Learning the Evolution of Statistical Moments and Extreme Values from Data

Ryan Shìjié Dù

February 6, 2023

1 Introduction

Assuming that the time-series data x follow random variables at each time, we are often interested in the evolution of the shape of the conditional probability distribution $p(x|t)$. Learning the evolution of statistical moments is a useful proxy. Extreme values of a heavy tail random variable often correspond to significant real-world events and their changes are also worth learning from data. In this note, we propose a simple modification from linear regression that could learn the evolution of statistical moments and extreme values from data. We will explore the benefits and shortcomings of this method, and compare this method with existing methods in the literature.

2 Linear regression for learning changes in arbitrary expectations

Suppose in a time period $t \in [0, T]$ there is a random variable $X_t \in \mathbb{R} \times [0, T]$ that evolves in time. It has the probability density function $p(x, t)$. The expected value of $f(X_t)$ conditioned on t is:

$$g(t) = \mathbb{E}_t[f(X_t)] = \int_{\mathbb{R}} f(x) p(x|t) \, \mathrm{d}x. \tag{1}$$

We have X_i: samples of X_t at time $t = t_i$. In this note, we will show that we can extract information about the time-change of $g(t)$ from applying least-squares fit to $f(X_i)$. Specifically, by choosing specific f functions, we could diagnose the evolution of statistical moments and extreme events. For example, $f(x) = x$ gives the mean $\overline{x}(t)$ and $f(x) = (x - \overline{x}(t))^2$ gives the variance $k_2(t)$ (second cumulant) as a function of time.

Consider the function $g(t) \in L^2([0, T])$ and a number of linearly independent basis functions $e_j(t) \in L^2([0, T])$. Linear regression finds $\sum_j \alpha_j e_j(t)$, an approximation of $g(t)$, by minimizing

$$\inf_{\alpha_j} \int_0^T \left[g(t) - \sum_j \alpha_j e_j(t) \right]^2 \mathrm{d}t. \tag{2}$$

Since the optimization is convex, it is equivalent to setting the derivative to zero:

$$\int_0^T \left[g(t) - \sum_j \alpha_j e_j(t) \right] e_i(t) \, \mathrm{d}t = 0 \qquad \text{for all } i. \tag{3}$$

We can interpret this formulation as the requirement that the residual of the fit is orthogonal to the span of the basis functions $e_j(t)$.

Now we combine (1) and (3) and have that if we were to apply linear regression to approximate probability expectation of $f(X_t)$ (i.e.: $\mathbb{E}[f(X_t)]$), we need to satisfy

$$\int_0^T \int_{\mathbb{R}} \left[f(x) - \sum_j \alpha_j e_j(t) \right] e_i(t) p(x,t) \, \mathrm{d}x\mathrm{d}t = 0 \qquad \text{for all } i \tag{4}$$

where we used $\displaystyle\int_{\mathbb{R}} p(x,t) \, \mathrm{d}x = 1$ for all t and $e_i(t)$'s are independent of x. This is equivalent to the minimization problem

$$\inf_{\alpha_j} \int_0^T \int_{\mathbb{R}} \left[f(x) - \sum_j \alpha_j e_j(t) \right]^2 p(x,t) \, \mathrm{d}x\mathrm{d}t. \tag{5}$$

Therefore to fit the time-change of $\mathbb{E}[f(X_t)] = g(t)$, we could apply the method of least-squares to the data $f(X_i)$.

To make this explicit, as long as the data is sampled uniformly in time[1], the double integral could be approximated as a sum. In principle, we use Monte-Carlo to approximate the probability integral and Riemann sum for the time integral. We have the minimization problem

$$\inf_{\boldsymbol{\alpha}} \left[f(X_i) - \boldsymbol{A}\boldsymbol{\alpha} \right]^T \left[f(X_i) - \boldsymbol{A}\boldsymbol{\alpha} \right] \tag{6}$$

where $f(X_i)$ is the vector of f evaluated at X_i, $\boldsymbol{A}$ is the Vandermonde matrix with the j^{th} columns as $e_j(t_i)$, and $\boldsymbol{\alpha}$ is the vector with entries α_j. Taking the gradient of the above minimization problem with respect to $\boldsymbol{\alpha}$, we get the discrete version of (4)

$$\boldsymbol{A}^T \boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{A}^T f(X_i). \tag{7}$$

This is equivalent to the over-determined linear algebra problem

$$\boldsymbol{A}\boldsymbol{\alpha} = f(X_i). \tag{8}$$

We are in the familiar territory of linear regression.

