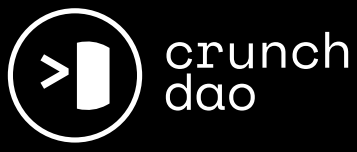


Machine Learning meets Statistical Physics: a Web3 perspective

CrunchDAO

Matteo Manzi, Utkarsh Pratiush, Enzo Caceres

2023/01/18

Machine Learning meets Statistical Physics: a Web3 perspective

abstract

CrunchDAO's Machine-Learning-enabled ensemble framework builds on top of traditional econometric risk models, requiring a number of steps in the data preparation: features orthogonalization, standardization, model order reduction, and data obfuscation will be discussed. It is discussed how, in the context of ensemble learning and bagging in particular, combining a variety of orthogonal models yields more accurate estimates of expectations. Moreover, the statistics of the set of predictions can be used to infer a measure of risk in the portfolio management process. We discuss how to integrate this in modern portfolio theory. We briefly discuss the necessary relation between these design choices and the ergodic hypothesis on financial data.

Econometrics

In a multi-factor framework of M factors (Sharpe 1964), (Fama, French, and French 1993), the total excess return over the risk-free rate of asset $i \in [1, ..., N]$ is given by

$$r_i = \sum_{j=1}^{M} \beta_{ij} F_j + \epsilon_i \tag{1}$$

with $M \ll N$. β_{ij} (the exposure of asset i to factor j) identifies entries of the **B** matrix, F_j (the return of factor j) of the **F** vector and ϵ_i of the **E** vector: this is part of the excess return not explained by common factors, called specific return (also idiosyncratic return).

Factor Neutral Portfolio

In this context, the expected return is, with a set of positions represented by the vector ω:

$$\mathbb{E}(r) = (\mathbf{B}\omega)^T \mathbb{E}(\mathbf{F}) + \omega^T \mathbb{E}(\mathbf{E}) \approx \omega^T \mathbb{E}(\mathbf{E}) \tag{2}$$

$$\sum_{i=1}^{N} \beta_{ij} \omega_j \approx 0 \quad i = 1, ..., M \tag{3}$$

Machine Learning can be used to obtain nonlinear models (Chan 2022), (Bonne, Wang, and Zhang 2021), (Prado 2019).

Features

For each cross section, a feature Matrix $\mathbf{X}_{N\times F}$ is our current representation of the state of the system.

In a supervised learning framework, multiple targets are associated with such feature matrix, which for us represent the compound return orthogonal unexplained by the factor model. As we can work with multiple targets independently here, we will focus on one: **Y**.

Orthogonalization

One reason for which predictions $\hat{Y}$ with good linear correlation:

$$corr(\hat{Y}, Y) \gg 0 \tag{4}$$

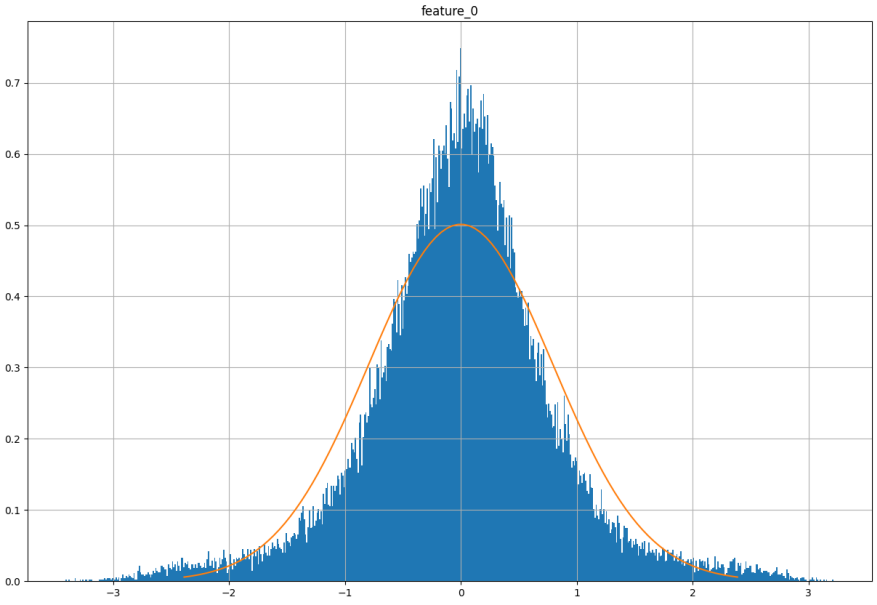
are not desirable is that, using Taylor expansion:

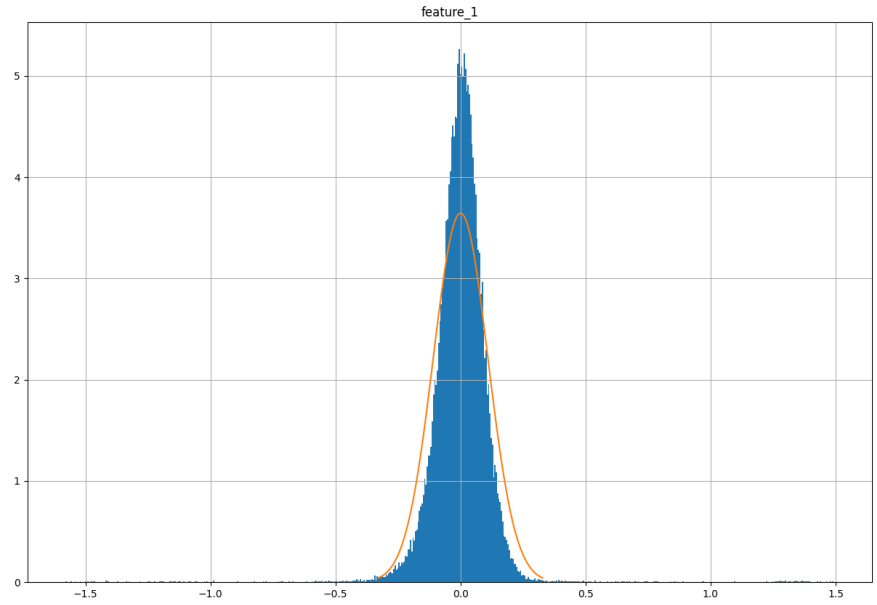
$$\hat{Y} = f(X) = AX + \sum_{n=2}^{\infty} A_n X^n = AX + g(X) \tag{5}$$

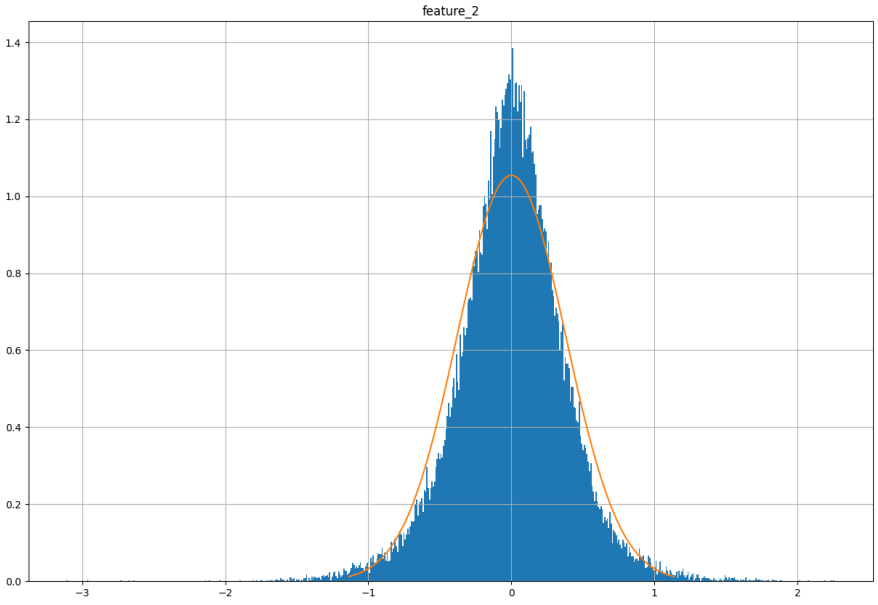
$$corr(X, B) \gg 0 \Rightarrow \left(f(X) \approx AX \Rightarrow (\mathbf{B}\omega)^T \mathbf{F} \gg 0 \right) \tag{6}$$

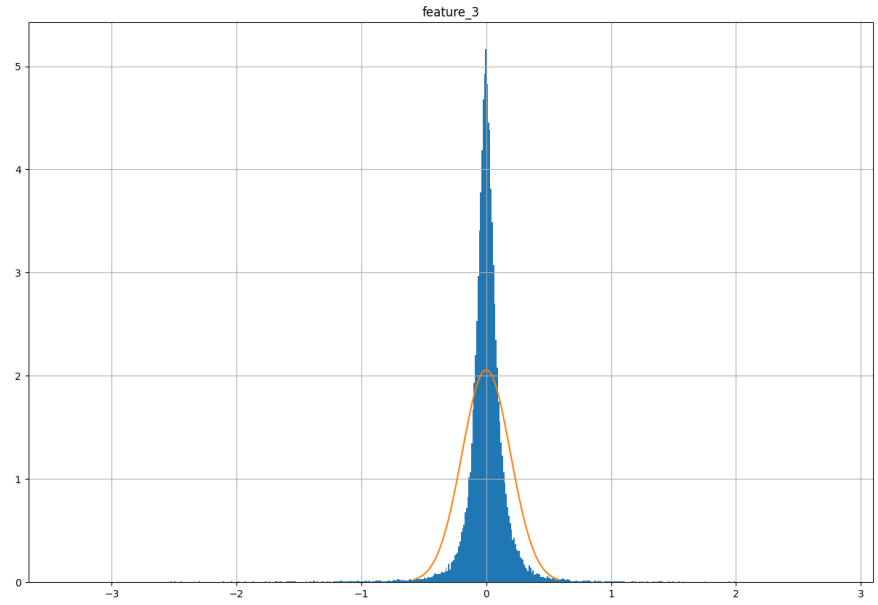
in which the last implication is due to the fact that, in a portfolio management framework, ω and $\hat{Y}$ are naturally highly correlated. We cannot use Gram-Schmidt (iteratively orthogonalize against all factors), as factors in general do not define an orthogonal basis. We project each feature in a least-square sense:

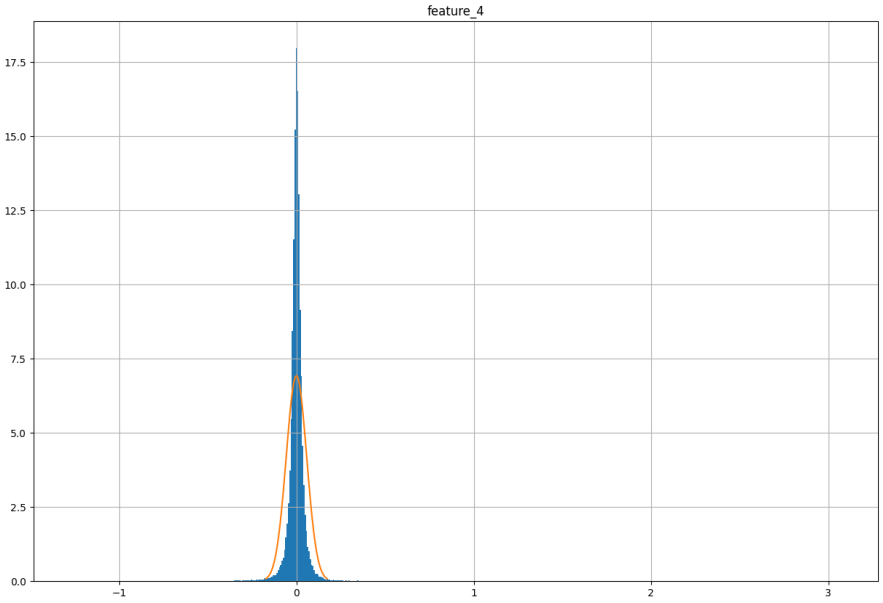
$$h(X) = X - B\left(B^T B\right)^{-1} B^T X \tag{7}$$

