Distribution of Score-Based Test Statistic for Gaussian Model Fails in the High-Dimensional Regime

Ulrik Unneberg*

June 6, 2023

Abstract

Unnormalized statistical models play an important role in physics and machine learning. Score-matching techniques have in recent years demonstrated powerful capabilities in making inference about models without the computational expense of obtaining the normalization constant of multidimensional models. One recently developed technique is the score-based hypothesis testing method, where a test statistic is based on the difference between Hyvärinen scores corresponding to the null and alternative hypotheses. When the dimension of the parameter p is fixed, the null distribution of the test statistic converges in distribution to a quadratic form of multivariate Gaussian vectors. In this paper, we provide numerical simulations which suggests that this result does not hold in general in the high-dimensional regime, where $p/n \overset{n\to\infty}{\longrightarrow} \kappa > 0$.

1 Introduction

Consider a statistical model $\{f_{\boldsymbol{\theta}}(\mathbf{x}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$, $\mathbf{x} \in \mathbb{R}^p$, and let $q_{\boldsymbol{\theta}}$ be its unnormalized density, defined as

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{Z} q_{\boldsymbol{\theta}}(\mathbf{x}), \tag{1}$$

where the partition function $Z = \int_{\mathbb{R}^p} q_{\boldsymbol{\theta}}(\mathbf{s}) d\mathbf{s}$ is the normalization constant, which might depend on $\boldsymbol{\theta}$. Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n \overset{\text{iid}}{\sim} f_{\boldsymbol{\theta}}$ be the observed data. The goal is to estimate the parameter $\boldsymbol{\theta}$. M-estimation is a general estimation framework formulated by Huber [1], where the estimator is an extremum, often the minimum, of the sample average of an objective funciton $\rho(\boldsymbol{\theta}, \mathbf{x})$,

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}\in\Theta} \frac{1}{n} \sum_{i=1}^{n} \rho(\boldsymbol{\theta}, \mathbf{x}_i).$$

A classical example is the maximum likelihood estimator (MLE) where $\rho_{\text{MLE}}(\boldsymbol{\theta}, \mathbf{x}) = -\log f_{\boldsymbol{\theta}}(\mathbf{x})$. Hyvärinen proposed in 2005 [2] the objective function

$$\rho_{\text{sm}}(\boldsymbol{\theta}, \mathbf{x}) = \frac{1}{2} ||\nabla_{\mathbf{x}} \log f_{\boldsymbol{\theta}}(\mathbf{x})||_2^2 + \Delta_{\mathbf{x}} \log f_{\boldsymbol{\theta}}(\mathbf{x}), \tag{2}$$

*ulrikmu@stud.ntnu.no

where $\Delta_{\mathbf{x}} \equiv \sum_{i=1}^{p} \frac{\partial^2}{\partial x_i^2}$. One can easily verify that this objective function requires no information about the normalization constant of $f_{\boldsymbol{\theta}}$, as

$$\frac{1}{2}||\nabla_{\mathbf{x}} \log f_{\boldsymbol{\theta}}(\mathbf{x})||_2^2 + \Delta_{\mathbf{x}} \log f_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{2}||\nabla_{\mathbf{x}} \log q_{\boldsymbol{\theta}}(\mathbf{x})||_2^2 + \Delta_{\mathbf{x}} \log q_{\boldsymbol{\theta}}(\mathbf{x}).$$

The M-estimator using this objective function has become known as the score-matching estimator $\hat{\boldsymbol{\theta}}_{\text{sm}}$, explicitly defined as

$$\hat{\boldsymbol{\theta}}_{\text{sm}} = \text{argmin}_{\boldsymbol{\theta}\in\Theta} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left(\frac{1}{2}||\nabla_{\mathbf{x}} \log q_{\boldsymbol{\theta}}(\mathbf{x})||_2^2 + \Delta_{\mathbf{x}} \log q_{\boldsymbol{\theta}}(\mathbf{x}) \right)}_{\rho_{\text{sm}}(\boldsymbol{\theta},\mathbf{x})} \Bigg|_{\mathbf{x}=\mathbf{X}_i}, \tag{3}$$

which Hyvärinen proved is a consistent estimator[2].

