# Time Series Analysis and Forecasting with ADAM

Ivan Svetunkov

2020-06-23

# Contents

# Preface

This textbook uses two packages from R, namely `greybox`, which focuses on forecasting using regression models, and `smooth`, which implements Single Source of Error (SSOE) state space models for purposes of time series analysis and forecasting. The textbook focuses on explaining how ADAM ("ADAM is Dynamic Adaptive Model" - recursive acronym), one of the `smooth` functions (introduced in v3.0.0) works, also showing how it can be used in practice with examples from R. ADAM is a state space model based on exponential smoothing in ETS format and ARIMA. It encompasses both models and is expanded by introducing:

1. Explanatory variables (including time varying parameters);
2. Multiple frequencies;
3. Handling of intermittent data (data with natural zeroes);
4. Handling of missing data;
5. Variables and components selection and combination;
6. Analysis of parameters of the model;
7. And other minor features.

All these extentions are needed in order to solve specific real life problems, so we will have examples and case studies later in the book, in order to see how all of this can be used.

If you want to run exampels from the textbook, two packages are needed (Svetunkov, 2020a,b):

```
install.packages("greybox")
install.packages("smooth")
```

Some explanations of functions from the packages are given in my blog: Package greybox for R, Package smooth for R.

A very important thing to note is that this textbook **does not use `tidyverse` packages**. I like base R, and, to be honest, I am sure that `tidyverse` packages are great, but I have never needed them in my research. So, I will not use pipeline operators, `tibble` or `tsibble` objects and `ggplot2`. It is assumed throughout the textbook that you can do all those nice tricks on your own if you want to.

If you want to get in touch with me, there are lots of ways to do that: comments section on any page of my website, my Russian website, vk.com, Facebook, Linkedin, Twitter.

You can also find me on ResearchGate, StackExchange and StackOverflow, although I'm not really active there. Finally, I also have GitHub account.

# Chapter 1

# Introduction

I have started writing this book in 2020 during the COVID-19 pandemic, having figured out that it has been more than 10 years since the publishing of the fundamental textbook of (Hyndman et al., 2008), who discuss ETS (Error-Trend-Seasonality) framework in the Single Source of Error (SSOE) form. If you are interested in knowing more about exponential smoothing, then this is a must read material on the topic. However, there has been some progress in the area since 2008, and I have developed some models and functions based on ETS, making the framework a bit more flexible and general. Given that the publication of all the aspects of these models in peer-reviewed journals is difficult and challenging, I have decided to summarise all the progress in the book, showing what happens inside the models and how to use the functions in different cases.

Before we move to nitty gritty details of the models, it is important to agree what we are talking about. So, here is a couple of definitions:

- **Statistical model** (or 'stochastic model', or just 'model' in this textbook) is a 'mathematical representation of a real phenomenon with a complete specification of distribution and parameters' (Svetunkov and Boylan, 2019). Very roughly, the statistical model is something that contains a structure (defined by its parameters) and a noise that follows some distribution.
- **True model** is the idealistic statistical model that is correctly specified (has all the necessary components in correct form), applied to the data in population. By this definition, true model is never reachable in reality, but it is achievable in theory if for some reason we know what components and variables should definitely be in the model and have all the data in the world.
- **Estimated model** (aka 'used model' or 'applied model') is the statistical model that was constructed and estimated on the available sample of data.

This typically differs from the true model, because the latter is not known. Even if the specification of the true model is known for some reason, the parameters of the estimated model will differ from the true parameters due to sampling randomness.

- **Data generating process** (DGP) is an artificial statistical model, showing how the data could be generated in theory. This notion is utopic and can be used in simulation experiments in order to check, how the selected model with the specific estimator behave in a specific setting. In real life, the data is not generated from any process, but is usually based on complex interactions between different agents in a dynamic environment. Note that I make a distinction between DGP and true model, because I do not think that the idea of something being generated using a mathematical formula is helpful. Many statisticians will not agree with me on this distinction.
- **Forecasting method** is a mathematical procedure that generates point and / or interval forecasts, with or without a statistical model (Svetunkov and Boylan, 2019). Very roughly, forecasting method is just a way of producing forecasts that does not explain how the components of time series interact with each other. It might be needed in order to filter out the noise and extrapolate the structure.

Later in this book, we will see several examples of statistical models, forecasting methods, DGPs and other notions.

Note that this textbook assumes that the reader is familiar with introductory statistics and knows forecasting principles. (Hyndman and Athanasopoulos, 2018) can be a good start if you do not know either. We will also use elements of linear algebra to explain some modelling parts, but this will not be the main focus of the textbook and you will be able to skip the more challenging parts without jeopardising the main understanding of the topic.

## 1.1 Forecasting process and forecasts evaluation

Before we move to the discussion of models and their estimation it makes sense to discuss the forecasting process and how the forecasts should be evaluated.

