



DEPARTMENT OF
STATISTICS

Chi-square Approximations with Stein's Method

Gesine Reinert

Department of Statistics
University of Oxford

Symposium in Memory of Charles Stein, Singapore

1 Chi-Square Statistics

2 Stein's method and $h(x) - \chi_p^2 h = xf''(x) + \frac{1}{2}(p-x)f'(x)$

3 Three proofs of a simple fact, using
 $h(x) - \chi_p^2 h = xf''(x) + \frac{1}{2}(p-x)f'(x)$

Based on joint papers with Andreas Anastasiou, Robert Gaunt and Alastair Pickett

Chi-Square Statistics

Fan, Hung, Wong (1997) "One of the most celebrated folklores in statistics is the theory of maximum likelihood ratio statistics"

Under smoothness conditions, the null distribution of $-2 \log \Lambda$ tends to a chi-square distribution as the sample size tends to infinity

Proof:

asymptotic normality of the MLE
multivariate normal approximation

Problems:

- Bound on the rate of convergence?
- Chi-square approximation may hold even when no underlying normality (*Luk (1994), Fan, Hung, Wong (1997)*)

Stein's method

Stein's Method for standard normal

Stein (1972) $Z \sim \mathcal{N}(0, 1)$ if and only if for all smooth functions f

$$\mathbb{E}f'(Z) = \mathbb{E}Zf(Z)$$

For any smooth function h there is a smooth function $f = f_h$:

$$h(x) - Nh = f'(x) - xf(x) \quad \text{Stein equation}$$

Here Nh is the standard normal expectation of h .

Hence for any random variable W , smooth h , we have

$$\mathbb{E}h(W) - Nh = \mathbb{E}f'(W) - \mathbb{E}Wf(W).$$

Other distributions

Stein's method has been extended to many other distributions:

- Poisson (Chen 1975)
- multivariate normal (Goetze 1992, Barbour 1990)
- compound Poisson (Barbour, Chen and Loh 1992)
- Binomial (Ehm 1991)
- Gamma (Luk 1994)
- Beta (Goldstein and R. 2013)
- variance Gamma (Gaunt 2013)
- Laplace (Pyke and Ren 2015)
- ...

Stein's method in a nutshell

For μ a target distribution, with support \mathcal{I} :

1. Find a suitable operator \mathcal{A} and a wide class of functions $\mathcal{F}(\mathcal{A})$ such that $X \sim \mu$ if and only if for all functions $f \in \mathcal{F}(\mathcal{A})$,

$$\mathbb{E}\mathcal{A}f(X) = 0.$$

2. Let $\mathcal{H}(\mathcal{I})$ be a measure-determining class on \mathcal{I} and $X \sim \mu$. For each $h \in \mathcal{H}$ find a solution $f = f_h \in \mathcal{F}(\mathcal{A})$ of the

$$h(x) - \mathbb{E}h(X) = \mathcal{A}f(x).$$

Then for any random element W ,

$$\mathbb{E}h(W) - \mathbb{E}h(X) = \mathbb{E}\mathcal{A}f(W).$$

Chisquare distributions

Operator for χ_p^2 :

$$Af(x) = xf''(x) + \frac{1}{2}(p-x)f'(x)$$

(Luk 1994: *Gamma*(r, λ)). Stein equation:

$$(\chi_p^2) \quad h(x) - \chi_p^2 h = xf''(x) + \frac{1}{2}(p-x)f'(x)$$

where $\chi_p^2 h$ is the expectation of h under the χ_p^2 -distribution

Thus, for any random variable W , smooth test function h ,

$$Eh(W) - \chi_p^2 h = EWf''(W) + \frac{1}{2}E(p-W)f'(W).$$

Bounds on the solutions of the Stein equation

Stein equation:

$$(\chi_p^2) \quad h(x) - \chi_p^2 h = xf''(x) + \frac{1}{2}(p-x)f'(x)$$

If h is bounded and has three bounded derivatives, then the Stein equation (χ_p^2) has solution f which satisfies

$$\|f^{(k)}\| \leq \frac{4}{p+k-1} (3\|g^{(k-1)}\| + \|g^{(k-2)}\|)$$

for $k = 2, 3, 4$.