1.1 Hypothesis testing

Wu et al. developed in 2022 [3] how this estimator can be used in hypothesis testing when considering unnormalized statistical models. Let $\boldsymbol{\theta}_\star$ be the true parameter, and consider the two-sided test

$$H_0 : \boldsymbol{\theta}_\star = \boldsymbol{\theta}_0, \qquad \text{vs.} \qquad H_1 : \boldsymbol{\theta}_\star \neq \boldsymbol{\theta}_0.$$

They provided the test statistic

$$T = 2(S(\boldsymbol{\theta}_0) - S(\hat{\boldsymbol{\theta}}_{\text{sm}})), \tag{4}$$

where $S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \rho_{\text{sm}}(\boldsymbol{\theta}, \mathbf{X}_i)$. They further used the multivariate Wilk's theorem and asymptotic normality of a consistent estimator to show that under the null hypothesis,

$$T \xrightarrow{\mathcal{L}} \mathbf{z}^\top \mathbf{H} \mathbf{z},$$

where $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$, $\mathbf{V} \equiv \mathbf{H}^{-1}\mathbf{K}\mathbf{H}^{-1} \in \mathbb{R}^{p\times p}$, and

$$\mathbf{H} \equiv \mathbb{E}_\star[\nabla^2_{\boldsymbol{\theta}} \rho_{\text{sm}}(\boldsymbol{\theta}, \mathbf{X})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\star}],$$

$$\mathbf{K} \equiv \mathbb{E}_\star[\nabla_{\boldsymbol{\theta}} \rho_{\text{sm}}(\boldsymbol{\theta}, \mathbf{X}) \nabla^\top_{\boldsymbol{\theta}} \rho_{\text{sm}}(\boldsymbol{\theta}, \mathbf{X})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\star}].$$

Here, $\mathbb{E}_\star[\,\cdot\,]$ is expectation with respect to $\mathbf{X} \sim f_{\boldsymbol{\theta}_\star}$.

2 Experimental design

In this paper, we will demonstrate numerical simulations suggesting how the asymptotic distribution of T under the null appear to break down in high dimensions for a particular model. The high-dimensional regime should in this paper be interpreted as when $n, p \to \infty$ in such a way that

$p/n \to \kappa > 0$. As the example of study, consider the Gaussian model of the form

$$f_{\boldsymbol{\theta}}(\mathbf{x}) \propto \exp\Big\{ -\frac{\beta}{p}\sum_{j=1}^{p}\sum_{k=1}^{p} x_j x_k - \sum_{j=1}^{p}\theta_j x_j^2 \Big\} = \exp\Big\{ -\frac{1}{2}\mathbf{x}^\top \underbrace{2(\Lambda_{\boldsymbol{\theta}} + \frac{\beta}{p}\mathbf{1}\mathbf{1}^\top)}_{\Sigma^{-1}}\mathbf{x} \Big\},$$

where $\Lambda_{\boldsymbol{\theta}} = \mathrm{diag}(\theta_1, ..., \theta_p)$. This model has the explicit score-matching estimator

$$\hat{\boldsymbol{\theta}}_{\mathrm{sm}} = \Big(-\frac{\beta}{p}\frac{1}{n}\sum_{i\in[n]} \mathbf{X}_i(\mathbf{1}^\top \mathbf{X}_i) + \frac{1}{2}\mathbf{1} \Big) \oslash \frac{1}{n}\sum_{i\in[n]} \mathbf{X}_i^{\otimes 2}. \tag{5}$$

Here, $\otimes$ and $\oslash$ denotes point-wise multiplication and division respectively, $\mathbf{X}^{\otimes c} = \overbrace{\mathbf{X}\otimes\mathbf{X}\otimes\cdots\otimes\mathbf{X}}^{c \text{ times}} \in \mathbb{R}^p$, and $\mathbf{1} = [1, 1, ..., 1]^\top \in \mathbb{R}^p$. The model is chosen as it is less trivial than an estimator for a mere sample mean, but simple enough to yield an explicit expression for both the estimator and its corresponding covariance matrix $\mathbf{V}$, which drastically enhance computation.