3 Some numerical tests for statistical moments and extreme values

We apply this method to synthetic data with known probability distributions for all time and learn the trend in moments and extreme values. But the method is very general, we could choose any function f and any set of basis functions e_j's.

[1] If the data is not uniformly sampled in time, we should consider appropriately weighting the sum (see Dù 2023).

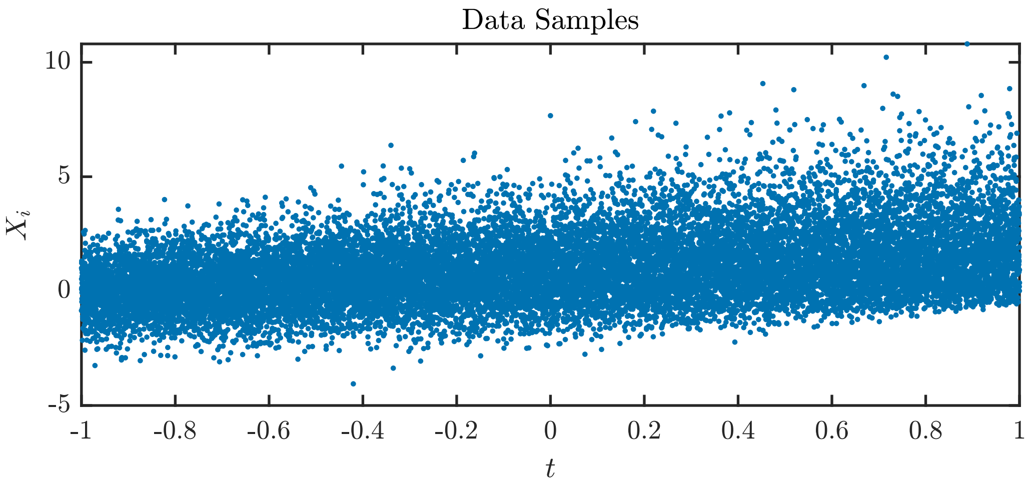

Figure 1: Synthetic data generated using the Pearson system. The four moments change linearly, they follow the black lines in Figure 2.

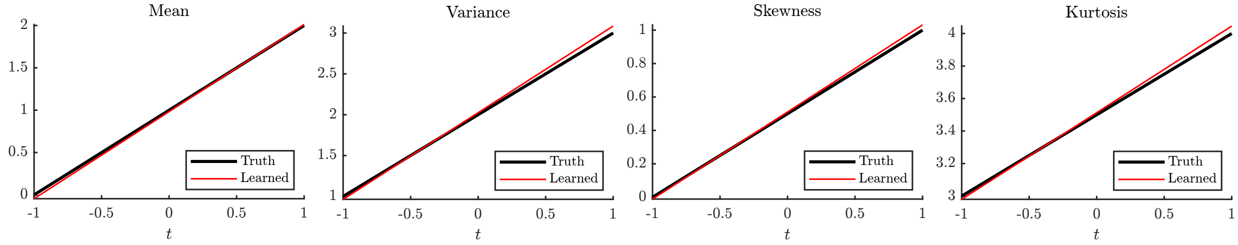

Figure 2: The true evolution (black lines) and learned evolution from data (red lines) for the four moments of interest.

