Is more data always better?
Optimal data usage in non-stationary systems

Jakob Krause *

School of Business and Economics, Martin Luther University of Halle-Wittenberg

November 1, 2017

Abstract

Past Financial Crisis have shown that contemporary risk management models provide an unjustified sense of security and fail miserably in situations in which they are needed the most. In this paper we start from the assumption that risk is a notion that changes over time and therefore past datapoints only have limited explanatory for the current situation. Our objective is to derive the optimal amount of representative information by optimizing between the two adverse forces of estimator convergence, incentivising us to use as much data as possible, and the aforementioned non representativeness doing the opposite. In this endeavour a corner stone assumption of probability theory is weakened, identically distributed random variables and substituted by the assumption that the law of the data generating process changes over time. Hence, in this paper we give a quantitative theory on how to perform statistical analysis in non-ergodic systems. As an application we discuss the impact of a paragraph in the last iteration of proposals by the Basel Committee on Banking Regulation.

Classification: Mathematical Finance, Non-Ergodicity, Semimartingale Modeling, Risk Management, Banking Regulation

1 Introduction

At the beginning of the 2007 financial crisis Goldman Sachs' former CFO David Viniar told the Financial Times

> *We were seeing things that were 25-standard deviation moves, several days in a row.* [1]

As shown in [2] the probability of a single $25-\sigma$ event is simply absurd and one has to rely on cosmic quantities in order to 'make sense' of the corresponding probabilities.

However, the realisation that risk management models can fail on such a cosmic scale is by no means a thought that has emerged after the most recent financial crisis. In an account of the events that brought down 'Long-Term Capital Management' (a nobel-laureate staffed hedge fund) in the wake of the Russia crisis in 1998 it is literary said that

> *Theoretically, the odds against a loss such as August's [1998] had been prohibitive; an event so freakish as to be unlikely to occur even once over the entire life of the Universe and even over numerous repitions of the Universe.* [3] [4]

*jakob.krause@wiwi.uni-halle.de

[1] Financial Times, (2007). *Goldman pays the price of being big* https://www.ft.com/content/d2121cb6-49cb-11dc-9ffe-0000779fd2ac

[2] Kevin Dowd, John Cotter, Chris Humphrey and Margaret Woods *How Unlucky is 25-Sigma?* available on arXiv: https://arxiv.org/ftp/arxiv/papers/1103/1103.5672.pdf

[3] Roger Lowenstein, (2001). *When Genius Failed*, Fourth Estate. (p. 159)

[4] Kolman, (1999). *LTCM speaks*, Derivatives Strategy, April 1999.)

Given those considerations one can conclude that contemporary risk management models were not particularly helpful in the situations for which they were designed and more importantly suggested an unjustified level of security. It remains to be seen whether the modest adaptions regulators have proposed in the mean time are sufficient. Until then some healthy skepticism is advised and given the severity of the error contemporary risk management models produced it is reasonable to target fundamental properties of those models.

To this end, we question one of the fundamental assumptions of contemporary stochastics, identically distributed random variables. Most of modern statistical methods are based on the assumption of having access to identically distributed and independent (i.i.d.) random variables/observations. Those two properties are appropriate idealisations when describing physical systems. Economists have realised that it is not appropriate to carry over the assumption of independence when describing economic systems and therefore dependence concepts like correlation and copulas are omnipresent in the description of economic systems. Large parts of classical probability theory can be recast in a world with (suitable) dependence including law of large number like statements which are called 'ergodic theorems'.

The concept of ergodicity targets the question whether the history of a process is representative for the current state of a process (in an expectation sense) and lies at the heart of the other part of the i.i.d. assumption, identical distribution. The severity of this assumption can not be overstated and is fundamentally at odds with how our society is organised (see [5], [6], and [7]). In effect, it is equivalent to saying that data from the past will always be representative for the current situation. In Physics this approach is unproblematic since the underlying laws governing the system do not change over time. When turning to economic problems, however, we cannot assume that this is the case since we react to our economic environment and change our environment at the same time (see [8]). Assumptions should correspond to the system they describe. More specifically, the severity of assumptions should correspond to the robustness of the system they describe.

The realisation that Ergodicity is both, the cornerstone of the description of dynamics in systems with dependent increments (via Ergodic Theory) which is currently state of the art in risk management and absolutely inappropriate should give us a hint of how limited our current understanding of Economics is.

The objective of this article is to employ modeling methods that are able to capture some features of the instability one should encounter when studying economic problems. Data from the past is not truly representative for the current situation. However, data becoming non-representative is a gradual process. A picture of your favourite place on earth that has been made a minute ago will be almost representative for the current state of this place (concerning time of day, wheather, objects, people) whereas a picture made a month ago will be representative for only parts of the places properties. This gradual deterioration of representativeness will be the structure that is exploited in the remainder of the paper.

When acknowledging that the underlying rules of a system can change we have to specify what questions are meaningful in this context. In this paper we target the question of estimating the current state of an unstable system in some optimal way. This question has been investigated in the field of 'locally stationary processes' under the name of 'optimal segment length' (see [9]). The methodology in the aforementioned paper, however, is based on Fourier techniques and Cramer Representations of time series and therefore deals, naturally, with the autocorrelation profile of time series whereas in this paper a highly different methodology is used.

