Permutation Inference for the General Linear Model and the *G*-statistic

Anderson M. Winkler[1,2,3], Gerard R. Ridgway[1,4], Matthew A. Webster[1], Stephen M. Smith[1], Thomas E. Nichols[1,5]

1. Oxford Centre for Functional MRI of the Brain, University of Oxford, UK. 2. Global Imaging Unit, GlaxoSmithKline, Brentford, UK. 3. Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA. 4. Wellcome Trust Centre for Neuroimaging, UCL Institute of Neurology, London, UK. 5. Department of Statistics & Warwick Manufacturing Group, University of Warwick, Coventry, UK.

FMRIB
Oxford Centre for Functional MRI of the Brain

1 Introduction

Permutation methods provide exact control of false positives and allow the use of non-standard statistics, making only weak assumptions about the data. However, even under these weak assumptions, the standard statistics may behave erratically in some circumstances. For example, while a single permutation test may be valid, the null distribution may vary from voxel to voxel. This variation in null distribution is known as a violation of pivotality; when pivotality does not hold, even though the familywise error rate may be controlled overall, the risk of false positives may be greater in some places in the image, and smaller in others. Here we introduce to neuroimaging a new statistic, the *G*-statistic, a generalisation of the *F*-statistic that is robust to heteroscedasticity, that can be used in various useful cases, and preserves pivotality.

2 The *G*-statistic

Consider the common analysis of a neuroimaging experiment. At each voxel, vertex, face or edge (or any other imaging unit), we have a linear model expressed as:

$$\mathbf{Y} = \mathbf{M}\psi + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ contains the experimental data, $\mathbf{M}$ contains the regressors, $\boldsymbol{\psi}$ the regression coefficients, which are to be estimated, and $\boldsymbol{\epsilon}$ the residuals. For a linear null hypothesis that $\mathbf{C}'\boldsymbol{\psi} = \mathbf{0}$, where $\mathbf{C}$ is a contrast vector, if rank($\mathbf{C}$) = 1, the Student's *t* statistic can be calculated as:

$$t = \hat{\boldsymbol{\psi}}'\mathbf{C}\left(\mathbf{C}'(\mathbf{M}'\mathbf{M})^{-1}\mathbf{C}\right)^{-\frac{1}{2}} \Bigg/ \sqrt{\frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{N - \mathsf{rank}\,(\mathbf{M})}}$$

If rank($\mathbf{C}$) ≥ 1, the *F* statistic can be calculated as:

$$F = \frac{\hat{\boldsymbol{\psi}}'\mathbf{C}\left(\mathbf{C}'(\mathbf{M}'\mathbf{M})^{-1}\mathbf{C}\right)^{-1}\mathbf{C}'\hat{\boldsymbol{\psi}}}{\mathsf{rank}\,(\mathbf{C})} \Bigg/ \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{N - \mathsf{rank}\,(\mathbf{M})}$$

If, however, the variances for the observations differ, i.e., in the presence of heteroskedasticity, we demonstrate in the results that *t* and *F* are no longer pivotal and not guaranteed to be the same throughout the image, therefore creating difficulties when controlling the familywise error-rate (FWER). This is the case even if the FWER p-values are assessed via permutation tests. To solve this issue, we propose the use of a robust statistic that is invariant with respect to heteroskedasticity, *G*, which can be computed as:

$$G = \frac{\hat{\boldsymbol{\psi}}'\mathbf{C}\left(\mathbf{C}'(\mathbf{M}'\mathbf{W}\mathbf{M})^{-1}\mathbf{C}\right)^{-1}\mathbf{C}'\hat{\boldsymbol{\psi}}}{\Lambda \cdot s}$$

where $\mathbf{W}$ is a diagonal matrix that has elements:

