A Bayesian take on the family-wise error rate

Justin Chumbley[1,2] & Klaas Stephan[2]

Affiliation

Author Note

Correspondence concerning this article should be addressed to Justin Chumbley, Jacobs Center for Productive Youth Development, Andreasstrasse 15, CH-8050 Zurich. E-mail: justin.chumbley@pm.me

Abstract

Newcomers to neuroimaging quickly face a choice between frequentist and Bayesian statistical parametric mapping. Having been told that frequentist methods are only as good as their control over the all-important spatial multiple comparisons problem, they typically worry if and how Bayesians solve this central problem too. The answer may appear frustratingly indirect, being grounded in a different philosophy and expressed in different terms: shrinkage, hierarchical regularization, adaptive smoothing, etc. Our short tutorial soothes this concern by explaining in simple terms how Bayesian methods understand and control a quintissentially frequentist notion of multiple comparisons, the gold-standard family-wise error rate (FWER). To do this we define a Bayesian counterpart to this family-wise error probability, and show that it is small in the absence of real experimental effects in the brain. We explain how Bayesian tools achieve control by different means: adapting parameter estimation rather than the test threshold. The work here aims to complement other work connecting frequentist and Bayesian notions of FDR control.

A Bayesian take on the family-wise error rate

In the context of statistical tests, frequentist error probabilities reflect a simple thought experiment: if I were to repeat this test an infinite number of times under the same conditions what fraction of them would give the wrong answer. This classical test theory emphasizes that false positive error probability increases with the number of tests. In it's simplest form, this so-called "multiple testing problem" can be explained by analogy. Let a 20-sided black die with just one red face represent a classical test: there is a 0.05 percent of (red) error. If you throw a bucket of dice, the chance of one or more red faces - the so-called family-wise error rate - is much higher: it soon approaches 1. We then say a type-I error has occured in the "family" of die. For a more realistic example take a simple linear model, say ANOVA with a single, univariate outcome and one experimental factor taking n levels: it is then possible to do n tests of the levels themselves, $\binom{n}{2}$ pairwise comparisons etc. There are clearly many die.

This problem is so basic to statistics that many solutions have been proposed. One might bypass multiple comparisons by conducting a single "omnibus" test, a standard F-test in this example. Alternatively, one might choose to increase the threshold of the multiple constituent tests, using say Bonferonni or Benjamini-Hochberg adjustments, making positive test results less common. This effectively increases the number of sides per dice while keeping just one bad (red) face. An alternative solution is to reduce error in the parameter estimation itself, rather than in the test. This traditionally comes under the rubric of "bias-variance tradeoff", because error reduction entails reducing random variability by "biasing" parameter estimation somehow - typically shrinking parameters toward one another or toward zero. This corresponds to using some magic or other to ensure that the dice do not fall independently, but are coupled. In the extreme that this "bias" is very strong - you insist that all die land on the same side - you are really just performing a single test and the multiplicity problem is gone.

Another interesting solution, which seems to have emerged from neuroimaging, is based on the frequency distribution of the *largest test statistic* across all parameters. The gold standard frequentist correction for multiple, spatially-correlated tests in neuroimaging is based on this frequency distribution of the maximal test statistic over all voxels in the brain image (Friston et al., 1994)[1]. This maxima fully characterizes the FWER, here the chance of declaring 1 or more false positive voxels in the whole image, because it exceeds test threshold $h > 0$ exactly when 1 or more voxels exceed threshold. By choosing h, the test threshold common to all voxels, high enough to ensure that the maximal voxel only exceeds h with say 0.05 probability, we control FWER to 0.05. Intuitively, if I repeated this experiment an infinite number of times under the same conditions how much would my global maxima - and therefore my "family-wise error" - vary? As we discuss next, we believe that this focus on stochastic maxima offers one bridge for Bayesians to understand frequentist FWER on their own terms.