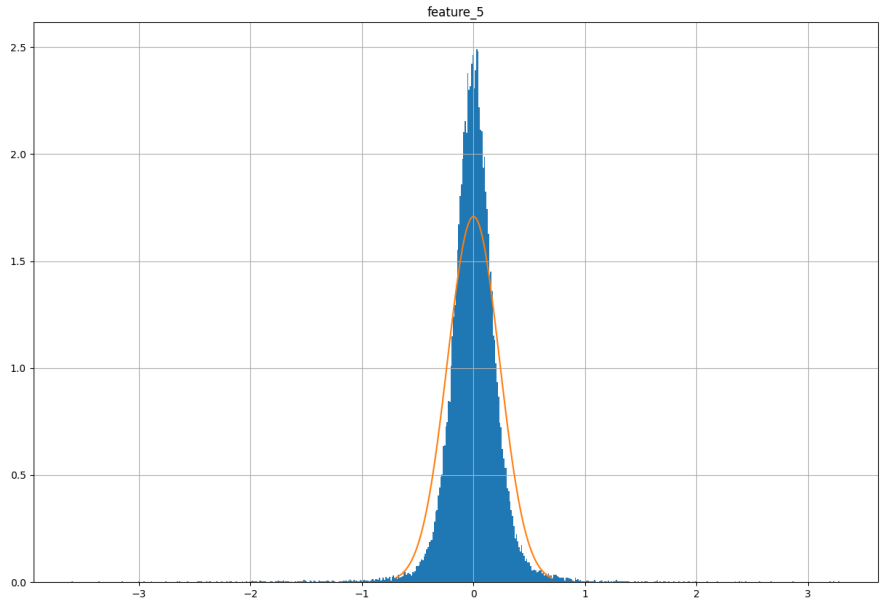

feature_0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−3
−2
−1
0
1
2
3

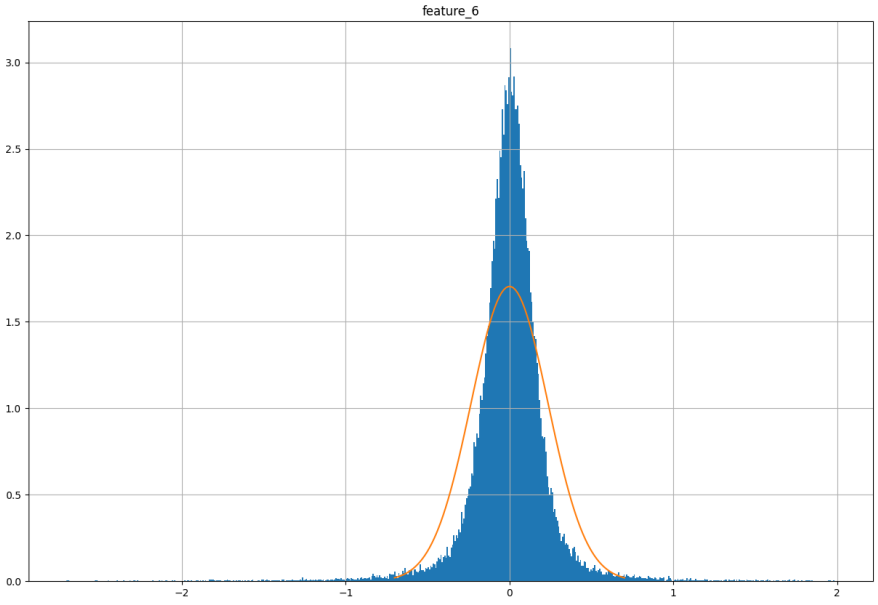

feature_1
0
1
2
3
4
5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


feature_2
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2


feature_3
5
4
3
2
1
0
−3
−2
−1
0
1
2
3

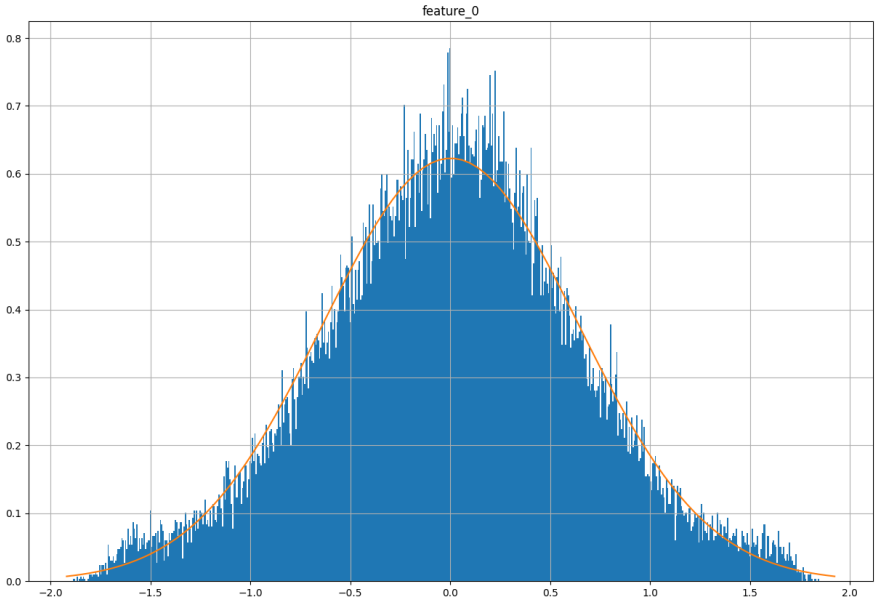

feature_4
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−1
0
1
2
3

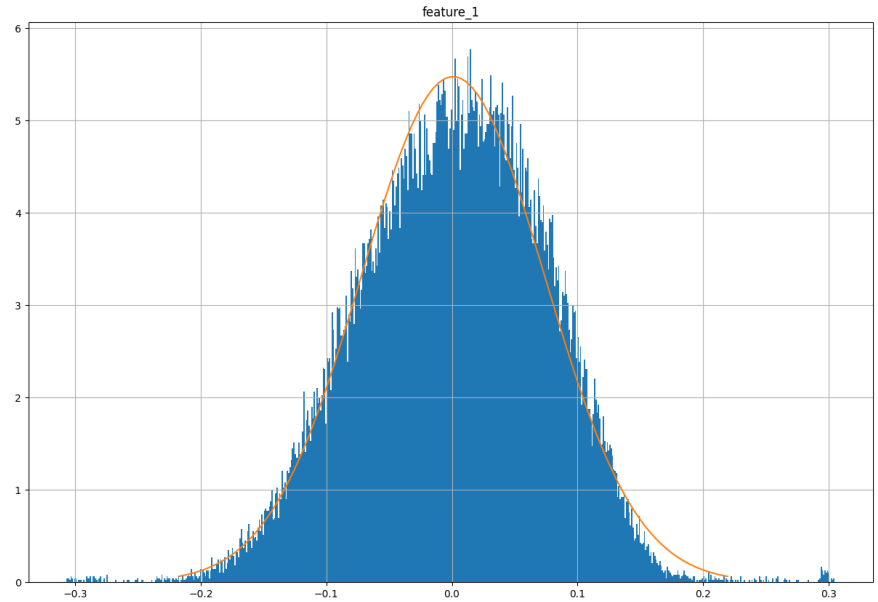

feature_5
2.5
2.0
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
3

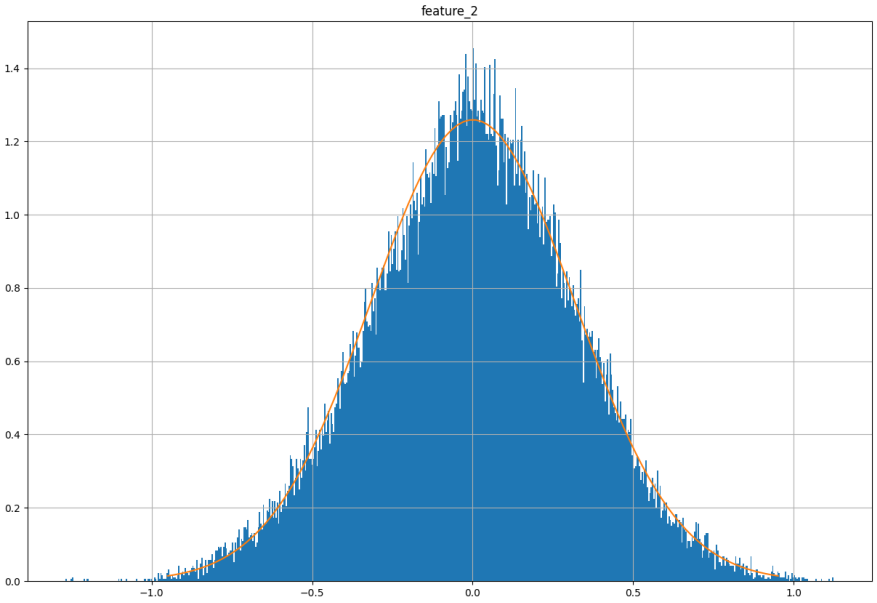


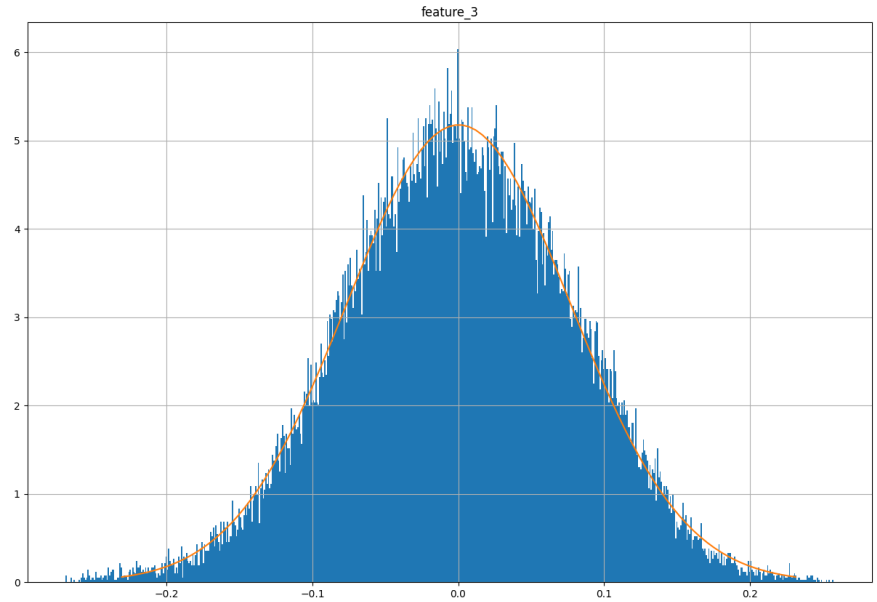
Gaussianization

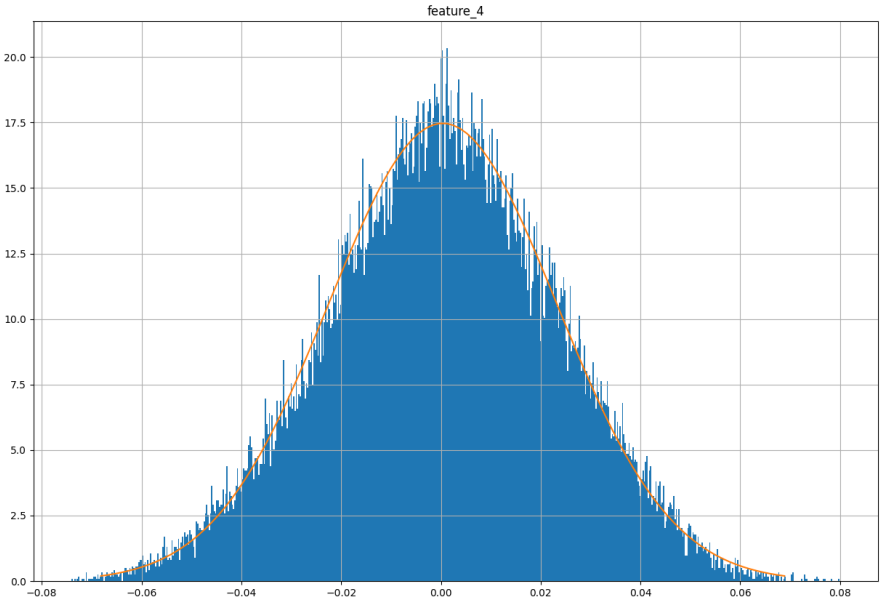
At this point features are strongly non-Gaussian (Taleb 2020), while the volatility of the first four statistical moments is small enough for us to define an invariant measure to Gaussianize them (Goerg 2010), (Arbabi and Sapsis 2019), (Marti et al. 2016). This step also further reduces the non-stationarity of the features.

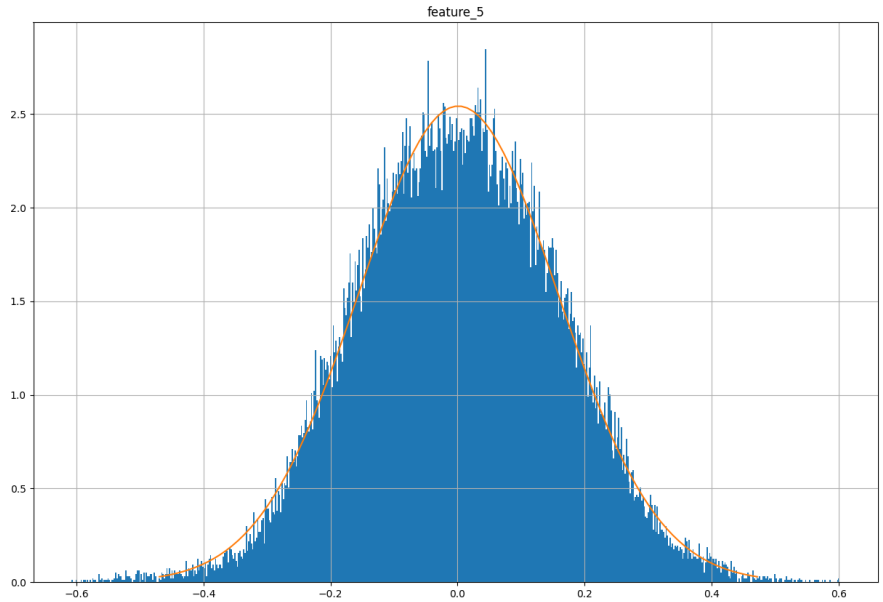

feature_0
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

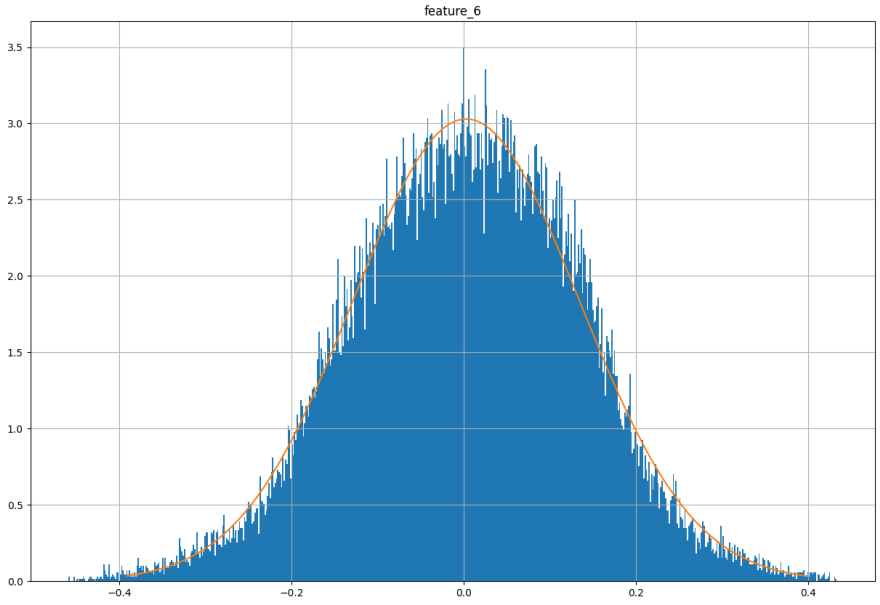

feature_1
6
5
4
3
2
1
0
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3


feature_2
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0


feature_3
6
5
4
3
2
1
0
−0.2
−0.1
0.0
0.1
0.2


feature_4
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08

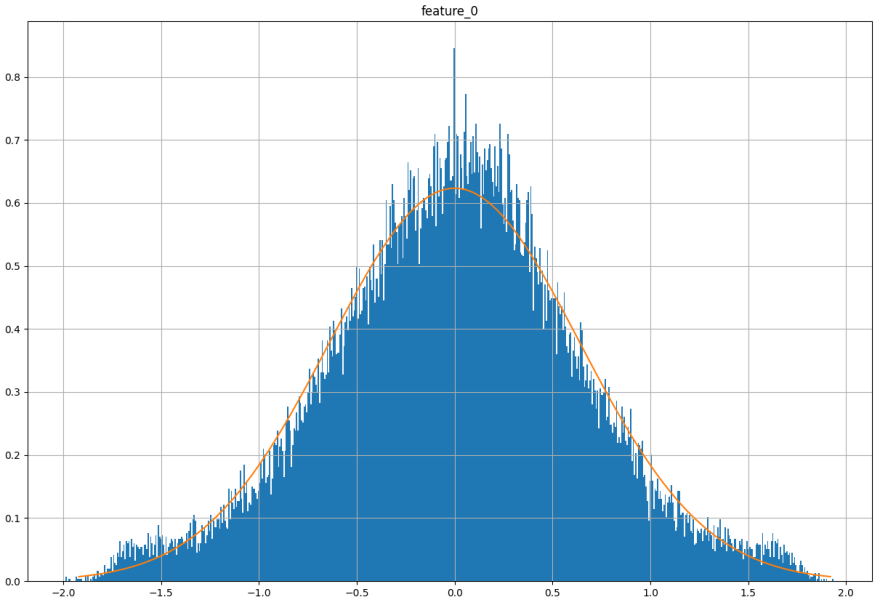

feature_5
2.5
2.0
1.5
1.0
0.5
0.0
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

