Approximations to the distribution of a large quadratic form

Tong Chen Thomas Lumley

Department of Statistics
University of Auckland

NZSA, 2018

What is a quadratic form?

- If X is an n dimensional vector with mean vector μ and covariance matrix Σ. A quadratic form in $\mathbb{R}$ is a function of $Q(X)$: $\mathbb{R}^n \to \mathbb{R}$, which could be expressed as:

$$Q(X) = X^T A X$$

  where A is an $n \times n$ symmetric and non-negative definite matrix..

- The question of interest is to estimate $Pr(Q(X) > q)$.

Why is it interesting?

- The distribution of $Q(X)$ is a linear combination of noncentral χ^2_1 variables.
- These quadratic forms are used when a set of asymptotically Normal test statistics are combined using a weight matrix other than the inverse of their covariance matrix.
- Null distribution: a quadratic form in Gaussian variables

$$Q(X) = X^T A X = Y^T \Lambda Y = \sum \lambda_i \chi^2_1$$

where $\lambda_1 \ldots \lambda_n$ are the eigenvalues of ΣA.

Large quadratic forms in genetics studies

- Previous study mainly focused on
  1. small quadratic forms ($n < 10$)
  2. large p-value ($p > 10^{-2}$)
- In genetics studies, we have
  1. large quadratic forms ($n > 1000$)
  2. very small p-value ($p < 10^{-4}$)
- Existing methods may have problems in both accuracy and computational time.

Existing methods

- 'Exact' methods: the approximation error can be arbitrarily small if calculations are done to arbitrary precision.
- Approximations based on matching moments
- A saddlepoint approximation

'Exact' methods

- Davies (1980) approximated the distribution of $Q(X)$ based on the characteristic function of $Q(X)$.
- Farebrother (1984) wrote $Pr(Q(X) > q)$ as an infinite series of central chi-square distribution. This is done by writing the linear combination as a mixture (Robbins and Pitman, 1949).
- Bausch (2013) showed that a linear combination of gamma densities form an algebra under convolution.

Moment based methods

- Liu et al. (2009) (four-moment approximation):
  1. Using $\chi^2_l(\delta)$ to approximate $Q(x)$
  2. l and δ are obtained by equalling the skewnesses and minimizing the difference between kurtosis of $Q(x)$ and $\chi^2_l(\delta)$.

- Satterthwaite approximation:
  1. Using $a\chi^2_v$ to approximate $Q(x)$.

$$a = (\sum_{i=1}^{n} \lambda_i^2)/(\sum_{i=1}^{n} \lambda_i) \qquad v = (\sum_{i=1}^{n} \lambda_i)^2/(\sum_{i=1}^{n} \lambda_i^2).$$

  2. Give correct mean and variance
  3. $\sum \lambda = \text{trace}(\Sigma A)$ and $\sum \lambda^2 = \text{trace}((\Sigma A)^2) = \sum_{i,j} a_{i,j}^2$. Thus they are available in $O(n^2)$ time.

A saddlepoint approximation

- Kuonen (1999) proposed a form of saddlepoint approximation based on a Normal approximation to an exponentially-shifted density.
- To obtain $Pr(X > z)$ for a density $f_X(x)$, create an exponential family $g(x;\theta) = cf(x)\exp(x\theta)$. Then choosing θ so that z is the mean of density $g(x;\theta)$ (Normal approximation to the sum works well near the mean).
- It's known that the first-order term in the error is uniformly bounded.

Accuracy - the distribution of $Q(X)$

Lemma

$Pr(\sum \lambda_i z_i^2 > q) \simeq c e^{-\frac{1}{2\lambda_1} q}$ for large q.

- The density for χ^2_k is

$$f(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})}$$

  When x is big, only exponential matters.

- If X and Y have exponential tails with different tail rates, $X + Y$ has an exponential tail and tail rate is dominated by the one with larger multiplier (Berman et al., 1992).

Accuracy - Exact methods

- Davies's and Farebrother's method would break down when p-value is close to or beyond machine epsilon
- They compute $1 - (1 - Pr(Q > q))$

Method	The upper tail probability for different quantiles				
	50	100	150	300	400
Farebrother	1.037×10^{-04}	1.716×10^{-08}	3.348×10^{-12}	6.661×10^{-16}	6.661×10^{-16}
Davies	1.037×10^{-04}	1.716×10^{-08}	3.348×10^{-12}	0	0

Table 1: Upper tail probability of $Q = 3\chi_1^2 + 2\chi_1^2 + \chi_1^2$

Accuracy - Exact methods

- Farebrother's method is not usable if the smallest eigenvalue is much smaller than the leading terms for large n. Because it would cause a_0 in the algorithm to underflow.

$$a_0 = exp\left(\frac{1}{2}(n \log \lambda_n - \sum_i^n \log \lambda_i)\right)$$

- Bausch's method has rounding error especially in the left tail with double precision and is slow with multiple precision.