A typical Bayesian interprets probability very differently: it quantifies their belief about effects in the single observed data-set. The *credibility*, not the *frequency*, of a meaningful positive effect at any given voxel is quantified by the posterior mass exceeding fixed threshold $h > 0$. The crucial thing we emphasize in this tutorial is this: posterior belief about multiple parameters is inherently multivariate but we can derive any scalar consequence of this belief by simple application of the laws of probability. In particular, we can derive our posterior belief about the *largest* effect in the brain[2]. Ironically, this is best understood by a sampling analogy. Say we draw 1000 samples from our multivariate posterior probability map. Each realized sample is then just a spatial array or map of numbers, and we can record it's maximal value: the height of it's highest peak. Repeat this procedure for the 1000 samples and build a 1d histogram of these values. This histogram

[1] Assuming zero experimental effect anywhere in the brain, this maxima distribution can be approximated using ideas from random field theory.

[2] or in a parameter vector, matrix, etc.

approximates our uncertainty about the largest effect in the map: the credible size of the peak effect in the brain. Furthermore, our posterior subjective belief that there is any non-neglible, positive effect in the brain is exactly the fraction of this histogram exceeding h. Note that, *assuming there are actually zero effects in the data generating process - i.e. assuming an omnibus null hypothesis of no effect anywhere in the brain - this latter is interpretable as a Bayesian counterpart to the frequentist FWER*. It amounts to false belief in some positive effect, when all effects actually equal zero by definition. Having translated FWER into Bayesian terms, the next question is whether and how Bayesian inference controls this "error".

Here we focus on the hierarchical Bayesian approach, which partly evolved from the idea of a "bias-variance" tradeoff introduced above. Briefly, the approach is to jointly infer our multiple parameters and their relationship to one another. If the latter suggests they share a similar value, posterior belief gravitates or "shrinks" towards that common value. The many technicalities are readily available in any Bayesian textbook or paper. It suffices for us that, *assuming all effects in the data generating process are zero* this hierarchical bias will shrink all component parameters towards their common value of zero. Our derived belief about the largest effect will follow, and the fraction of our histogram (see previous paragraph) exceeding h will diminish. Thus a hierarchical Bayesian approach will indeed control (shrink) our FWER counterpart. The numerical details of this control depend on our specific model of which and how parameters relate to one another: our Bayesian spatial model of the brain responses over space, be it inspired by a discretely indexed, exchangeable random process, a Markov random field or a continuously-indexed, non-stationary process. For illustration here we consider the simplest possible case.

A simple intuitive simulation

We start by examining a useful hierarchical model with the simplest assumptions. In neuroimaging this example falls under the rubric of "posterior probability mapping with global shrinkage priors", where i indexes voxels of an image.

Assume that we have spatio-temporal measurements on a regular lattice in response to an occasional auditory stimulus - say a "beep" - which occurs at random time points. We would like to contrast brain responses to this stimulus against baseline BOLD activity across the entire brain (or some volume of interest). This is conventionally achieved using linear regression of observed BOLD on the predicted BOLD, the latter derived by convolving the stimulus impulse function with the BOLD response function. Omiting unnecessary details, our model is $V_{ij} = \beta_i + x_j \cdot U_i + \epsilon_{ij}$ where V_{ij} is the observed, normalized smoothed BOLD observation at time j and voxel i, x_j is the convolved BOLD prediction, ϵ_{ij} is measurement error, U_i is the effect of auditory stimulus on voxel i and β_i is a voxel-specific intercept (these latter two are parameters and often elsewhere denoted θ_0, θ_1, etc). We will sometimes omit the time subscript j in order to concisely denote the whole time-series at voxel i with $V_i \equiv \{V_{ij} : j \in J\}$. To illustrate the statistical concepts with minimal mathematical distraction we simplify our example still further to $V_{ij} = U_i + \epsilon_{ij}$, where U_i still encodes the voxel-specific effect to our auditory stimulus. This emphasizes there is really only one parameter of interest in our example, whose value varies over voxels.

