) = \frac{1}{\tau}\int_0^\tau f^{[\mathrm{NS}]}(u;\theta)\mathrm{d}u$$

$$= \underbrace{\beta_0}_{\text{Level}} + \underbrace{\beta_1 \frac{1-e^{-\lambda\tau}}{\lambda\tau}}_{\text{Slope}} + \underbrace{\beta_2\left[\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right]}_{\text{Hump}}.$$

Adding more flexibility: the Svensson (1994) extension

A problem: While the Nelson-Siegel model can fit shorter maturities well, it struggles with longer maturities (say, 15+ years). The forward rates simply asymptote too quickly to be able to capture the convexity at longer maturities.

The solution: Svensson (1994) proposed an extension:

$$f^{[\text{SV}]}(\tau) = \underbrace{\beta_0 + \beta_1 e^{-\lambda_1 \tau} + \beta_2 \lambda_1 \tau e^{-\lambda_1 \tau}}_{\text{Nelson-Siegel}} + \underbrace{\beta_3 \lambda_2 \tau e^{-\lambda_2 \tau}}_{\text{Extra term}}$$

for parameters $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \lambda_1, \lambda_2)$, where $\lambda_1, \lambda_2 > 0$. This additional hump is generally needed to accurately fit to longer maturity yields.

We could go on with adding more of these terms. In fact, both Nelson and Siegel (1987) and Svensson (1994) are specific instances of the so-called exponential polynomial family

$$f^{[\text{EP}]}(\tau) = \sum_{i=1} p_i(\tau) e^{-\kappa_i \tau},$$

where $p_i(\tau)$ are polynomials in τ for all i.

Overview of US Treasury debt securities

US Treasury debt securities

Name	Coupon	Maturities	Principal
Treasury Bills	None	A few weeks to a year	Fixed
Treasury Notes	Semi-annual	Between two and ten years	Fixed
Treasury Bonds	Semi-annual	Thirty years[1]	Fixed
TIPS[2]	Semi-annual	Between five and twenty years	Inflation-adjusted

- Largest in the world at $\approx$ \$28 trillion.
- TIPS are special securities that are protected from inflation risk. The yields from TIPS are therefore not directly comparable to the yields from the other main types of treasury securities.
- Other types of securities have also been issued, including securities with option-like features.

We will next consider the yield curve construction method employed by Gürkaynak et al. (2007) for US Treasury securities.

[1] Before 1985 it was twenty years.

[2] Short for *Treasury Inflation-Protected Securites*.

Data selection methodology from Gürkaynak et al. (2007)

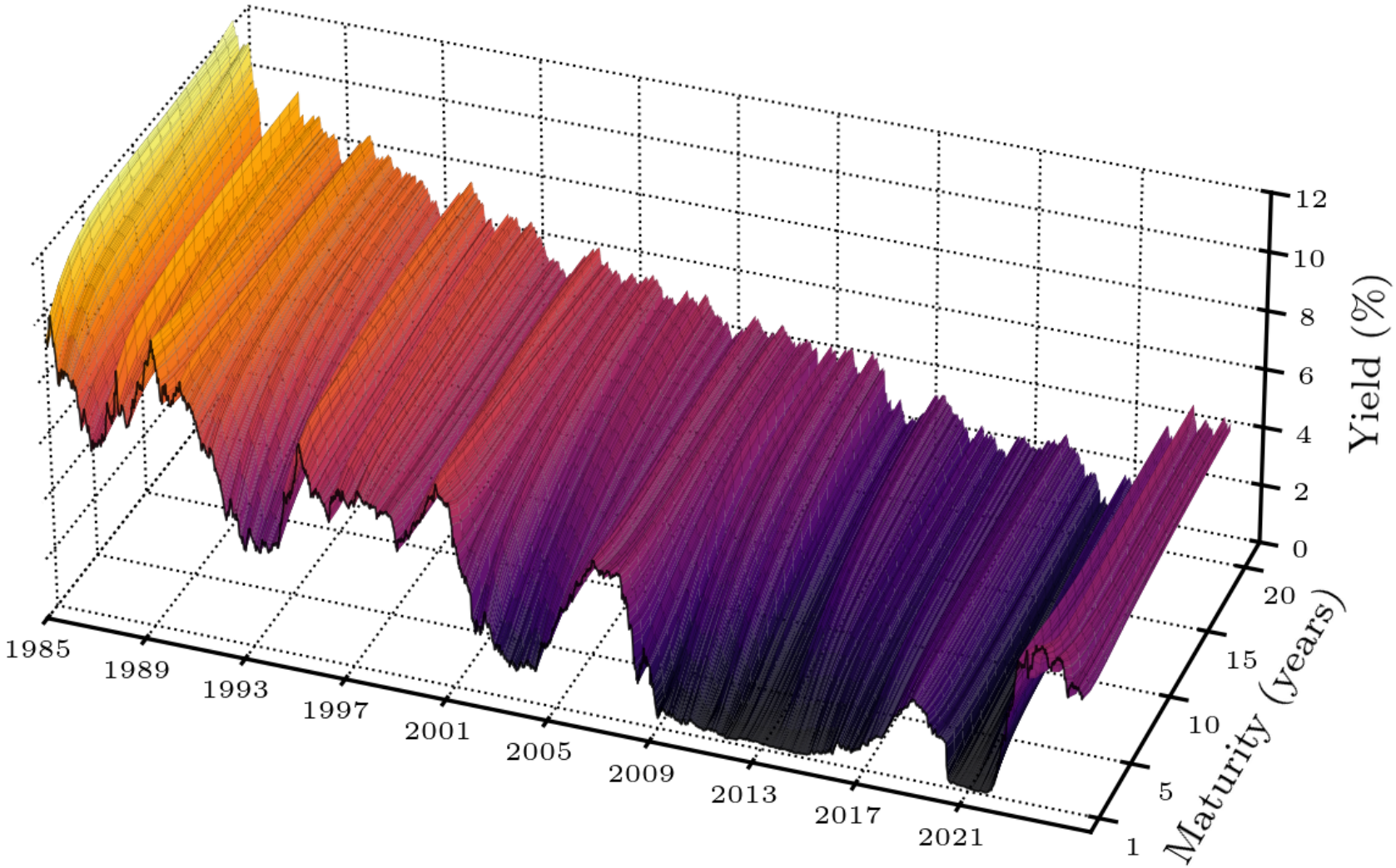
Starting with end-of-day quotes for all U.S. Treasury debt securities, the authors arrive at their final dataset by

1. excluding all securities with option-like features.
2. excluding all securities with less than three months to maturity due to liquidity and market segmentation concerns.
3. excluding all treasury bills because of market segmentation concerns.
4. excluding twenty-year bonds in 1996 because of liquidity concerns and tax-related reasons.
5. excluding the most recently issued securities for a wide range of maturities. These *on-the-run* and *first off-the-run* secuirites are typically traded at a premium because of high liquidity and their unique role in the repo market.
6. excluding some securities on an ad hoc basis because of specific market circumstances.

Furthermore, they use weights equal to the inverse of the duration of the securities.

The point being: A lot of financial understanding is needed to select a representative set of securities.

US Treasury yield curve (1985 to today)



Dimensionality reduction

- Denote by $\mathbf{y}_t$ the d-dimensional vector containing the time-t yields with yearly maturities between one and twenty years, such that $d = 20$.
- Furthermore, denote by $\mathbf{Y}$ the $d \times N$-dimensional data matrix constructed by stacking all the N daily curves. Using 1985 as our start date, this would mean that $N > 10000$.
- Can we transform the daily yield curves into a smaller-dimensional representation $\mathbf{y}_t \rightarrow \mathbf{x}_t$, while retaining (almost) all the information?
- Many methods exist for this type of problem, including recent advances in deep learning using (variational) autoencoders.
- For tractability, we will consider linear dimension reduction techniques, such that each latent variable is given as a linear combination of the observed yields.
- A very good candidate solution to our problem is then *principal component analysis*, which we will consider now.

Principal component analysis: preliminaries

Assume we have N different observations of some d-dimensional random variable $\mathbf{y}$, that we will (without loss of generality) assume has sample mean zero. Denote by $\mathbf{Y}$ the $d \times N$ matrix whose columns are the observations $\mathbf{y}$. The sample covariance matrix is given by

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i \mathbf{y}_i^\top .$$

$\mathbf{S}$ is a symmetric $d \times d$ matrix (which we assume has full rank). We can therefore use the spectral decomposition

$$\mathbf{S} = \mathbf{E} \mathbf{\Lambda} \mathbf{E}^\top = \sum_{i=1}^{d} \lambda_i \mathbf{e}_i \mathbf{e}_i^\top ,$$

where:

- Λ is the diagonal matrix containing the eigenvalues in decreasing order, i.e., $\Lambda = \mathrm{Diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$ with $\lambda_1 > \lambda_2 > \cdots > \lambda_n$.
- $\mathbf{E}$ is the matrix whose columns are the eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ in the order corresponding to the eigenvalues in Λ. Note that $\mathbf{E}$ is an orthogonal matrix such that $\mathbf{E}\mathbf{E}^\top = \mathbf{E}^\top \mathbf{E} = \mathbf{I}_d$.

Principal component analysis: first steps

Note: Informal proof.[3]

- We wish to find a unit vector[4] $\mathbf{u}_1$ such that $\mathbf{x}_1 = \mathbf{u}_1^\top \mathbf{Y}$ captures as much of the total variance as possible. Notice that the covariance of $\mathbf{x}_1$ is given by

$$\frac{1}{N}\mathbf{x}_1\mathbf{x}_1^\top = \frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_1^\top \mathbf{y}_n\mathbf{y}_n^\top \mathbf{u}_1 = \mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1,$$

  which means that we can formalize the problem as

$$\mathbf{u}_1 = \underset{\mathbf{u}:\|\mathbf{u}\|=1}{\operatorname{argmax}}\ \mathbf{u}^\top \mathbf{S}\mathbf{u}.$$

- Now use the spectral decomposition of $\mathbf{S}$ to rewrite the expression as

$$\mathbf{u}^\top \mathbf{S}\mathbf{u} = \sum_{i=1}^{d}\mathbf{u}^\top \mathbf{e}_i\mathbf{e}_i^\top \mathbf{u}\lambda_i = \sum_{i=1}^{d}\left(\mathbf{u}^\top \mathbf{e}_i\right)^2\lambda_i,$$

  such that

$$\mathbf{u}_1 = \underset{\mathbf{u}:\|\mathbf{u}\|=1}{\operatorname{argmax}}\ \sum_{i=1}^{d}\left(\mathbf{u}^\top \mathbf{e}_i\right)^2\lambda_i.$$

[3] See e.g., `https://web.math.ku.dk/~richard/courses/statlearn2011/pca.pdf` for a more formal introduction.

[4] A vector $\mathbf{v}$ is a unit vector if $\|\mathbf{v}\| = \mathbf{v}^\top \mathbf{v} = 1$

Principal component analysis: continued

- Since $\mathbf{e}_1, \ldots, \mathbf{e}_d$ forms an orthonormal basis and $\mathbf{u}$ is a unit vector, we have that

$$\sum_{i=1}^{d} \left(\mathbf{u}^\top \mathbf{e}_i \right)^2 = \|\mathbf{u}\|^2 = 1.$$

  Thus the expression we are trying to maximize is just a linear combination of the eigenvalues. So what $\mathbf{u}$ do we want? The one that makes the sum equal to the largest eigenvalue λ_1. This is done by selectiong $\mathbf{u}_1 = \mathbf{e}_1$, such that

$$\sum_{i=1}^{d} \left(\mathbf{u}_1^\top \mathbf{e}_i \right)^2 \lambda_i = \left(\mathbf{e}_1^\top \mathbf{e}_1 \right)^2 \lambda_1 = \lambda_1.$$

- Now, lets select a second $\mathbf{u}$ such that $\mathbf{x}_2 = \mathbf{u}_2^\top \mathbf{Y}$. We want this new principal factor to capture new information not contained in the previous principal component. The sample covariance between $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{u}_2^\top \mathbf{y}_i \mathbf{y}_i^\top \mathbf{e}_1 = \mathbf{u}_2^\top \mathbf{S} \mathbf{e}_1 = \lambda_1 \mathbf{u}_2^\top \mathbf{e}_1.$$

  Thus our second principal components $\mathbf{u}_2$ has to be orthogonal to the first eigenvector to ensure zero correlation, in addition to being the solution of the variance maximization problem stated above. The solution is $\mathbf{u}_2 = \mathbf{e}_2$, such that $\mathbf{e}_2^\top \mathbf{S} \mathbf{e}_2 = \lambda_2$. And so on.

Principal component analysis: overview

Principal component analysis

Consider N different observations of a d-dimensional vector $\mathbf{y}$ with sample covariance matrix $\mathbf{S}$. Denote by $\mathbf{Y}$ the data matrix given by stacking all observations of $\mathbf{y}$ to a $d \times N$ matrix.

- The jth principal component is given by the eigenvector $\mathbf{e}_j$ corresponding to the jth largest eigenvalue λ_j of $\mathbf{S}$.
- Denote by $\mathbf{E}_r$ the $d \times r$ matrix whose columns are given by the r eigenvectors corresponding to the largest r eigenvalues, then the $r \times N$-dimensional representation of the data is given by

$$\mathbf{X} = \mathbf{E}_r^\top \mathbf{Y}.$$

  The vector $\mathbf{x}_j = \mathbf{e}_j^\top \mathbf{Y}$ is called the jth *principal factor*.
- The sample covariance of $\mathbf{X}$ is given by $\frac{1}{N}\mathbf{X}\mathbf{X}^\top = \mathrm{Diag}\,(\lambda_1, \ldots, \lambda_r)$.