Another interesting question is how long we can consider findings in the social sciences and economics to be 'true'. This question is tightly connected to the 'reproducibility crisis' and raises the question whether reproducibility is actually an appropriate scientific standard

[5] Mark Kirstein, (2015). *From the Ergodic Hypothesis in Physics to the Ergodic Axiom in Economics*, Prepared for the 7. Wintertagung des ICAE Linz

[6] Ole Peters, Alexander Adamou, (2017). *Ergodicity Economics*, lecture notes, available at: `https://ergodicityeconomics.files.wordpress.com/2017/03/ergodicity_economics_20170824.pdf`

[7] George Soros, (2013). *Fallibility, Reflexivity, and the human uncertainty principle* Journal of Economic Methodology , 2013 Vol. 20, No. 4, 309-329 `http://dx.doi.org/10.1080/1350178X.2013.859415`

[8] John R. Doyle, Catherine Huirong Chen *The wandering weekday effect in major stock markets* Journal of Banking & Finance, Volume 33, Issue 8, August 2009, Pages 1388-1399

[9] Rainer Dahlhaus, Liudas Giraitis, (1998). *On the optimal segment length for parameter estimates for locally stationary time series*, Journal of Time Series Analysis, Vol. 19, No. 6, Blackwell Publishers Ltd.

for 'reflexive systems' like economic systems (see [10]).

The remainder of the paper is structured as follows: The objective of section 2 is to introduce an appropriate class of semimartingales that is compatible with the intuition layed out above. In addition, we target the question of how stationary estimators behave when applied to non-stationary observations. In this context the 'non-representativeness' error is derived. Under appropriate assumptions this result will yield that one should use as few data as possible when interested in estimating the current state of the system.

Section 3 has the objective of analysing the problem under the assumption of a finite data density. Under this assumption there is an incentive to use as much data as possible in order to achieve sufficient estimator convergence. Hence, considering the results of section 2 and 3 it is natural to ask what is the optimal tradeoff between the two forces estimator convergence and non-representativeness in order to find a minimally biased estimator. This is the objective of section 4.

Subsequently, we turn to applications and further weaken our assumptions. Until now we assumed that we know the dynamic of the change in the distributional properties. Now proxies are employed in order to gauge how stable our system is. We use those quantities as a proxy for the dynamic and introduce them in the form of 'information metrics', a concept to ensure that new information overwrites old information. One example would be to use a severe change in trading volume as a proxy for a change in volatility.

In section 6 we apply our framework to analyse a statute in the Basel III accord in order to assess the impact that a straightforward application of our thinking would have on risk management models. We also analyse some trading strategies.

In Section 7 we conclude.

2 Estimator Dynamics and Non-representativeness

Starting from the premise that the past is only partially representative for the present the objective of this section is to quantify the behavior of (classical) statistical estimators when applied to non-stationary data. In the context of this section we assume perfect knowledge, i.e. we know the current state of the system and would like to quantify the error we make when applying classical statistical estimators. Additionally, we know how the system changes over time, i.e. we know to which degree past observations are distorted. This implies that we know the representativeness of past observations. Under suitable assumptions on how the system changes over time the representativeness for the current situation will be declining the further we go into the past. Hence, when applying statistical techniques it will be useful to choose a connected set of observations to estimate the current situation.

In the first subsection we will introduce a setup in which we can set up a version of the representativeness problem. In subsection 2 we sidetrack a little and elaborate on the connection of our setup to the realm of 'model risk'. Subsequently, we turn to the description of the non-representativeness error.

2.1 Stochastic Formulation

In order for 'representativeness' to have any meaning we need a class of stochastic processes that is able to exhibit non-stationary behavior. For this purpose a convenient class of semimartingales is used. In addition, we also need a structure to track the non-representativeness of the past, e.g. through learning. Concerning their stochastic properties returns might look the same in 2005 and 2015. However, the financial crisis in between certainly had an influence on how we behave and therefore the 2005 period is not representative for the situation in 2015. However, subjective reaction due to new experiences is only one way for representativeness breaking down especially when considering technological progress as a transformative force.

2.1.1 Mathematical Structures

In order to proceed we need a flexible class of stochastic processes as well as a structure carrying the representativeness.

[10] Jakob Krause, (2017). *Normative Sciences and the Reproducibility Crisis* (September 9, 2017). Available at SSRN: `https://ssrn.com/abstract=3034660`

For the former we use a convenient class of semimartingales. For the latter filtrations are utilised and endowed with an additional metric structure capturing their growth. The two interact by assuming a connection between the growth rate of the filtration and the rate of change of the characteristics of the underlying semimartingale.

To this end, let $(\Omega, \mathbb{F}, \mathcal{F}, \mathbb{P})$ be a filtered complete prob. space fulfilling the usual hypothesis. Let X_t be a semimartingale with characteristics of the form $(0, \sigma(t), 0)$, where $\sigma(t)$ is assumed to be a continuously differentiable deterministic function. This choice corresponds to using a 'time-inhomogeneous brownian motion'. Assuming $\sigma(t)$ to be determinstic yields that the increments of $(X_t)_t$ are independent. For more details concerning this class of processes, see [11], Ch.2.

Trivially, this class of processes is, in general, non-stationary and the underlying state of the stochastic system is given by $\sigma(t)$. If $\sigma(t)$ would be constant the state of the stochastic system would not change and therefore the past would be fully representative for the present. Hence, in order to formalise the concept of representativeness we now have to establish a connection between the dynamic of $\frac{\partial \sigma(t)}{\partial t}$ and the structure carrying the representativeness. To this end we introduce the concept of an 'information metric':

Definition A map

$$I : \mathbb{R} \times \mathbb{R} \to [0, \infty), \quad (t_1, t_2) \mapsto I(t_1, t_2) \tag{1}$$

is called an **information metric** when it fulfills

i) $I(t, t) = 0$ for all $t \in \mathbb{R}$.

ii) $I(t_1, t_2) = I(t_2, t_1)$ for all $t_1, t_2 \in \mathbb{R}$.

iii) $I(t_1, t_3) \leq I(t_1, t_2) + I(t_2, t_3)$ for all $t_1, t_2, t_3 \in \mathbb{R}$.

iv) $\frac{\partial I(x,y)}{\partial y} > 0$ for a fixed x and $x < y$.