3.1 Learn the linear trend of statistical moments

In $t \in [-1, 1]$, we use the Pearson system of random numbers to generate samples X_i of size 50×365 (50 years of daily data, see Figure 1). We prescribe the four statistical moments — mean, variance, skewness, and kurtosis — to change linearly (black lines in Figure 2 and 3). We then apply the described method to these data, using the basis functions $e_1(t) = 1$, $e_2(t) = t$. For the four moments, we use

$$f_1(x) = x, \qquad f_2(x) = (x - \overline{x})^2, \tag{9}$$

$$f_3(x) = \frac{(x - \overline{x})^3}{\kappa_2^{3/2}} \qquad f_4(x) = \frac{(x - \overline{x})^4}{\kappa_2^2} \tag{10}$$

where $\overline{x}$ is the fitted mean, and κ_2 is the fitted variance. Note that the higher moments depend on the fitted first two moments, therefore to have good fits for the higher moments, it is important to learn the mean and variance well. However, this is not a necessary feature of our method. Different choices of f could avoid this issue. For example in the next section, we will show an f that does not depend on any fitted quantity.

Figure 2 shows a set of results of our method. We see that the learned changes match the true evolution very well. Figure 3 shows a set where the results do not look as perfect. This is because the magnitudes of changes for the higher statistical moments are an order of magnitude less than

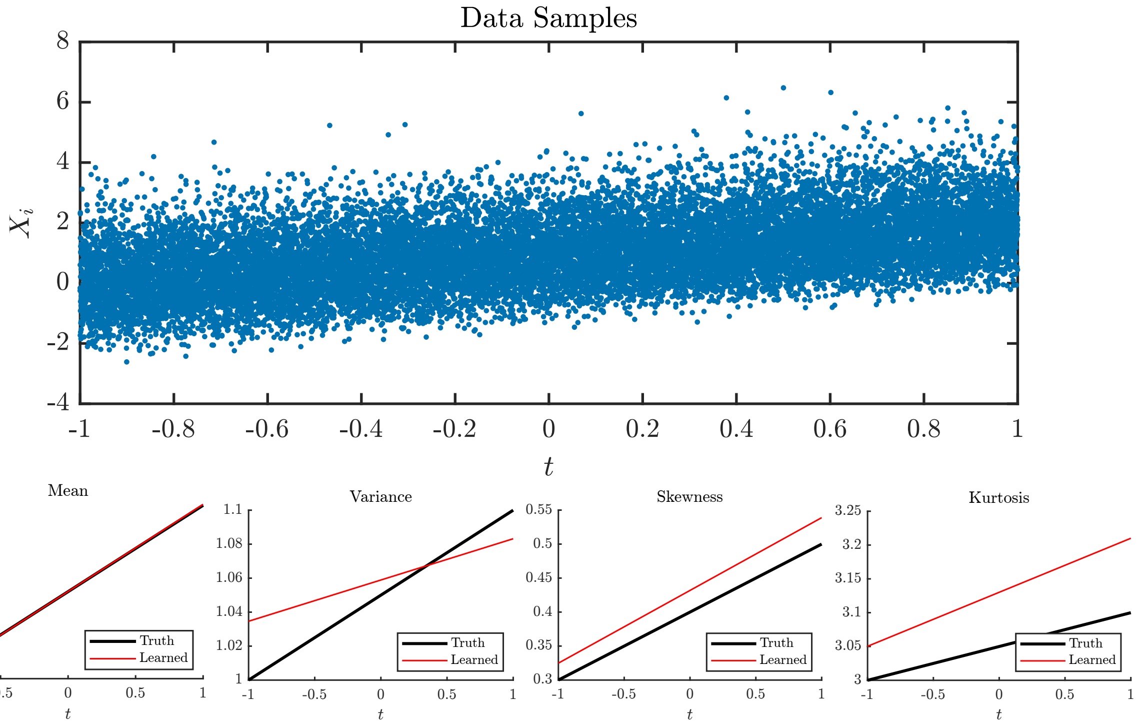

Figure 3: The top and bottom figures are similar to Figure 1 and 2 respectively.

the change in the mean. Therefore the error in the fit for the mean is relatively more influential for the fits for the higher statistical moments. From this example we see that it might not be a great practice to choose f that depends on other fitted quantities.

We comment that it is simple to apply some kind of significant testing (e.g.: the bootstrapping procedure described in Falasca et al. 2022) to check whether there are statistically significant changes in the moments (i.e. non-zero slope). Preliminary results (not shown) indicated that it works well.