Principal component analysis: a few comments

If we denote by $\text{trace}\,(\mathbf{S}) = \sum_{i=1}^{d} \lambda_i$ the total variability of the data, then the first r principal components expains a fraction

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

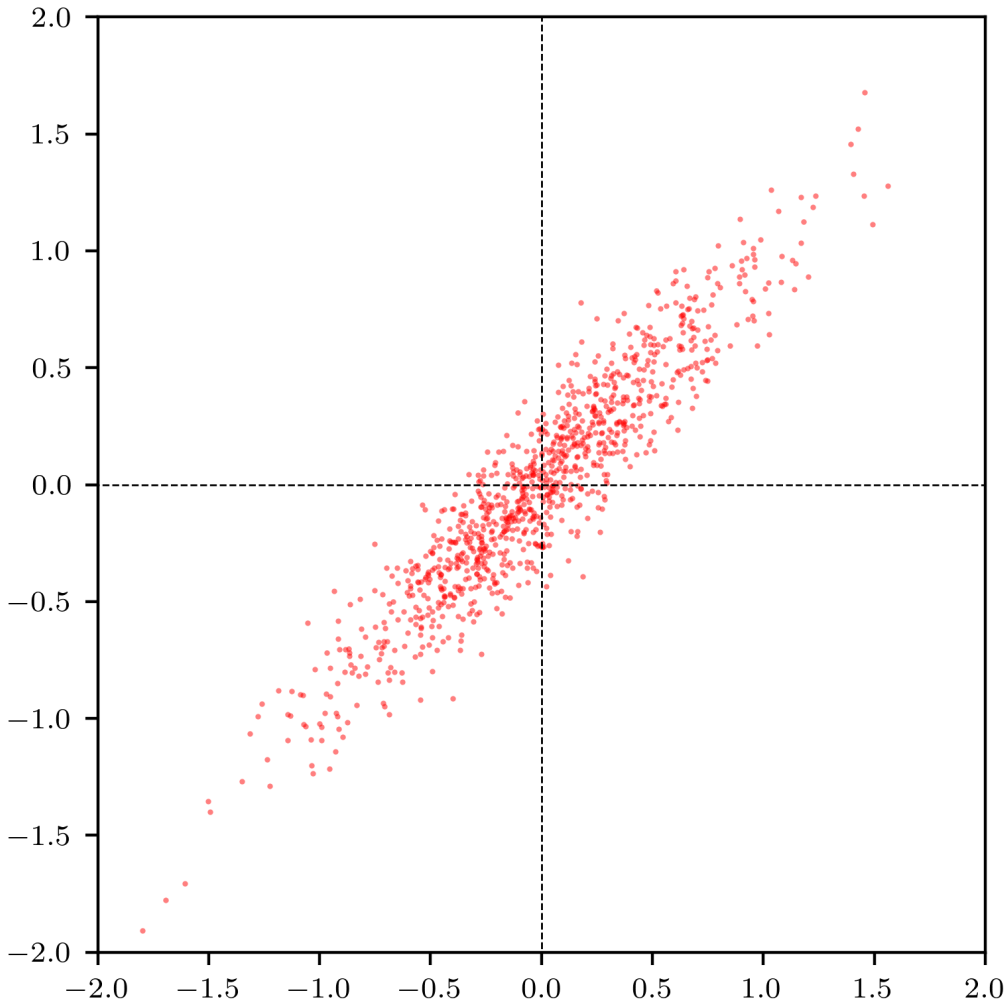
of this variability. The hope is that the above fraction is (sufficiently) close to one for $r \ll d$. Note that in general this will result in a loss of total variance.

Cookbook for Principal component analysis[5]

1. Compute the sample covariance matrix $\mathbf{S}$ of your dataset $\mathbf{Y}$.
2. Perform eigendecomposition on $\mathbf{S}$, which results in the eigenvalues $\lambda_1 > \cdots > \lambda_d$ and eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$.
3. Select the r eigenvectors that correspond to the r largest eigenvalues. These are the principal components.
4. Combine the eigenvectors into the $d \times r$ matrix $\mathbf{E}_r = (\mathbf{e}_1, \ldots, \mathbf{e}_r)$. The principal factors $\mathbf{X}$ is a $r \times N$ matrix given by $\mathbf{X} = \mathbf{E}_r^\top \mathbf{Y}$.

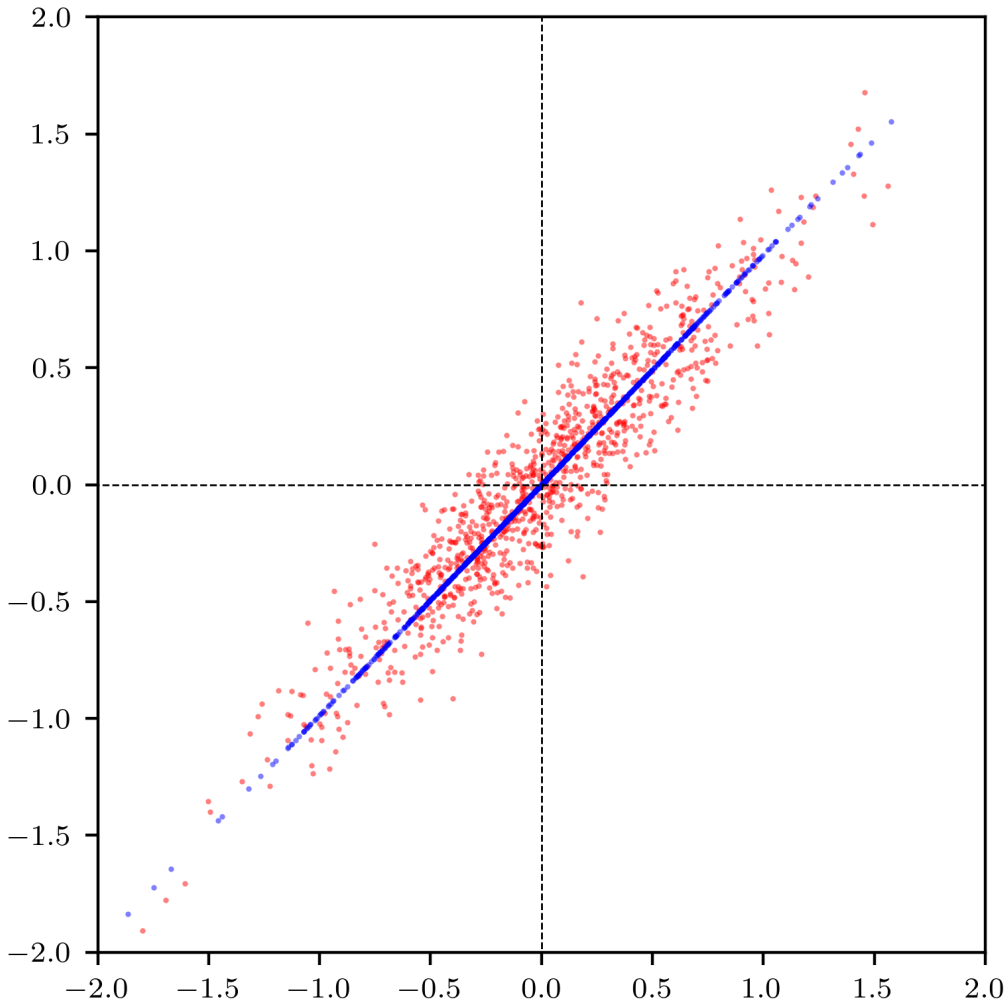
[5] Or you can just use a standard implementation.

Principal component analysis: simple example



Principal component example for two-dimensional correlated Gaussian samples

Principal component analysis: simple example



Principal component example for two-dimensional correlated Gaussian samples

PCA on US Treasury yields

Questions

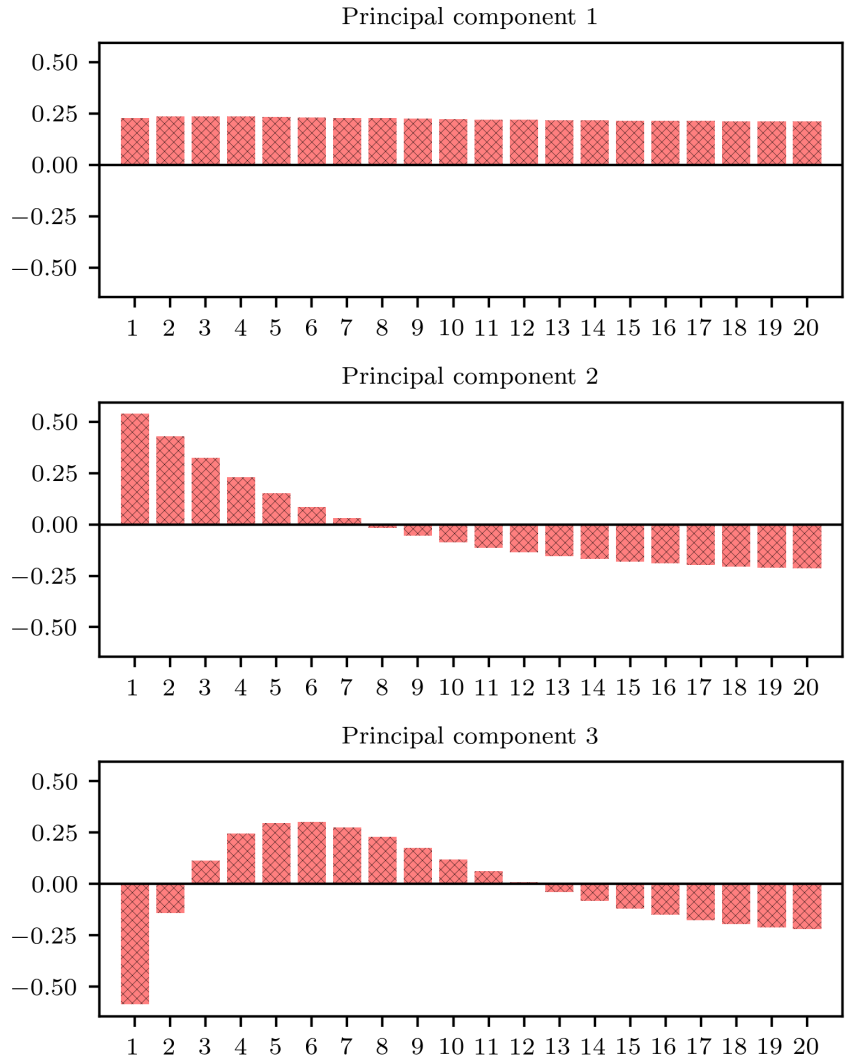
1. How many factors do we need to explain the yield curve?
2. What do the principal components look like?
3. What does the time series of the principal factors look like?

Principal component	1	2	3	4
Explained variance	96.98%	2.89%	0.1%	0.025%
Cumulate explained variance	96.98%	99.87%	99.971%	99.996%

Answer to question 1: Two or three should be sufficient.

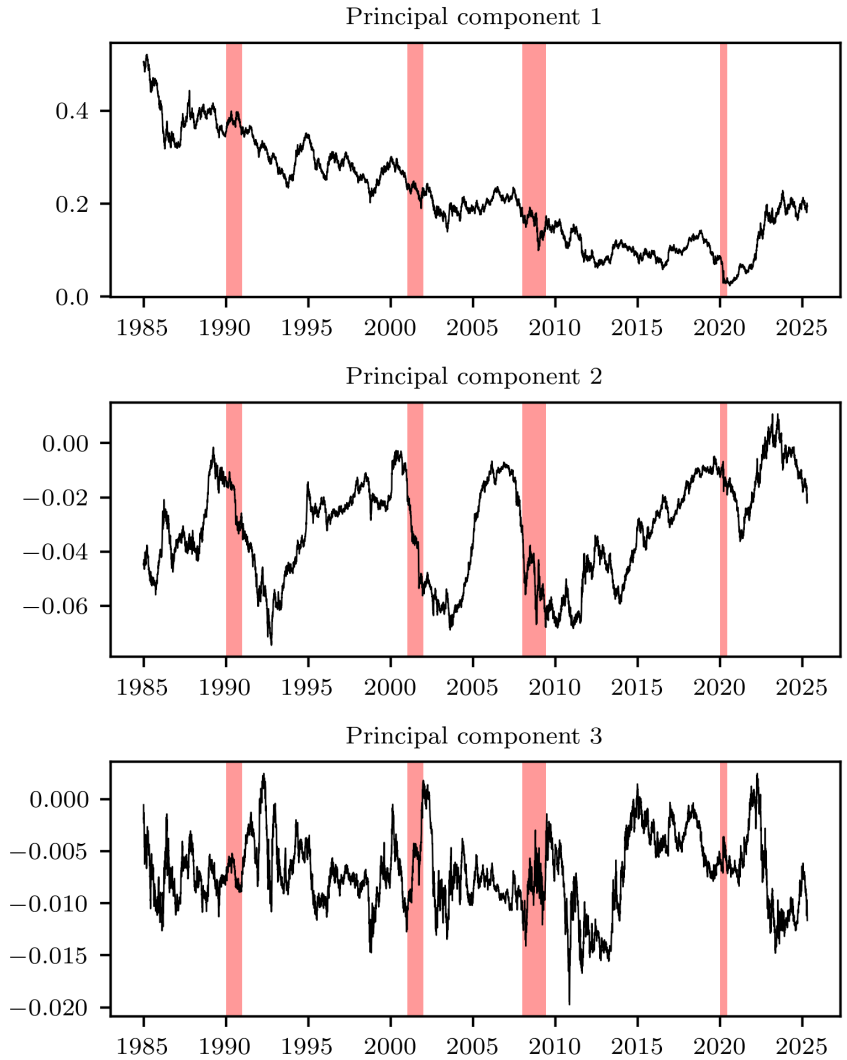
First three principal components

- The first principal components is easily interpretable as *level*.
- . . . the second as *slope*.
- . . . and the third as *hump* (or *curvature*).
- Decomposing the yield curve into a level, slope and hump factor sounds familiar. Where have seen it before?