The intuition behind I is to capture differences in the state of the world between two points in time. In this context property iv) keeps track of our learning progress of *the underlying dynamics of the system*. This concept quantifies the growth of the information and therefore can be seen as the formalisation of the increase in information and is thus tightly connected to the underlying filtration. An more formal alternative would be to use the bracket purpose for the purpose of carrying the information metric [12]

For the specific choice of stochastic process we now tie the two concepts together by assuming that the rate of change of $\sigma(t)$ locally defines the information metric. More specifically:

$$I(t_1, t_2) = \int_{t_1}^{t_2} |\frac{\partial \sigma(s)}{\partial s}| ds \ . \tag{2}$$

For a physical system, the metric is zero since the underlying laws of nature do not change (that much). In a system involving thinking inhabitants it symbolises the learning effect and therefore we assume that the metric is always increasing (new information overwrites old information). This happens, however, not always with the same speed. Formally, we are interested to determine the optimal amount of data in situation where the velocity $\frac{\partial}{\partial t_2} I(t_1, t_2)$ is high but not infinite. Note, that an infinite velocity in $I(\cdot, \cdot)$ would correspond to a strict regime switch. To sum up the quintessential point: The information metric does not only measure the inflow of new information but also represents a measure of how much the underlying rules of the system changed since in social system new knowledge is always used to change the system since every social science can be assumed to have a normative inclination! This backlash is the most crucial difference between the natural sciences and the social sciences and the information metric allows us to incorporate the robustness of the underlying rules in the axiomatisation of the stochastic system. Obviously our modeling allows for a much more freewheeling dynamics. Hence it is implausible to assume that we can answer similar questions than with the more restrictive assumptions. Another implication is that statistical

[11] Jean Jacod, Albert N. Shiryaev, (2003). *Limit Theorems for Stochastic Processes* Springer-Verlag Berlin Heidelberg

[12] Jacod, Shiryaev, Thm II.4.4., pg. 102

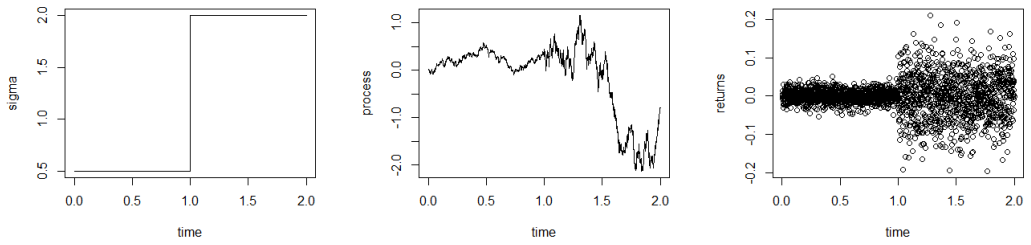

Figure 1: **time inhomogeneous Brownian Motion** left: time characteristics, middle: one sample path, right: corresponding returns. $\sigma_1 = 0.5$, $\sigma_2 = 2$, I = [0,2], script: 20171025_graphics_nonrepdata.r

truth is not a static concept. Performing similar experiments in a system with learning inhabitants can naturally yield different results which offers an angle on the replicability crisis in the social sciences and especially psychology [13].

2.1.2 Fitting stationary processes to non-stationary data on fixed intervals

The objective of this section is to find some stationary process that 'best' approximates some non-stationary process on a fixed interval under the assumption that we know the real dynamics of $\sigma(t)$.

To this end, let us fix the interval $\mathcal{I} = [0, 2)$ and let us assume that

$$\sigma(t) = \begin{cases} \sigma_1 & \text{for } t \in [0, 1) \\ \sigma_2 & \text{for } t \in [1, 2) \end{cases} . \tag{3}$$

This process is non-stationary and has a strict regime switch at $t = 1$ (see Figure 1). This process violates the assumption that the regime switches are smooth. However, by usual algebraic induction arguments (approximating smooth functions via step functions) this assumption will later be recovered. We now consider the question which stationary process approximates this process in some optimal way on the given interval. Stationarity means that time does not play a role and that within the interval every point in time is treated independently. Hence, we can ask what is the expected local characteristic of the non-stationary process at some random point $t \in [0, 2)$. We then can take the process with the (time-independent) local characteristic to be the approximation of the non-stationary process.

For the process above the result, naturally, is

$$\tilde{\sigma} = \frac{1}{2}\sigma_1 + \frac{1}{2}\sigma_2 . \tag{4}$$

More generally, we can rewrite this into

$$\tilde{\sigma} = \frac{\lambda(\{\sigma(t) = \sigma_1\} \cap \mathcal{I})}{\lambda(\mathcal{I})}\sigma_1 + \frac{\lambda(\{\sigma(t) = \sigma_2\} \cap \mathcal{I})}{\lambda(\mathcal{I})}\sigma_2 , \tag{5}$$

where $\lambda(\cdot)$ is a standard Lebesgue measure. We observe that this characteristics is equivalent to the one we would get when using two stationary brownian motions with characteristics σ_1 and σ_2 over the whole interval $[0, 2)$. Hence, the natural approximating process is given by

$$B_{app}(t) = \frac{1}{2}\left(B_1(t) + B_2(t)\right) . \tag{6}$$

This assertion can be checked via simulation (estimating and comparing the variance of the two processes (see figure below)). However, we can also argue formally on the basis of characteristic functions. In this argument we also reestablish the continuity property of $\sigma(t)$.

[13] Jakob Krause, (2017). *Normative Sciences and the Reproducibility Crisis (September 9, 2017)*. Available at SSRN: `https://ssrn.com/abstract=3034660`

2.2 Formal argument

Let us start with the situation above. Let $B_1(t)$ and $B_2(t)$ be two constricted Brownian Motions with respective volatility σ_1 and σ_2 defined as follows

$$\tilde{B}_i(t) = \begin{cases} 0 & \text{for } t<(i-1) \\ B_i(t) & \text{for } t \in [i-1, i) \\ B_i(i) & \text{for } t > i \end{cases} . \tag{7}$$

Without loss of generality we assume $\sigma_1 < \sigma_2$. Hence, the characteristics (as a function of time) is a step function. This corresponds to the first step in the usual algebraic induction argument one uses to approximate sufficiently smooth functions. Since we do consider non-stationarity to be an *interval dependent* property but treat observations within some interval on equal terms, since we would otherwise need additional assumptions on their stable connection *over time* which we want to strictly avoid. Hence, under suitable 'addition properties'(infinitely divisible distributions) we can use the *expected characteristics* of the non-stationary process for the stationary process.