3.2 Learn the time-change in extreme values

An interesting choice for $f(x)$ is a step function $f(x) = \chi_{x>a}$. We have

$$\int_{\mathbb{R}} \chi_{x>a} p(x|t) \, \mathrm{d}x = \mathbb{P}(X_t > a), \tag{11}$$

We are learning the time change in the probability of extreme values. For example, this could be the probability of dangerously high wet-bulb temperatures, or the probability of high sea levels in a city.

We use the same data from the last section as examples. Since we are using the Pearson system of distributions, we know the true $\mathbb{P}(X_t > a)$ theoretically. We pick the threshold of $a = 5$. Because the probability of extremes changes non-linearly in time, we use $e_1(t) = 1, e_2(t) = t$, and the exponential ($e_3(t) = e^t$) as basis functions. It is conceivable that we could obtain a better fit if we choose better basis functions, informed by the mathematical theory of extremes. We show the result using the data in Figure 1 on the left of Figure 4, and the result using the data in Figure

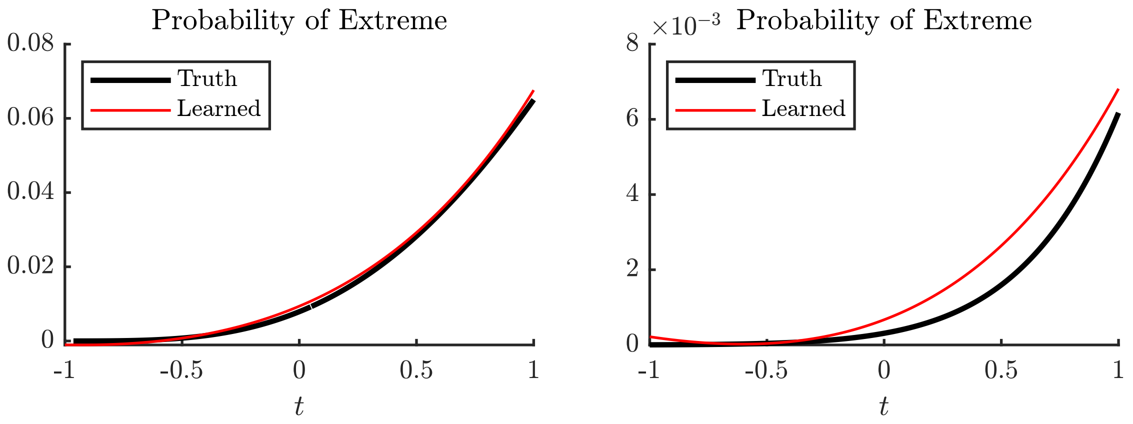

Figure 4: Left: the true and fitted (learned) time-change in the probability of extremes in the data in Figure 1. Right: the same for the data in Figure 3.

3 on the right. The method learns the change in extremely well. In particular, the result is not dependent on the fit of some other quantities.

We remark that $f(x)$ as a step function gives the probability of extreme values. Another choice of $f(x)$ can also account for the severity of the extremes as well. One example is the ReLU function.

4 Comments on the method, and comparison with existing methods

4.1 Benefits and deficiencies

The method presented in this note is a simple extension of linear regression. It is easy to understand and implement for a wide array of applications. This is its main strength. The method can be improved from its current form by incorporating improvements for nonparametric regression. For example, one could choose better basis functions for a more general class of functions (Tibshirani 2018), or incorporate regularization by using a Bayesian framework (Bishop 2006). In fact, Tibshirani 2018 mentioned our method for variance estimation in §8.

However, the method suffers from similar problems as regression. In particular, the convergence rate of the fit degrades with high-dimensional data. There are many ideas to improve the convergence behavior in the literature on high-dimensional regression that we can incorporate into our method. For example, when the parameters can be assumed to be “sparse”, we could use LASSO (Bishop 2006).

To use our method to diagnose the change in the probability of the extreme, we need enough samples in the extreme regime so that the Monte Carlo approximation of the integral is accurate. This might not be the case for the given data set when extreme events are very rare. For these situations, one could resort to sampling techniques that are geared towards sampling the probability tails (e.g.: important sampling).

4.2 Compare with methods in the literature

There are many methods to diagnose variance from data, and many are discipline specific. Here we will compare our method to a few that we have seen. The list is by no means extensive.

