HERA Memo 132
Received 19 Sept 2024

χ^2 with correlated Gaussian random variables

Mike Wilensky

July 2024

1 Introduction

The purpose of this document is to explore the first two moments of reduced χ^2 statistics of Gaussian random variables, with inverse variance and inverse covariance weighting. There are two main results, which are not novel to humanity but I reckon are useful to be looked at in tandem.

1. When correlations are present and inverse-*co*variance weighting is chosen, the expected value is 1 and the variance is $2/N$. Since we derive this with an arbitrary covariance matrix, it also holds for the uncorrelated, inverse-variance weighted case.

2. When correlations are present but inverse variance weighting is chosen, the expected value of the statistic is still 1, but the variance is larger than $2/N$. This means failing to account for correlations and choosing inverse-variance weighting can produce χ^2 values that appear in some instances highly anomalous but in actual fact have relatively low significance when the data is modeled correctly.

2 Definitions

Suppose $\{X_i : i = 1 \dots N\}$ are jointly Gaussian random variables with[1] $\langle X_i \rangle = 0$ and $\langle X_i X_j \rangle = C_{ij}$ (consider these as matrix elements of a matrix, C). I define two statistics: inverse-covariance weighted reduced χ^2,

$$\chi^2_c \equiv \frac{1}{N} \sum_{i,j} X_i X_j (C^{-1})_{ij}, \tag{1}$$

and inverse-variance weighted reduced χ^2,

$$\chi^2_v \equiv \frac{1}{N} \sum_i \frac{X_i^2}{C_{ii}}, \tag{2}$$

[1] This is just a convenient choice essentially reflecting the assumption that you've modeled the means of your variables correctly. Exploration of incorrect mean modeling is left as future work.

3 Result 1

First I show that $\langle \chi_c^2 \rangle = 1$.

$$\langle \chi_c^2 \rangle = \frac{1}{N} \sum_{i,j} (C^{-1})_{ij} \langle X_i X_j \rangle \tag{3}$$

$$= \frac{1}{N} \sum_{i,j} (C^{-1})_{ij} C_{ij} \tag{4}$$

$$= \frac{1}{N} \mathrm{tr}(C^{-1}C) \tag{5}$$

$$= \frac{1}{N} \mathrm{tr}(I) \tag{6}$$

$$= 1. \tag{7}$$

Here $\mathrm{tr}(\cdot)$ denotes the trace, and I is the identity matrix of dimension equal to N. I've used the fact that C is symmetric and

$$\mathrm{tr}(AB) = \sum_{i,j} A_{ij} B_{ij}^T. \tag{8}$$

Now I compute $\mathrm{Var}[\chi_c^2] = \langle (\chi_c^2)^2 \rangle - \langle \chi_c^2 \rangle^2$. First,

$$\langle (\chi_c^2)^2 \rangle = \frac{1}{N^2} \sum_{i,j,k,l} (C^{-1})_{ij} (C^{-1})_{kl} \langle X_i X_j X_k X_l \rangle. \tag{9}$$

Assuming Gaussianity, we can relate the four-point correlation function of a Gaussian to its two-point function by Isserlis' theorem,

$$\langle X_i X_j X_k X_l \rangle = C_{ij} C_{kl} + C_{ik} C_{jl} + C_{il} C_{jk}. \tag{10}$$

If we plug Equation 10 into Equation 9, we get

$$\langle (\chi_c^2)^2 \rangle = \frac{1}{N^2} \left(\mathrm{tr} \left(C^{-1} C \right)^2 + 2 \mathrm{tr} \left(C^{-1} C C^{-1} C \right) \right) \tag{11}$$

$$= \frac{1}{N^2} (N^2 + 2N) \tag{12}$$

$$= 1 + \frac{2}{N} \tag{13}$$

Using what we derived above, namely that $\langle \chi_c^2 \rangle = 1$, we have

$$\mathrm{Var}[\chi_c^2] = \langle (\chi_c^2)^2 \rangle - \langle \chi_c^2 \rangle^2 = \frac{2}{N}. \tag{14}$$

4 Result 2

It is easy to show that $\langle \chi_v^2 \rangle = 1$:

$$\langle \chi_v^2 \rangle = \frac{1}{N} \sum_i \frac{\langle X_i^2 \rangle}{C_{ii}} = \frac{1}{N} \sum_i \frac{C_{ii}}{C_{ii}} = \frac{1}{N} \sum_i 1 = 1. \tag{15}$$

For the variance, again using Isserlis' theorem,

$$\langle(\chi_v^2)^2\rangle = \frac{1}{N^2}\sum_{i,j}\frac{\langle X_i^2 X_j^2\rangle}{C_{ii}C_{jj}} \tag{16}$$

$$= \frac{1}{N^2}\sum_{i,j}\frac{C_{ii}C_{jj}+2(C_{ij})^2}{C_{ii}C_{jj}}. \tag{17}$$

Note that in terms of the correlation coefficient, ρ_{ij},

$$C_{ij} = \rho_{ij}\sqrt{C_{ii}C_{jj}}. \tag{18}$$

This implies

$$\langle(\chi_v^2)^2\rangle = \frac{1}{N^2}\left(N^2 + 2\sum_{i,j}\rho_{ij}^2\right), \tag{19}$$

whence

$$\mathrm{Var}[\chi_v^2] = \frac{2}{N^2}\sum_{i,j}\rho_{ij}^2. \tag{20}$$

Noting that $\rho_{ii}^2 = 1$ by definition, and that $\rho_{ij}^2 \leq 1$ for $i \neq j$ is required for C to be positive semi-definite, we have

$$N \leq \sum_{i,j}\rho_{ij}^2 \leq N^2. \tag{21}$$

The lower bound is for totally independent X_i while the upper bound occurs when all the X_i are perfectly degenerate. This means

$$\frac{2}{N} \leq \mathrm{Var}[\chi_v^2] \leq 2. \tag{22}$$

In particular, for N quite large but with strongly correlated X_i (or anticorrelated, since this is independent of the sign of ρ_{ij}), there can be a large spread in χ_v^2 values despite modeling the means and variances correctly compared to what is expected in the independent case.