As usual the approximation of continuous functions is done via step functions. Let us start with two independent brownian motions $B_1(t)$ and $B_2(t)$, with characteristics $(0, \sigma_1, 0), (0, \sigma_2, 0)$ over the whole interval [0,2]. The average of the two is given by

$$B_{app}(t) = \frac{1}{2}\left(B_1(t) + B_2(t)\right) .$$

Naturally, the characteristic function of B_{app} is given by

$$\varphi_{\frac{1}{2}B_1(t)}(x) \cdot \varphi_{\frac{1}{2}B_2(t)}(x) = \exp\{-\frac{\sigma_1^2 t \frac{1}{2} x^2}{2}\} \cdot \exp\{-\frac{\sigma_2^2 t \frac{1}{2} x^2}{2}\} \tag{8}$$

$$= \exp\{-\frac{\frac{\sigma_1^2+\sigma_2^2}{2} t x^2}{2}\} = \varphi_{B_{app}(t)}(x) \tag{9}$$

implying that B_{app} is a brownian motion with variance $\frac{1}{2}(\sigma_1^2 + \sigma_2^2)$ on the given interval. When $\sigma(t)$ is a function of time, the same logic applies and since we are only interested in the stochastic properties of the distribution convergence in distribution is sufficient. This is achieved by employing Levy's continuity theorem (see, e.g., [14])

Formally,

$$\lim_{n\to\infty} \varphi_{\sum_{i=1}^n p_i X_i} = \varphi_{\int_0^T x \cdot f(x) d\lambda(x)} \tag{10}$$

where the X_i are stochastic processes, i.e. random variables on the path space, over paths of the interval $[0, T]$. As mentioned earlier we assume that we do not have any additional information on the path, i.e. we need to assume $p_i = \frac{1}{n}$ for all $i \in \{1, ..., n\}$. In the limit this corresponds to using a uniform distribution, i.e. $f(x) = \frac{1}{T}$ in the interval $[0, T]$.

Note, that the random variable here is the underlying law of the distribution. Here we defined everything for the interval $[0, T]$. Since we will make the interval dynamic we now introduce a notation for more generic intervals. For a continuous function $\sigma(t) : I \to \mathbb{R}$ we define accordingly

$$\varphi^I_{\mathcal{N}(0,\sigma(t))}(x) := \exp\left\{-\frac{\frac{1}{\lambda(I)}\int_I \sigma(t)dt\ t\ x^2}{2}\right\} . \tag{11}$$

This choice realises the intuition above. In the next subsection we explore the connection to model uncertainty and show that the methodology introduced here is equivalent to hierarchical Bayesian modelling.

2.3 Model Risk or Time Dynamics?

Especially in Macroeconomics a lot of effort has been put into building models involving 'model uncertainty' [15]. This effort banks a lot on Bayesian statistics methodology and therefore

[14] D. Williams, 1991. *Probability with Martingales*, Cambridge University Press.

[15] Lars Peter Hansen, Thomas J. Sargent. (2015). *Uncertainty within Economic Models.* World Scientific Series in Economic Theory - Vol. 6

we want to briefly point out that Bayesian models and models of the type introduced in the last subsection are closely related. One natural implication is that instead of parameter uncertainty it could also be that the underlying parameters in macroeconomic systems exhibit a certain dynamics. A point that is natural for the economist that is not dogmatic concerning the stability of the underlying laws of the economic system (see, e.g., [16]).

To this end, consider the hierachical model given by

$$\sigma_1{}^2 \approx Uni[0,1], \quad \sigma_2 \approx Uni[0,1] \tag{12}$$

$$X_i|\sigma_i^2 \approx Norm(0,\sigma_i^2)\ . \tag{13}$$

Our objective is to predict the (unconditional) sample variance of X_i based on the distribution of σ_i.

Observations of this model are equivalent to a model of the last subsection in the following way: Consider a stochastic process on $[0,1]$ for which $\sigma_1(t)^2 = t$, $\sigma_2(t)^2 = t^2$. The hierarchical structure above can be replicated with the help of this process by drawing a random point in the interval $[0,1]$ according to a uniform distribution, say u. Since the time dynamics of σ_i have been determined, we know $\sigma_1 = \sigma_1(u) = u$ and $\sigma_2 = \sigma_2(u) = u^2$, respectively. Hence, the corresponding X_i is distributed according to $Norm(0,u)$ and $Norm(0,u^2)$, respectively. This is exactly the hierarchical structure from before.

In view of the results from the last subsection we now expect that the sample variance of X_i is given by the weighted variance

$$\int_0^1 \sigma_i^2 ds\ ,$$

i.e. $\mathbb{V}(X_1) = \int_0^1 s^2 ds = \frac{1}{3}$ and $\mathbb{V}(X_2) = \int_0^1 s\ ds = \frac{1}{2}$.

This relationship can easily be confirmed via simulation.

The crucial point here is easily missed: Here we showed that observations from a hierarchical bayesian model (at one point in time) are indistinguishable from observations of a process whose underlying rules change over time. Hierarchical Bayesian Models are used in Macroeconomics in order to incorporate model risk. However, in the interpretation of this section it could perfectly well be that it is a time dynamics of the model that causes the same empirical results.