Three proofs of a simple fact

Example: squared sum

$X_i, i = 1, \dots, n$ i.i.d. mean zero, variance one, existing 8th moment

$$S = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \text{ and } W = S^2$$

Theorem (R. 2005)

For h smooth,

$$|\mathbb{E}h(W) - \chi_{(1)}^2 h| \leq \frac{\mathbb{E}X^8}{n} (428 + 2375|\mathbb{E}X^3|) \sum_{\ell=0}^3 \|h^{(\ell)}\|.$$

Sketch of the first proof: relate to standard normal

Want

$$2\mathbb{E}Wf''(W) + \mathbb{E}(1-W)f'(W).$$

Put $g(s) = sf'(s^2)$, then

$$g'(s) = f'(s^2) + 2s^2f''(s^2)$$

and

$$2\mathbb{E}Wf''(W) + \mathbb{E}(1-W)f'(W) = \mathbb{E}g'(S) - \mathbb{E}Sg(S).$$

Now proceed as in $\mathcal{N}(0, 1)$:

Put $S_i = \frac{1}{\sqrt{n}} \sum_{j \neq i} X_j$. Then by Taylor expansion,

$$\begin{aligned}
 \mathbb{E} S g(S) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E} X_i g(S) \\
 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E} X_i g(S_i) + \frac{1}{n} \sum_{i=1}^n \mathbb{E} X_i^2 g'(S_i) + R_1 \\
 &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} g'(S_i) + R_1 \\
 &= \mathbb{E} g'(S) + R_1 + R_2.
 \end{aligned}$$

Key for bounding the remainder terms: symmetry

Note that g'' is antisymmetric, $g''(-s) = -g''(s)$, so for $Z \sim \mathcal{N}(0, 1)$ we have

$$\mathbb{E}g''(Z) = 0.$$

(Almost) routine now to show that

$$|\mathbb{E}g''(S)| \leq \frac{c(f)}{\sqrt{n}}$$

for some $c(f)$.

Combining these bounds show a bound on the distance to $\text{Chisquare}(1)$ for smooth test functions which is of order $\frac{1}{n}$.

We can calculate the constant $c(f)$ explicitly.

Application: Pearson's Chi-square Statistic

n balls thrown into m boxes independently; $p_i = \mathbb{P}(X \text{ in box } i)$;
 $U_i =$ number of balls in box i , $i = 1, \dots, m$; then

$$W = \sum_{i=1}^m \frac{(U_i - np_i)^2}{np_i} \approx \chi_{m-1}^2 \text{ as } n \rightarrow \infty.$$

Rule of thumb: $np_i \geq 5$ for at least 80 % of the cells.

Theorem (Pickett 2005; Gaunt, Pickett and R. 2017)

Let (U_1, \dots, U_m) represent the multinomial vector of $n \geq 2$ observed counts, where $m \geq 2$, and suppose that $np_j \geq 1$ for all $j = 1, \dots, m$. Denote the Pearson statistic by W . Let $h \in C_b^5(\mathbb{R}^+)$. Then

$$\begin{aligned} & |\mathbb{E}h(W) - \chi_{(m-1)}^2 h| \\ & \leq \frac{4}{(m+1)n} \left(\sum_{j=1}^m \frac{1}{\sqrt{p_j}} \right)^2 \{19\|h\| + 366\|h'\| + 2016\|h''\| \\ & \quad + 5264\|h^{(3)}\| + 106965\|h^{(4)}\| + 302922\|h^{(5)}\|\}. \end{aligned}$$

Notes

The theorem applies for all n .

The dependence on n is optimal.

The numerical constants may not be optimal.

A key argument in our proof is that $t(s) = s^2$ is symmetric. Hence $g(s) = sf'(s^2)$ is antisymmetric.

More generally we can consider test functions $h(t)$ where t is a symmetric function.

We could have used a multivariate normal approximation with symmetry arguments.