### 1.1.1 Fixed origin versus rolling origin

### 1.1.2 Measuring accuracy of point forecasts

### 1.1.3 Measuring uncertainty

# Chapter 2

# A short introduction to main statistical ideas

Before moving forward and discussing distributions and models, it is also quite important to make sure that we undersatnd what **bias**, **efficiency** and **consistency** of estimates of parameters mean. Although there are strict statistical definitions of the aforementioned terms (you can easily find them in Wikipedia or anywhere else), I do not want to copy-paste them here, because there are only a couple of important points worth mentioning in our context.

**Bias** refers to the expected difference between the estimated value of parameter (on a specific sample) and the "true" one (in the true model). Having unbiased estimates of parameters is important because they should lead to more accurate forecasts (at least in theory). For example, if the estimated parameter is equal to zero, while in fact it should be 0.5, then the model would not take the provided information into account correctly and as a result will produce less accurate point forecasts and incorrect prediction intervals. In inventory context this may mean that we constantly order 100 units less than needed only because the parameter is lower than it should be.

**Efficiency** means, if the sample size increases, then the estimated parameters will not change substantially, they will vary in a narrow range (variance of estimates will be small). In the case with inefficient estimates the increase of sample size from 50 to 51 observations may lead to the change of a parameter from 0.1 to, let's say, 10. This is bad because the values of parameters usually influence both point forecasts and prediction intervals. As a result the inventory decision may differ radically from day to day. For example, we may decide that we urgently need 1000 units of product on Monday, and order it just to realise on Tuesday that we only need 100. Obviously this is an exaggeration, but no one wants to deal with such an erratically behaving model, so we need to have efficient estimates of parameters.

**Consistency** means that our estimates of parameters will get closer to the stable values (true value in the population) with the increase of the sample size. This is important because in the opposite case estimates of parameters will diverge and become less and less realistic. This once again influences both point forecasts and prediction intervals, which will be less meaningful than they should have been. In a way consistency means that with the increase of the sample size the parameters will become more efficient and less biased. This in turn means that the more observations we have, the better.

**Remark.** *There is a prejudice in the world of practitioners that the situation in the market changes so fast that the old observations become useless very fast. As a result many companies just through away the old data. Although, in general the statement about the market changes is true, the forecasters tend to work with the models that take this into account (e.g. Exponential smoothing, ARIMA, discussed in this book). These models adapt to the potential changes. So, we may benefit from the old data because it allows us getting more consistent estimates of parameters. Just keep in mind, that you can always remove the annoying bits of data but you can never un-throw away the data.*

Another important aspect to cover is what the term **asymptotic** means in our context. Here and after in this book, when this word is used, we refer to an unrealistic hypothetic situation of having all the data in the multiverse, where the time index $t \to \infty$. While this is impossible, the idea is useful, because asymptotic behaviour of estimators and models is helpful on large samples of data.

Finally, we will use different estimation techniques throughout this book, one of the main of which is **Maximum Likelihood Estimate** (MLE). We will not go into explanation of what specifically this is at this stage, but a rough understanding should suffice. In case of MLE, we assume that a variable follows a parametric distribution and that the parameters of the model that we use can be optimised in order to maximise the respective probability density function. The main advantages of MLE is that it gives consistent, asymptotically efficient and normal estimates of parameters.

Now that we have a basic understanding of these statistical terms, we can move to the next topic, distributions.

## 2.1 Theory of distributions

There are several probability distributions that will be helpful in the further chapters of this textbook. Here, I want to briefly discuss those of them that will be useful.

### 2.1.1 Normal distribution

Every statistical textbook has normal distribution. It is that one famous bell-curved distribution that every statistician likes because it is easy to work with and it is an asymptotic distribution for many other well-behaved distributions in some conditions (so called "Centrel Limit Theorem"). Here is the probability density function (PDF) of this distribution:

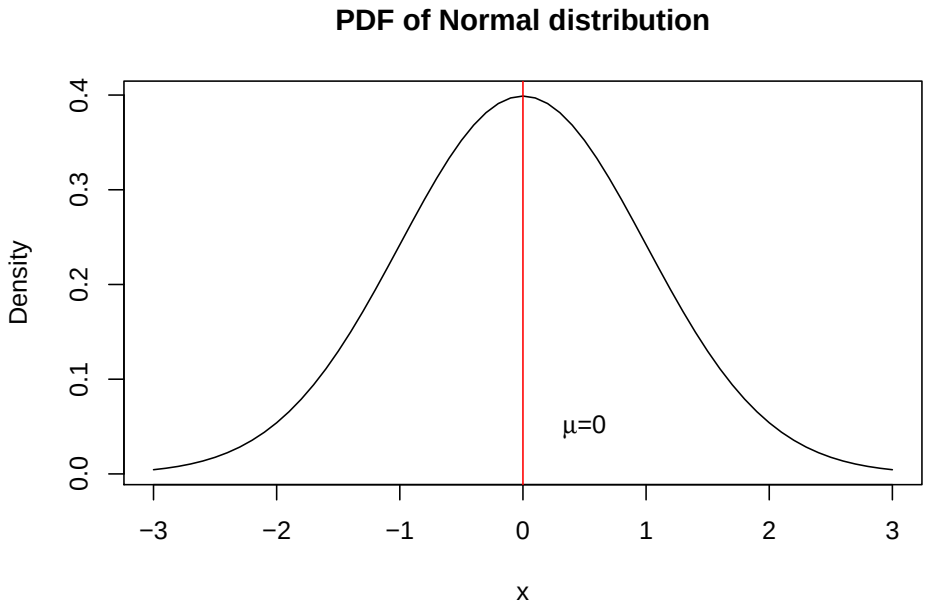
$$f(y_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \mu_t)^2}{2\sigma^2}\right), \tag{2.1}$$

where $y_t$ is the value of the response variable, $\mu_t$ is the mean on observation $t$ and $\sigma^2$ is the variance of the error term. The maximum likelihood estimate of $\sigma^2$ is:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \mu_t)^2, \tag{2.2}$$

which coincides with Mean Squared Error (MSE), discussed in the section 1.

And here how this distribution looks:




What we typically assume in the basic time series models is that a variable is random and follows normal distribution, meaning that there is a central tendency (in our case - the mean *mu*), around which the concentration of values is the highest and there are other potential cases, but their probabilitu of appearance reduces proportionally to the distance from the centre.

The normal distribution has skewness of zero and kurtosis of 3 (and excess kurtosis, being kurtosis minus three, of 0).

Additionally, if normal distribution is used for the maximum likelihood estimation of a model, it gives the same parameters as the minimisation of MSE would give.

### 2.1.2 Laplace distribution

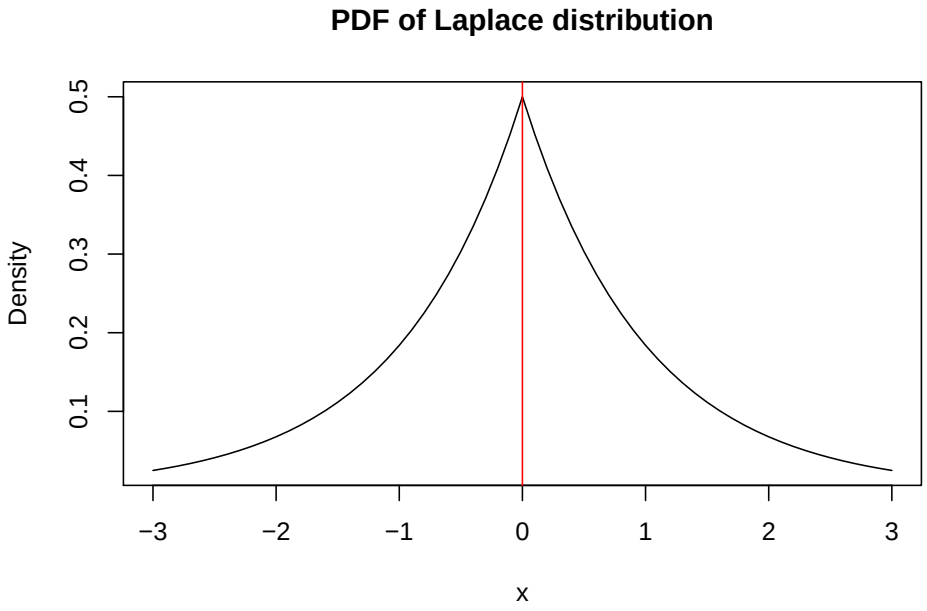
A more exotic distribution is Laplace, which has some similarities with Normal, but has higher excess. It has the following PDF:

$$f(y_t) = \frac{1}{2s} \exp\left(-\frac{|y_t - \mu_t|}{s}\right), \tag{2.3}$$

where $s$ is the scale parameter, which, when estimated using likelihood, is equal to the Mean Absolute Error (MAE):

$$\hat{s} = \frac{1}{T}\sum_{t=1}^{T} |y_t - \mu_t| . \tag{2.4}$$

It has the following shape:




Similar to the normal distribution, the skewnes of Laplace is equal to zero. However, it has fatter tails - its kurtosis is equal to 6 instead of 3.