2.4 The non-representative error

The results from section 2.2 allow us to quantify the impact of local stationarity (in the above sense) on the variance for gaussian processes in terms of an average over a fixed time interval, i.e. it allows us to quantify the mistake we make when using data from the past by comparing the average over the interval and the endpoint of the interval. For the remainder of the paper we always position ourselfs at the end of the interval because we want to look backwards in order to figure out how much data we should use from the past. This subsection has the objective of introducing the *interval dependent structure* carrying the notion of representativeness which will play a role in our optimization.

Let $\sigma(b)$ be the true variance of the process $\{X_t\}_{t\in\mathcal{T}}$ in $t = b$, assuming that $b \in \mathcal{T}$. In addition, assume that the distribution parameter of $\{X_t\}_{t\in\mathcal{T}}$ is given by the (differentiable) function $\sigma(t)$, i.e. $X_t \approx N(0,\sigma(t))$. Then, following the arguments in section 2.1 and 2.2 the average variance of the process on the segment $[a,b]$ is

$$\sigma^2_{[a,b]} = \frac{1}{b-a}\int_a^b \sigma^2(u)du\ .$$

This value can also be interpreted as the sample variance of the observations in $[a,b]$ assuming an infinite data density. With in infinite data density we minimize this error by using the data in an infinitesimal interval $[b-\varepsilon, b]$, $\varepsilon > 0$.

[16] George Soros, (2013). *Fallibility, Reflexivity, and the human uncertainty principle* Journal of Economic Methodology , 2013 Vol. 20, No. 4, 309-329 http://dx.doi.org/10.1080/1350178X.2013.859415

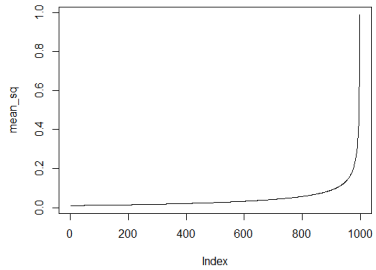

Figure 2: **Figure description**: We are positioned at Index 1000 and depict the average error out of 1000 simulations that is made when using only the observations in $[index, 1000]$. Naturally, the average error is large for a small number of observations and the smaller the more data is used.

Accordingly, if we are interested in the current variance σ_b^2, the error that stems from taking the observations in the interval $[a, b]$, the **non-representativeness error** $\beta([a, b], [b])$, is given by

$$\beta([a, b], [b]) = \sigma_b^2 - \frac{1}{b-a} \int_a^b \sigma^2(u) du \ .$$

In a stationary system this error is trivially zero.

3 Convergence Error

In the last section we derived the non-representativeness error under the assumption of an infinite data density. This notion alone incentivizes to use as small an interval as possible. Applied to frequency data this means to use as few data as possible. However, in classical statistics we usually also want as much data as possible in order to achieve sufficient estimator convergence. In this section we look at the estimator convergence in our non-stationary system, i.e. our objective is to quantifying the impact of a finite data density.

Here we encounter the problem that the convergence rate of the variance estimator for a normally distributed random variable is dependent on the variance itself $\sqrt{\frac{\sigma}{n}}$. The natural emulation of this concept in the situation we are in is to use $\sqrt{\frac{s^2_{[a,b]}}{\#[a,b]}}$, where $\#[a, b]$ denotes the number of data points in the interval $[a, b]$ and $s^2_{[a,b]}$ is the sample variance of the observations in the interval $[a, b]$. For few observations in the interval this will yield, on average, a big error and for a high number of observations this would yield, on average, a low error, a premise that can be tested by simulation (see Figure 2 below, [a,b] = [0,1], axis = [0,1000], average from a number of 1000 observations).

Let

$$\alpha_{[a,b]} = |\sigma^2_{[a,b]} - s^2_{[a,b]}| = \sqrt{\frac{s^2_{[a,b]}}{\gamma \cdot (b-a)}}$$

be the average **estimator convergence error**, i.e. the difference between the true value of the variance $\sigma^2_{[a,b]}$ of observations in an interval $[a, b]$ and the sample variance. Here, for convenience, we introduced the parameter γ denoting the data density. This simplification is only possible when assuming that the time between observations is constant. For a particular data set the quantity $\alpha_{[a,b]}$ is not known, i.e. we can only rely on the expected difference and have to tie the expectation to the local stationarity.

4 Minimal bias estimates

Under assumption of knowledge of the nature of the characteristics function, $\sigma(t)$, we can determine the optimal amount of data by minimizing the sum of the two errors described

above in the following way. We now consider ourselfs at the endpoint of the interval $[0,T]$ and we are looking for the optimal $t^*, 0 \leq t^* \leq T$ that minimizes the sum of the two errors, i.e. t^* is determined by the condition

$$t^* = \min_{t\in[0,T]} |\alpha_{[t,T]} + \beta_{[t,T]}| = \min_{t\in[0,T]} |\sqrt{\frac{\sigma^2_{[t,T]}}{\gamma \cdot (T-t)}} + \left(\sigma^2_T - \sigma^2_{[t,T]}\right)| \, .$$

Based on the arguments above we use

$$\sigma^2_{[t,T]} = \frac{1}{T-t}\int_t^T \sigma(u)du \, .$$

For t close to T the estimator convergence error will be very high. The more data is used, the more the estimator will have converged towards the moving target $\sigma_{[t,T]}$, i.e. for this target the estimator convergence error will decline with t getting smaller. However, in the same spirit the non-representativeness error will increase with t getting smaller. Trivially, the problem can be rewritten in terms of one unknown variable by setting $x = T - t$. Then the problem reads as

$$x^* = \min_{x\in[0,T]} |\tilde{\alpha}(x) + \tilde{\beta}(x)| = \min_{x\in[0,T]} |\sqrt{\frac{\tilde{\sigma}^2(x)}{\gamma \cdot x}} + \left(\sigma^2_T - \tilde{\sigma}^2(x)\right)| \, ,$$

where $\tilde{\alpha}(x) = \alpha_{[t,T]}$, $\tilde{\beta}(x) = \beta_{[t,T]}$ and $\tilde{\sigma}^2(x) = \sigma^2_{[t,T]}$.