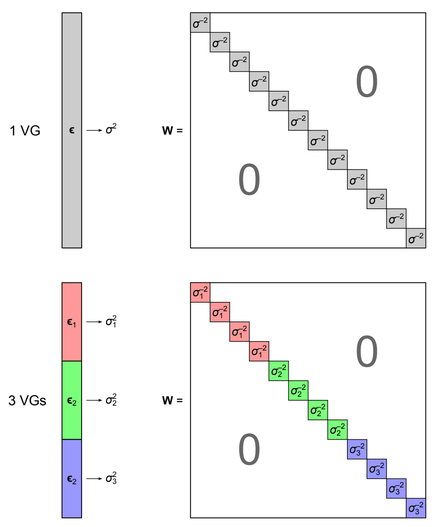
$$W_{nn} = \frac{\sum_{n' \in g_n} R_{n'n'}}{\hat{\boldsymbol{\epsilon}}'_{g_n}\hat{\boldsymbol{\epsilon}}_{g_n}}$$

and where $R_{n'n'}$ are the n' diagonal elements of the residual forming matrix, and g_n is the variance group to which the n-th observation belongs. The remaining denominator term, Λ, is given by:

$$\Lambda = 1 + \frac{2(s-1)}{s(s+2)}\sum_g \frac{1}{\sum_{n\in g} R_{nn}}\left(1 - \frac{\sum_{n\in g} W_{nn}}{\mathsf{trace}(\mathbf{W})}\right)^2$$

where s = rank($\mathbf{C}$). The matrix $\mathbf{W}$ can be seen as a weighting matrix, the square root of which normalises the model such that the errors have then unit variance and can be ignored. It can also be seen as being itself a variance estimator.

Figure 1: The $\mathbf{W}$ matrix is constructed from the estimated variances of the error terms.



The *G*-statistic can be used with the general linear model (GLM), and only requires the definition of the variance groups, i.e., the sets of observations that share the same variance. The exact value of the variance is not assumed to be known. The *G*-statistic is a generalisation of various well known statistics, including the *F*-statistic itself.

Table 1: *G* is a generalisation of other well-known statistics.

	rank ($\mathbf{C}$) = 1	rank ($\mathbf{C}$) > 1
Homoscedastic errors, unrestricted exchangeability	Square of Student's t	F-ratio
Homoscedastic within VG, restricted exchangeability	Square of Aspin–Welch v	Welch's v^2

3 Permutation strategies

When some of the covariates are nuisance effects, more than one permutation strategy is possible. Various different methods have been proposed, in which the design matrix $\mathbf{M}$ is partitioned into effects of interest, $\mathbf{X}$, and nuisance effects, $\mathbf{Z}$, and parts of the design are shuffled differently (**Table 2**).

Table 2: A number of methods are available to obtain parameter estimates and construct a reference distribution in the presence of nuisance variables.

Method	Model
Draper–Stoneman	$\mathbf{Y} = \mathbf{P}\mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$
Still–White	$\mathbf{P}\mathbf{R}_\mathbf{Z}\mathbf{Y} = \mathbf{X}\beta + \epsilon$
Freedman–Lane	$(\mathbf{P}\mathbf{R}_\mathbf{Z} + \mathbf{H}_\mathbf{Z})\,\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$
Manly	$\mathbf{P}\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$
ter Braak	$(\mathbf{P}\mathbf{R}_\mathbf{M} + \mathbf{H}_\mathbf{M})\,\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$
Kennedy	$\mathbf{P}\mathbf{R}_\mathbf{Z}\mathbf{Y} = \mathbf{R}_\mathbf{Z}\mathbf{X}\beta + \epsilon$
Huh–Jhun	$\mathbf{P}\mathbf{Q}'\mathbf{R}_\mathbf{Z}\mathbf{Y} = \mathbf{Q}'\mathbf{R}_\mathbf{Z}\mathbf{X}\beta + \epsilon$
Smith	$\mathbf{Y} = \mathbf{P}\mathbf{R}_\mathbf{Z}\mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$
Parametric	$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon,\ \ \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$

Table 3: The eight different simulation scenarios, each with its own same sample sizes and different variances, along with the proportion of error type I and power (%) for the statistics F and G. Confidence intervals (95%) are shown in parenthesis. The letters in the last column (marked with a star, ⋆) indicate the variance configurations represented in the pairwise comparisons shown in **Figure 2**.