Stein's method for multivariate normal approximation

(Götze's (1993)) Let $t : \mathbb{R}^d \rightarrow \mathbb{R}$ be continuous and let Z have standard d -dimensional normal distribution. Let Σ be non-negative definite. A multivariate normal Stein equation with test function $h(t(\cdot))$ is

$$\nabla^T \Sigma \nabla f(w) - w^T \nabla f(w) = h(t(w)) - \mathbb{E}h(t(\Sigma^{1/2}Z)).$$

We can bound $|\mathbb{E}h(t(W)) - \mathbb{E}h(t(\Sigma^{1/2}Z))|$ by solving the MVN Stein equation for f and then bounding

$$\mathbb{E}[\nabla^T \Sigma \nabla f(W) - W^T \nabla f(W)].$$

Problem: Depending on t the derivatives of f may not be bounded.

Example: squared sum

$X_i, i = 1, \dots, n$ i.i.d. mean zero, variance one, $S = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$
and $W = S^2$ so $W = t(X_1, \dots, X_n)$ with

$$t(x_1, \dots, x_n) = \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2.$$

Then

$$\frac{\partial}{\partial x_j} t(x_1, \dots, x_n) = \frac{2}{n} \sum_{i=1}^n x_i$$

is not bounded. Note $\frac{\partial^2}{\partial x_i \partial x_j} t(x_1, \dots, x_n) = \frac{2}{n}$ and all higher derivatives vanish.

Let $m \geq 1$, $A \geq 0$, $B_1, \dots, B_d \geq 0$, $r_1, \dots, r_d \geq 0$ and $P(w) := A + \sum_{i=1}^d B_i |w_i|^{r_i}$. Take $t : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\forall w \in \mathbb{R}^d$,

$$\left| \frac{\partial^k t(w)}{\prod_{j=1}^k \partial w_{i_j}} \right|^{m/k} \leq P(w), \quad k = 1, \dots, m.$$

We say then that $t \in C_P^m(\mathbb{R}^d)$.

Write $h_m = \sum_{j=1}^m \|h^{(j)}\| \frac{1}{j!} \sum_{i=0}^j (-1)^{j-i} \binom{j}{i} i^m$.

Suppose that $h \in C_b^m(\mathbb{R})$. Let $Z_i \sim N(0, \sigma_{ii})$. Then, for all $w \in \mathbb{R}^d$, the solution f of the MVN Stein equation for $h(t)$ satisfies

$$\left| \frac{\partial^m f(w)}{\prod_{j=1}^m \partial w_{i_j}} \right| \leq \frac{h_m}{m} \left[A + \sum_{i=1}^d 2^{r_i} B_i (|w_i|^{r_i} + \mathbb{E}|Z_i|^{r_i}) \right].$$

Application: Complete block design

This approach generalises to local dependence. Take X_{ij} , $i = 1, \dots, n$, $j = 1, \dots, d$ mean zero random variables; suppose that $X_{1,j}, \dots, X_{n,j}$ are independent for a fixed j , but that the random variables $X_{i,1}, \dots, X_{i,d}$ may be dependent for any fixed i .

For $j = 1, \dots, d$, let $W_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{ij}$ and denote $W = (W_1, \dots, W_d)^T$ with covariance matrix Σ .

Theorem (GR 2016)

Assume that $t \in C_p^6(\mathbb{R}^d)$ is a symmetric function and that $\mathbb{E}|X_{ij}|^{r_k+4} < \infty$ for all i, j and $1 \leq k \leq d$. Then, for $h \in C_b^6(\mathbb{R})$,

$$|\mathbb{E}h(t(W)) - \mathbb{E}h(t(\Sigma^{1/2}Z))| \leq \frac{h_4}{n}c_1(P) + \frac{h_6}{n}c_2(P),$$

where $c_1(P)$ contains mixed absolute moments up to order $\max_k(r_k) + 4$ and $c_2(P)$ contains mixed absolute moments up to order $\max_k(r_k) + 3$ and can be given explicitly.

Example: the power divergence family

Let (U_1, \dots, U_r) represent the multinomial vector of $n \geq 2$ observed counts, where $r \geq 2$, and suppose that $np_j \geq 1$ for all $j = 1, \dots, r$. The power divergence statistic with index $\lambda \in \mathbb{R}$ is given by

$$T_\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^r U_j \left[\left(\frac{U_j}{np_j} \right)^\lambda - 1 \right];$$

$\lambda = 1$ gives the Pearson statistic.