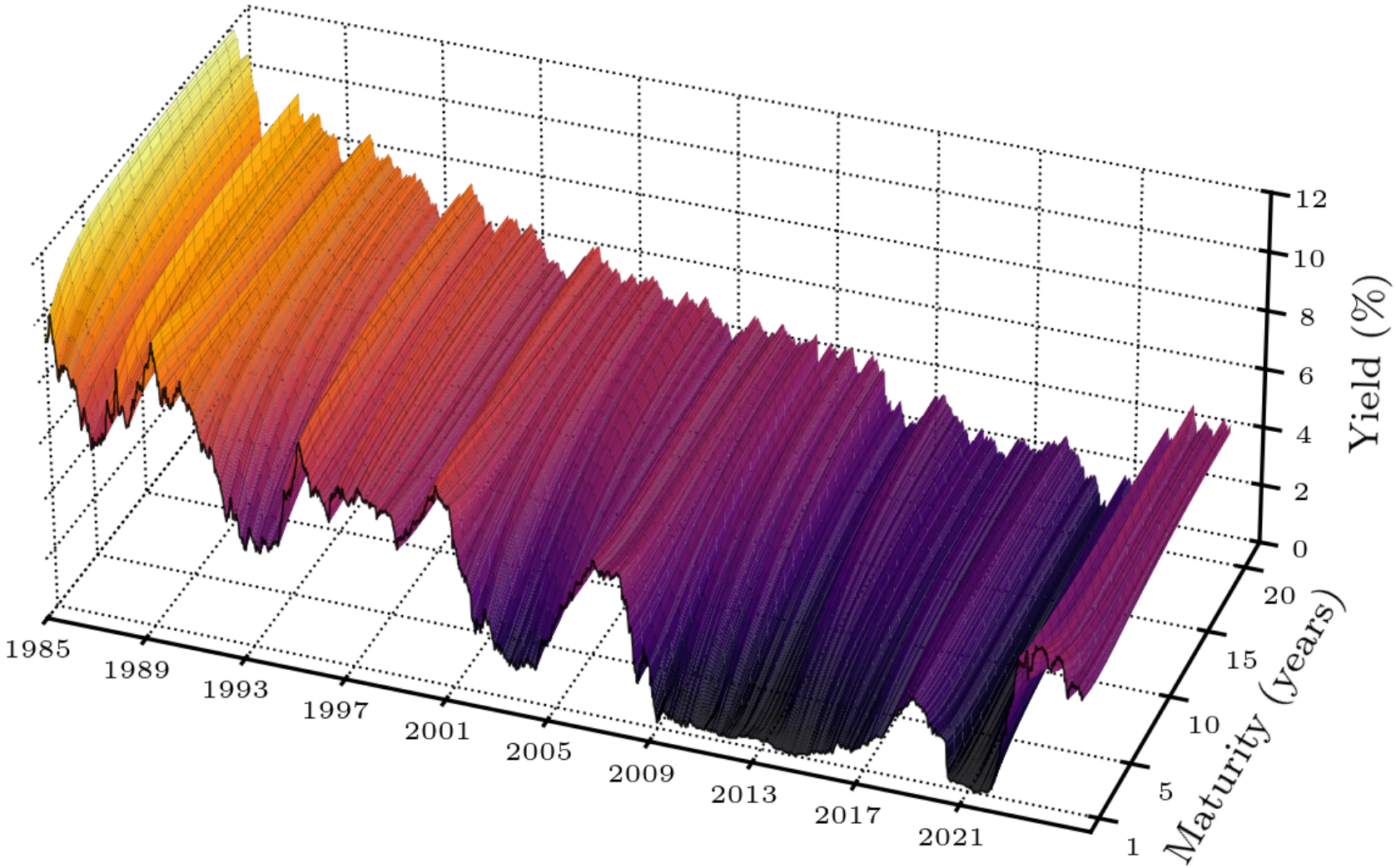


Time series of principal factors

- The level factor has been slowly trending downwards since the 1980s, but going upwards since the end of the Covid pandemic.
- Periods with recession (red) typically start with high slope ($\approx$ *inverted yield curve*), followed by a sharp drop (why?)
- Visually, the level factor appears to be highly persistent, while the slope and hump factors appear quite mean-reverting.

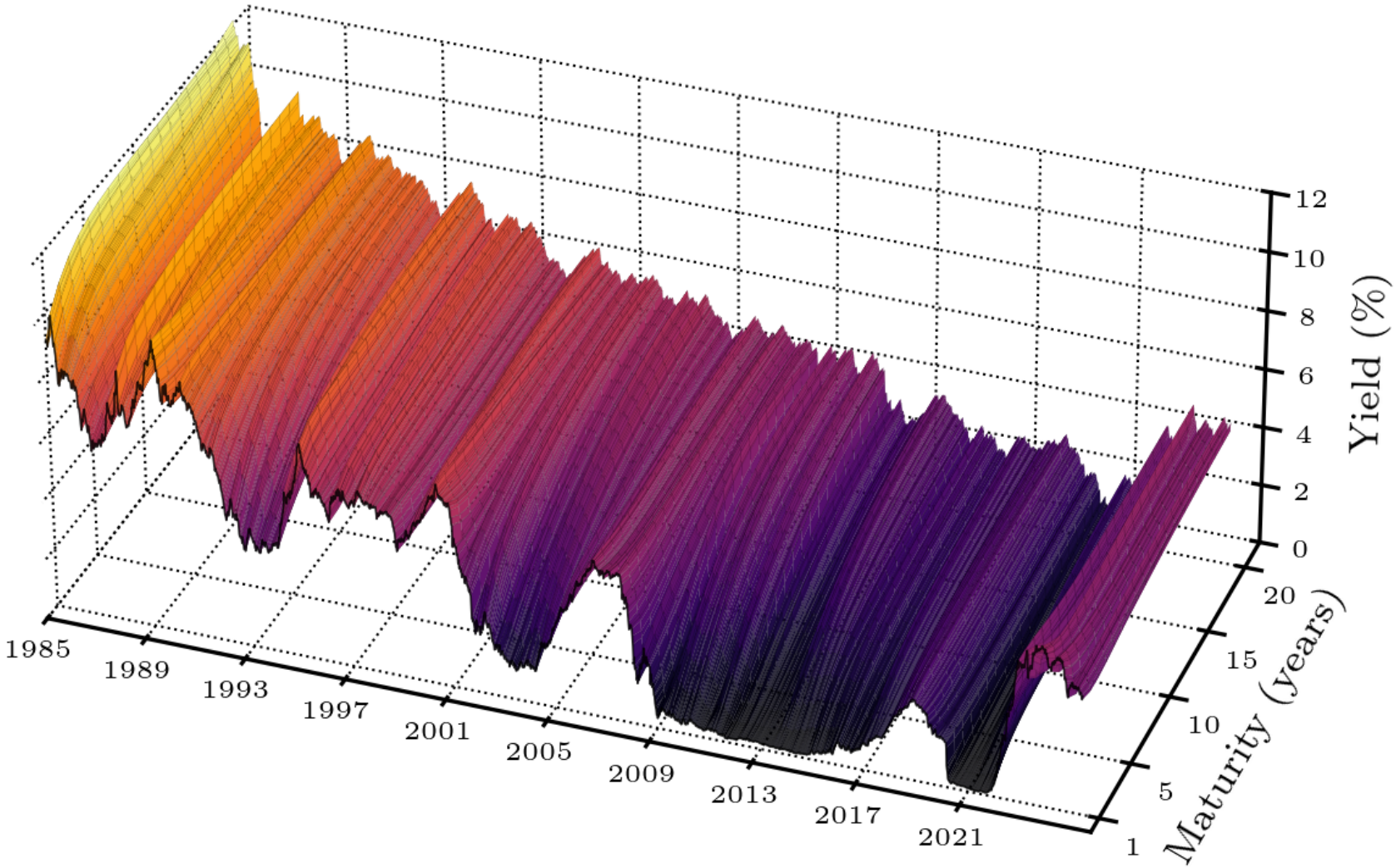


US Treasury yield curve (1985 to today)



Reconstructed from three principal components.

US Treasury yield curve (1985 to today)



True yields from Gürkaynak et al. (2007)

Stylized facts: preliminaries

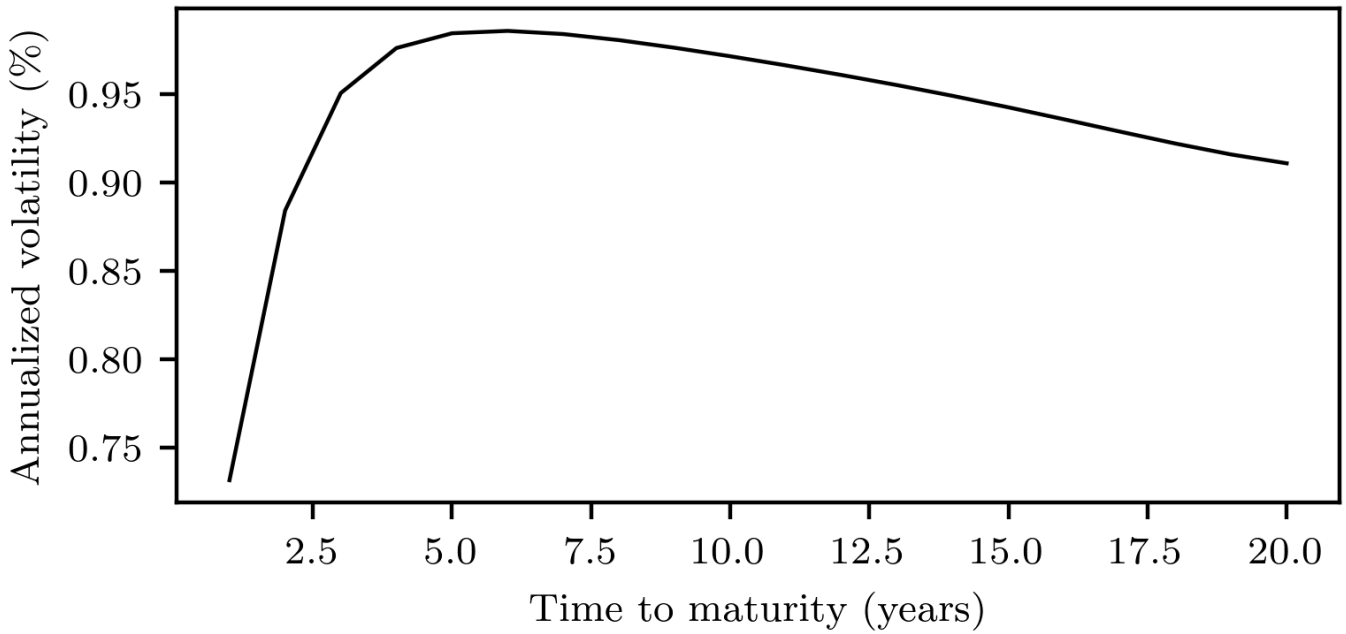
Consider the (scaled) daily changes in yields given by

$$\Delta y(t_j, \tau_i) = \frac{y(t_j, \tau_i) - y(t_{j-1}, \tau_i)}{\sqrt{\Delta t}}$$

where $\Delta t = 1/252$ and for maturities $\tau = 1, 2, \ldots, 20$ years.

Now, ignoring small drift terms, we can consider the sample covariance matrix of $\Delta \mathbf{y}(t_j) = (\Delta y(t_j, \tau_1), \ldots, \Delta y(t_j, \tau_M),)^\top$ for $t_j = 2, \ldots, T$. How does this look?

Term structure of volatility

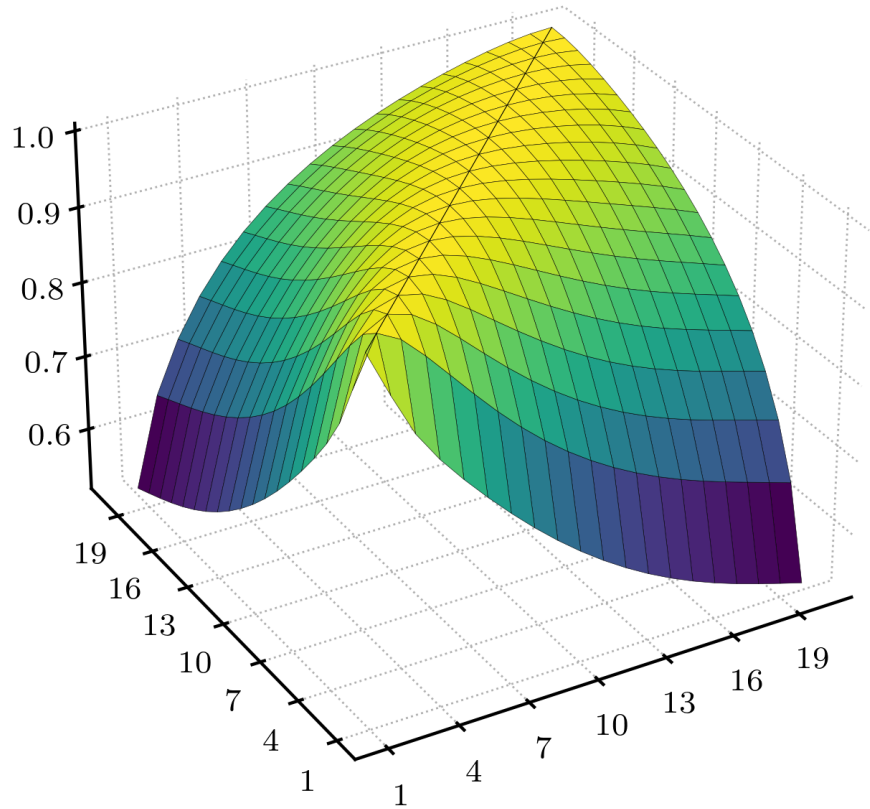


Stylized facts.

1. The volatility of yields peaks at ≈ 5 years.
2. Short-maturity yields are typically the least volatile. (Why?)

(The empirical term structure of forward rate volatility is also typically hump-shaped).

Empirical correlation matrix



Stylized facts. For two yields with maturities T_1 and T_2 (where $T_1 \leq T_2$), we observe that

1. Correlation between yields declines in $T_2 - T_1$ in a "near-exponential" manner towards some asymptote.
2. The rate of decay and the asymptote both seem to depend on T_1. Small T_1 means higher decay and lower asymptote.

Summary and next steps

Summary

1. Yield curve construction handles the problem of turning a representative set of securities into a full yield curve.
2. The Nelson and Siegel (1987) model and its' extension, the Svensson (1994) model, are two ways to do this.
3. From PCA, we have seen that a low-dimensional (say, 2 or 3) representation of the data is sufficient to capture almost all of the variability.
4. Empirically, we have found that
   - The term structure of yield volatility is hump-shaped, with a peak at around five years to maturity. Yields with a very short time to maturity are not as volatilie as longer maturities.
   - All yields with long maturities are highly correlated. Short maturity yields are less so.

Next steps

1. Until now we have not truly used the information from the time series aspect of the data. Our next objective is thus to develop methods to do so.
2. We want to combine the knowledge gained so far with a dynamic model of the yield curve.