Example

Let us assume $T = 5$, $\sigma^2_t = t, t \in [0,5]$. It follows that

$$\tilde{\sigma}^2(x) = \frac{1}{T-x}\int_x^T \sigma_s ds = \frac{x+5}{2} \, .$$

Correspondingly the optimization problem above reads as

$$\min_{x\in[0,5]} |\sqrt{\frac{\frac{x+5}{2}}{\gamma \cdot (5-x)}} + \left(5^2 - \frac{x+5}{2}\right)| \, ,$$

here γ is the number of observations in a one time period. It is not possible to solve this problem analytically, but numerical solutions are easily obtained. Below we depict the error function for a variety of data densities(black: 1000 observations per time period, blue: 100, green: 10, red: 1). Correspondingly, the optimal amount of datapoints that one should use in those situations is given by

171 ($\gamma = 1000$) from the time interval $[4.82, 5]$ yielding an error of 0.25550 in the estimator,
37 ($\gamma = 100$) from the time interval $[4.63, 5]$ yielding an error of 0.5457,
8 ($\gamma = 10$) from the interval $[4.2, 5]$ yielding an error of 1.1581, and
2 (sic!) ($\gamma = 1$) from the interval $[3, 5]$ yielding an error of 2.4090.
In the Figure 3 below the results above are depicted.

5 Discrete Information Metrics and Implementation

The objective of the last section was to show that there is an optimal amount of data in a system whose distributional properties change over time. In order to uncover this result we worked under the assumption of total information, i.e. the knowledge of the nature of the change in the distributional parameter and derived the maximization problem under this assumption together with the necessary simulations that back up the results. In reality the true signal is not observable, hence the change in the signal is unknown. In order to apply our thinking to a real world situation we need a reasonable proxy for the rate of change of the parameter of interest. In different fields of application this proxy can have different forms. In general the proxy should carry information about a change in the underlying mechanic. In Finance, one of the indicators that can constitute a 'change of rules' is liquidity. Market liquidity is the first victim of every crisis [17]. Hence, it is reasonable to use liquidity metrics

[17] Markus K. Brunnermeier, (2008). *Deciphering the 2007-08 Liquidity and Credit Crunch* The Journal of Economic Perspectives, Vol. 23, No. 1 (Winter, 2009), pp. 77-100

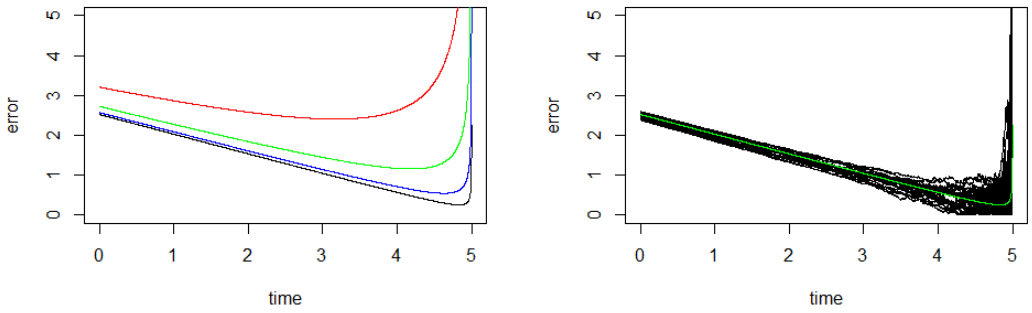

Figure 3: **Theoretical and Empirical non-representative errors for finite data densities:** left: Depicted are the error functions for different data densities of the theoretical error function given in the example. right: In green the theoretical error function is depicted. In addition, 50 empirical sample paths have been simulated and the corresponding empirical non-representative errors are depicted in the graphic. (script: 20170531_linearvariance_minimalerror.r)

as a proxy for the change of market conditions. This approach implements the intuition that once liquidity significantly changes, one cannot expect things to behave like usual, i.e. once liquidity changes data from the past is not representative to the current situation.

E.g. it is known that there is tight connection between trading volume, one liquidity metric, and volatility [18]. Hence, it is reasonable to assume that when the trading volume changes, the volatility changes as well and it is of interest to estimate what is the current level of volatility.

In addition, it has been shown that high frequency volatility (realized variance) can be used to improve lower frequency volatility forecasts [19]. These sources suggest that it is reasonable to use lower frequency volatility as well as trading volume as an *information metric* for higher frequency volatility. By an information metric we here mean a proxy for the (absolute value of the) dynamics of the parameter in question. Formally, this concept has already been introduced in Chapter 2. Here, however, we have to account for the frequency nature of data as well. The definition of the information metric and its usage for data analysis are the main contributions of this paper and offer a tool that offers a clear methodological discinction between statistics for Natural Science experiment like Physics where the underlying laws of nature are stable and the analysis of data in the context of Economics.

5.1 Definition

Let $X_1, ..., X_T$ be a series of random variables. In this context, the index is interpreted as time. By an **information metric** we mean the distance in distribution between the two (subsequent) random variables $d_I(X_t, X_{t+1}) = d_D(\mathcal{L}(X_t), \mathcal{L}(X_{t+1}))$, where $d_D(\mathcal{L}(X_t), \mathcal{L}(X_{t+1}))$ is a distance in distribution $d_D(\cdot)$ between the distributions of X_t, $\mathcal{L}(X_t)$, and X_{t+1}, $\mathcal{L}(X_{t+1})$.