Let I denote the index set over discrete space (the voxel indices) and J the index set of all discrete time points. The hierarchical approach starts by specifying a model of the data-generating process. Our example therefore starts by assuming that our data $V \equiv \{V_{ij} : i \in I, j \in J\}$ was generated as follows. The hidden signal U_i are independent mean-zero Gaussian samples, drawn from a common distribution given unknown

parameters ψ_U. Then independent mean-zero Gaussian noise replicates ϵ_{ij} are added, giving the observed $V_{ij} = U_i + \epsilon_{ij}$. A parameter ψ_V specifies the scale of this observation noise and is shared over all the ϵ_{ij}. By assuming this independent, spatially homogenous error model we intentionally brush aside details of the more realistic scenario, i.e. where observation noise is spatially and temporally correlated. We write the model

$$U_i|\psi_U \overset{iid}{\sim} N(U_i|0, \psi_U)$$
$$V_{ij}|U_i, \psi_V \overset{iid}{\sim} N(V_{ij}|U_i, \psi_V).$$

We define parameters $\psi = (\psi_U, \psi_V)$, which control the randomness of hidden signal U_i and observation noise ϵ_{ij} respectively, as Gaussian precision or inverse-variance parameters. Inspired by the frequentist tradition, we provisionally define a *family null model* as the case where our observed data V_{ij} was generated from the above model with $\psi_U \to \infty$ for any $\psi_V < \infty$, a because it implies $U_i \approx 0$ for all i. The latter is just the family null hypothesis familiar from the conventional, non-hierarchical hypothesis testing framework. We denote this null data, and inferences based on it, with an asterix, for example V^*.

If we knew the values of ψ_U, ψ_V, then it can easily be shown by Bayes theorem that uncertain posterior inference $U_i|V, \psi_U, \psi_V$, is also Gaussian, with precision parameter equal to $\psi_U + n\psi_V$ and mean parameter equal to $\frac{n\psi_V \hat{u}_i}{\psi_U + n\psi_V}$, where $\hat{u}_i = \frac{1}{J}\sum_{j=1}^{J} v_{ij}$ is the observed sample mean. These equations show that a large ψ_U causes the posterior mean inference to shrink towards zero and away from the observed sample mean $\hat{u}_i$. Posterior uncertainty given by variance $1/\psi_U$ also shrinks to zero, effectively trapping all posterior mass close to zero. When the family null is true, even the simplest hierachical model will naturally cause such shrinkage towards zero. Take an empirical Bayes method which first estimates $\hat{\psi}_U$ and $\hat{\psi}_V$ from data, before using these estimates to infer U_i via the above equations for the posterior mean and precision. Then if data V are generated under the family null hypothesis, such a method will clearly yeild a large estimate for $\hat{\psi}_U$ and large shrinkage will

follow. Crucially, this shrinkage acts on all parameters and therefore shrinks our belief about the largest effect in the brain: the credible size of the peak effect in the brain shrinks. We have offered the simplest illustration of the more general idea in the introduction: hierarchical Bayesian models offer some control on the (Bayesian) family-wise error by adjusting the estimation process rather than any test threshold.

It is by this mechanism, not the law of large numbers, that hierarchical methods suppress noise from infered signal under the family null hypothesis. In what follows we use the shorter notation $U_i|V^*$ to denote $U_i|V^*, \hat{\psi_U}, \hat{\psi_V}$. This compact notation is ultimately more suitable for describing full Bayesian hierarchical inference.

It is important for understanding this paper to note that because all posterior components $U_i|V^*$ converge in probability to zero with bigger infered $\hat{\psi_U}$ under family null data, so does the posterior maximum, denoted $U_{max}|V^*$. We emphasize that this $U_{max}|V^*$ refers to the distribution of the posterior maxima and *not* the familiar maxima of the posterior distribution, i.e. *not* to the maximum *a posteriori* or MAP estimate.

Results

Discussion

Statistical errors accumulate whenever one estimates or tests multiple unknowns from data. This problem plagues science and engineering, but is particularly acute in imaging biology and omics, where there may be millions of unknowns. A conventional solution is to adjust test thresholds to limit aggregate error over tests, often defined as the "family-wise error rate" (FWER). Modern hierarchical Bayesian solutions instead adapt hyper-parameters in order to shrink error over estimators, not tests. Despite the enduring popularity of both approaches, it remains unclear how exactly they relate to one another, i.e. does shrinking estimation error automatically limit test error? Here we illustrate that