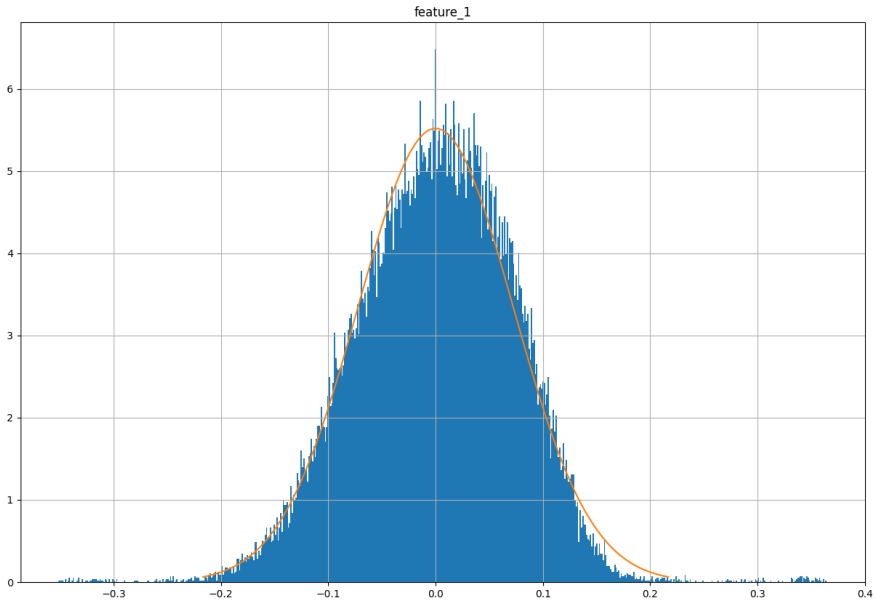


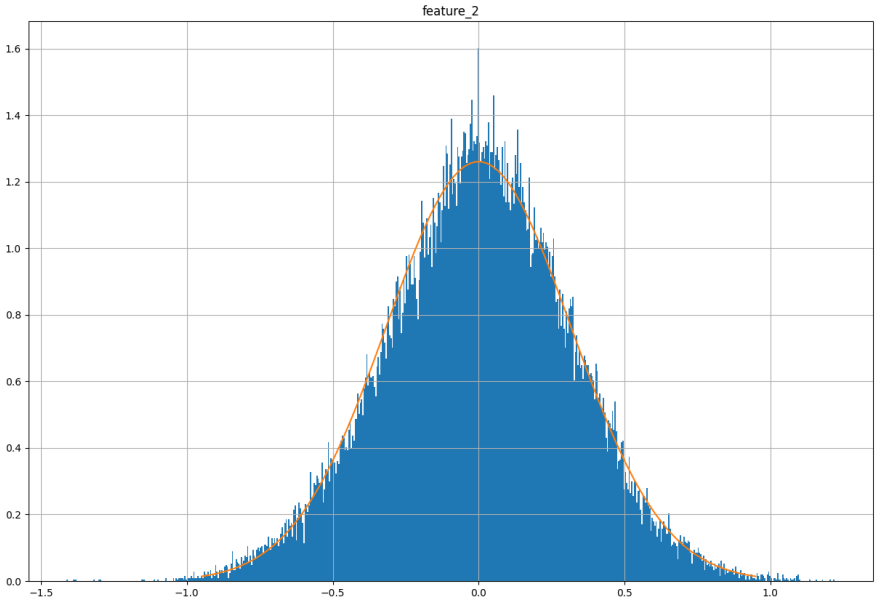
It is worth underlying, at this step, an analogy between this procedure and the use of fractional calculus for financial time series, discussed in (Prado 2018) and the use of the fractional Fokker-Plank equation to model anomalous diffusion (i.e., noise models with non-Gaussian statistics) (Metzler, Barkai, and Klafter 1999). As fractional calculus for nonlocal dynamics is associated with nonlocal operators in time (Li and Rosenfeld 2021) and can be used to transform observations to a desired statistics (maximizing correlation between the original data), so this Gaussianization step implicitly introduces a cross-sectional nonlinear metric. The two analogous perspective are equivalent in the context of the ergodic hypotesis (Poitras, Wong, and Heaney 2015).

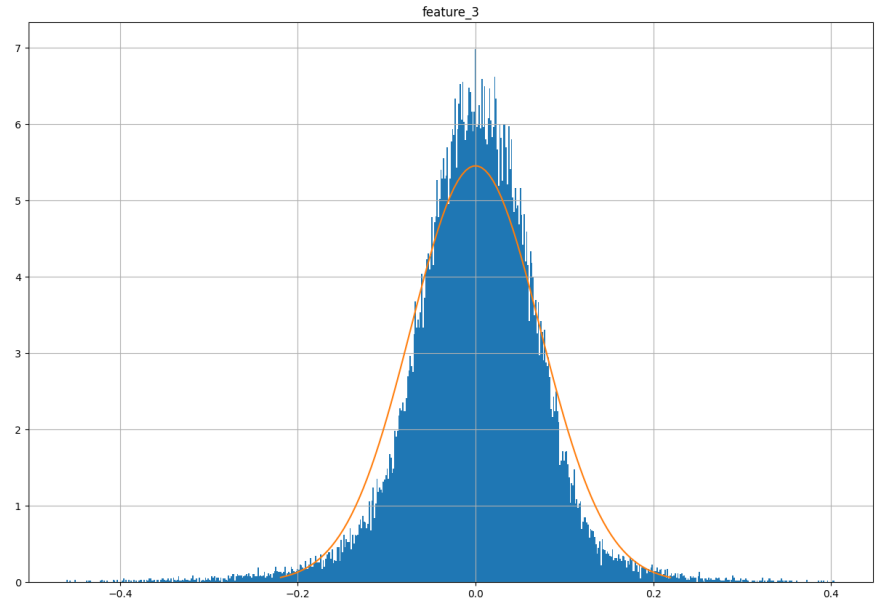
As the Gaussianization step is nonlinear, the orthogonality condition is destroyed: we can however perform again orthogonalization and obtain orthogonal, Gaussian features.

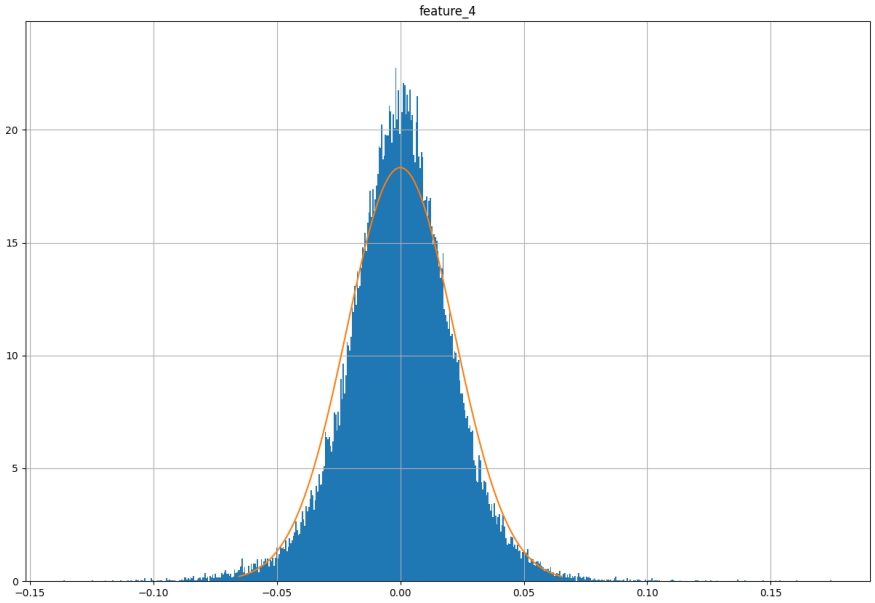
Moreover, performing the three steps OGO, compared to only the first O step, leads to features which are always more than 98.09% Spearman rank correlated.

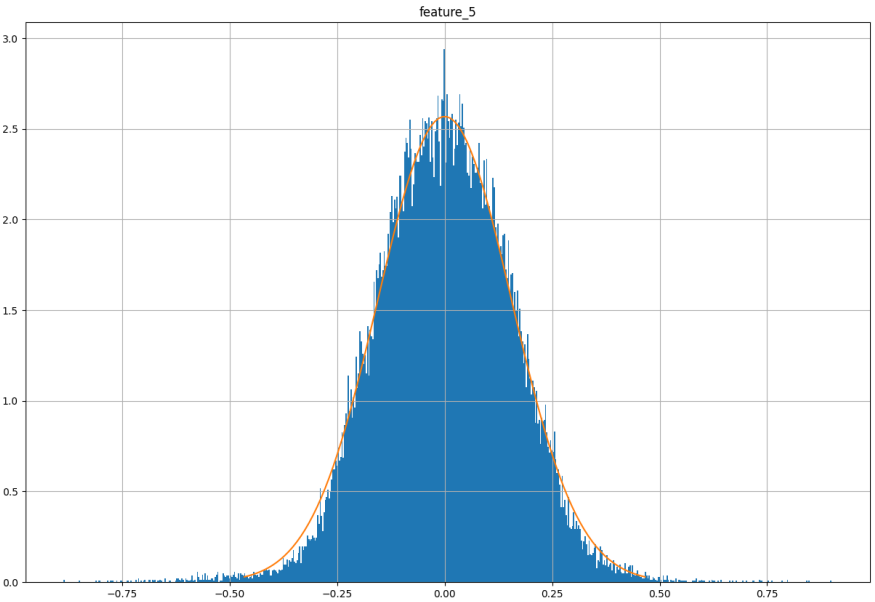

feature_0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

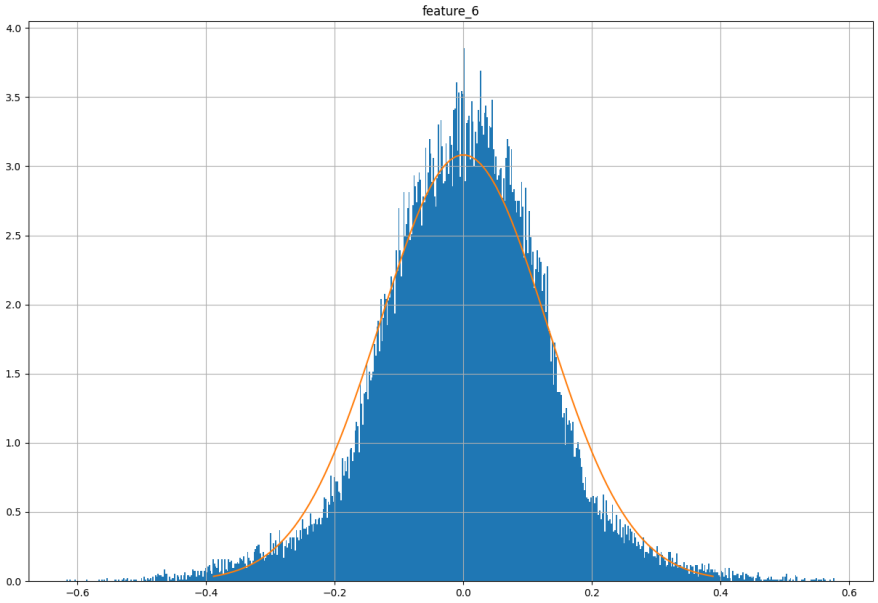

feature_1
0
1
2
3
4
5
6
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4


feature_2
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−1.5
−1.0
−0.5
0.0
0.5
1.0


feature_3
7
6
5
4
3
2
1
0
−0.4
−0.2
0.0
0.2
0.4


feature_4
20
15
10
5
0
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15


feature_5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75



Standardization

Given the degree of stationarity, we can standardize using a global transformation.