With the model chosen above, we compute 10^5 realizations of both T and $\mathbf{z}^\top\mathbf{H}\mathbf{z}$, with $\boldsymbol{\theta}_\star = \boldsymbol{\theta}_0$ chosen randomly such that $(\boldsymbol{\theta}_\star)_j \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$, $j \in [p]$. We further let $n = 150$, $\beta = 1$, and let $\kappa = p/n$ take the values $\{0.01, 0.1, 1, 5\}$ to prominently demonstrate the trend. Hence, p will take the values $\{2, 15, 150, 750\}$. The histograms are demonstrated in Figure 1. The x-axis is also rescaled such that $\mathbf{z}^\top\mathbf{H}\mathbf{z}$ has unit variance in each plot, to more easily compare the plots.

3 Results

According to classical theory, $T \xrightarrow{\mathcal{L}} \mathbf{z}^\top\mathbf{H}\mathbf{z}$. However, the plots suggest that this might not hold in high dimensions. The distribution of $\mathbf{z}^\top\mathbf{H}\mathbf{z}$ is plotted with the distribution of T for different κ. The two distributions seems to coincide for small values of κ, but as it keeps increasing, the distribution of T is right-shifted and stretched out compared to that of $\mathbf{z}^\top\mathbf{H}\mathbf{z}$.

While this difference might seem negligible with no implications in practical applications, a perhaps more illuminating demonstration is that of its corresponding p-values. As known, hypothesis tests are often performed by calculating the p-value of the test statistic. That is, the probability of obtaining a value equal to, or more extreme than, the obtained value for the test statistic, under the null hypothesis. If the p-value is less than the set significance level, one rejects the null hypothesis.

By running 10^5 experiments getting realizations of the test statistic when the null hypothesis actually is true, we expect the p-values to be uniformly distributed. As shown in Figure 2, the p-values appear increasingly right skewed for increasing κ.