Accuracy

- Moment methods approximate the distribution of $Q(X)$ with a single χ^2_k. They are anti-conservative in extreme tails.
- The saddlepoint approximation has the correct exponential tail rate in the extreme right tail.

Time complexity

- Finding all the eigenvalues are slow for large n.
- Computing all λ, the time complexity is $O(n^3)$.
- Computing the leading k λ_i, the time complexity is $O(kn^2)$

A leading eigenvalue approximation

- Lumley et al. (2018) proposed a leading eigenvalue approximation
  1. Only compute the largest k eigenvalues
     1. using Subsampled Random Hadamard Transform (Tropp, 2011) to get a linear transformed H.
     2. use the QR decomposition to the matrix $(H\Sigma A)^T$ to get an orthonormal matrix Q.
     3. the eigenvalue decomposition of $Q\Sigma A Q^T$
  2. Using Satterthwaite approximation to approximate the rest

- The reference distribution is

$$T \sim \left(\sum_{i=1}^{k} \lambda_i \chi_1^2\right) + a\chi_v^2$$

where $\lambda_1, ..., \lambda_k$ are the largest k eigenvalues of ΣA.

A leading eigenvalue approximation

- The time complexity is $O(kn^2)$
  1. Compute the largest k eigenvalues using random matrix theory.
  2. Remainder term is available in $O(n^2)$ time.
- Accuracy: the tail rate is correct.
- Approximate $Q(X)$ with leading terms using:
  1. exact method when p-value ranges from 10^{-13} to 1
  2. the saddlepoint approximation when p-value ranges from 0 to 10^{-13}.

Data generation

- Using the Markov Coalescent simulator (Chen et al., 2009)
  1. fix n to obtain $m \approx n$
  2. discard variants if minor allele frequency $> 5\%$
  3. result in a large sparse matrix

- Generate 5 datasets:

Dataset	Case 1	Case 2	Case 3	Case 4	Case 5
No. of People (n)	1000	2000	3000	4000	5000
No. of Variants	770	1955	2864	3826	4028

Table 2: Data

Empirical comparisons

- Compare a leading eigenvalue approximation with a full eigendecomposition of Davies's method.

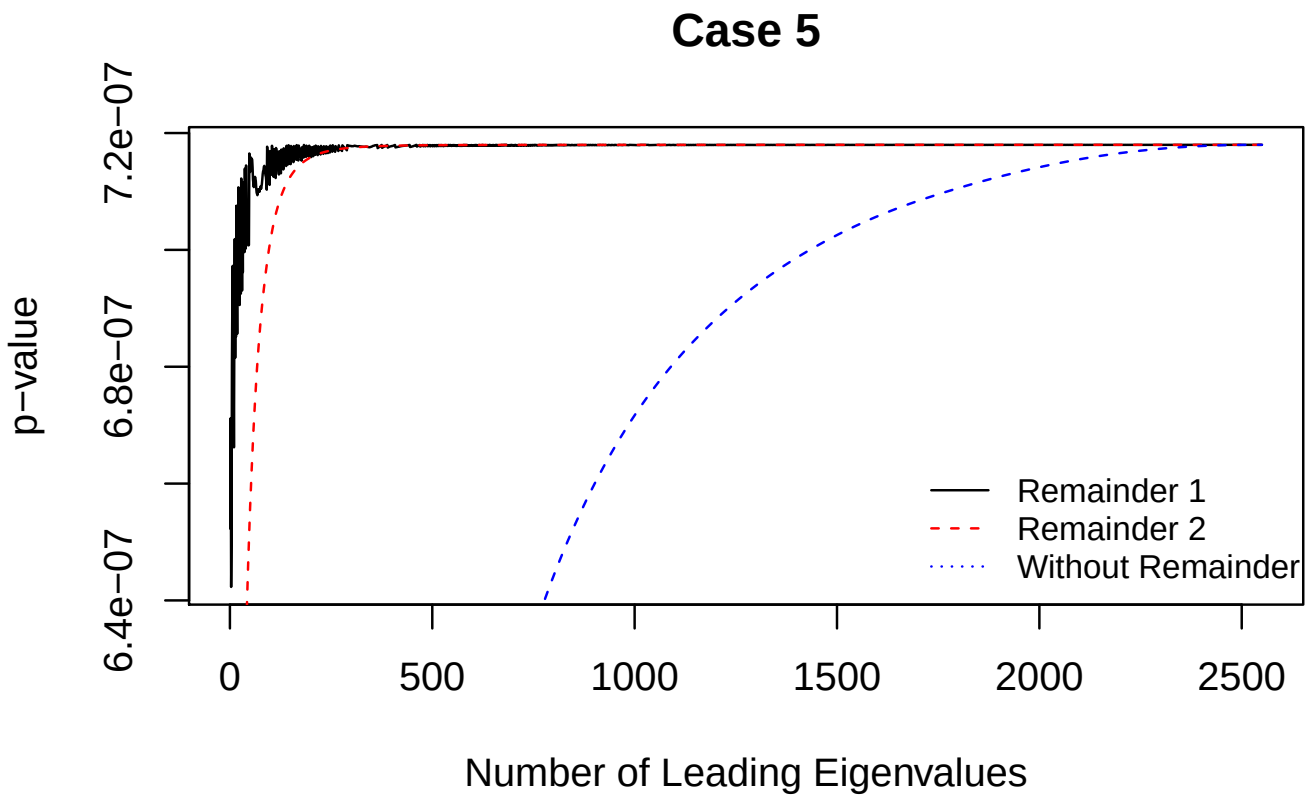
Data	N	Time Difference(s)	Relative Error (%) for P-value Near				
			10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
Case 1	770	0.8	.31	.31	.28	.09	.08
Case 2	1955	17.7	.76	.72	.70	.70	.81
Case 3	2864	65.7	.43	.42	.42	.41	.43
Case 4	3826	155.9	.54	.52	.51	.49	.50
Case 5	4028	231.1	.56	.55	.54	.64	.65