Simulation scenario	Sample sizes for each VG	Variances for each VG	⋆	Proportion of error type I		Power	
				F	*G*	*F*	*G*
1	8, 4	5, 1	(*a*)	5.9 (4.6-7.5)	6.1 (4.8-7.8)	20.1 (17.7-22.7)	23.8 (21.3-26.5)
		1.2, 1	(*b*)	4.9 (3.7-6.4)	5.3 (4.1-6.9)	28.3 (25.6-31.2)	31.9 (29.1-34.9)
		1, 1	(*c*)	4.7 (3.6-6.2)	4.5 (3.4-6.0)	29.3 (26.6-32.2)	32.6 (29.8-35.6)
		1, 1.2	(*d*)	4.9 (3.7-6.4)	4.6 (3.5-6.1)	29.9 (27.1-32.8)	32.0 (29.2-35.0)
		1, 5	(*e*)	3.9 (2.9-5.3)	4.1 (3.0-5.5)	14.0 (12.0-16.3)	14.1 (12.1-16.4)
2	20, 5	5, 1	(*a*)	6.7 (5.3-8.4)	6.6 (5.2-8.3)	29.1 (26.4-32.0)	38.3 (35.3-41.4)
		1.2, 1	(*b*)	5.0 (3.8-6.5)	4.6 (3.5-6.1)	42.4 (39.4-45.5)	48.8 (45.7-51.9)
		1, 1	(*c*)	5.0 (3.8-6.5)	5.8 (4.5-7.4)	44.6 (41.6-47.7)	48.9 (45.8-52.0)
		1, 1.2	(*d*)	6.1 (4.8-7.8)	6.2 (4.9-7.9)	42.3 (39.3-45.4)	46.7 (43.6-49.8)
		1, 5	(*e*)	5.9 (4.6-7.5)	6.2 (4.9-7.9)	19.5 (17.2-22.1)	19.0 (16.7-21.6)
3	80, 30	5, 1	(*a*)	5.2 (4.0-6.8)	5.0 (3.8-6.5)	90.4 (88.4-92.1)	92.3 (90.5-93.8)
		1.2, 1	(*b*)	4.9 (3.7-6.4)	5.1 (3.9-6.6)	99.7 (99.1-99.9)	99.8 (99.3-100)
		1, 1	(*c*)	6.3 (5.0-8.0)	6.2 (4.9-7.9)	99.8 (99.3-100)	99.8 (99.3-100)
		1, 1.2	(*d*)	4.4 (3.3-5.9)	4.4 (3.3-5.9)	99.6 (99.0-99.8)	99.6 (99.0-99.8)
		1, 5	(*e*)	4.4 (3.3-5.9)	4.4 (3.3-5.9)	72.9 (70.1-75.6)	72.9 (70.1-75.6)
4	40, 30, 20, 10	15, 10, 5, 1	(*a*)	6.4 (5.0-8.1)	5.7 (4.4-7.3)	10.2 (8.5-12.2)	19.4 (17.1-22.0)
		3.6, 2.4, 1.2, 1	(*b*)	5.3 (4.1-6.9)	5.6 (4.3-7.2)	37.8 (34.9-40.9)	45.6 (42.5-48.7)
		1, 1, 1, 1	(*c*)	5.7 (4.4-7.3)	4.9 (3.7-6.4)	72.2 (69.3-74.9)	74.9 (72.1-77.5)
		1, 1.2, 2.4, 3.6	(*d*)	3.1 (2.2-4.4)	3.7 (2.7-5.1)	34.6 (31.7-37.6)	44.6 (41.6-47.7)
		1, 5, 10, 15	(*e*)	4.5 (3.4-6.0)	4.2 (3.1-5.6)	9.7 (8.0-11.7)	15.7 (13.6-18.1)
5	4, 4	1, 1	(*a*)	4.3 (3.2-5.7)	4.3 (3.2-5.7)	29.9 (27.1-32.8)	29.9 (27.1-32.8)
		1, 1.2	(*b*)	4.3 (3.2-5.7)	4.3 (3.2-5.7)	30.6 (27.8-33.5)	30.6 (27.8-33.5)
		1, 5	(*c*)	6.9 (5.5-8.6)	6.9 (5.5-8.6)	14.5 (12.5-16.8)	14.5 (12.5-16.8)
6	20, 20	1, 1	(*a*)	3.3 (2.4-4.6)	3.3 (2.4-4.6)	92.6 (90.8-94.1)	92.6 (90.8-94.1)
		1, 1.2	(*b*)	4.4 (3.3-5.9)	4.4 (3.3-5.9)	90.5 (88.5-92.2)	90.5 (88.5-92.2)
		1, 5	(*c*)	4.4 (3.3-5.9)	4.4 (3.3-5.9)	53.7 (50.6-56.8)	53.7 (50.6-56.8)
7	4, 4, 4, 4	1, 1, 1, 1	(*a*)	5.6 (4.3-7.2)	5.5 (4.3-7.1)	11.0 (9.2-13.1)	8.8 (7.2-10.7)
		1, 1.2, 2.4, 3.6	(*b*)	5.2 (4.0-6.8)	4.4 (3.3-5.9)	6.5 (5.1-8.2)	7.8 (6.3-9.6)
		1, 5, 10, 15	(*c*)	5.7 (4.4-7.3)	4.8 (3.6-6.3)	5.8 (4.5-7.4)	6.9 (5.5-8.6)
8	20, 20, 20, 20	1, 1, 1, 1	(*a*)	4.6 (3.5-6.1)	4.5 (3.4-6.0)	78.7 (76.1-81.1)	78.1 (75.4-80.6)
		1, 1.2, 2.4, 3.6	(*b*)	4.6 (3.5-6.1)	5.6 (4.3-7.2)	40.7 (37.7-43.8)	45.5 (42.4-48.6)
		1, 5, 10, 15	(*c*)	4.7 (3.6-6.2)	4.8 (3.6-6.3)	11.6 (9.8-13.7)	19.3 (17.0-21.9)