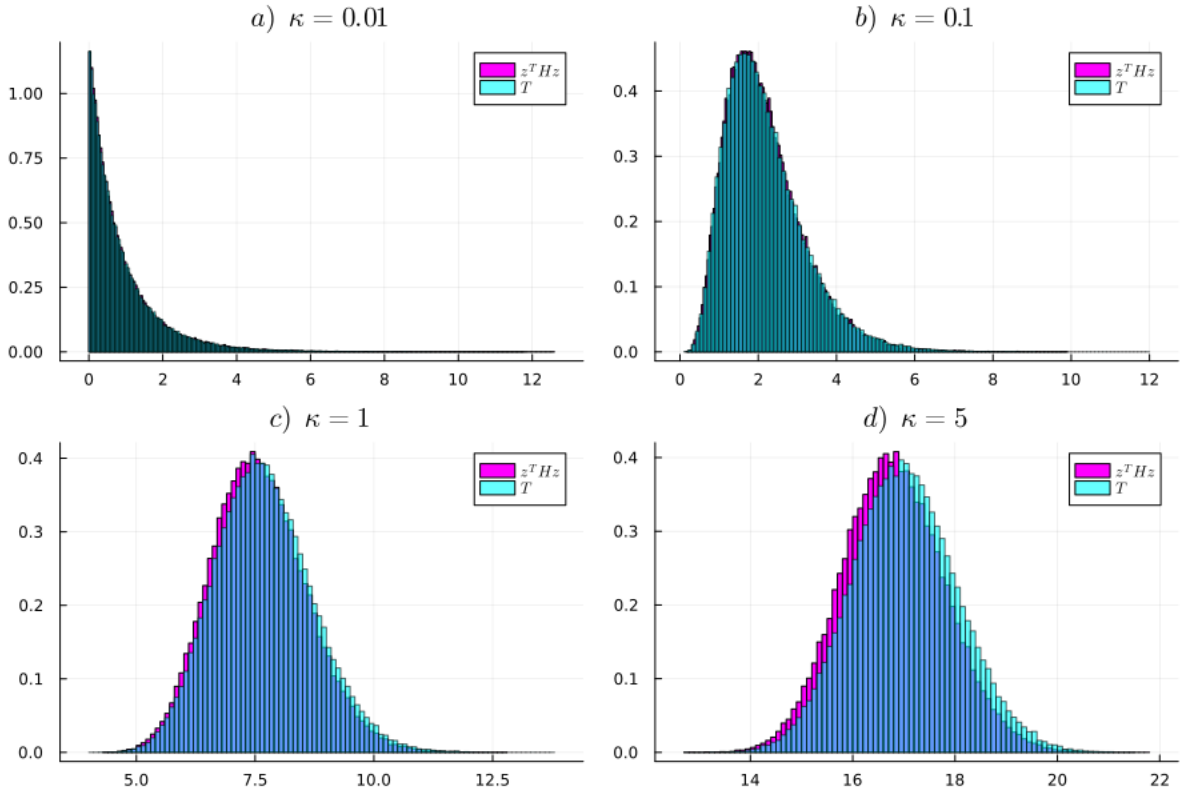

Figure 1: Histograms of $\mathbf{z}^\top \mathbf{Hz}$ and T with for different κ. An apparent right-shift of the distribution of T grows prominent for increasing κ

4 Discussion

The fact that empirical p-values based on the algorithm provided by Wu [3] has the distribution as shown in Figure 2 under the null hypothesis, results in far more false rejections than the significance level indicates, and thus induces a false confidence when rejecting.

4.1 Prior work

The properties of estimators in diverging parameter dimensions have been studied extensively in the past. In particular, the asymptotic properties of the regression coefficient $\boldsymbol{\theta} \in \mathbb{R}^p$ in a linear regression problem of the form

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\theta} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} f_\epsilon$$

$i \in [n]$, as $p/n \to \kappa > 0$ was demonstrated heuristically by El Karoui et al. [4] and proved using approximate message passing by Donoho et al. [5]. They find that under some regularity assumptions, the asymptotic distribution of the MLE of $\boldsymbol{\theta}$, denoted $\hat{\boldsymbol{\theta}}_{MLE}$, has an additional gaussian noise, which does not exist when p is fixed. The variance of this gaussian noise can further be found by solving a system of two equations involving the proximal operator used on a convex objective