Principal Component Analysis

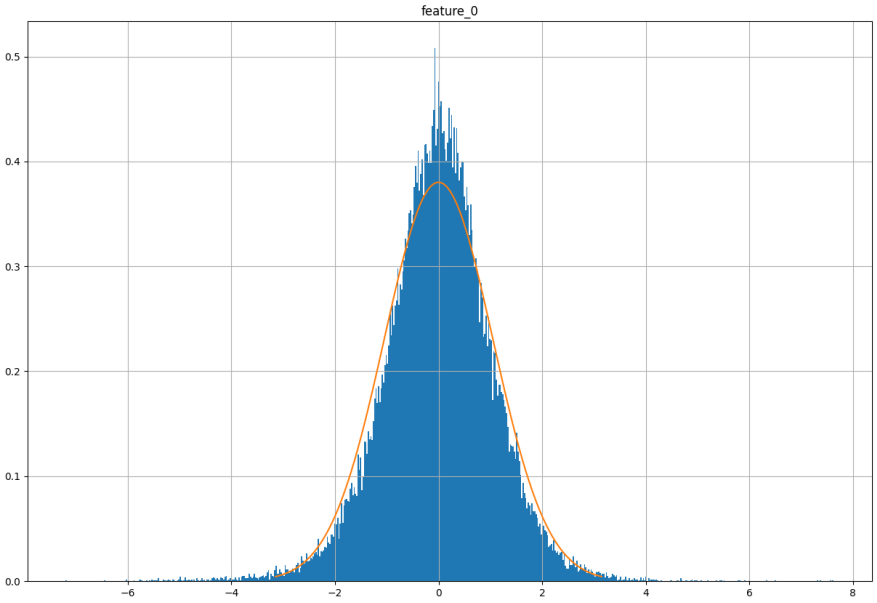
We perform Principal Component Analysis (PCA) to decorrelate the features as much as possible. Again, we perform a global basis transformation thanks to the fact that features are close to stationary.

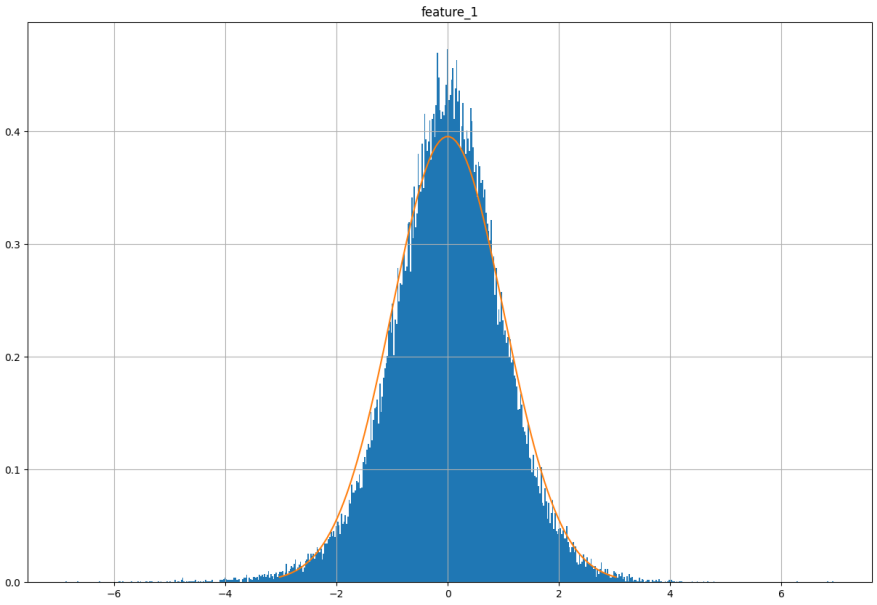
The Gaussianization step is implicitly introducing a kernel, so that this procedure can be though of a specific case of Kernel PCA (Nateghi and Manzi 2016).

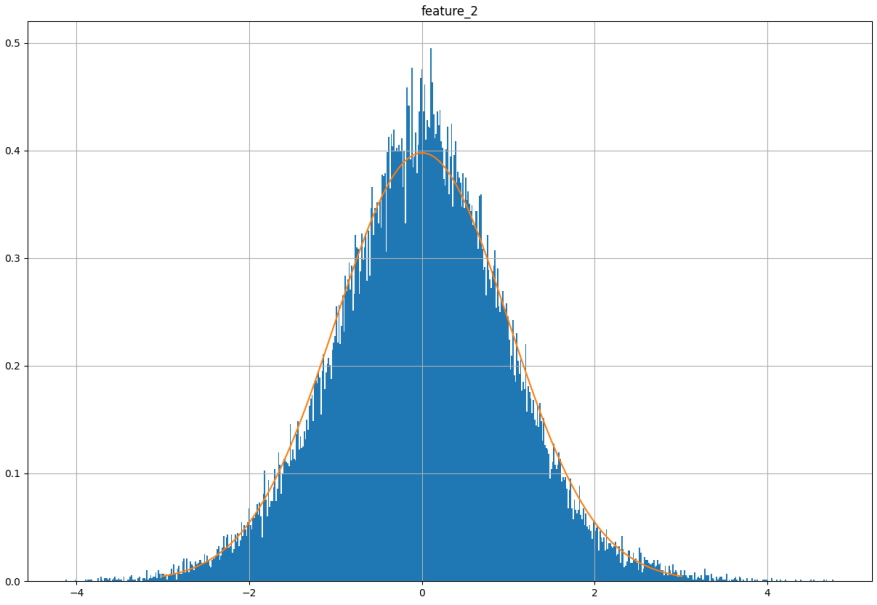
We here obtain new features linearly combining them. The linear combination coefficients come from Single Value Decomposition (SVD): we can hence standardize again.

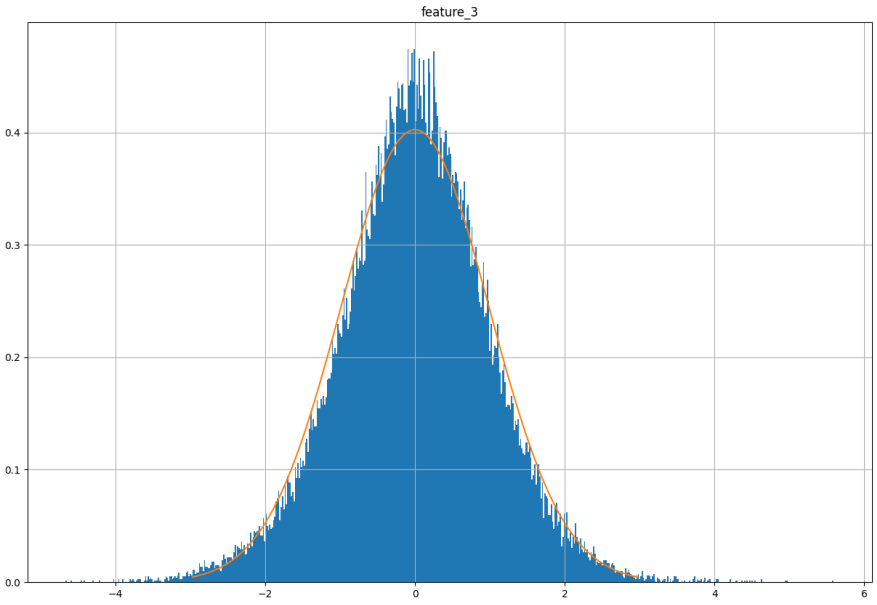
$$X_j^{out} = \frac{1}{\sigma_{j2}} \sum_i \alpha_{ij} \frac{1}{\sigma_{j1}} \cdot \left(k_j \left(X_i^{in} - A_{i1} \right) - A_{i2} \right) \tag{8}$$

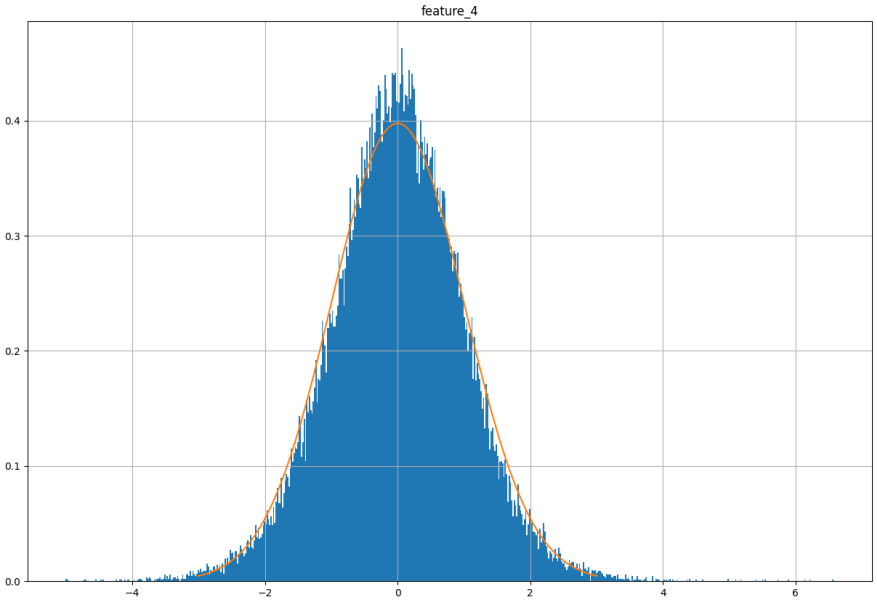
All the steps after the second orthogonalization are linear transformations: output features remain orthogonal to the factor space.