Regarding the different permutation strategies, we found that the Freedman-Lane and Smith methods produced the best control over error rates and power (**Table 4**).

Table 4: A summary of the results for the 1536 simulations with different parameters. The amount of error type I is calculated for the 768 simulations without signal. Confidence intervals (CI) at 95% were computed around the nominal level α = 0.05, and the observed amount of errors for each regression scenario and for each method was compared with this interval. Methods that mostly remain within the CI are the most appropriate. Methods that frequently produce results below the interval are conservative;those above are invalid. Power was calculated for the remaining 768 simulations, which contained signal.

Method	Proportion of error type I			Average power
	Within CI	Below CI	Above CI	
Draper–Stoneman	86.33%	8.20%	5.47%	72.96%
Still–White	67.84%	14.58%	17.58%	71.82%
Freedman–Lane	88.67%	8.46%	2.86%	73.09%
ter Braak	83.59%	11.07%	5.34%	73.38%
Kennedy	77.60%	1.04%	21.35%	74.81%
Manly	73.31%	15.89%	10.81%	73.38%
Smith	89.32%	7.81%	2.86%	72.90%
Huh–Jhun	85.81%	9.24%	4.95%	71.62%
Parametric	77.47%	14.84%	7.68%	72.73%

4 Evaluation method

To assess the impact of pivotality, eight scenarios of heteroscedasticity and imbalance were simulated (**Table 3**). The resulting distributions of the *G* and *F* statistic were compared using the two-sample Kolmogorov-Smirnov test. By comparing the distributions of the same statistic obtained in different variance settings, this evaluation strategy mimics what is observed when the variances for each voxel varies across space in the same imaging experiment. Typically, the variances are not assumed to the same. The same eight scenarios were used to evaluate control over error rates and power.