$T_\lambda(W)$ is asymptotically $\chi_{(r-1)}^2$ distributed for all $\lambda \in \mathbb{R}$.

We obtain an explicit $\frac{1}{n}$ -bound for $\lambda \in \{1, 2, \dots\}$ or $\lambda \geq 5$.

The proof uses that $T_\lambda(W)$ is a sum of squares plus a second term, and the second term is small.

Sketch of the third proof: no recourse to normal

Let $Z_{i,j}$, $i = 1, \dots, r, j = 1, \dots, n$ be independent mean zero variance 1 random variables such that $Z_{i,j}$ has the same distribution as $Z_{i,\ell}$, for $i = 1, \dots, r$, and all mixed moments up to order 8 exist.

Let

$$T = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^n \sum_{\ell=1}^n Z_{i,j} Z_{i,\ell} = Z^T Z.$$

Then we can assess the distance of T to χ_r^2 .

Sketch of the argument

First note that $\mathbb{E}[T] = r$. As $Z_{i,j}$ is independent of $Z_{k,\ell}$ for $\ell \neq j, k = 1, \dots, r$, T is a sum of locally dependent summands. For $j, \ell = 1, \dots, n$, set

$$T^j = \frac{1}{n} \sum_{i=1}^r \sum_{s \neq j} \sum_{t \neq j} Z_{i,s} Z_{i,t}.$$

Then T^j is independent of $Z_{a,j}$ for all $a = 1, \dots, r$.

From Stein's method,

$$\mathbb{E}[g(T)] - \mathbb{E}[g(\chi_r)] = \mathbb{E}\left[Tf''(T) + \frac{1}{2}(r-T)f'(T)\right].$$

Taylor expansion gives

$$\begin{aligned} & \mathbb{E}[Tf'(T)] \\ &= \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^n \sum_{\ell=1}^n \mathbb{E}[Z_{i,j}Z_{i,\ell}\{f'(T^j) + (T - T^j)f''(T^j)\}] + R_1 \\ &= r\mathbb{E}[f'(T)] + \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^n \sum_{\ell=1}^n \mathbb{E}[Z_{i,j}Z_{i,\ell}(T - T^j)f''(T^j)] + R_1 + R_2, \end{aligned}$$

with R_1 and R_2 explicit remainder terms.

Now,

$$T - T^j = \frac{1}{n} \sum_{k=1}^r Z_{k,j} \left(Z_{k,j} + 2 \sum_{s \neq j} Z_{k,s} \right)$$

and so

$$\begin{aligned} & \mathbb{E} [Z_{i,j} Z_{i,\ell} (T - T^j) f''(T^j)] \\ &= \frac{1}{n} \sum_{k=1}^r \mathbb{E} [Z_{i,j} Z_{i,\ell} Z_{k,j}^2 f''(T^j)] \\ & \quad + \frac{2}{n} \sum_{k=1}^r \sum_{s \neq j} \mathbb{E} [Z_{i,j} Z_{i,\ell} Z_{k,j} Z_{k,s} f''(T^j)]. \end{aligned}$$

For $\frac{1}{n} \sum_{k=1}^r \mathbb{E} \left[Z_{i,j} Z_{i,\ell} Z_{k,j}^2 f''(T^j) \right],$

$\mathbb{E} \left[Z_{i,j} Z_{i,\ell} Z_{k,j}^2 \right] = 0$ unless $\ell = j$ or $i = k$

and for $\frac{2}{n} \sum_{k=1}^r \sum_{s \neq j} \mathbb{E} \left[Z_{i,j} Z_{i,\ell} Z_{k,j} Z_{k,s} f''(T^j) \right],$

$$\begin{aligned} \mathbb{E} \left[Z_{i,j} Z_{i,\ell} Z_{k,j} Z_{k,s} f''(T^j) \right] &\approx 1(i = k) \mathbb{E} \left[Z_{i,\ell} Z_{i,s} f''(T^j) \right] \\ &\approx 1(i = k) \mathbb{E} \left[Z_{i,\ell} Z_{i,s} f''(T) \right]. \end{aligned}$$

These approximations can be quantified.