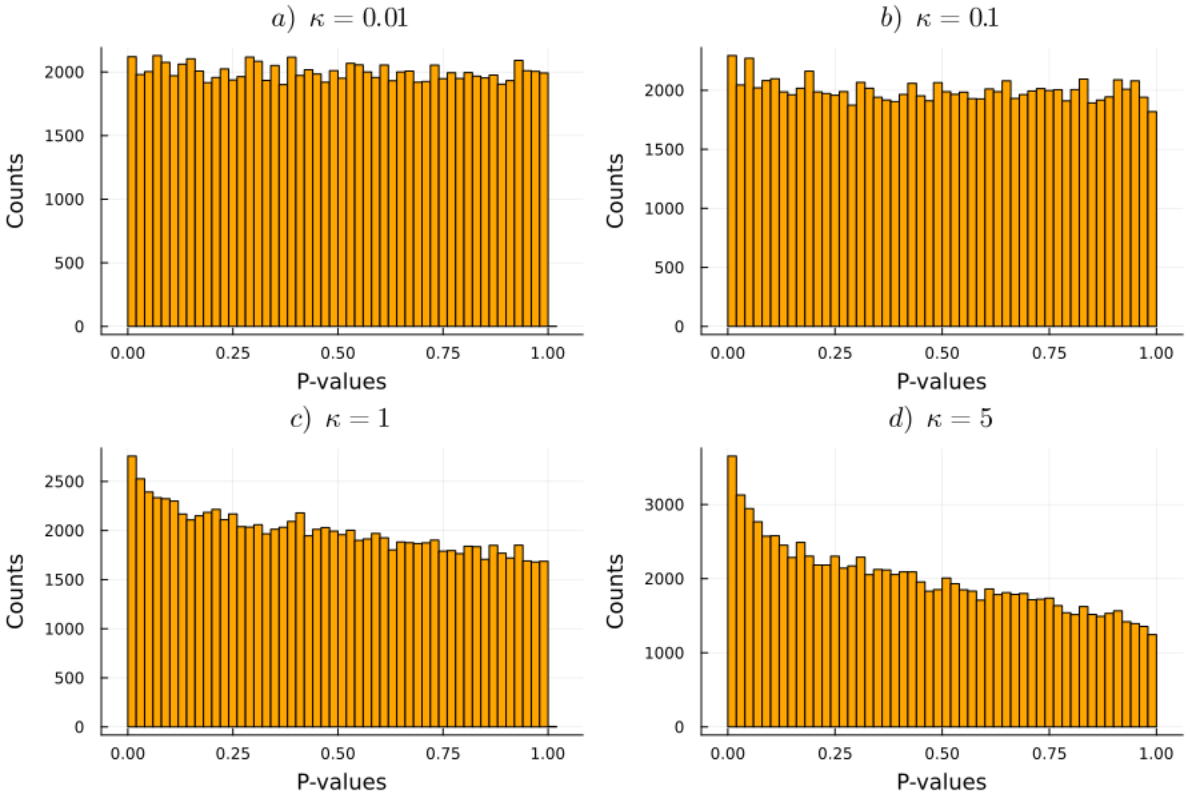

Figure 2: Histogram of the p-values under true null hypothesis.

function $\rho(Y_i - \mathbf{X}_i^\top \boldsymbol{\theta})$. This additional noise results in the likelihood ratio test (LRT) statistic

$$2\Lambda = 2(\ell(\boldsymbol{\theta}_0) - \ell(\hat{\boldsymbol{\theta}}_{\text{MLE}}))$$

having an asymptotically rescaled χ^2-distribution in the high-dimensional regime [6], where the scaling factor is greater than one. A similar tendency as shown in Figure 2 is apparent in [6], page 4. In their case, the p-values is plotted under the null hypothesis $\theta_j = 0$ for $j \in [p]$.

While the problems have similarities, a major difference is that [4–7] study a regression problem where the density f_ϵ is that of a scalar $Y_i - \mathbf{X}_i^\top \boldsymbol{\theta}$. While $\boldsymbol{\theta}$ is high-dimensional, it only appears as a scalar sum in the inner product $\mathbf{X}_i^\top \boldsymbol{\theta}$. The theory provided by [4–6] relies on one-dimensional input in the objective function ρ, and does not by the author seem easily generalizable to an objective function of the general form $\rho(\boldsymbol{\theta}, \mathbf{x})$, which might or might not be convex in $\boldsymbol{\theta}$.

4.2 Limitations of results

A serious limitation in the experimental setup is the computational time complexity's growth with n, which made it computationally intractable to make n as large as wanted to demonstrate the trend

in a closer-to-asymptotic regime. However, it's worth noting that in several practical applications, a sample size of no greater than $n \leq 150$ and $\kappa > 5$ is not so unrealistic. Examples to mention is in genomics, bio-statistics and climate science models. Based on Figure 2, any hypothesis in such a scenario would be utmost inaccurate, regardless of the asymptotic results. The error in distribution of the score-based test statistic should be explored also in the finite-sample case, to further investigate its limitations.

Another limitation to keep in mind is the specific choice of model. Several models should be tested, to assess if similar trends occurs. As always is the case in simulations in the asymptotic domain, experiments can never simulate unbounded n, and theory is ultimately needed to make conclusions. The purpose of this paper was to point out a trend which resembles ones that have been found in previous papers, but on a slightly different type of inference problem.

References

1. Huber PJ. Robust Statistics. 2nd ed. John Wiley & Sons Inc., 2009.
2. Hyvärinen A. Estimation of Non-Normalized Statistical Models by Score Matching. Journal of Machine Learning Research 2005;2005:695–709.
3. Wu S, Diao E, Elkhali K, Ding J, and Tarokh V. IEEE Access: Score-Based Hypothesis Testing for Unnormalized Models. IEEE Access 2022;2022.
4. El Karoui N, Bean D, Bickel PJ, Lim C, and Yu B. On robust regression with high-dimensional predictors. PNAS 2013;110:14557–62.
5. Donoho D and Montanari A. High Dimensional Robust M-Estimation: Asymptotic Variance via Approximate Message Passing. 2013.
6. Sur P, Chen Y, and Candès EJ. The Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a Rescaled Chi-Square. 2017.
7. Bean D, Bickel PJ, El Karoui N, and Yu B. Optimal M-estimation in high-dimensional regression. PNAS 2013;110:14563–8.