PART 2:
LINEAR STATE SPACE MODELS

Motivation

- We have a time series of measurements $\mathbf{y}_n$ at discrete times $n = 1, \ldots, N$. In our case, $\mathbf{y}_n$ is a vector of yields for different maturities, and each n is one business day.
- We argued using the results from the principal component analysis that a low-dimensional representation of the yield curve seems plausible, such that

$$y_t(\tau) = h(\mathbf{x}_t, \tau) \text{ for } \tau > 0,$$

  for latent states $\mathbf{x}_t$.
- We also suspect that the yields are not perfectly observed, such that each measurement is corrupted by some degree of measurement noise. (Why?)
- Our aim is to estimate a dynamical model, where the latent states $\mathbf{x}_t$ are described by some stochastic differential equation

$$\mathrm{d}\mathbf{x}_t = \mu(\mathbf{x}_t)\mathrm{d}t + \sigma(\mathbf{x}_t)\mathrm{d}W_t^{\mathbb{P}}.$$

- Notice that we wish to estimate not only the latent states $\mathbf{x}_t$, but also the parameters of μ, σ and h, given the measurements $\mathbf{y}_n$ from $n = 1, \ldots, N$.
- This next section will build up the necesssary tools to do so.

The Kalman filter

- Let us for now focus on estimating the latent states $\mathbf{x}_t$ assuming everything else is known.
- Putting everything finance-related to the side for a moment, this problem is a particular example of the so-called filtering problem:
  Given measurements $\mathbf{y}_1, \ldots, \mathbf{y}_n$*, what is the best estimate of the latent state* $\mathbf{x}_n$ *based on these measurements?*
- It turns out that our problem can be solved using the Kalman filter, an extremely versatile algorithm with a lot of applications (inside and outside finance).
- What follows is only a *hands-on* introduction to Kalman filtering. If you're curious, see Särkkä and Svensson (2023) for a great introduction based on Bayesian filtering, or Harvey (1991) for a classical treatment.

Linear Gaussian state space model

The assumptions:

1. We assume that our measurements $\mathbf{y}_n$ for $n = 1, \ldots, N$ are related to the latent state $\mathbf{x}_n$ by the *measurement equation*

$$\mathbf{y}_n = \mathbf{H}\mathbf{x}_n + \mathbf{d} + \epsilon_n,$$

   where ϵ_n is a vector of measurement noise that we will assume is iid zero-mean multivariate Gaussian with covariance matrix $\mathbf{R}$.

2. We assume that the dynamic model of our latent states $\mathbf{x}_n$ is given by the *transition equation*

$$\mathbf{x}_{n+1} = \mathbf{A}\mathbf{x}_n + \mathbf{b} + \omega_n,$$

   where ω_n is iid zero-mean multivariate normal with covariance matrix $\mathbf{Q}$.

3. We furthermore assume that we are given an initial distribution of $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{P}_1)$.

4. Finally, we assume that the process noise ω_n, the measurement noise ϵ_k as well as the initial state $\mathbf{x}_1$ are all independent of each other for all n, k.

Preliminary thoughts: What to expect?

- Both the transition and measurement equations in the linear state space model are combinations of
  - affine transformations,
  - adding (independent) Gaussian noise.
- Remember that
  - the Gaussian distribution is closed under affine transformations.
  - if two random variables $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian, then the sum $\mathbf{X} + \mathbf{Y}$ is also Gaussian.
- We start out with a multivariate Gaussian random variable $\mathbf{x}_1$.

Intuitively, we would therefore expect the estimates of the latent state to remain Gaussian. Gaussian distributions are highly tractable.

Conditioning on measurements 1/2

We start with a Gaussian $\mathbf{x}_1$ and want to find the best estimate for $\mathbf{x}_1$ *conditional* on our first measurements $\mathbf{y}_1$. First consider the joint distribution of $(\mathbf{x}_1, \mathbf{y}_1)$:

- The mean of the measurements is given by

$$\mathbb{E}\{\mathbf{y}_1\} = \mathbf{H}\mathbb{E}(\mathbf{x}_1) + \mathbf{d} = \mathbf{H}\mathbf{m}_1 + \mathbf{d}.$$

- The variance is given by

$$\begin{aligned}\mathbb{V}\{\mathbf{y}_1\} &= \mathbb{E}\left\{[\mathbf{H}(\mathbf{x}_1 - \mathbf{m}_1) + \epsilon_1]\,[\mathbf{H}(\mathbf{x}_1 - \mathbf{m}_1) + \epsilon_1]^\top\right\} \\ &= \mathbf{H}\mathbb{E}\left\{[\mathbf{x}_1 - \mathbf{m}_1]\,[\mathbf{x}_1 - \mathbf{m}_1]^\top\right\}\mathbf{H}^\top + \mathbb{E}\left\{\epsilon_1\epsilon_1^\top\right\} \\ &= \mathbf{H}\mathbf{P}_1\mathbf{H}^\top + \mathbf{R}.\end{aligned}$$

- The covariance between $\mathbf{x}_1$ and $\mathbf{y}_1$ is given by

$$\mathrm{Cov}\,(\mathbf{y}_1, \mathbf{x}_1) = \mathbb{E}\left\{([\mathbf{H}(\mathbf{x}_1 - \mathbf{m}_1) + \epsilon_1]\,[\mathbf{x}_1 - \mathbf{m}_1]^\top\right\} = \mathbf{H}\mathbf{P}_1.$$

Thus the joint distribution is multivariate Gaussian with

$$\begin{pmatrix}\mathbf{x}_1 \\ \mathbf{y}_1\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{m}_1 \\ \mathbf{H}\mathbf{m}_1 + \mathbf{d}\end{pmatrix}, \begin{pmatrix}\mathbf{P}_1 & \mathbf{P}_1\mathbf{H}^\top \\ \mathbf{H}\mathbf{P}_1 & \mathbf{H}\mathbf{P}_1\mathbf{H}^\top + \mathbf{R}\end{pmatrix}\right)$$

Conditioning on measurements 2/2

A convenient result

Assume that $(\mathbf{x}, \mathbf{y})$ are jointly Gaussian such that

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}, \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B} \end{pmatrix} \right),$$

then the conditional distribution of $\mathbf{x}|\mathbf{y}$ is Gaussian with

$$\mathbf{x}|\mathbf{y} \sim N\left(\mathbf{a} + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \mathbf{b}), \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top \right).$$

Since we observe $\mathbf{y}_1$, we can condition on it, giving us a new estimate of the latent state

$$\mathbf{x}_1|\mathbf{y}_1 \sim N(\mathbf{m}_{1|1}\mathbf{P}_{1|1}).$$

The mean and covariance are given by

$$\mathbf{m}_{1|1} = \mathbf{m}_1 + \mathbf{P}_1\mathbf{H}^\top \left[\mathbf{H}\mathbf{P}_1\mathbf{H}^\top + \mathbf{R}\right]^{-1} \left[\mathbf{y}_t - (\mathbf{H}\mathbf{m}_1 + \mathbf{d})\right],$$

$$\mathbf{P}_{1|1} = \mathbf{P}_1 - \mathbf{P}_1\mathbf{H}^\top \left[\mathbf{H}\mathbf{P}_1\mathbf{H}^\top + \mathbf{R}\right]^{-1} \mathbf{H}\mathbf{P}_1.$$

Propagating the dynamic model

- We now have the distribution of our latent state at time $n = 1$ conditional on the information from the first measurement $\mathbf{y}_1$.
- Our next step is to find the best estimate of $\mathbf{x}_2$ conditional on the previous information. It is common in (Kalman) filtering litterature to write this as $\mathbf{x}_{i|i-1} \triangleq \mathbf{x}_i|\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}$.
- From the transition model we know that

$$\mathbf{x}_2 = \mathbf{A}\mathbf{x}_1 + \mathbf{b} + \omega_1,$$

straightforward calculations give us that $\mathbf{x}_2|\mathbf{y}_1$ is Gaussian with

$$\begin{aligned}
\mathbf{m}_{2|1} &\triangleq \mathbb{E}\left\{\mathbf{x}_2|\mathbf{y}_1\right\} = \mathbf{A}\mathbb{E}\left\{\mathbf{x}_1|\mathbf{y}_1\right\} + \mathbf{b} = \mathbf{A}\mathbf{m}_{1|1} + \mathbf{b}, \\
\mathbf{P}_{2|1} &\triangleq \mathbb{V}\left\{\mathbf{x}_2|\mathbf{y}_1\right\} = \mathbb{E}\left\{\left[\mathbf{A}(\mathbf{x}_{1|1} - \mathbf{m}_{1|1}) + \omega_1\right]\left[\mathbf{A}(\mathbf{x}_{1|1} - \mathbf{m}_{1|1}) + \omega_1\right]^\top\right\} \\
&= \mathbf{A}\mathbf{P}_{1|1}\mathbf{A}^\top + \mathbf{Q}.
\end{aligned}$$

Of course, we can now condition on the measurements $\mathbf{y}_2$, giving us $\mathbf{x}_{2|2}$, and predict the next state $\mathbf{x}_{3|2}$, condition on $\mathbf{y}_3$ to get $\mathbf{x}_{3|3}$ and ...

The Kalman filter recursions

For simplicity, we can write the prediction error as $\mathbf{v}_t \triangleq \mathbf{y}_t - (\mathbf{H}\mathbf{m}_1 + \mathbf{d})$ and its' covariance as $\mathbf{S}_t \triangleq \mathbf{H}\mathbf{P}_1\mathbf{H}^\top + \mathbf{R}$. Then we can write the so-called *Kalman gain* matrix as $\mathbf{K}_t \triangleq \mathbf{P}_1\mathbf{H}^\top\mathbf{S}_t^{-1}$.

Kalman filter recursions

Let $\mathbf{m}_{1|0} = \mathbf{m}_1$ and $\mathbf{P}_{1|0} = \mathbf{P}_1$. Then for $n = 1, \ldots, N$:

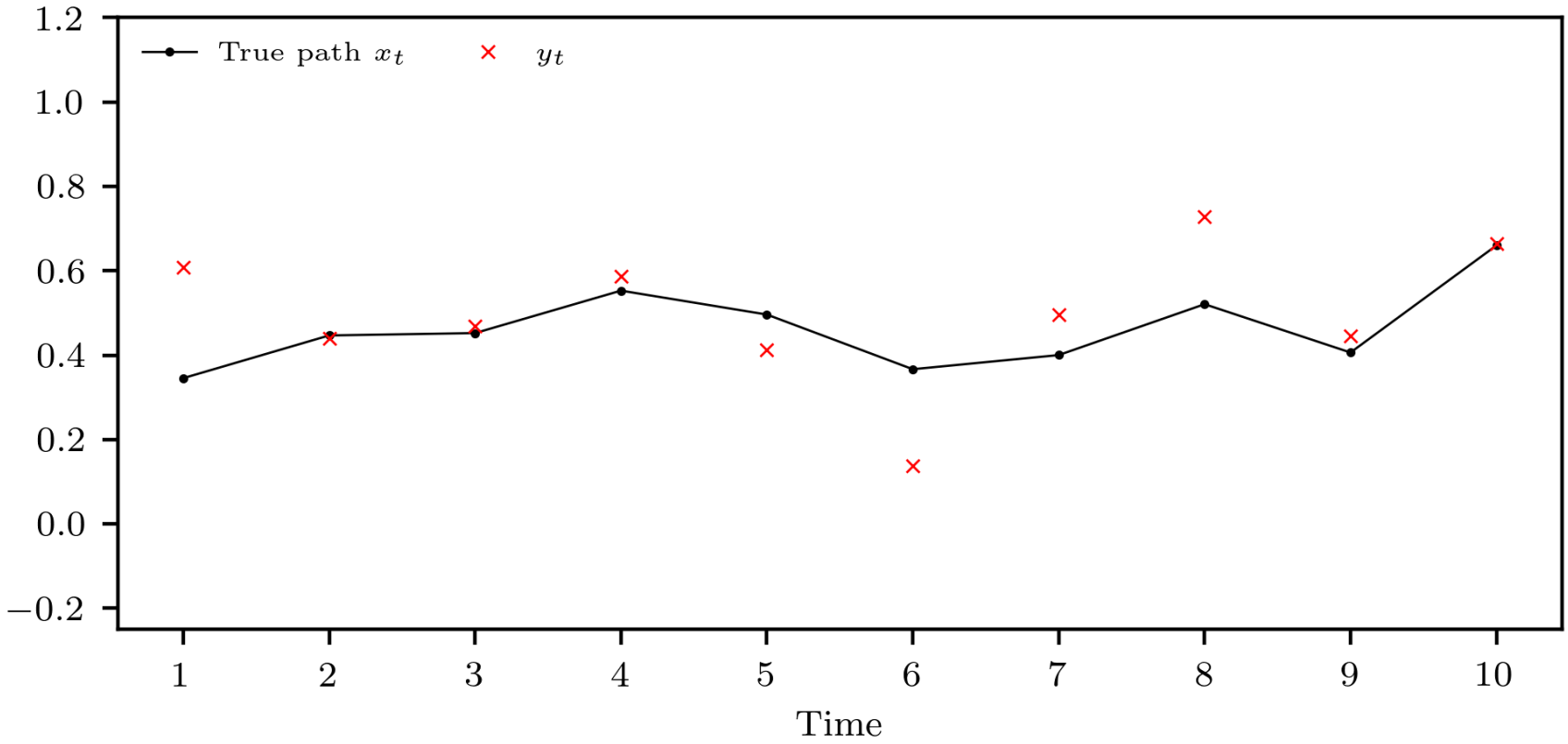
Update step:

$$\begin{aligned}
\mathbf{v}_n &= \mathbf{y}_n - \left(\mathbf{H}\mathbf{m}_{n|n-1} + \mathbf{d}\right) \\
\mathbf{S}_n &= \mathbf{H}\mathbf{P}_{n|n-1}\mathbf{H}^\top + \mathbf{R} \\
\mathbf{K}_n &= \mathbf{P}_{n|n-1}\mathbf{H}^\top\mathbf{S}_t^{-1} \\
\mathbf{m}_{n|n} &= \mathbf{m}_{n|n-1} + \mathbf{K}_n\mathbf{v}_n \\
\mathbf{P}_{n|n} &= \mathbf{P}_{n|n-1} - \mathbf{K}_n\mathbf{S}_n\mathbf{K}_t^\top
\end{aligned}$$

Prediction step:

$$\begin{aligned}
\mathbf{m}_{n+1|n} &= \mathbf{A}\mathbf{m}_{n|n} + \mathbf{b} \\
\mathbf{P}_{n+1|n} &= \mathbf{A}\mathbf{P}_{n|n}\mathbf{A}^\top + \mathbf{Q}
\end{aligned}$$

An example: scalar random walk



We now consider a simple scalar random walk with the state space model

$$y_n = x_n + \epsilon_n,$$
$$x_{n+1} = x_n + \omega_n,$$

and $x_1 \sim N(m_1, P_1)$.