A chisquare approximation for quadratic forms

The proof generalises to quadratic forms. Let

$Z_{i,j}, i = 1, \dots, r, j = 1, \dots, n$ be mean zero random variables such that $Z_{i,j}$ is independent of $\{Z_{k,\ell}, k = 1, \dots, r, \ell \neq j\}$ and $Z_{i,j}$ has the same distribution as $Z_{i,\ell}$, for $i = 1, \dots, r$. Assume that

$$\beta(I) = \mathbb{E} \left(\prod_{i \in I} Z_{i,1} \right),$$

for I a multiset of indices in $\{1, \dots, r\}$, exists for $|I| \leq 8$. Let

$$Z_i = \frac{1}{\sqrt{n}} \sum_{j=1}^n Z_{i,j}; \quad i = 1, \dots, r, \quad \text{and} \quad Z = (Z_1, \dots, Z_r)^T.$$

Let $\tau_{i,k} = \text{Cov}(Z_{i,1}, Z_{k,1})$ and assume that the $r \times r$ matrix $\tau = (\tau_{i,k})_{i,k=1,\dots,r}$ is invertible. Let $U = \tau^{-1}$ and let $T = Z^T U Z$.

Theorem

For all functions $g \in C_b^3(\mathbb{R})$,

$$|\mathbb{E}[g(T)] - \mathbb{E}[g(\chi_r)]| \leq \frac{16 \|g\|_3}{r\sqrt{n}} R(r)$$

with R an explicit remainder term.

For fixed r the overall bound is of order $n^{-\frac{1}{2}}$.

If $\mathbb{E}(W^k)$ is of order r^k for $k = 2, 3, 4$, then the overall bound tends to 0 as $n \rightarrow \infty$ if $r = o(n^{\frac{3}{8}})$.

Application: log likelihood ratio statistics

Let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ be independent and identically distributed (i.i.d.) observations from a distribution with probability density function $f(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^T \in \Theta \subset \mathbb{R}^d$. The test problem is

$$H_0 : \theta_{0,j} = 0, \quad j = 1, \dots, r$$

against the general alternative $H_1 : \boldsymbol{\theta} \in \Theta$. Assume that $\dim(\Theta) = d$; then $\Theta_0 = \{\boldsymbol{\theta} \in \Theta : \theta_{0,j} = 0 \text{ for } j = 1, \dots, r\}$ has dimension $d - r$. Writing $\boldsymbol{\theta} = (\boldsymbol{\theta}_{[1:r]}, \boldsymbol{\theta}_{[r+1:d]})^T$ where $\boldsymbol{\theta}_{[1:r]}$ is the vector of the first r components of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{[r+1:d]}$ is the vector of the remaining $d - r$ components of $\boldsymbol{\theta}$, the null hypothesis translates to $H_0 : \boldsymbol{\theta}_{0,[1:r]} = 0$.

Let $L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^n f(\mathbf{x}_i | \boldsymbol{\theta})$ denote the likelihood function which is assumed to be regular, so that the maximum likelihood estimate exists and is unique, and having derivatives of up to third order with respect to $\boldsymbol{\theta}$. Set

$$\hat{\boldsymbol{\theta}}^{\text{res}}(\mathbf{x}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}; \mathbf{x}) = \left(0_{[1:r]}, \hat{\boldsymbol{\theta}}^*_{[r+1:d]}(\mathbf{x}) \right)^T = \left(0_{[1:r]}, \hat{\boldsymbol{\theta}}^*(\mathbf{x}) \right)$$

$$\hat{\boldsymbol{\theta}}_n(\mathbf{x}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x}).$$

The log-likelihood ratio statistic is

$$-2 \log \Lambda = 2 \log \left(\frac{T_1}{T_2} \right) \quad \text{with} \quad T_1 = \frac{L(\hat{\boldsymbol{\theta}}_n(\mathbf{x}); \mathbf{x})}{L(\boldsymbol{\theta}_0; \mathbf{x})} \quad \text{and} \quad T_2 = \frac{L(\hat{\boldsymbol{\theta}}^{\text{res}}(\mathbf{x}); \mathbf{x})}{L(\boldsymbol{\theta}_0; \mathbf{x})}$$

with $\boldsymbol{\theta}_0$ the unknown true parameter.