More generally, we define the distance between two arbitrary observations X_t and X_s, $1 \leq s, t \leq T$ as

$$d_I(X_t, X_s) := \sum_{i=t}^{s-1} d_I(X_i, X_{i+1}) \, .$$

Note, that even though X_t and X_s could have the same distribution, the distance between them in the information metric could be positive. This intuition is drawn from the usual

[18] Torben G. Andersen. (1996). *Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility.* The Journal of Finance, Vol. 51, No. 1, March 1993

[19] Andersen, T.; Bollerslev, T.; Lange, S. (1999). *Forecasting financial market volatility: Sample frequency vis-a-vis forecast horizon.* Journal of Empirical Finance 6 457-477.

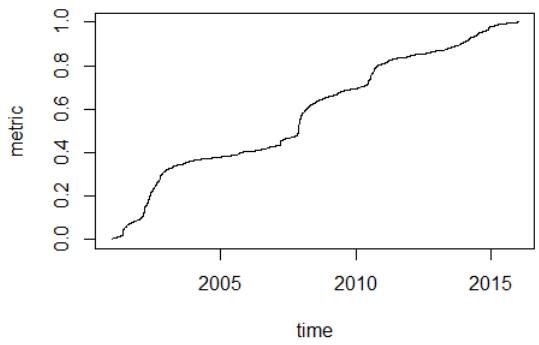

Figure 4: **Information Metric**: Depicted are the (normalized) cumulative squared returns of the German Stock Index (DAX) in the period from 2001 to 2016. It is clearly visible that the propagation is not homogeneous. (script: Dax_trading.r)

assumption of filtrations of stochastic processes. In this definition it is conveyed that information is always increasing and the intuition we want to implement is that even though two random variables might come from a 'calm' period, there is a difference between a calm period in 2005 and 2015 due to a difference in understanding of some aspects of financial markets. Hence, we need to put some distance in the information metric between 2005 and 2015. The definition above achieves this goal in a natural way. In Figure 4 we show an example of an information metric (volatility) for the German stock index DAX in the period from 2001 to 2016 which will be used in Chapter 6.

5.2 Algorithmic determination of the optimal amount of data

Information Metrics are a measure for how fast the distribution changes. Hence, if the long-term velocity in this metric is different than the short term velocity, then the window of data that is used should be shrinked. Hence, in order to determine the optimal amount of data we have to compare the 'speed' in the information metric for different data windows. If there is an incompatibility between the long-term speed and the short-term speed we have to shrink our data window. However, the smaller the window gets the more difference we tolerate since the smaller the window the less estimator convergence.

In order to implement this thinking we need to start with values for the long-term window and the short-term window. We arbitrarily choose the long-term window to start at 1000 data points and the short-term window to start at 5.

By default we start with the longest possible window, compare the velocity and shrink the window whenever the difference between the two is intolerable. Intolerability is tied to the estimator convergence error as shown below. Then, for every data point t we proceed as follows:

1. $i = 5$, $D = 1000$
2. While
$$\frac{\frac{1}{i} d_I(X_{t-i}, X_t)}{\frac{1}{D} d_I(X_{t-D}, X_t)} \notin [1 - \varepsilon_D, 1 + \varepsilon_D]$$
and $D > i$ reduce D, the optimal amount of data.

Here, ε_D represents the tolerance for difference in velocity. i is fixed and with D decreasing, ε_D should increase. However, the functional dependence of ε_D depends on the convergence rate of the estimator. For the locally normal situation above one can use $\varepsilon_D = \frac{1}{\sqrt{D}}$ if one renormalises the information metric over the interval $[T - D, T]$.

6 Application - Banking Regulation, optimal rolling window size

The issue of how much data should be used in order to estimate a risk management model is answered very insatisfactory in the literature on risk management models. Different sources recommend different amounts of data, e.g. in [20] it is recommended to take 500-2000 data points, whereas in [21] it is recommended to take upwards of 2000 daily observations for the estimation of a 1% VaR. In addition, the recommendation of using some number in an interval, like *use between 500-2000* is not really helpful since in [22] it is shown that the size of the rolling window matters.

The only sensible question to answer is, based on the best guess of our current state, what can we say about tomorrow? And this is a situation not necessarily covered by, say, conditionally independent $GARCH$-models since this class of models also relies on a stable heritage regime between today and tomorrow. The heritage regime is dependent on the data that is used in order to estimate the $GARCH$ model. Hence it is appropriate to ask what is the data that one calibrates the $GARCH$-model with? Context matters for the decision of people and it is questionable whether the current level of volatility is *enough* context in order to reflect human behavior in a model.

Interestingly, the thinking that a crisis might bring about a structural break in the set of rules and the logical conclusion that one should therefore shorten the period of data that is used to estimate the model has been recognized by the Basel Committee on Banking Supervision: In their regulatory framework [23], the following statement can be found:

'The supervisory authority may also require a bank to calculate its Expected Shortfall using a shorter observation period if, in the supervisors judgement; this is justified by a significant upsurge in price volatility. In this case, however, the period should be no shorter than 6 months.'

This statement is interesting for a variety of reasons and raises the following questions:

1) Can we use the method developed in this paper to give the regulator a tool that yields the 'optimal amount of data' which should not be shorter than 6 months?
2) This statute is in the currently active regulations. Typically, a 'significant upsurge in price volatility' is associated with the beginning of a severe crisis. We are not aware of an impact study of this statute and want to give a crude assessment of what would happen when one shortens the representative data frame at the beginning of a crisis.
3) Is the 6 months lower bound a reasonable time frame ?
4) Here, the 'information metric' is given by 'price volatility' which is not a well-defined term. This could mean implied volatility, realized volatility, daily volatility, among others. In addition, the 'prices' it refers to are not clear as well.
5) In general, this statute has all the ingredients that are utilised in this paper.