Kalman filter recursion for scalar random walk

Algorithm: Kalman filtering scalar random walk
Set $m_{1|0} = m_1$ and $P_{1|0} = P_1$. Then for $n = 1, \ldots, N$:

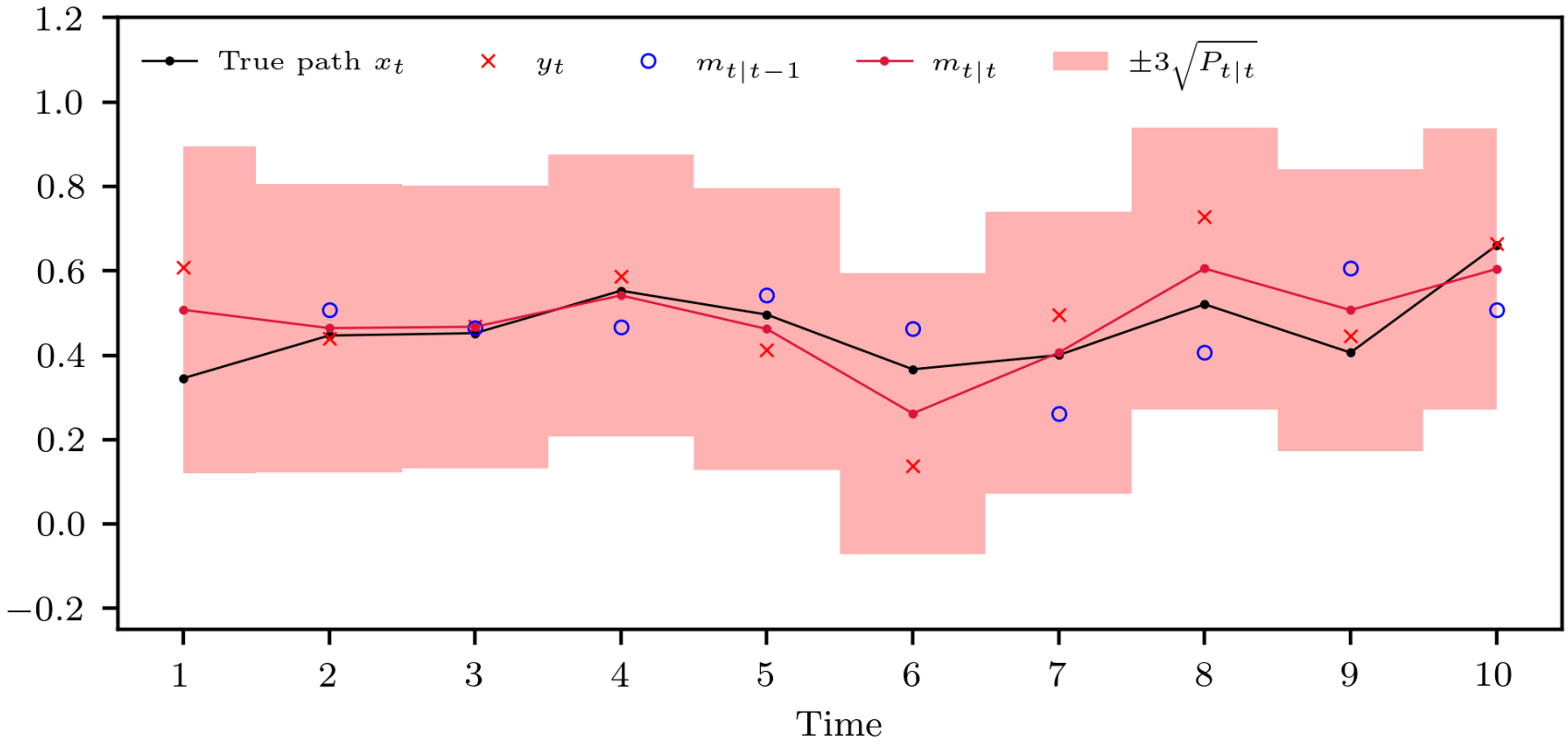
1. Update the estimate using the new measurements y_n:

$$
\begin{aligned}
v_n &= y_n - m_{n|n-1}, \\
S_n &= P_{n|n-1} + R, \\
K_n &= \frac{P_{n|n-1}}{P_{n|n-1} + R}, \\
m_{n|n} &= m_{n|n-1} + \frac{P_{n|n-1}}{P_{n|n-1} + R} v_n = m_{n|n-1} + K_n v_n, \\
P_{n|n} &= P_{n|n-1} - \frac{\left(P_{n|n-1}\right)^2}{P_{n|n-1} + R} = P_{n|n-1} - K_n^2 S_n.
\end{aligned}
$$

2. Predict the distribution of the latent states at the next step:

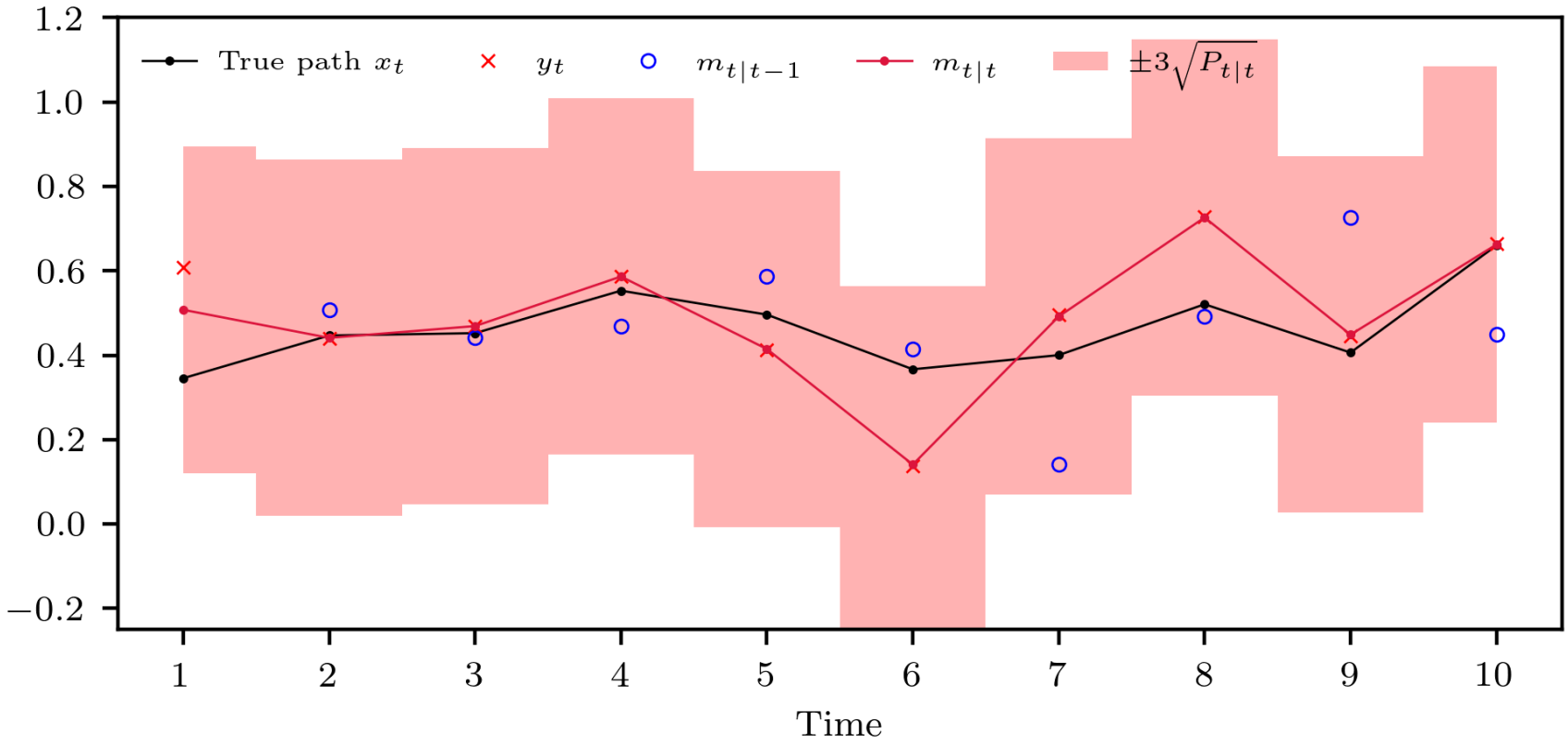
$$
\begin{aligned}
m_{n+1|n} &= m_{n|n}, \\
P_{n+1|n} &= P_{n|n} + Q.
\end{aligned}
$$

Visualizing the filtering process



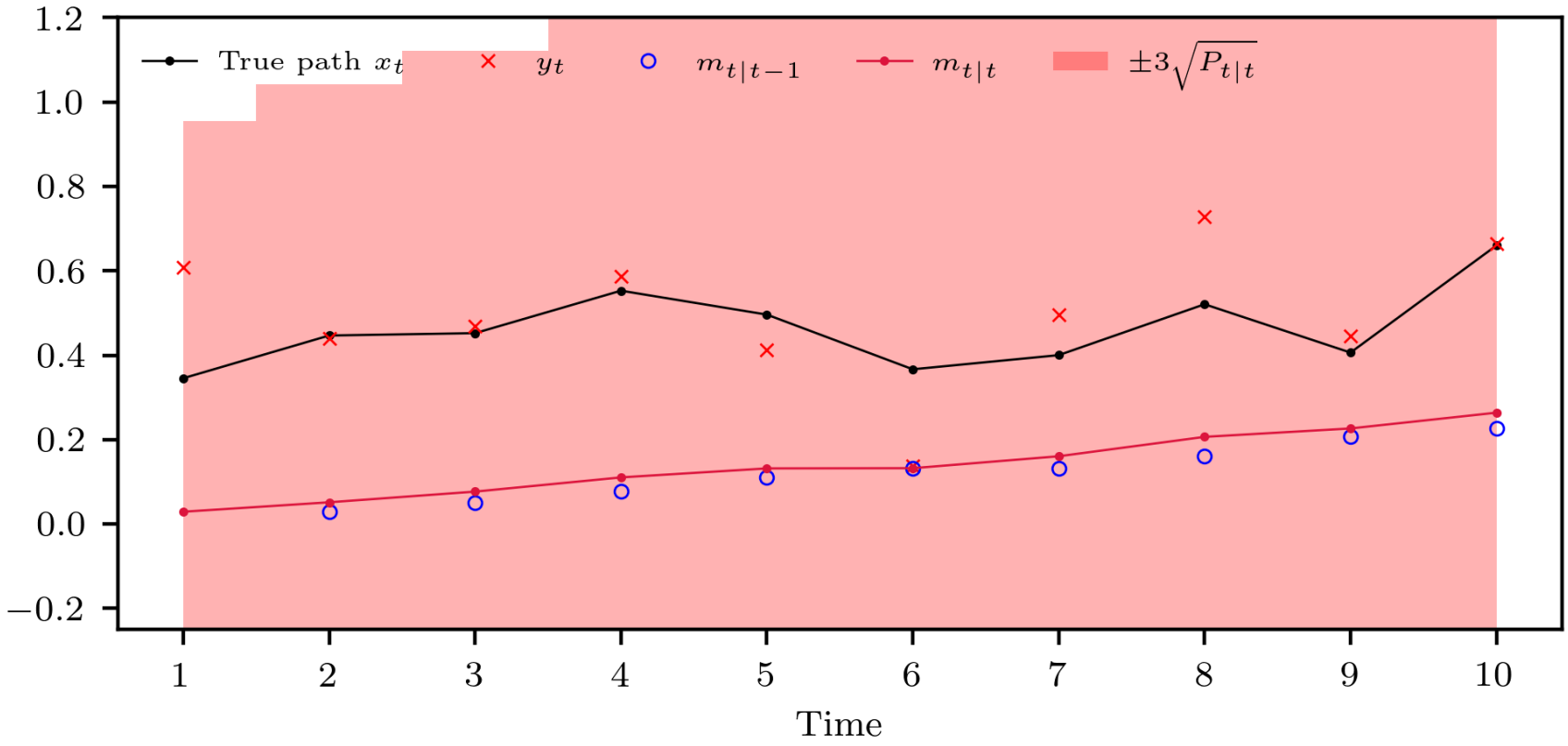
Kalman filter with true parameters, $R \approx Q$

Visualizing the filtering process



Kalman filter with $R \ll Q$, or:
When we don't trust our predictions

Visualizing the filtering process



Kalman filter with $Q \ll R$, or:
When we don't trust our measurements

Dual estimation

- We now know how to estimate the latent states given the measurements.
- But more importantly, we wish to estimate the parameters of the transition and measurement equations

$$\begin{aligned} \mathbf{y}_n &= \mathbf{C}\mathbf{x}_n + \mathbf{d} + \epsilon_n & \epsilon_n &\sim N(0, \mathbf{R}), \\ \mathbf{x}_{n+1} &= \mathbf{A}\mathbf{x}_n + \mathbf{b} + \omega_n & \omega_n &\sim N(0, \mathbf{Q}). \end{aligned}$$

  That is: can we estimate $(\mathbf{C}, \mathbf{d}, \mathbf{Q}, \mathbf{A}, \mathbf{b}, \mathbf{R})$ as well as the latent states from just the measurements?
- It turns out that we can, and that we only need one more observation to understand how.
- We need to derive the likelihood

$$\mathcal{L} = \sum_{n=1}^{N} p(\mathbf{y}_n | \mathbf{y}_1, \ldots, \mathbf{y}_{n-1})$$

- **Notice** that $\mathbf{y}_n | \mathbf{y}_1, \ldots, \mathbf{y}_{n-1}$ is Gaussian with mean and covariance

$$\mathbf{y}_n | \mathbf{y}_1, \ldots, \mathbf{y}_{n-1} \sim N(\mathbf{C}\mathbf{m}_{n|n-1} + \mathbf{d}, \mathbf{C}\mathbf{P}_{n|n-1}\mathbf{C}^\top + \mathbf{R})$$

The prediction error decomposition of the likelihood

Thus we can evaluate the likelihood directly from prediction errors $\mathbf{v}_n$ and their covariance matrices $\mathbf{S}_n$ using

$$\log \mathcal{L}(\Theta; \mathbf{y}_1, \ldots, \mathbf{y}_N) = \sum_{n=1}^{N} \left[-\frac{N}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{S}_n| - \frac{1}{2} \mathbf{v}_n^\top \mathbf{S}_n^{-1} \mathbf{v}_n \right],$$

notice that we have to compute both of these terms during the Kalman filter recursions anyway. The maximum likelihood estimate is then the parameters $\hat{\Theta}$ that minimize the negative loglikelihood

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left[-\log \mathcal{L}(\Theta; \mathbf{y}_1, \ldots, \mathbf{y}_N) \right].$$

Skipping a few technical requirements[6] standard errors of our parameter estimates can be found using the formula

$$\hat{\Omega}(\hat{\Theta}) = \frac{1}{N} \sum_{t=1}^{N} \left[\frac{1}{N} \frac{\partial \log \mathcal{L}_n(\hat{\Theta})}{\partial \Theta} \left(\frac{\partial \log \mathcal{L}_n(\hat{\Theta})}{\partial \Theta} \right)^\top \right]^{-1}$$

where $\mathcal{L}_n$ is the likelihood of the prediction error at time n.