Notation

The Fisher information matrix for one random vector is denoted by

$$I(\theta_0) = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}.$$

The score function for θ_0 is

$$S(\theta_0) = S(\theta_0, \mathbf{x}) = \nabla \log L(\theta_0; \mathbf{x}) = \sqrt{n} \begin{pmatrix} \xi(\theta_0, \mathbf{x}) \\ \eta(\theta_0, \mathbf{x}) \end{pmatrix}$$

with column vectors $\xi \in \mathbb{R}^r$ and $\eta \in \mathbb{R}^{d-r}$. Under regularity,

$$\mathbb{E}[S(\theta_0)S(\theta_0)^T] = nI(\theta_0).$$

We write

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \log(L(\boldsymbol{\theta}; \mathbf{x})) = \sum_{i=1}^n \log(f(\mathbf{x}_i | \boldsymbol{\theta})) = \sum_{i=1}^n \ell_{\mathbf{x}_i}(\boldsymbol{\theta}).$$

Let $Y_j = \nabla \ell_{\mathbf{X}_j}(\boldsymbol{\theta}) = (Y_{i,j}, j = 1, \dots, d)$ so that

$$Y_{i,j} = \frac{\partial}{\partial \theta_i} \log(f(\mathbf{X}_j | \boldsymbol{\theta})).$$

Key: $Y_j, j = 1, \dots, n$ are i.i.d. vectors of possibly dependent entries. Under regularity, for every fixed i the sum $\sum_{j=1}^n Y_{i,j}$ satisfies a LLN as well as a CLT.

Heuristics for Wilks' Theorem

As $\frac{\partial}{\partial \theta_i} \ell(\hat{\theta}_n(\mathbf{x}); \mathbf{x}) = 0$, for $i = 1, 2, \dots, d$, by Taylor expansion

$$2 \log T_1 \approx n \left(\hat{\theta}_n(\mathbf{x}) - \theta_0 \right)^T I(\theta_0) \left(\hat{\theta}_n(\mathbf{x}) - \theta_0 \right).$$

By Taylor expansion,

$$S(\theta_0) = \sqrt{n} \begin{pmatrix} \xi(\theta_0, \mathbf{x}) \\ \eta(\theta_0, \mathbf{x}) \end{pmatrix} \approx n I(\theta_0) (\hat{\theta}_n(\mathbf{x}) - \theta_0).$$

Re-arranging,

$$2 \log T_1 \approx \begin{pmatrix} \xi \\ \eta \end{pmatrix}^T [I(\theta_0)]^{-1} \begin{pmatrix} \xi \\ \eta \end{pmatrix}.$$

Similarly we obtain an expression for $2 \log T_2$. As $\xi = 0$ under H_0 , only the lower right corner C of $I(\theta_0)$ enters, so that

$$2 \log T_2 \approx \eta^T C^{-1} \eta.$$

For the likelihood ratio statistic, thus under H_0 ,

$$-2 \log \Lambda \approx \begin{pmatrix} \xi \\ \eta \end{pmatrix}^T [I(\theta_0)]^{-1} \begin{pmatrix} \xi \\ \eta \end{pmatrix} - \eta^T C^{-1} \eta = Z^T \tau^{-1} Z$$

with $\tau = A - BC^{-1}B^T$ and

$$Z_{i,j} = \frac{\partial}{\partial \theta_i} \log(f(\mathbf{X}_j | \theta_0)) - \sum_{k=1}^{d-r} (BC^{-1})_{i,k} \frac{\partial}{\partial \theta_{k+d}} \log(f(\mathbf{X}_j | \theta_0)).$$

Now

$$-2 \log \Lambda \approx Z^T \tau^{-1} Z$$

and the chisquare approximation theorem for quadratic forms can be applied.