### 2.1.3 S distribution

This is something relatively new, but not groun braking. I have derived S distribution few years ago, but have never written a paper on that. It has the

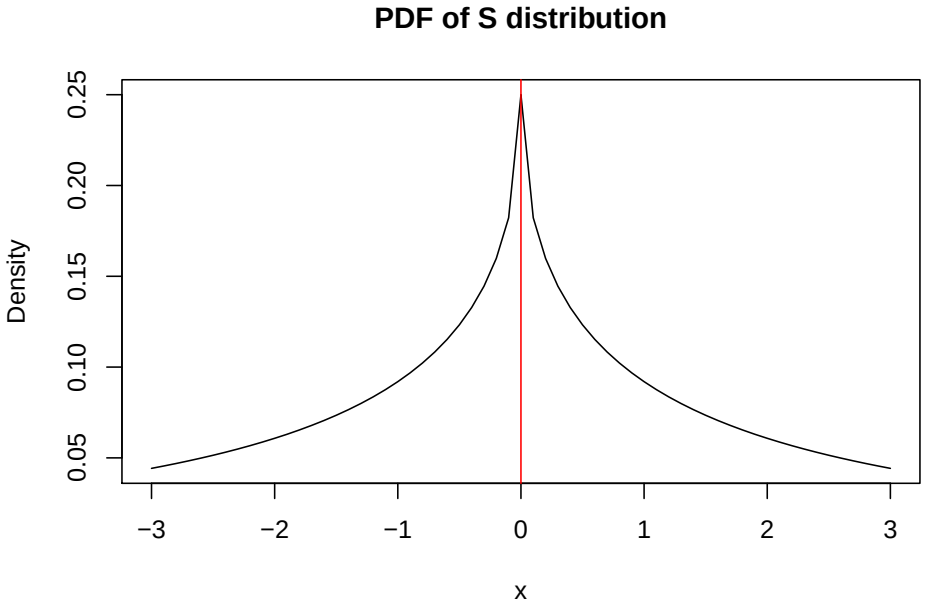
following density function:

$$f(y_t) = \frac{1}{4s^2} \exp\left(-\frac{\sqrt{|y_t - \mu_t|}}{s}\right), \tag{2.5}$$

where $s$ is the scale parameter. If estimated via maximum likelihood, the scale parameter is equal to:

$$\hat{s} = \frac{1}{2T} \sum_{t=1}^{T} \sqrt{|y_t - \mu_t|}, \tag{2.6}$$

which corresponds to the minimisation of a half of "Mean Root Absolute Error" or "Half Absolute Moment" (HAM). This is a more exotic type of scale, but the main benefit of this distribution is sever heavy tails - it has kurtosis of 25.2. It might be useful in cases of randomly occurring incidents and extreme values (Black Swans?).




### 2.1.4 Generalised normal distribution

Generalised normal distribution (as the name says) is a generalisation for normal distribution, which also includes Laplace and S as special cases. There are two versions of this distribution, we are mainly interested in the first one, which has the following PDF:

$$f(y_t) = \frac{\beta}{2s^{\beta^{-1}}\Gamma(\beta^{-1})} \exp\left(-\frac{|y_t - \mu_t|^{\beta}}{s}\right), \tag{2.7}$$

where $\beta$ is the shape parameter, and $s$ is the scale of the distribution, which, when estimated via MLE, is equal to:

$$\hat{s} = \frac{\beta}{T} \sum_{t=1}^{T} |y_t - \mu_t|^{\beta}, \tag{2.8}$$

which has MSE, MAE and HAM as special cases, when $\beta$ is equal to 2, 1 and 0.5 respectively. The parameter $\beta$ influences the kurtosis directly, it can be calculated for each special case as $\frac{\Gamma(5/\beta)\Gamma(1/\beta)}{\Gamma(3/\beta)^2}$. The higher $\beta$ is, the lower the kurtosis is.

The advantage of GN distribution is it's flexibility. In theory, it is possible to model extremely rare events with this distribution, if the shape parameter $\beta$ is fractional and close to zero. Alternatively, when $\beta \to \infty$, the distribution converges pointwise to the uniform distribution on $(\mu_t - \sqrt[\beta]{s}, \mu_t + \sqrt[\beta]{s})$.

Note that the estimation of $\beta$ is a difficult task, especially, when it is less than 2 - the MLE of it loose properties of consistency and asymptotic normality.

# Chapter 3

# Methods

We describe our methods in this chapter.

# Chapter 4

# Applications

Some *significant* applications are demonstrated in this chapter.

## 4.1 Example one

## 4.2 Example two

# Chapter 5

# Final Words

We have finished a nice book.

# Bibliography

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice, 2nd edition. Accessed on 01.04.2020*. OTexts: Melbourne, Australia.

Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing*. Springer Berlin Heidelberg.

Svetunkov, I. (2020a). *greybox: Toolbox for Model Building and Forecasting*. R package version 0.6.1.41007.

Svetunkov, I. (2020b). *smooth: Forecasting Using State Space Models*. R package version 2.6.0.

Svetunkov, I. and Boylan, J. E. (2019). Multiplicative State-Space Models for Intermittent Time Series.