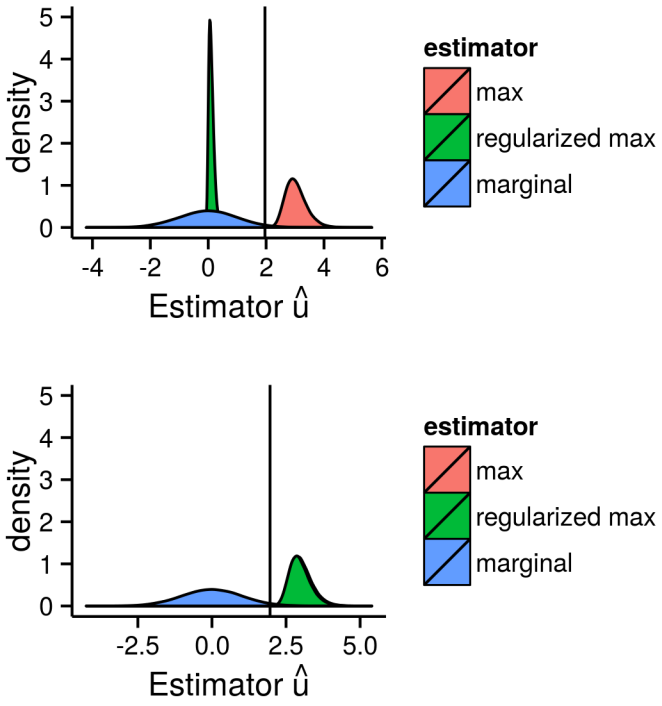

Figure 1. The marginal distribution of one element $\hat{U}_i|u^*$ of a classical signal estimator $\hat{U}|u^*$ in blue, the distribution of the maximum of that classical estimator $\hat{U}_{max}|u^*$ in red, and the distribution of the maximum hierarchical Bayesian posterior $U_{max}|v, u^*$ in green (see text for details of notation and simulation). The vertical bar illustrates the (uncorrected) test threshold h, that would only be correct if our signal had exactly one element, i.e. $|I| = 1$. Instead we assume $|I| = 100$. Under the family null $u_i = 0, \forall i$ (upper plot), the maximum signal estimator in a classical, non-hierarchical model demonstrates extreme over-estimating, deviating far from the true setting of zero. Classically, this requires increasing h to correct for multiple tests. In contrast, the posterior signal maximum of our hierarchical Bayesian model automatically shrinks to beneath h. Critically, this restriction on posterior inference is adaptive to the data. It is automatically lifted when the underlying signal is not null (lower plot), i.e. when latent target signal u is more complex and $u_i \neq 0$ but in fact varies randomly over $i \in I$. In the lower panel it was sampled from a non-null signal distribution on U with $\psi_U = (\mu_0, \tau_0) = (0, 1/10)$. This is reminiscent of adaptive testing, in which the test threshold - rather than estimator - changes according to inferred signal variation.

under the family null hypothesis, hierarchical estimation automatically limits FWER to an acceptable level. This makes conventional test adjustments unnecessary. Our results strengthen previous work which emphasizes hierarchical control of the, less conservative, False Discovery Rate.

This joint posterior may be, for example, defined on a set of a priori exchangeable random coefficients in a multilevel model: it's maxima just encodes my posterior belief in the magnitude of the largest of those coefficients (which "should" be zero for this data) and can be estimated for example by MCMC. The idea is that hierarchical Bayesian extreme values helpfully contract to zero with the number of coefficients in this setting, while non-hierarchical frequentist extreme values increase. The latter being more typically quantified by other "error" parameters such as FWER "multiple comparisons problem" or MSE "overfitting". Thus, this offers a clear way to show that hierarchical inference can automatically control the (weak) FWER, without Bonferroni-style adjustments to the test threshold. Mathematically, I imagine some asymptotic – in the number of coefficients – argument for this behavior of the maxima, that I would need time or collaboration to formalize (I am not a mathematician by any means). In any case, the intuition is that because posterior coefficients are all increasingly shrunk, so is their maximum. I have chosen to study the maxima because it is applicable across the very different hierarchical and frequentist models used in practice in the fields I work on (imaging, genomics): spatial, cross-sectional, temporal, neither or both. For example, the posterior maximum is defined for a discretely indexed, exchangeable random process, or a continuously-indexed, non-stationary process. As a point of interest, frequentist distribution of spatial maxima is used for standard style multiple-comparisons adjusted p-values in mainstream neuroimaging, e.g. SPM.

References

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210.