Theorem

Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be i.i.d. \mathbb{R}^t -valued, $t \in \mathbb{Z}^+$, random vectors with pdf (or pmf) $f(\mathbf{x}_1 | \boldsymbol{\theta})$, for which Θ is an open subset of \mathbb{R}^d . Assume regularity on f and let $h \in \mathbb{C}_b^3(\mathbb{R})$. Then

$$|\mathbb{E}[h(-2 \log \Lambda)] - \mathbb{E}[h(\chi_r^2)]| \leq \frac{16 \|h\|_3}{r \sqrt{n}} R(r) + \frac{1}{\sqrt{n}} \|h'\| S(\boldsymbol{\theta}_0),$$

where $R(r)$ is given in the chisquare approximation for quadratic forms and $S(\boldsymbol{\theta}_0)$ relates to the approximation of the mle by $\boldsymbol{\theta}_0$.

Remarks

The bounds on the remainder terms use ideas from the bounds on distance to normal for maximum likelihood estimators, see *A. Anastasiou (2018), Anastasiou + R. (2017)*, but without using a normal approximation.

For fixed d , the upper bound is $\mathcal{O}(n^{-1/2})$.

The overall order of the bound is of order $d^{23/2}n^{-1/2}$ when both r and d are not fixed. and the chi-square approximation is justified when $d = o(n^{1/23})$.

The proof involves repeated applications of the Cauchy-Schwarz inequality, and hence we do not expect this bound to be tight.

Example: logistic regression

Assume $(\mathbf{X}_i, Y_i), i = 1, \dots, n$, are i.i.d. observations where $\mathbf{X}_i \in \mathbb{R}^d$ and $Y_i \in \{0, 1\}$, related through the model

$$\mathbb{P}_{\boldsymbol{\theta}}(Y_i = 1 | \mathbf{X}_i = \mathbf{x}) = \left(1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}\right)^{-1} \text{ for } \boldsymbol{\theta} \in \mathbb{R}^d.$$

Test $H_0 : \boldsymbol{\theta}_0 = 0$ against the general alternative: *Sur, Chen and Candès (2017)* show that when d grows linearly with n , the chisquare approximation is no longer valid.

Portnoy (1988) showed that the chi-square asymptotic is still valid when $d = o\left(n^{\frac{2}{3}}\right)$. Our criterion $d = o\left(n^{\frac{1}{23}}\right)$ is not as strong, but our bound is explicit and derived in a more general setting.

Related work

For quadratic forms of i.i.d. variables: Götze and Tikhomirov (2005): $O(\frac{1}{n})$ in Kolmogorov distance.

For Pearson's statistic: Mann (1998) uses Götze's (1993) bound on the multidimensional CLT to obtain an explicit explicit $O\left(\frac{1}{\sqrt{n}}\right)$ bound in Kolmogorov distance.

Götze and Ulyanov (2003): a non-explicit $O(\frac{1}{n})$ bound when $r \geq 5$.

For power divergence statistics: For $r \geq 4$, Ulyanov and Zubov (2009): $O(n^{(r-1)/r})$ in Kolmogorov distance and, for $r = 3$, Assylbekov, Ulyanov and Zubov (2011): $O(n^{-3/4+0.065})$ in Kolmogorov distance. They used number theory and differential geometry.

Chi-square approximations via Stein-Malliavin calculus:

Universality for homogeneous sums: *Nourdin, Peccati, R. (2010)*

Malliavin calculus for second Wiener chaos: *Nourdin and Poly (2012)*

Invariance principle for sequences of stochastic integrals in the second Wiener chaos: *Azmoodeh, Peccati and Poly (2014)*

Stein characterisation for centered random variables living in a finite sum of Wiener chaoses: *Arras, Azmoodeh, Poly and Swan (2016)*



Stein 1972: *"I regret that, in order to complete this paper in time for publication, I have been forced to submit it with many defects remaining."*

References

Anastasiou, A., & Reinert, G. (2018). Bounds for the asymptotic distribution of the likelihood ratio. arXiv preprint arXiv:1806.03666.

Gaunt, R. E., Pickett, A. M., & Reinert, G. (2017). Chi-square approximation by Stein's method with application to Pearson's statistic. *The Annals of Applied Probability*, 27(2), 720-756.

Gaunt, R. E., & Reinert, G. The rate of convergence of some asymptotically chi-square distributed statistics by Stein's method. arxiv.org/abs/1603.01889