[6] See Harvey (1991)

From SDE to transition equation

Goal: we wish to convert an affine SDE

$$\mathrm{d}\mathbf{X}(t) = \mathbf{K}\left(\theta - \mathbf{X}(t)\right)\mathrm{d}t + \Sigma \mathrm{d}W(t),$$

into a transition equation

$$\mathbf{X}_{t+\Delta} = \mathbf{A}\mathbf{X}_t + \mathbf{b} + \omega_t,$$

where ω_t is a zero-mean noise term.

We will assume $\mathbf{X}(t)$ to be stationary, such that the real component of the eigevalues of $\mathbf{K}$ are positive.

The matrix exponential

For a $n \times n$ matrix $\mathbf{A}$ we define $e^{\mathbf{A}}$ as the power series

$$e^{\mathbf{A}} = \sum_{k=1}^{\infty} \frac{A^k}{k!},$$

where $\mathbf{A}^0 = \mathbf{I}_n$ and $\mathbf{A}^2 = \mathbf{AA}$ and $\mathbf{A}^k = \mathbf{A}\mathbf{A}^{k-1}$.

Some useful properties

1. $e^{\mathbf{0}_{n\times n}} = \mathbf{I}_n$ and $\left(e^{\mathbf{A}}\right)^\top = e^{\mathbf{A}^\top}$.
2. For invertible $\mathbf{B}$:
$$e^{\mathbf{BAB}^{-1}} = \mathbf{B}e^{\mathbf{A}}\mathbf{B}^{-1}$$
3. If two matrices $\mathbf{A}$ and $\mathbf{B}$ commute, such that $mathbf{A}\mathbf{B} = \mathbf{BA}$, then
$$\mathbf{A}e^{\mathbf{B}} = e^{\mathbf{B}}\mathbf{A} \text{ and } e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}}$$
4. If $\mathbf{A}$ is diagonal such that $A = \text{Diag}(a_1, a_2, \ldots, a_n)$, then $e^{\mathbf{A}} = \text{Diag}\left(e^{a_1}, e^{a_2}, \ldots, e^{a_n}\right)$

Solving the affine SDE

We will use a multivariate version of the trick used to solve the Ornstein-Uhlenbeck SDE. Consider

$$\mathrm{d}\left(e^{\mathbf{K}t}\mathbf{X}_t\right) = e^{\mathbf{K}t}\mathrm{d}\mathbf{X}_t + \mathbf{K}e^{\mathbf{K}t}\mathbf{X}_t\mathrm{d}t = e^{\mathbf{K}t}\theta\mathrm{d}t + e^{\mathbf{K}t}\Sigma\mathrm{d}W_t$$

Integrating both sides and isolating $\mathbf{X}_t$ yields

$$\mathbf{X}_t = e^{-\mathbf{K}t}\mathbf{X}_0 + \int_0^t e^{-\mathbf{K}(t-u)}\theta\mathrm{d}t + \int_0^t e^{-\mathbf{K}(t-u)}\Sigma\mathrm{d}W_t$$

If $\mathbf{K}$ is nonsingular, which we assume it is, then

$$\int_0^t e^{-\mathbf{K}(t-u)}\mathrm{d}t\mathbf{K}\theta = \left(\mathbf{I} - e^{-\mathbf{K}t}\right)\theta.$$

Thus the expectation of $\mathbf{X}_{t+\Delta}$ given $\mathbf{X}_t$ can be written as

$$\mathbb{E}\left\{\mathbf{X}_{t+\Delta}|\mathbf{X}_t\right\} = e^{-\mathbf{K}\Delta}\mathbf{X}_t + \left(\mathbf{I} - e^{-\mathbf{K}\Delta}\right)\theta = \mathbf{A}\mathbf{X}_t + \mathbf{b}.$$

Note that $\lim_{\Delta\to\infty}\mathbb{E}\left\{\mathbf{X}_{t+\Delta}|\mathbf{X}_t\right\} = \theta$.

What about the covariance?

Note: This follows the technique outlined in Christensen et al. (2015). The conditional covariance is given by (use Itô isometry):

$$\mathbb{V}\left\{\mathbf{X}_{t+\Delta}|\mathbf{X}_t\right\} = \int_0^{\Delta} e^{-\mathbf{K}u}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{-\mathbf{K}^\top u}\mathrm{d}u.$$

Assume that $\mathbf{K}$ is diagonalizable, such that

$$\mathbf{K} = \mathbf{E}\Lambda\mathbf{E}^{-1}$$

where V is the matrix whose columns are the eigenvectors and Λ is a diagonal matrix containing the eigenvalues. This allows us to rewrite the expression as

$$e^{-\mathbf{K}u} = \mathbf{E}e^{-\Lambda u}\mathbf{E}^{-1} \text{ where } e^{-\Lambda u} = \text{Diag}\left(e^{-\lambda_1 u}, \ldots, e^{-\lambda_d u}\right).$$

such that the covariance matrix becomes

$$\mathbb{V}\left\{\mathbf{X}_{t+\Delta}|\mathbf{X}_t\right\} = \mathbf{E}\left[\int_0^{\Delta} e^{-\Lambda u}\mathbf{E}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\left(\mathbf{E}^{-1}\right)^\top e^{-\Lambda^\top u}\mathrm{d}u\right]\mathbf{E}^\top.$$

Now we define a new matrix

$$\bar{S} \triangleq \mathbf{E}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\left(\mathbf{E}^{-1}\right)^\top.$$

Covariance computation continued

From here we can rewrite the intergrand elementwise as

$$\left[e^{-\Lambda u}\bar{S}e^{-\Lambda^\top u}\right]_{ij} = \bar{S}_{ij}e^{-(\lambda_i+\lambda_j)u}.$$

Define another new matrix $\mathbf{V}(\Delta)$, such that

$$\mathbf{V}_{ij}(\Delta) \triangleq \int_0^\Delta \bar{S}_{ij}e^{-(\lambda_i+\lambda_j)u}\mathrm{d}u = \frac{\bar{S}_{ij}}{\lambda_i+\lambda_j}\left[1-e^{-(\lambda_i+\lambda_j)\Delta}\right].$$

Finally, *sandwiching* this term inbetween the eigenvector matrix results in

$$\mathbb{V}\{\mathbf{X}_{t+\Delta}|\mathbf{X}_t\} = \mathbf{E}\mathbf{V}(\Delta)\mathbf{E}^\top.$$

The unconditional covariance can be obtained using the limit

$$\mathbf{V}_{ij}(\infty) = \lim_{\Delta\to\infty}\mathbf{V}_{ij}(\Delta) = \frac{\bar{S}_{ij}}{\Lambda_i+\Lambda_j}$$

such that

$$\mathbb{V}\{\mathbf{X}_t\} = \mathbf{E}\mathbf{V}(\infty)\mathbf{E}^\top.$$

Example: Dynamic Nelson-Siegel

1. Inspired by Diebold and Li (2006), we will convert the static Nelson-Siegel model to a dynamic one
2. We will posit the dynamics of the latent states using our previous findings and derive the transition equation.
3. We will construct the measurement equation that determines the relationship between the latent states and the observed yields.
4. We then estimate the parameters using the yield data from Gürkaynak et al. (2007) by applying the Kalman filter in conjunction with the prediction error decomposition.
5. Finally, we will discuss some aspects of this model that are less desirable.

Constructing a dynamic Nelson-Siegel model

The standard Nelson-Siegel uses four parameters to define the yield curve

$$y(\tau) = \beta_0 + \beta_1 \frac{\left(1 - e^{-\lambda\tau}\right)}{\lambda\tau} + \beta_2 \left(\frac{\left(1 - e^{-\lambda\tau}\right)}{\lambda\tau} - e^{-\lambda\tau} \right).$$

We now consider $(\beta_0, \beta_1, \beta_2)$ as latent states $\mathbf{X}_t = (L_t, S_t, C_t)$

$$\begin{aligned} y_t(\tau) &= L_t + S_t \frac{\left(1 - e^{-\lambda\tau}\right)}{\lambda\tau} + C_t \left(\frac{\left(1 - e^{-\lambda\tau}\right)}{\lambda\tau} - e^{-\lambda\tau} \right) \\ &= B(\tau) \cdot \mathbf{X}_t. \end{aligned}$$

this leaves only λ as a "true" parameter. Notice that yields are linear functions of the latent states, so we can easily go from this to a measurement equation.

How should we model the dynamics of $\mathbf{X}_t$?

- Visual inspection of the time series of principal factors from the PCA gives us a clue.

Imposing dynamics on the latent states

Based on visual inspection of the time series of principal factors, it appears reasonable to posit multivariate Ornstein-Uhlenbeck type dynamics on the latent states, such that

$$\mathrm{d}\mathbf{X}_t = \mathbf{K}\left(\theta - \mathbf{X}_t\right)\mathrm{d}t + \Sigma \mathrm{d}W_t^{\mathbb{P}}$$

with (for simplicity) we assume that the matrices are diagonal:

$$\mathbf{K} = \begin{pmatrix} \kappa_1 & 0 & 0 \\ 0 & \kappa_2 & 0 \\ 0 & 0 & \kappa_3 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}$$

Let's apply the methods discussed previously to discretize this SDE into a transition equation:

$$\mathbf{X}_{t+\Delta} = \mathbf{A}\mathbf{X}_\mathbf{t} + \mathbf{b} + \omega_t, \qquad \omega_t \sim N(0, \mathbf{Q}).$$

For $\mathbf{A}$ and $\mathbf{b}$:

$$\mathbf{A} = \begin{bmatrix} e^{-\kappa_1\Delta} & 0 & 0 \\ 0 & e^{-\kappa_2\Delta} & 0 \\ 0 & 0 & e^{-\kappa_3\Delta} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} (1-e^{-\kappa_1\Delta})\theta_1 \\ (1-e^{-\kappa_2\Delta})\theta_2 \\ (1-e^{-\kappa_3\Delta})\theta_3 \end{bmatrix}$$

Our covariance

Remember that $\mathbf{K} = \text{Diag}(\kappa_1, \kappa_2, \kappa_3)$, thus the eigenvalues are given by $\lambda_1 = \kappa_1$, $\lambda_2 = \kappa_2$ and $\lambda_3 = \kappa_3$, and the eigenvector matrix is simply $\mathbf{I}_3$. This means that the conditional covariance is given by

$$\begin{aligned}\mathbb{V}\{\mathbf{X}_{t+\Delta}|\mathbf{X}_t\} &= \mathbf{E}\mathbf{V}(\Delta)\mathbf{E}^\top = \mathbf{V}(\Delta) \\ &= \begin{bmatrix} \frac{\sigma_1^2}{2\kappa_1}\left(1 - e^{-2\kappa_1\Delta}\right) & 0 & 0 \\ 0 & \frac{\sigma_2^2}{2\kappa_2}\left(1 - e^{-2\kappa_2\Delta}\right) & 0 \\ 0 & 0 & \frac{\sigma_3^2}{2\kappa_3}\left(1 - e^{-2\kappa_3\Delta}\right) \end{bmatrix}\end{aligned}$$

The unconditional covariance is

$$\mathbb{V}\{\mathbf{X}_t\} = \lim_{\Delta\to\infty} \mathbf{V}(\Delta) = \begin{bmatrix} \frac{\sigma_1^2}{2\kappa_1} & 0 & 0 \\ 0 & \frac{\sigma_2^2}{2\kappa_2} & 0 \\ 0 & 0 & \frac{\sigma_3^2}{2\kappa_3} \end{bmatrix}$$

The DNS transition equation: summary

We now have a complete dynamic model

$$\mathbf{X}_{t+\Delta} = \mathbf{A}\mathbf{X}_t + \mathbf{b} + \omega_t,$$

where $\omega_t \sim N(0, \mathbf{Q})$ with $\mathbf{Q} = \mathbb{V}\{\mathbf{X}_{t+\Delta}|\mathbf{X}_t\}$

Furthermore, as the initial distribution $\mathbf{X}_0$ we can use the unconditional distribution

$$\mathbf{X}_0 \sim N(\theta, \mathbb{V}\{\mathbf{X}_t\})$$

where the covariance was found on the previous slide.

So what are we missing?
The measurement equation, i.e. the system matrices **C**, **d** and **R**

Dynamic nelson-siegel measurement equation

Assume we observe yields with maturities $\tau_1, \tau_2, \ldots, \tau_M$.
From the Nelson-Siegel factor loadings we can readily construct the measurement equation

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{d} + \epsilon_t$$

with $\epsilon_t \sim N(0, \mathbf{R})$.

There's no constant in the yield formula, so $\mathbf{d} = 0$. Furthermore,

$$\mathbf{C} = \begin{pmatrix} 1 & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} - e^{-\lambda\tau_1} \\ \vdots & \vdots & \vdots \\ 1 & \frac{1-e^{-\lambda\tau_m}}{\lambda\tau_m} & \frac{1-e^{-\lambda\tau_m}}{\lambda\tau_m} - e^{-\lambda\tau_m} \end{pmatrix}.$$

We cannot determine $\mathbf{R}$ from our model, we have to choose a parametrization ourselves. Let's go for the simplest one:

$$\mathbf{R} = \sigma_R^2 \mathbf{I}_{m \times m}$$

for some scalar $\sigma_R > 0$.