Table 3 Computational time difference and the relative error for p-values evaluated near 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8} and 10^{-9} for different cases. N is the number of variants. Stochastic SVD with 100 leading eigenvalues are used for all cases.

Comparing versions of a leading eigenvalue approximation

- Remainder term 1: approximated by Satterthwaite approximation
- Remainder term 2: aprroximated by matching the mean.
- Version 3: has no remainder term

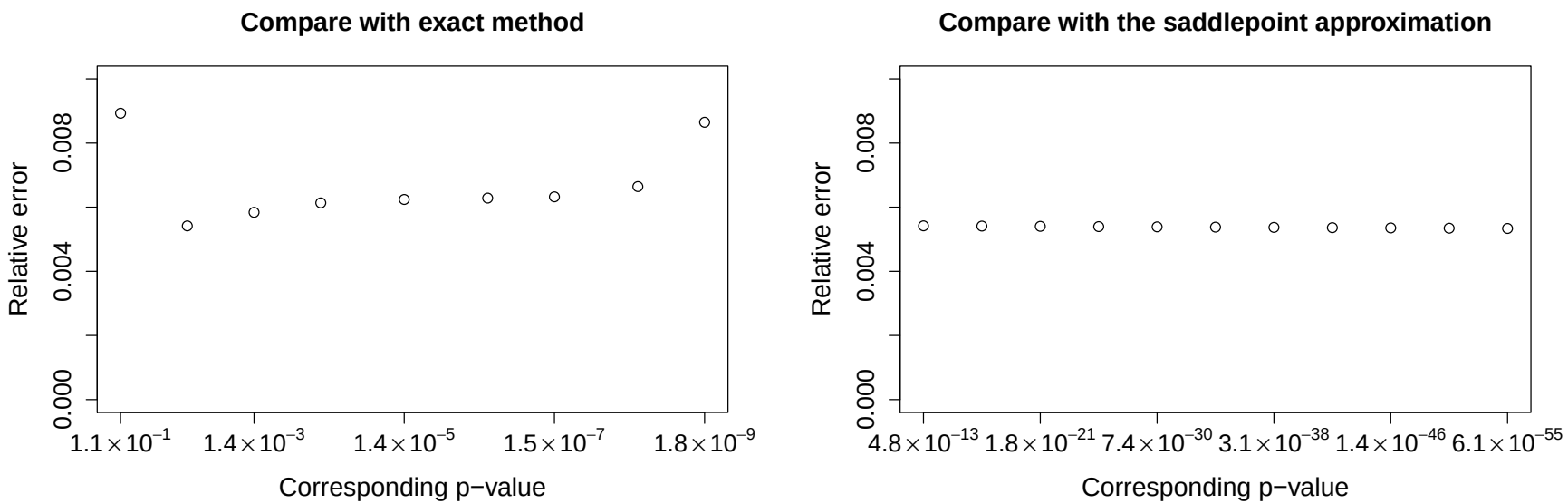
Comparing versions of a leading eigenvalue approximation



- The version with remainder term using the Satterthwaite approximation is not very sensitive to the choice of k.

Extreme tail comparisons

- Comparisons of a leading eigenvalue approximation with Davies's method (left panel) and the saddlepoint approximation (right panel).



Recommendation

- If $n > 1000$ and p-value $> 10^{-13}$, a leading eigenvalue approximation combined with Davies's method is optimal.
- If $n > 1000$ and p-value $< 10^{-13}$, a leading eigenvalue approximation combined with the saddlepoint approximation is optimal.
- If $n < 1000$ and p-value $> 10^{-13}$, a full eigendecomposition of Davies's method is optimal.
- If $n < 1000$ and p-value $< 10^{-13}$, a full eigendecomposition of the saddlepoint approximation is optimal.

Thank You!