Our method is superior to methods that bin the data to estimate statistical quantities first. For example, Tank and Können 2003 listed many quantities of practical interests in climate science in form of days per year. Calculating statistical quantities in a chosen time period is equivalent to performing the Monte Carlo approximation to the probability integral in our formula per bin. Linear regression on the data calculated from each bin is an approximation of the time integral. Our calculation shows that we can combine the two steps and approximate the double integral together. This makes use of the available data efficiently. We can use our method as long as the quantity of interests can be written as a function $f(x)$ on the data.

McKinnon et al. 2016 proposed a method of calculating linear changes in quantities close to statistical moments via quantile regression. Their method was improved by Falasca et al. 2022. Their method shows an interesting connection between linear changes in qualities to linear changes in moments-like quantities (both papers explicitly warned against interpreting their results as actual changes in moments, but called those changes as changes in mean, variance, skewness, and kurtosis). However, if one wants to diagnose the changes in moments directly, our method is numerically cheaper and more versatile. Our method can learn the nonlinear evolution of many statistical quantities not just moments, and it can be extended to higher dimensions.

Tabak and Trigila 2018 proposed a general framework that could be applied to our problem of moments estimation. They solved the problem of conditional density estimation (and sampling) using optimal transport in §6. For time-series data, their method would map the data X_t to $Y = Y(X_t; t)$, which follows a single random variable (the weighted barycenter) $\mu(y)$ that is independent of t. The map would minimize some cost measuring the introduced distortion. We can estimate or sample the conditional probability $\rho(x|t)$ by estimating or sampling $\mu(y)$ and using the inverse maps. Estimating moments is then an easy extension. Although the adaptation of their framework to estimate higher moments is not trivial, we believe that it has promise to outperform our method in high-dimension, data-poor regimes. Additionally, its unifying approach connects many common methods in an intellectually satisfying way and extends them. Anyone who is interested in estimating the evolution of probability distribution over parameters should consider their methods. However, we think our method is useful in its simplicity. It works well as a quick tool for proof of concepts.

Acknowledgment

This note was inspired by a talk by Andrew Brettin. We thank him and Fabrizio Falasca for explaining the methods in McKinnon et al. 2016 and Falasca et al. 2022. Many conversations with experts have improved this note. We thank Jonathan Niles-Weed for situating our work in the field of nonparametric estimation and pointing us to Tibshirani 2018. We thank Esteban Tabak for introducing us to Tabak and Trigila 2018.

References

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning.* Springer, August. ISBN: 978-0-387-31073-2.

Dù, Ryan Shìjié. 2023. “Plotting and Fitting Power-Laws.” *Personal Notes.*

Falasca, Fabrizio, Andrew Brettin, Laure Zanna, Stephen M. Griffies, Jianjun Yin, and Ming Zhao. 2022. *Exploring the Non-Stationarity of Coastal Sea Level Probability Distributions,* arXiv:2211.04608, November. https://doi.org/10.48550/arXiv.2211.04608. arXiv: 2211.04608 [physics].

McKinnon, Karen A., Andrew Rhines, Martin P. Tingley, and Peter Huybers. 2016. “The Changing Shape of Northern Hemisphere Summer Temperature Distributions.” *Journal of Geophysical Research: Atmospheres* 121 (15): 8849–8868. ISSN: 2169-8996. https://doi.org/10.1002/2016JD025292.

Tabak, Esteban G., and Giulio Trigila. 2018. “Explanation of Variability and Removal of Confounding Factors from Data through Optimal Transport.” *Communications on Pure and Applied Mathematics* 71 (1): 163–199. ISSN: 1097-0312. https://doi.org/10.1002/cpa.21706.

Tank, A. M. G. Klein, and G. P. Können. 2003. “Trends in Indices of Daily Temperature and Precipitation Extremes in Europe, 1946–99.” *Journal of Climate* 16, no. 22 (November): 3665–3680. ISSN: 0894-8755, 1520-0442. https://doi.org/10.1175/1520-0442(2003)016⟨3665:TIIODT⟩2.0.CO;2.

Tibshirani, Ryan. 2018. “Nonparametric Regression and Classification.”