A straightforward implementation of the algorithm outlined in 5.3. on a data set of daily returns of the DAX using $D = 1000$, $i = 125$ yields the representative datapoints depicted in Figure 5 below. It is easily identified that the data window collapses at the outbreak of the 2007 financial crisis, the 2011 debt crisis and the devaluation of the Chinese currency in 2015, i.e. our algorithm is helpful in at least a crisis indicator sense. In addition, we can give it to the regulator so that he has a tool to use if he is interested in applying the statute. However, we want to give a crude estimate of what applying this statute would do to some very basic notion of risk measures. Based on some crude Expected Shortfall estimations (equally weighted time frame, normal linear) we arrive at the results depicted in Figure 6.

[20] Christoffersen (2012), *Elements of Financial Risk Management*, Academic Press, Elsevier

[21] Alexander (2008), *Market Risk Analysis, Vol. IV*, Wiley

[22] Barbara Rossi and Atsushi Inoue (2012), *Out-of-Sample Forecast Tests Robust to the Window Size Choice*, working paper

[23] *Minimum capital requirements for market risk*, Bank for International Settlements, Jan 2016, section 3, par. 181 (e)

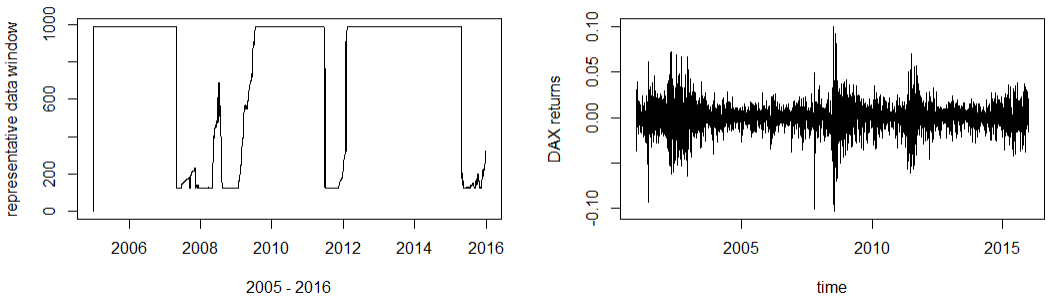

Figure 5: **Representative window size**: left: representative window size over the time frame from 2005 to 2016. At the end of the 'great moderation' period the window size is constantly high. This means that the short-term fluctuations in that period are not severe enough to trigger a shortening of the representative window. The representative window collapses in 2007 (Financial Crisis), 2011 (Debt Crisis), and 2015 (Chinese currency devaluation). right: Corresponding returns of the German stock index DAX.

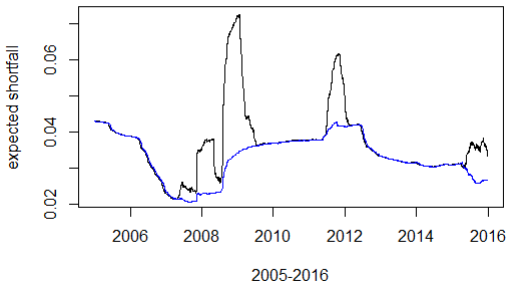

Figure 6: **Expected Shortfall:** In this figure we calculate the expected shortfall at the 97.5% level as follows: We estimate the Vakue at risk based on the variance that is estimated based on a fixed rolling window (blue) and based on the representative window (black). Assuming normally distributed returns we then multiply the Value at risk by an appropriate and level dependent factor to arrive at the (one day) expected shortfall.

This result is not surprising since the relative amount of 'crisis data' in a shrinked data interval is higher and therefore the risk measures are trained on a data set where the world looks a lot riskier. We also have to note that this result is based on a variety of very crude assumptions. However, I want to argue that the scenario layed out above is one the more modest ones. Above we assumed that returns are normally distributed. However, for fat tailed distributions that are more appropriate to use in the context of risk management, the quantiles corresponding to the tails are, by definition, further away from each other. Therefore, the increase in capital regulations would be even steeper.

In effect, this result could indicate that the regulator has a tool in his hands that he can use to shut down the financial system since increasing the capital requirements by the factor of two to three for all banks would in itself start a panic cycle that would lead to a tightening of financial conditions which would in turn yield even more necessary capital. The only times when *'there is a significant upsurge in [some definition of] volatility'* are the initial hits of a financial crisis. If the regulator would trigger this statute at the beginning of a crisis this would lead to a sharp increase of capital requirements in a situation where liquidity and confidence are already diminished. However, it has to be noted that the flipside of this argument is equally disturbing: If the only things that keep alive a bank (from a quantitative risk management perspective) are the smoothing effects of data prior to a crisis which arguably is not relevant to the current situation then this is not comforting. One could say that, at the beginning of a crisis, banks are only alive because the regulator lives in the past.

7 Conclusion

The objective of this paper was to highlight and to some extend remedy methodological flaws in how statistics is applied in Economics. The robustness of the underlying rules in Physics and Economics is highly different. Hence it is simply inappropriate to carry over the language that has been developed for applications in Physics to Economics and Finance. Typically, this argument is made for the formal tools used in the analysis of equilibrium models (neoclassical Economics). However, the same holds for the diagnostical, i.e. statistical tools, that have also been developed for physical systems and should be applied with caution to a situation for which they were not invented. This point is highlighted by the abysmal performance of risk management systems and economic forecasting in times of financial turmoil.

In this paper we approached the problem by introducing a tradeoff between representativeness and estimator convergence, where the former is something that is usually not looked at when economic systems are described. The main structure that is put into place to get hold of the notion of representativeness has been the 'information metric' which formaliZes a change of the underlying rules. Subsequently, the theory of locally stationary processes has been used and it has been shown that there is a significant overlap between locally stationary processes, a tool from the frequentistic world, and hierarchical bayesian models which can stimulate a discussion whether the notion of model risk is not just referring to the possibility that economic systems change over time.

Subsequently we turned to an application where at the very least the method introduced in this paper can be used as a crisis indicator and a tool for regulatory authorities. In addition, a crude estimation on risk metrics has been given. The results urge for follow ups.