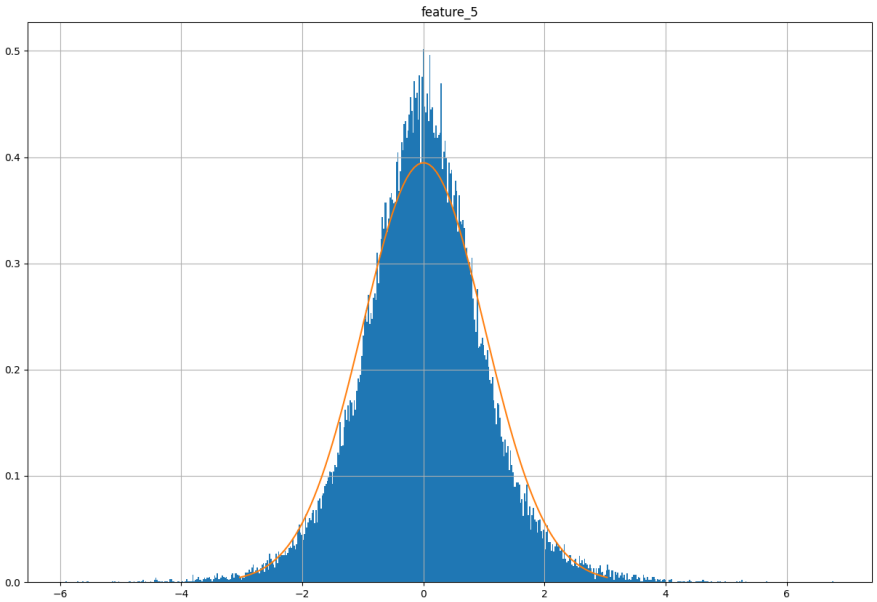

feature_0
0.5
0.4
0.3
0.2
0.1
0.0
−6
−4
−2
0
2
4
6
8

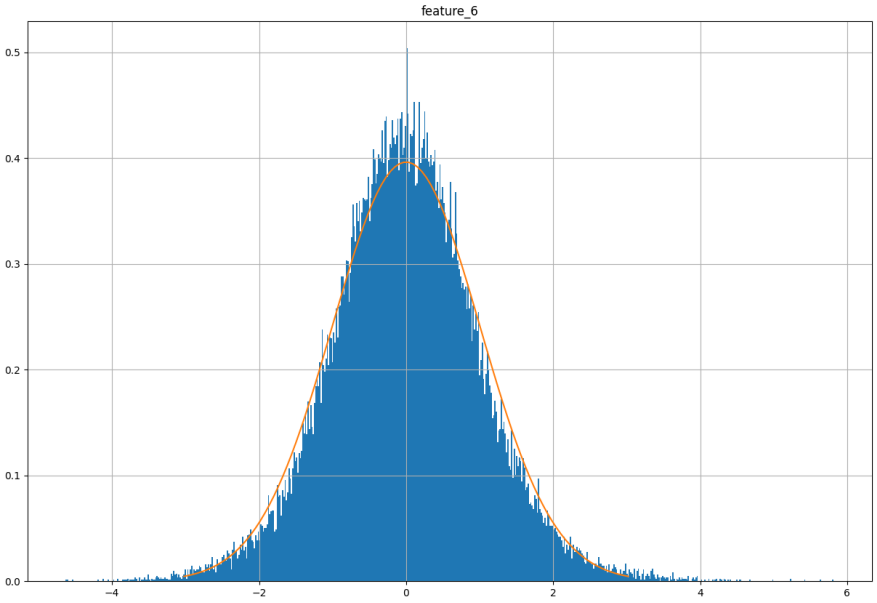

feature_1
0.4
0.3
0.2
0.1
0.0
−6
−4
−2
0
2
4
6

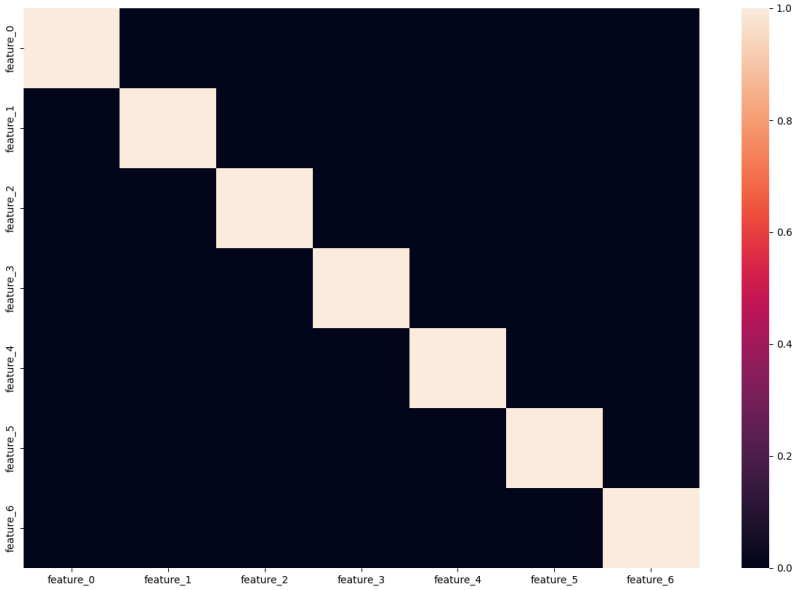

feature_2
0.5
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4


feature_3
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
6


feature_4
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
6

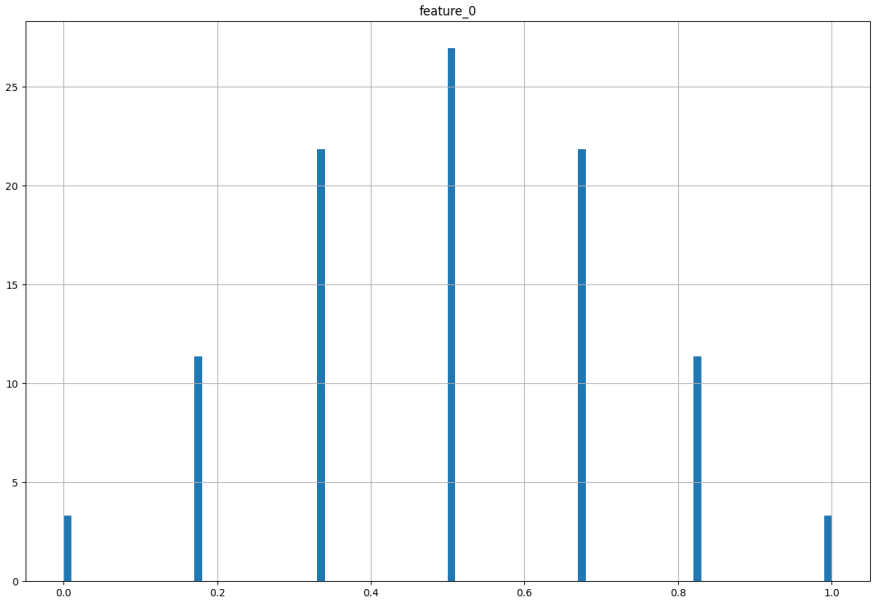

feature_5
0.0
0.1
0.2
0.3
0.4
0.5
−6
−4
−2
0
2
4
6

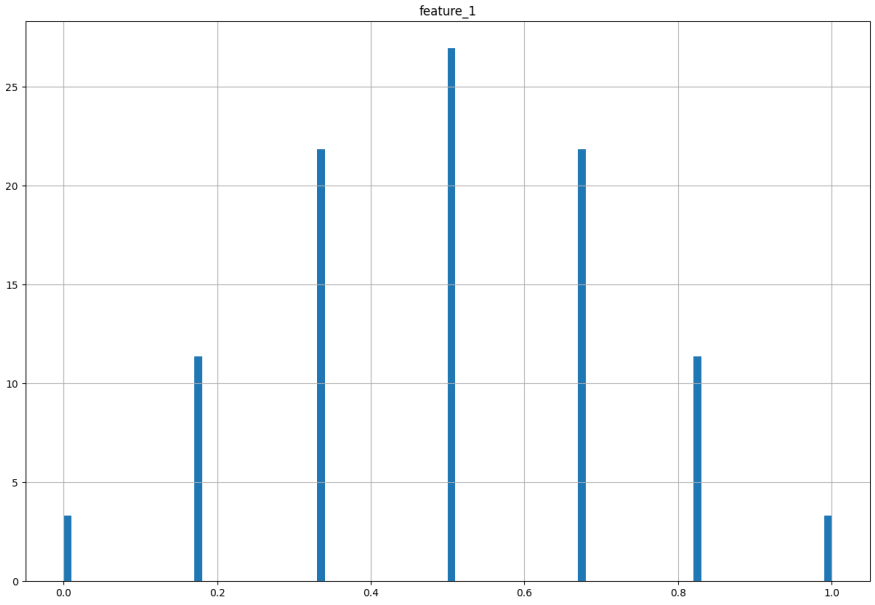

feature_6
0.5
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
6

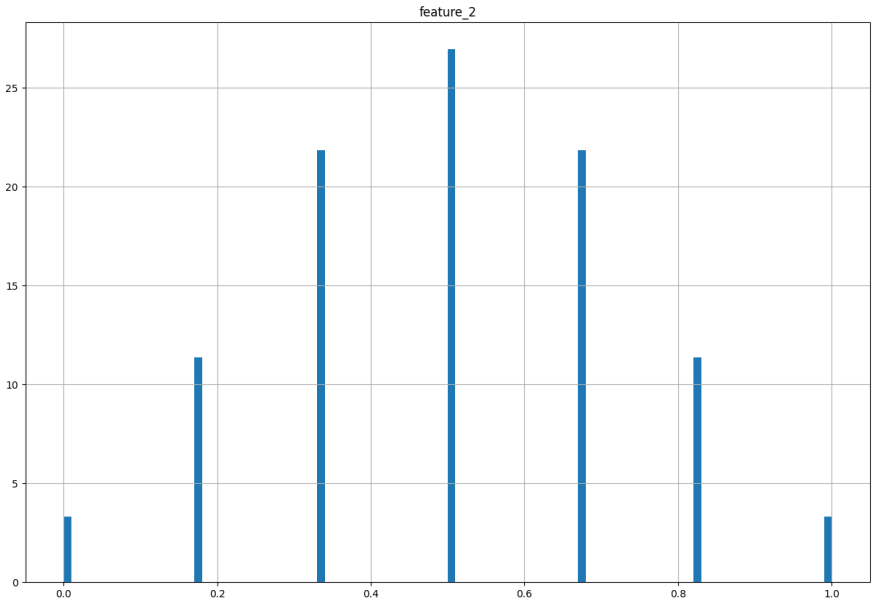

feature_0
feature_1
feature_2
feature_3
feature_4
feature_5
feature_6
1.0
0.8
0.6
0.4
0.2
0.0