Overview of the estimation technique

The estimation procedure takes as input

- the measurements $\mathbf{y}_1, \ldots, \mathbf{y}_N$ and their maturities $\tau_1, \ldots, \tau_m$,
- a time step Δ,
- an initial guess for the parameters $\Theta = (\lambda, \kappa_1, \kappa_2, \kappa_3, \theta_1, \theta_2, \theta_3, \sigma_1, \sigma_2, \sigma_3, \sigma_R)$.

1. For parameters Θ:
   - 1.1 Compute the system matrices of the transition equation $(\mathbf{A}, \mathbf{b}, \mathbf{Q})$ using the timestep Δ, as well as the unconditional mean and covariance $(\mathbf{m}_1, \mathbf{P}_1)$.
   - 1.2 Construct the system matrices of the measurement equation $(\mathbf{C}, \mathbf{d}, \mathbf{R})$ from the parameters and the yield maturites $\tau_1, \ldots, \tau_m$.
   - 1.3 Initialize the Kalman filter at $(\mathbf{m}_1, \mathbf{P}_1)$ and perform the recursion from $n = 1$ to N. Save $\mathbf{v}_n$ and $\mathbf{S}_n$.
   - 1.4 Evaluate the negative loglikelihood using the prediction error decomposition form of the likelihood.
2. If your termination criterion is not satisfied, construct a new guess for θ and repeat (1.1-1.4). If it is satisfied, then the final Θ corresponds to your ML parameters and the corresponding time series $\mathbf{m}_{n|n}$ is your estimated latent states.

DNS parameter estimates

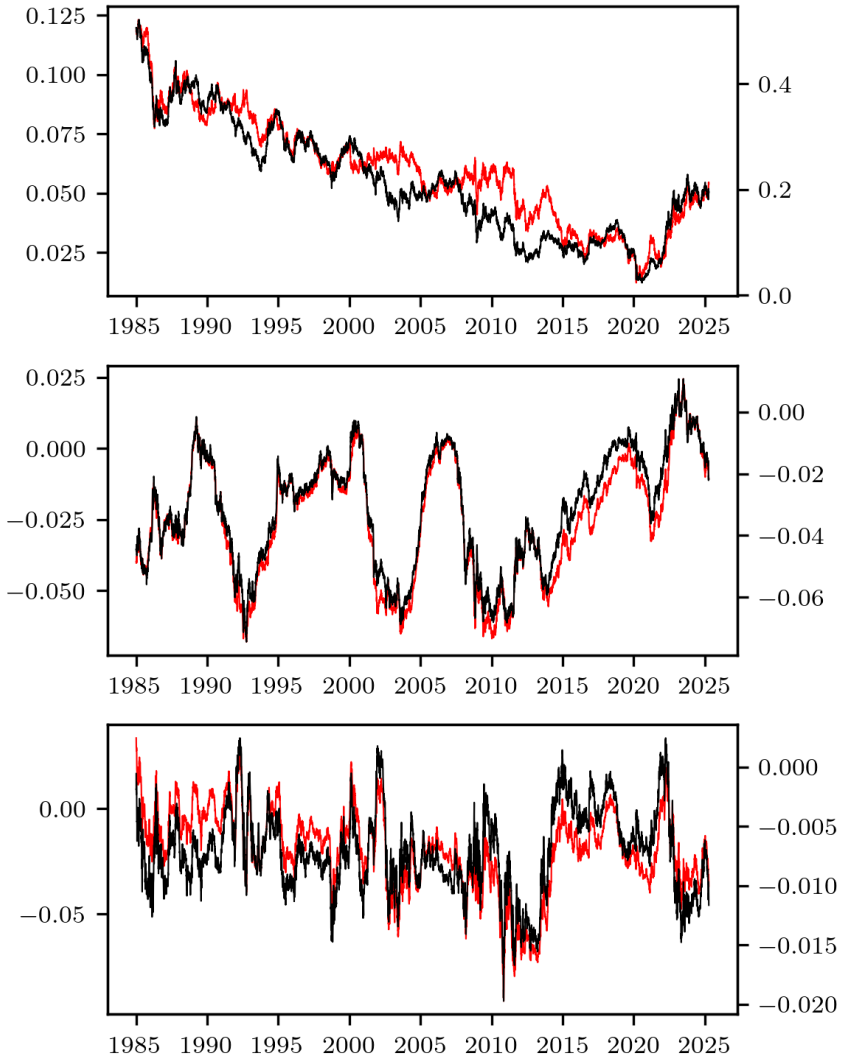
Data: daily yields with maturities $1, 2, \ldots, 20$ years.

The estimated dynamics of $\mathbf{X}_t$ are given by

$$\begin{aligned} \mathrm{d}\mathbf{X}_t = &\begin{bmatrix} 0.042 & 0 & 0 \\ 0 & 0.112 & 0 \\ 0 & 0 & 0.484 \end{bmatrix} \left[\begin{bmatrix} 0.074 \\ -0.025 \\ -0.018 \end{bmatrix} - \mathbf{X}_t \right] \mathrm{d}t \\ &+ \begin{bmatrix} 0.009 & 0 & 0 \\ 0 & 0.009 & 0 \\ 0 & 0 & 0.020 \end{bmatrix} \mathrm{d}W^{\mathbb{P}}(t), \end{aligned}$$

where $\lambda = 0.440$ and $\sigma_R = 0.0005$.

DNS states versus principal factors



Red lines are the estimated states of the DNS model, black lines are the principal factors

Bad news

1. Filipović (1999) shows that no matter what stochastic dynamics we impose on our latent states $\mathbf{X}(t)$, it's impossible to preclude arbitrage from the resulting yield curves.
2. The proof is rather abstract, but intuitively it should not suprise you too much. The model was constructed without at any point referencing any arbitrage theory.
3. The final part of this lecture will try to remedy this defiency. We do so by returning to affine term structure models.

PART 3:
DUAL MEASURE MODELS

What now?

- To ensure that our model is arbitrage-free, we will make use of affine term structure models.
- We will define the model under the risk-neutral measure $\mathbb{Q}$, and then consider specifications of the market price of risk that ensures that the physical dynamics remain tractable (affine).
- Ultimately, this allows us to construct an *arbitrage-free Nelson-Siegel* models proposed by Christensen et al. (2011).
- Note that these techniques can be used for any affine term structure model.

Reminder: Affine term structure models

Remember that a general affine term structure model can be represented using $r(t) = \delta_0 + \delta_1 \cdot \mathbf{X}_t$, where

$$\mathrm{d}\mathbf{X}_t = (\mathbf{K}_0 + \mathbf{K}_1\mathbf{X}_t)\,\mathrm{d}t + \Sigma S(\mathbf{X}_t)\mathrm{d}W^{\mathbb{Q}}(t).$$

The diagonal matrix $S(\mathbf{X})$ is given for all i by

$$(S(\mathbf{X}))_{ii} = \sqrt{\alpha_i + \beta_i \cdot \mathbf{X}}.$$

Zero-coupon bonds are then given by

$$P(t, T) = e^{A(T-t)+B(T-t)\cdot\mathbf{X}_t},$$

where $A(\cdot)$ and $B(\cdot)$ are solutions to the Ricatti equations. Using the expression for yields, this means that

$$y(t, T) = -\frac{A(T-t)}{T-t} - \frac{B(T-t)}{T-t} \cdot \mathbf{X}(t).$$

Yields are thus affine functions of the latent states.

Introducing the market price of risk

Remember that Girsanov (see first week) tells us that

$$\mathrm{d}W^{\mathbb{Q}}(t) = \mathrm{d}W^{\mathbb{P}}(t) - \Gamma_t \mathrm{d}t.$$

Notice that we start with risk-neutral dynamics

$$\mathrm{d}\mathbf{X}(t) = (\mathbf{K}_0 + \mathbf{K}_1 \mathbf{X}_t)\mathrm{d}t + \Sigma S(\mathbf{X}_t)\mathrm{d}W^{\mathbb{Q}}(t),$$

then this in turn implies that

$$\mathrm{d}\mathbf{X}(t) = (\mathbf{K}_0 + \mathbf{K}_1 \mathbf{X}_t - \Sigma S(\mathbf{X}_t)\Gamma_t)\,\mathrm{d}t + \Sigma S(\mathbf{X}_t)\mathrm{d}W^{\mathbb{P}}(t),$$

What do we want Γ_t to be like? Preferably, we would like

- a Γ_t that ensures that the process $\mathbf{X}_t$ is sufficently tractable under $\mathbb{P}$. The dynamics should be affine under both measures.
- a Γ_t that is sufficiently flexible to capture empirical reality.

Completely affine market price of risk

A first idea could be to set

$$\Gamma_t = S(\mathbf{X}_t)\gamma_1$$

for some vector γ_1. This specification is often denoted as the *completely affine market price of risk*.
The drift remains affine under $\mathbb{P}$ since

$$\begin{aligned} S(\mathbf{X}_t)S(\mathbf{X}_t)\gamma_1 &= \begin{bmatrix} (\alpha_1 + \beta_1 \cdot \mathbf{X}_t)\,\gamma_{1,1} \\ \vdots \\ (\alpha_d + \beta_d \cdot \mathbf{X}_t)\,\gamma_{1,d} \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1\gamma_{1,1} \\ \vdots \\ \alpha_d\gamma_{1,d} \end{bmatrix} + \begin{bmatrix} \gamma_{1,1}\beta_{1,1} & \cdots & \gamma_{1,1}\beta_{1,d} \\ \vdots & \ddots & \vdots \\ \gamma_{1,d}\beta_{d,1} & \cdots & \gamma_{1,d}\beta_{d,d} \end{bmatrix} \mathbf{X}_t = T_0 + T_1\mathbf{X}_t \end{aligned}$$

such that the drift under $\mathbb{P}$ becomes

$$\mu^{\mathbb{P}}(\mathbf{X}_t) = (\mathbf{K}_0 - \Sigma T_0) + (\mathbf{K}_1 - \Sigma T_1)\,\mathbf{X}_t$$

When we use this specification, Γ_t is driven exclusively by variations in $S(\mathbf{X}_t)$. Is this okay?

We need more flexiblity

> *The standard class of affine models produces poor forecasts of future Treasury yields. Better forecasts are generated by assuming that yields follow random walks. The failure of these models is driven by one their key features: Compensation for risk is a multiple of the variance of the risk.*

Duffee (2002)

$\Rightarrow$ Positing $\Gamma_t = S(\mathbf{X}_t)\gamma_1$ is ultimately not empirically justified. We need something more flexible.

Essentially affine market price of risk

Duffee (2002) proposes the extending form

$$\Gamma_t = S(\mathbf{X}_t)\gamma_1 + S^-(\mathbf{X}_t)\Gamma_2\mathbf{X}_t$$

where

$$\left(S^-(\mathbf{X}_t)\right)_{ii} = \begin{cases} \frac{1}{\sqrt{\alpha_i+\beta_i\cdot\mathbf{X}_t}} & \text{if } \inf\left(\alpha_i+\beta_i\cdot\mathbf{X}_t\right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Why is this necessary? If $\alpha_i + \beta_i \cdot \mathbf{X}_t$ goes to zero, its' inverse goes to infinity, and thus also Γ_t.
Notice that

$$\left(S(\mathbf{X_t})S^-(\mathbf{X}_t)\right)_{ii} = \begin{cases} 1 & \text{if } \inf(\alpha_i+\beta_i\cdot\mathbf{X}_t) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Define $\mathbf{I}^- \triangleq S(\mathbf{X_t})S^-(\mathbf{X}_t)$, then this means that the extra term in the drift becomes

$$\Sigma S(\mathbf{X}_t)\Gamma_t = \Sigma\left[S(\mathbf{X}_t)S(\mathbf{X}_t)\gamma_1 + \mathbf{I}^-\Gamma_1\mathbf{X}_t\right]$$

Variations in the MPR can now depend on both $S(\mathbf{X}_t)$ (volatility) as well as directly on the states $\mathbf{X}_t$.

What do we get in the Gaussian case?

Consider a Gaussian affine process again, where

$$\mathrm{d}\mathbf{X}_t = (\mathbf{K}_0 + \mathbf{K}_1\mathbf{X}_t)\,\mathrm{d}t + \Sigma\mathrm{d}W^{\mathbb{Q}}(t),$$

where $S(\mathbf{X}_t) = \mathbf{I}_d$ is simply the identity matrix.
This corresponds to setting $\alpha_i = 1$ and $\beta_i = 0$ for all i.
Completely affine gives us

$$\begin{aligned}\mathrm{d}\mathbf{X}_t &= (\mathbf{K}_0 + \mathbf{K}_1\mathbf{X}_t - \Sigma\gamma_1)\,\mathrm{d}t + \Sigma\mathrm{d}W^{\mathbb{P}}(t)\\ &= \left(\mathbf{K}_0^P + \mathbf{K}_1\mathbf{X}_t\right)\mathrm{d}t + \Sigma\mathrm{d}W^{\mathbb{P}}(t)\end{aligned}$$

with $\mathbf{K}_0^P = \mathbf{K}_0 - \Sigma\gamma_1$.
Essentially affine gives us

$$\begin{aligned}\mathrm{d}\mathbf{X}_t &= (\mathbf{K}_0 + \mathbf{K}_1\mathbf{X}_t - \Sigma\left(\gamma_1 + \Gamma_2\mathbf{X}_t\right))\,\mathrm{d}t + \Sigma\mathrm{d}W^{\mathbb{P}}(t)\\ &= \left(\mathbf{K}_0^P + \mathbf{K}_1^P\mathbf{X}_t\right)\mathrm{d}t + \Sigma\mathrm{d}W^{\mathbb{P}}(t)\end{aligned}$$

with $\mathbf{K}_0^P$ as before, but now $\mathbf{K}_1^P = \mathbf{K}_1 - \Sigma\Gamma_2\mathbf{X}_t$.