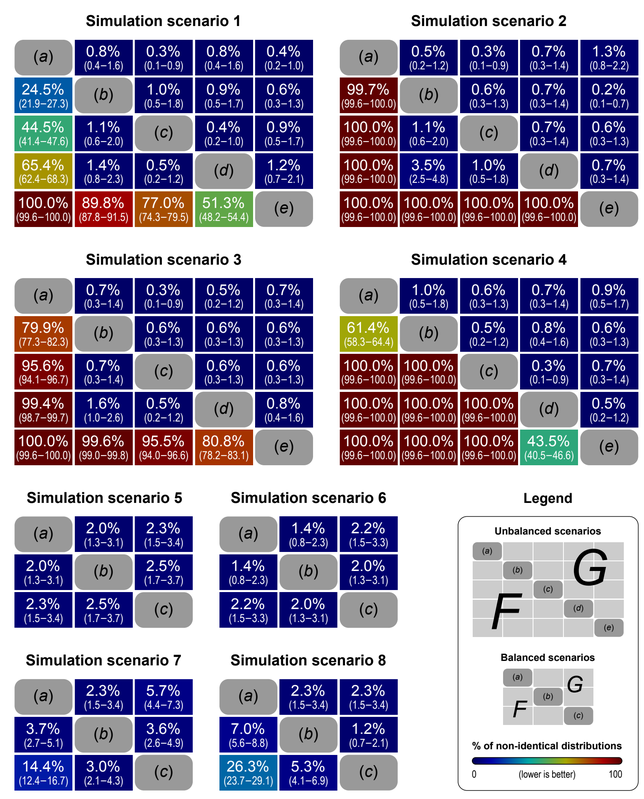
To evaluate the permutation strategies, we simulated various experimental designs with continuous and discrete regressors, different degrees of non-orthogonality, different error distributions, and different sample sizes. This diverse set parameters allowed a total of other 1536 different simulation scenarios, which were used to assess the methods in terms of their error rates and power.

5 Results

The distribution of the *G*-statistic was robust to heteroscedasticity, even with large variance differences across groups, in contrast to the distribution of the commonly used *F* statistic, which varied across tests in the presence of heteroscedasticity (**Figure 2**).

In addition, the *G*-statistic not only controlled the error rate, but it was also generally just as, or even more powerful than *F* (**Table 3**).

Figure 2: Heatmaps for the comparison of the distributions obtained under different variance settings for identical sample sizes. In each map, the cells below the main diagonal contain the results for the pairwise *F* statistic, and above, for the *G* statistic. The percentages refer to the fraction of the 1000 tests in which the distribution of the statistic for one variance setting was found different than for another in the same simulation scenario. Each variance setting is indicated by letters (a–e), corresponding to the same letters in **Table 3**. Smaller percentages indicate robustness of the statistic to heteroscedasticity. Confidence intervals (95%) are shown in parenthesis.



6 Conclusion

The distribution of the *G*-statistic was robust to heteroscedasticity, even with large variance differences across groups, in contrast to the distribution of the commonly used *F* statistic, which varied across tests in the presence of heteroscedasticity. Moreover, the *G*-statistic was generally more powerful than *F*. Regarding the different permutation strategies, we found that the Freedman-Lane and Smith methods produced the best control over error rates and power.

7 Publication in *NeuroImage*

All the results presented in this poster, as well as more (theory, restricted exchangeability, implementation of the **randomise** algorithm, and examples), have just been published in *NeuroImage* (2014;92:381-97).

http://bit.ly/1jKIBlq

8 References

[1] Welch B. **On the comparison of several mean values: an alternative approach.** *Biometrika.* 1951;38(3):330-336. [2] Horn S, Horn R, Duncan D. **Estimating heteroscedastic variances in linear models.** *J Am Stat Assoc.* 1975;70(350):380-385. [3] Freedman D, Lane D. **A nonstochastic interpretation of reported significance levels.** *J Bus Econ Stat.* 1983;1(4):292-8. [4] Nichols TE, Holmes AP. **Nonparametric permutation tests for functional neuroimaging: a primer with examples.** *Hum Brain Mapp.* 2002;15 (1):1-25. [5] Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. **Permutation inference for the general linear model.** *NeuroImage.* 2014;92:381-97.

UNIVERSITY OF OXFORD · THE UNIVERSITY OF WARWICK · UCL · Yale · gsk GlaxoSmithKline · MARIE CURIE ACTIONS