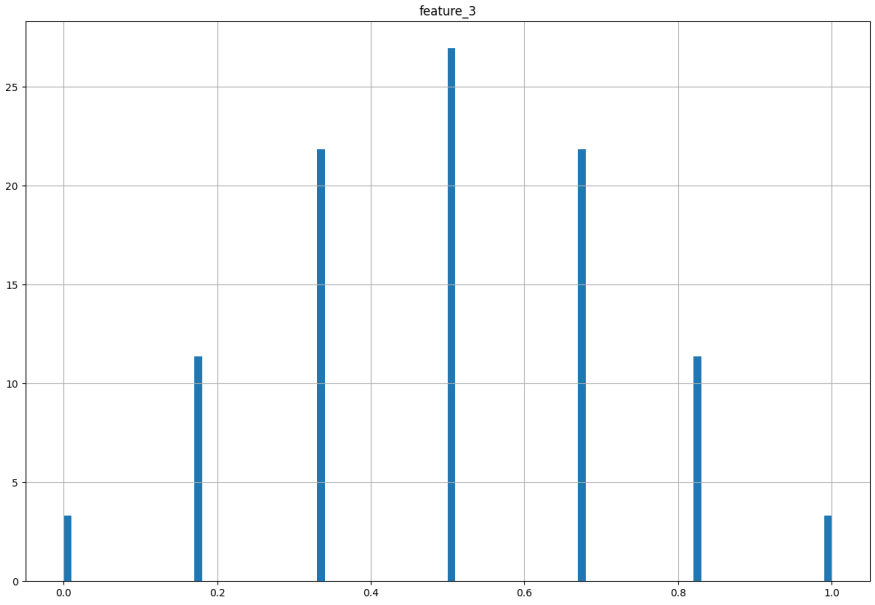
Quantization

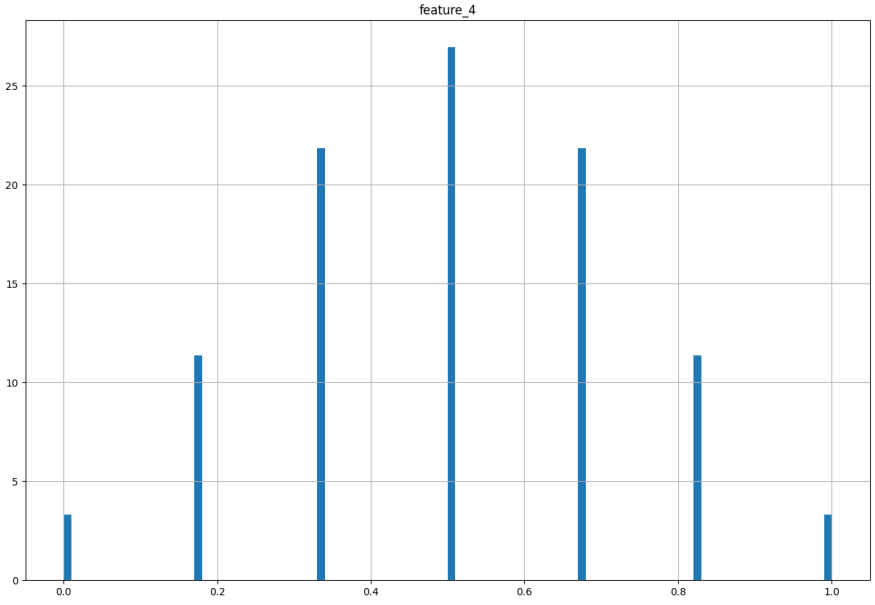
We perform Lloyd-Max Quantization (Lloyd, Lloyd, and Lloyd 1982) to solve a classification problem in a least-square sense.