Affine dual-measure term structure model: summary

The dynamics of $\mathbf{X}_t$ are given under $\mathbb{Q}$ and $\mathbb{P}$ as

$$\mathrm{d}\mathbf{X}_t = \mathbf{K}\left[\theta - \mathbf{X}_t\right]\mathrm{d}t + \Sigma S(\mathbf{X}_t)\mathrm{d}W_t^{\mathbb{Q}}$$
$$\mathrm{d}\mathbf{X}_t = \mathbf{K}\left[\theta - \mathbf{X}_t\right]\mathrm{d}t - \Sigma S(\mathbf{X}_t)\Gamma_t\mathrm{d}t + \Sigma S(\mathbf{X}_t)\mathrm{d}W_t^{\mathbb{P}}$$

where the market price of risk Γ_t is either essentially or completely affine.

$$\text{Pricing measure } \mathbb{Q} \underset{\text{Market price of risk } \Gamma_t}{\longleftrightarrow} \text{Physical measure } \mathbb{P}$$

- $\mathbb{Q}$-dynamics are estimated from the crossection of yields.
- $\mathbb{P}$-dynamics are estimated from the time series.

Constructing the AFNS model 1/3

Returning to the example problem, we wanted to find a model as close as possible to the dynamic Nelson-Siegel model, that is also arbitrage-free.
Let's start with the risk-neutral measure $\mathbb{Q}$.

$$\mathrm{d}\mathbf{X}_t = \mathbf{K}\left[\theta - \mathbf{X}_t\right]\mathrm{d}t + \Sigma \mathrm{d}W_t^{\mathbb{Q}}$$

where $r(t) = \rho_0 + \rho_1 \cdot \mathbf{X}_t$.
We know that yields will be given as

$$y_t(\tau) = -\frac{A(\tau)}{\tau} - \frac{B(\tau)}{\tau} \cdot \mathbf{X}_t,$$

where $A(\cdot)$ and $B(\cdot)$ are solutions to the Ricatti ODEs. Remember that B is given by

$$\frac{\mathrm{d}B(\tau)}{\partial \tau} = -\mathbf{K}^\top B(\tau) - \rho_1$$

What should the $\mathbb{Q}$-dynamics and (ρ_0, ρ_1) be to ensure that the yields have the same factor loadings as in the DNS?

Constructing the AFNS model 2/3

DNS factor loadings are given by:

$$y_t(\tau) = L_t + S_t \frac{\left(1 - e^{-\lambda\tau}\right)}{\lambda\tau} + C_t \left(\frac{\left(1 - e^{-\lambda\tau}\right)}{\lambda\tau} - e^{-\lambda\tau} \right)$$

- The short rate in the static Nelson-Siegel model is given as $\beta_0 + \beta_1$. Inspired by this, we can set $\rho_0 = 0$ and $\rho_1 = (1, 1, 0)^\top$.
- The level term is just a constant 1, suggesting that $B_1(\tau) = -\tau$. Notice that this is already fulfilled by the ODE

  $$\frac{\partial B_1(\tau)}{\partial \tau} = -1$$

  with $B_1(0) = 0$.
- $B_2(\tau)$ has a form similar to the Vasicek model, thus maybe

  $$\frac{\partial B_2(\tau)}{\partial \tau} = -\lambda B_2(\tau) - 1$$

  with $B_2(0) = 0$ works?
- The third one is less easy to see. Lets try

  $$\frac{\partial B_3(\tau)}{\partial \tau} = \lambda B_2(\tau) - \lambda B_3(\tau) - 1$$

  for $B_3(0) = 0$.

Constructing the AFNS model 3/3

Collecting the coefficients of the ODEs, we see that

$$\mathbf{K} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda & -\lambda \\ 0 & 0 & \lambda \end{bmatrix}$$

results in the desired factor loadings.

Making the model identifiable
The next step is to realize that the model is not identified when θ^Q is free. We can just fix it to $\theta^Q = 0$. Furthermore, Σ must be lower/upper triangular.

We still need to consider the $A(\cdot)$ term, which is

$$\frac{\partial A(\tau)}{\partial \tau} = -\frac{1}{2} B^\top(\tau) \Sigma \Sigma^\top B(\tau),$$

with $A(0) = 0$. Notice that the right side does not depend on $A(\cdot)$, thus this has the trivial solution

$$A(\tau) = \frac{1}{2} \int_0^\tau B^\top(u) \Sigma \Sigma^\top B(u) \mathrm{d}u$$

See Christensen et al. (2011) for the closed-form solution.

Summary: risk-neutral measure

Thus the risk-neutral dynamics are given by

$$\mathrm{d}\mathbf{X}_t = -\mathbf{K}\mathbf{X}_t\mathrm{d}t + \Sigma\mathrm{d}W_t^{\mathbb{Q}},$$

where

$$\Sigma = \begin{bmatrix} \sigma_{11} & 0 & 0 \\ \sigma_{21} & \sigma_{22} & 0 \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

Notice that Σ now impacts the prices through $A(\tau)$. This is the convexity effect. For $\mathbf{X}_t = (L_t, S_t, C_t)^\top$ we then get

$$\begin{aligned} y(\mathbf{X}_t, \tau) &= -\frac{B(\tau)\cdot\mathbf{X}_t}{\tau} - \frac{A(\tau)}{\tau} \\ &= \underbrace{L_t + \frac{1-e^{-\lambda\tau}}{\lambda\tau}S_t + \left[\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right]C_t}_{\text{Dynamic Nelson Siegel}} + \underbrace{\left(-\frac{A(\tau)}{\tau}\right)}_{\text{Convexity adjustment}} \end{aligned}$$

Thus this additional convexity adjustment $-\frac{A(\tau)}{\tau}$ ensures that the resulting yield curve is arbitrage-free.

The physical measure

Inspired by Duffee (2002), we use the essentially affine market price of risk

$$\begin{aligned}\Gamma(\mathbf{X}_t) &= S(\mathbf{X}_t)\lambda_0 + S^{-}(\mathbf{X}_t)\gamma_1\mathbf{X}_t \\ &= \gamma_1 + \Gamma_2\mathbf{X}_t.\end{aligned}$$

The physical dynamics are thus given by

$$\begin{aligned}\mathrm{d}\mathbf{X}_t &= \left(-\mathbf{K}\mathbf{X}_t - \Sigma\gamma_1 - \Sigma\Gamma_2\mathbf{X}_t\right)\mathrm{d}t + \Sigma\mathrm{d}W^{\mathbb{P}}(t) \\ &= \left(-\Sigma\gamma_1 - \left(\mathbf{K} + \Sigma\Gamma_2\right)\mathbf{X}_t\right)\mathrm{d}t + \Sigma\mathrm{d}W^{\mathbb{P}}(t) \\ &= \mathbf{K}^P\left(\theta^P - \mathbf{X}_t\right)\mathrm{d}t + \Sigma\mathrm{d}W^{\mathbb{P}}(t).\end{aligned}$$

Notice that the very flexible form of Γ ensures that we can find a γ_1 and Γ_2 for **any** $\left(\mathbf{K}^P, \theta^P\right)$.

An AFNS model

Let's consider a simple AFNS model given by requiring diagonal Σ and $\mathbf{K}^P$.

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_1 & 0 \\ 0 & 0 & \sigma_1 \end{bmatrix} \qquad \mathbf{K}^P = \begin{bmatrix} \kappa_1 & 0 & 0 \\ 0 & \kappa_2 & 0 \\ 0 & 0 & \kappa_3 \end{bmatrix}$$

This means that the state-space model is identitical to the dynamic Nelson-Siegel model **except** that the $\mathbf{d}$ vector is now given by the convexity adjustment

$$\mathbf{d} = \begin{bmatrix} -\frac{A(\tau_1)}{\tau_1} \\ \vdots \\ -\frac{A(\tau_m)}{\tau_m} \end{bmatrix}.$$

Let's estimate this model!

AFNS parameter estimates

AFNS

$$\mathrm{d}\mathbf{X}_t = \begin{bmatrix} 0.09 & 0 & 0 \\ 0 & 0.011 & 0 \\ 0 & 0 & 0.573 \end{bmatrix} \left[\begin{bmatrix} 0.086 \\ -0.029 \\ -0.025 \end{bmatrix} - \mathbf{X}_t \right] \mathrm{d}t + \begin{bmatrix} 0.004 & 0 & 0 \\ 0 & 0.009 & 0 \\ 0 & 0 & 0.021 \end{bmatrix} \mathrm{d}W^{\mathbb{P}}(t),$$

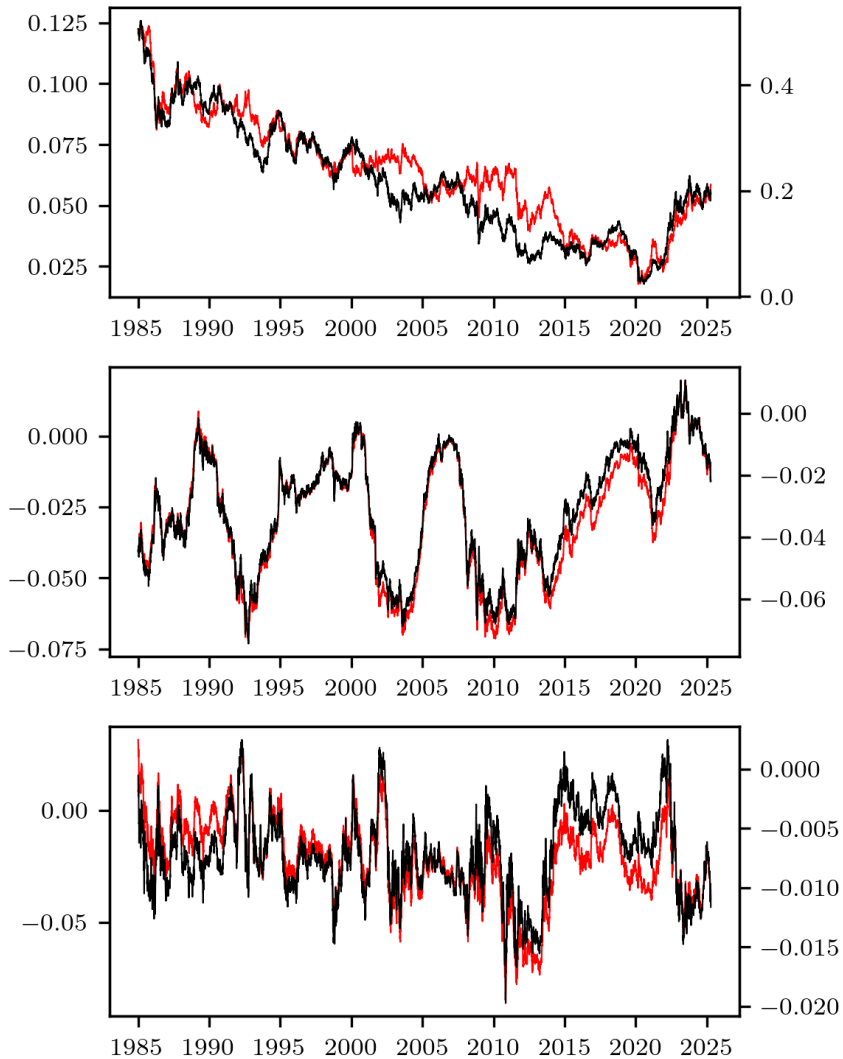
where $\lambda = 0.379$ and $\sigma_R = 0.0005$.

DNS

$$\mathrm{d}\mathbf{X}_t = \begin{bmatrix} 0.042 & 0 & 0 \\ 0 & 0.112 & 0 \\ 0 & 0 & 0.484 \end{bmatrix} \left[\begin{bmatrix} 0.074 \\ -0.025 \\ -0.018 \end{bmatrix} - \mathbf{X}_t \right] \mathrm{d}t + \begin{bmatrix} 0.009 & 0 & 0 \\ 0 & 0.009 & 0 \\ 0 & 0 & 0.020 \end{bmatrix} \mathrm{d}W^{\mathbb{P}}(t),$$

where $\lambda = 0.440$ and $\sigma_R = 0.0005$.

AFNS states versus principal factors



Red lines are the estimated states of the AFNS model, black lines are the principal factors

Extending the estimation procedure

- **What if we wanted a general affine process, not just a Gaussian one?** Formula for yields remains affine, but the noise in the transition equation

$$\mathbf{X}_{t+\Delta} = \mathbf{A}\mathbf{X}_t + \mathbf{b} + \omega_t,$$

  is no longer Gaussian. Use Quasi-maximum likelihood and approximate the distribution of the noise as a Gaussian with the same covariance as the true ω_t.
- **What if we wanted to use measurements that are given by a nonlinear function of $\mathbf{X}_t$, like e.g., option prices?** We can approximate the distribution of a general

$$\mathbf{y}_t = h(\mathbf{X}_t) + \epsilon_t$$

  as Gaussian. This is done by using techniques like the *extended Kalman filter* or the *unscented* Kalman filter. Both are more demanding to implement.
- **What if we have missing observations, or the measurement equation is time-inhomogenous?** This just makes the implementation a bit harder from a programming point of view, but it works the same.

The point being: Although the classical Kalman filter assumptions are rather strict, there's an enormous litterature on extending it to nonlinear / non-Gaussian cases. It's not guaranteed to work, of course.

References I

Christensen, J. H., Diebold, F. X., and Rudebusch, G. D. (2011). The affine arbitrage-free class of nelson–siegel term structure models. *Journal of Econometrics*, 164(1):4–20.

Christensen, J. H., Lopez, J. A., and Rudebusch, G. D. (2015). Analytical formulas for the second moment in affine models with stochastic volatility. Technical report.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364.

Duffee, G. R. (2002). Term premia and interest rate forecasts in affine models. *The Journal of Finance*, 57(1):405–443.

Filipović, D. (1999). A note on the nelson–siegel family. *Mathematical Finance*, 9(4):349–359.

Gürkaynak, R. S., Sack, B., and Wright, J. H. (2007). The u.s. treasury yield curve: 1961 to the present. *Journal of Monetary Economics*, 54(8):2291–2304.

Harvey, A. C. (1991). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge, England.

Nelson, C. R. and Siegel, A. F. (1987). Parsimonious modeling of yield curves. *The Journal of Business*, 60(4):473.

References II

Särkkä, S. and Svensson, L. (2023). *Bayesian filtering and smoothing*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge, England, 2 edition.

Svensson, L. E. (1994). *Estimating and Interpreting Forward Interest Rates: Sweden 1992 - 1994*.