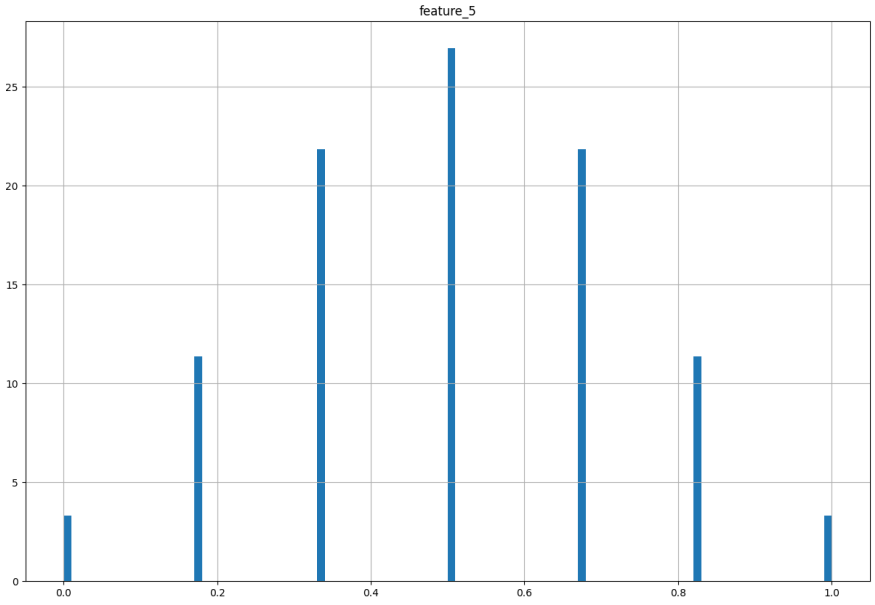


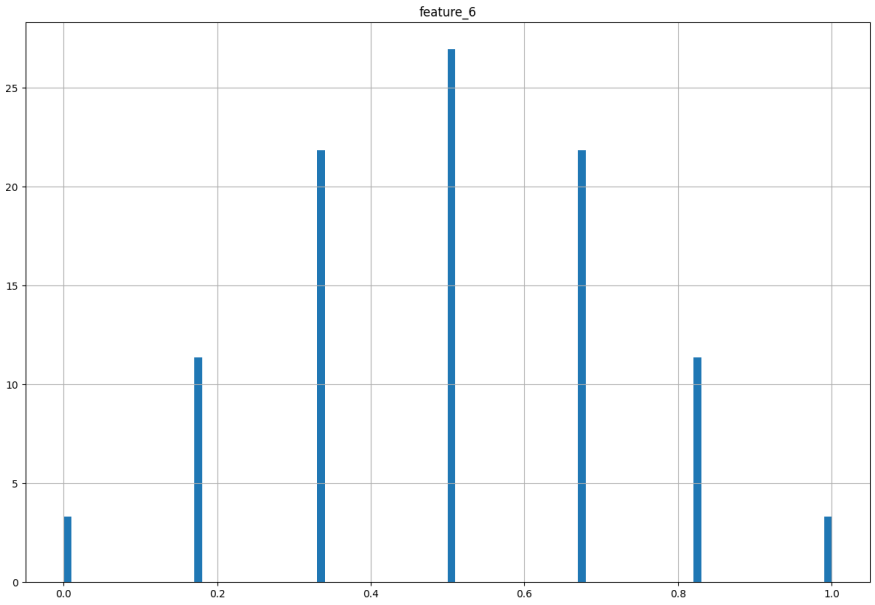

feature_1
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0

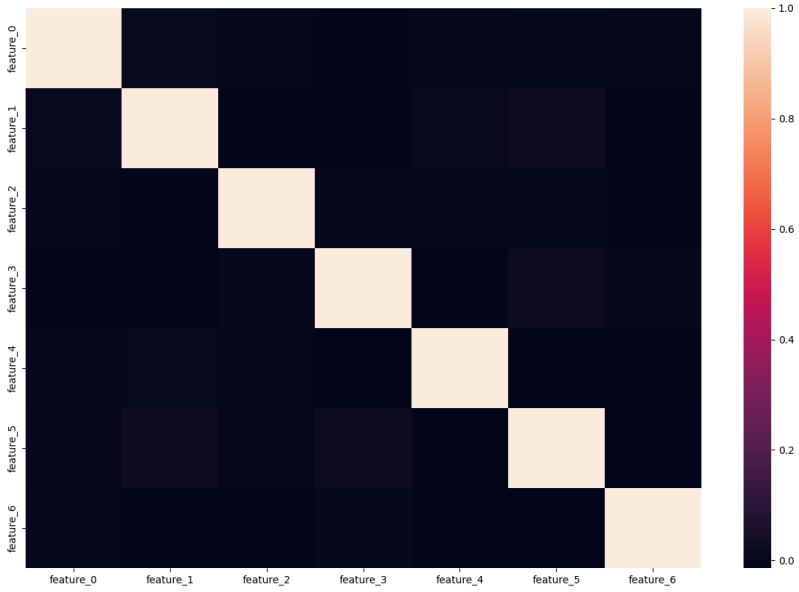

feature_2
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0


feature_3
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0


feature_4
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0

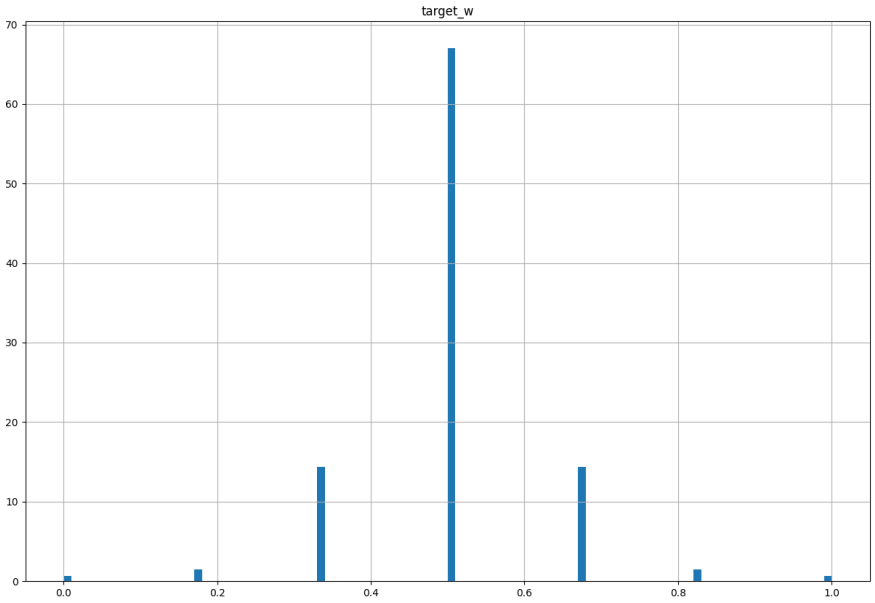

feature_5
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0

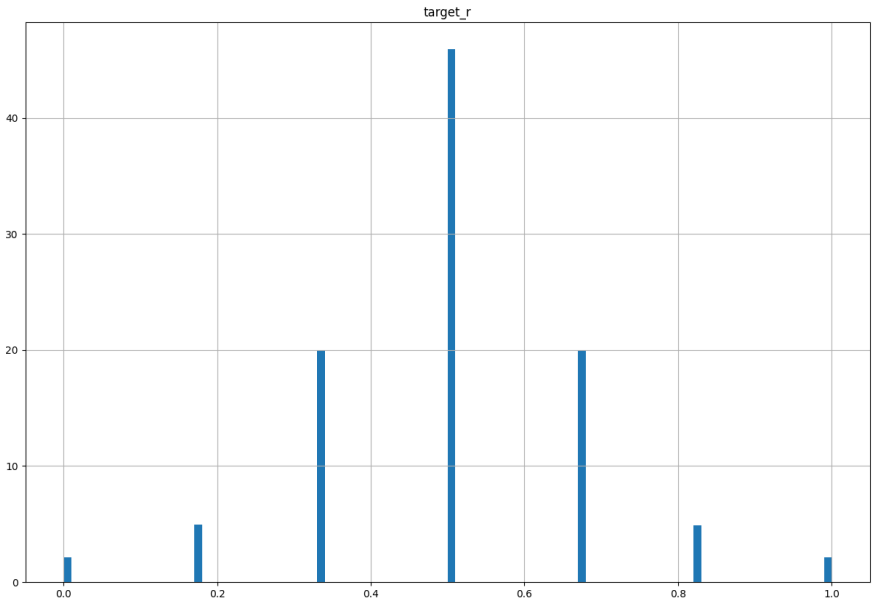

feature_6
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0

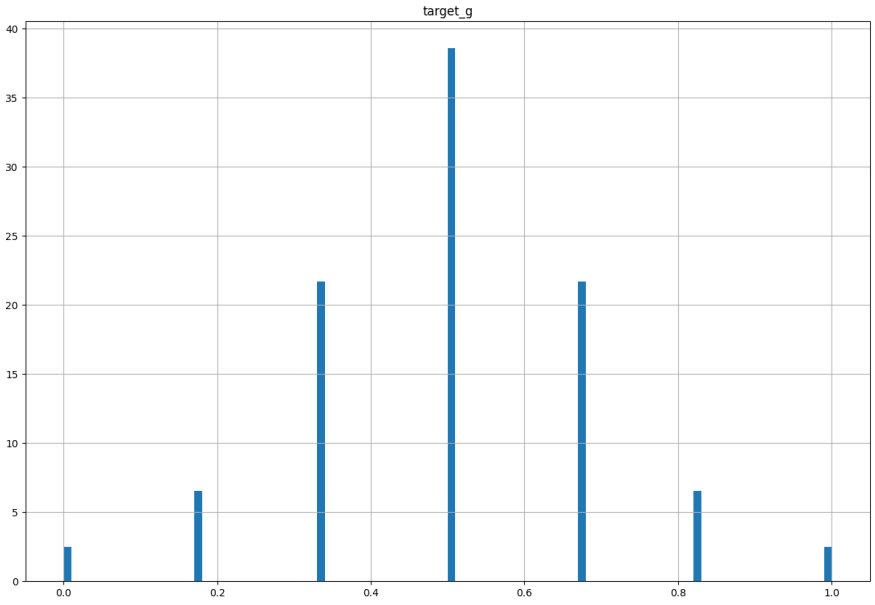


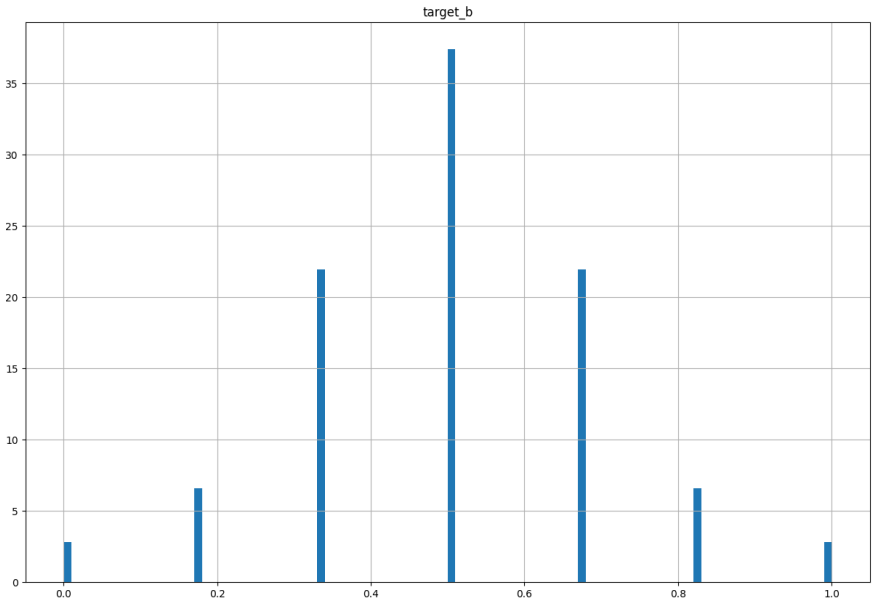
Targets

Targets are simply quantized. This is done maximizing the explained variance of the quantization scheme assuming a median distribution across all historical observations.


target_w
70
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0


target_r
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0


target_g
40
35
30
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0


target_b
35
30
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0

References

Arbabi, Hassan, and Themistoklis P. Sapsis. 2019. “Generative Stochastic Modeling of Strongly Nonlinear Flows with Non-Gaussian Statistics.” *SIAM/ASA J. Uncertain. Quantification*. https://doi.org/10.1137/20m1359833.

Bonne, Wang, and Zhang. 2021. “Machine Learning Factors: Capturing Non Linearities in Linear Factor Models.” *MSCI Research Insights*.

Chan, Matyas. 2022. “Econometrics with Machine Learning.” *Advanced Studies in Theoretical and Applied Econometrics*. https://doi.org/10.1007/978-3-031-15149-1.

Fama, Eugene F., Kenneth R. French, and Kenneth R. French. 1993. “Common Risk Factors in the Returns on Stocks and Bonds.” *Journal of Financial Economics*. https://doi.org/10.1016/0304-405x(93)90023-5.

Goerg, Georg M. 2010. “The Lambert Way to Gaussianize Heavy Tailed Data with the Inverse of Tukey’s H as a Special Case.” *arXiv: Statistics Theory*.

Li, Xiuying, and Joel A Rosenfeld. 2021. “Fractional Order System Identification with Occupation Kernel Regression.” In *2021 American Control Conference (Acc)*, 4001–6. https://doi.org/10.23919/ACC50511.2021.9483209.

Lloyd, S. P., S. P. Lloyd, and Seth Lloyd. 1982. “Least Squares Quantization in Pcm.” *IEEE Transactions on Information Theory*. https://doi.org/10.1109/tit.1982.1056489.

Marti, Gautier, Sébastien Andler, Frank Nielsen, and Philippe Donnat. 2016. “Exploring and Measuring Non-Linear Correlations: Copulas, Lightspeed Transportation and Clustering.” *NIPS Time Series Workshop*. https://doi.org/null.

Metzler, Ralf, Eli Barkai, and Joseph Klafter. 1999. “Anomalous Diffusion and Relaxation Close to Thermal Equilibrium: A Fractional Fokker-Planck Equation Approach.” *Phys. Rev. Lett.* 82 (18): 3563–7. https://doi.org/10.1103/PhysRevLett.82.3563.

Nateghi, Vahid, and Matteo Manzi. 2016. “Machine Learning Methods for Nonlinear Reduced-Order Modeling of the Thermospheric Density Field.” *Advances in Space Research*.

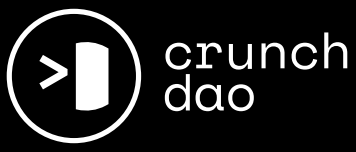
Poitras, Geoffrey, Wing K. Wong, and John Heaney. 2015. “Classical Ergodicity and Modern Portfolio Theory.” *American Physical Society*. https://doi.org/10.1155/2015/737905.

Prado, Marcos Lopez de. 2018. *Advances in Financial Machine Learning*. Wiley, 1st edition.

———. 2019. “Beyond Econometrics: A Roadmap Towards Financial Machine Learning.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3365282.

Sharpe, William F. 1964. “CAPITAL Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk.” *Journal of Finance*. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x.

Taleb, Nassim Nicholas. 2020. “Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications.” arXiv. https://doi.org/10.48550/ARXIV.2001.10488.


crunch
dao