The Bayesian approach to family-wise error revisited

August 2016

1 Abstract

Statistical errors accumulate whenever one estimates or tests multiple unknowns from data, as is common in biological imaging and -omics. One conventional frequentist solution seeks to limit the aggregate error over all tests, notably the "family-wise" error popular in genome- and brain-wide studies. To better compare this conventional solution with more recent hierarchical Bayesian alternatives, we define a Bayesian family-wise error. We find that this Bayesian error is much smaller at any given test threshold under the family null hypothesis. Our results build on previous work on the Bayesian False Discovery Rate.

2 Introduction

Statistical methods often aim to identify some parameter or signal u, such as the reliable causal effect of an experimental manipulation, despite unexplained variation or "noise" in the data. Because of this noise, inference on u varies randomly over seemingly identical measurement conditions, which exhagerates or "overfits" the estimated magnitude u. This problem grows with the dimensionality of the inference, for example when $u \equiv \{u_i : i \in I\}$ is a large collection of unknown conditions/voxels/genes/time-points with index set I. This is because noise can separately corrupt each estimated component u_i. When we test hypotheses about the individual elements of u this overfitting is known as the multiple testing problem. There are therefore different ways to quantify this overfitting/overtesting problem and different ways to contain the problem by reducing the impact of noise on inference. As we discuss next, such corrections themselves carry a risk of underfitting.

Conventional hypothesis testing procedures seek to avoid overfitting by favouring null hypotheses, such as $u_i = u_j$, which constrain the number of free parameters: they set a high threshold for rejecting this null, so naturally have low sensitivity or power to detect small differences $u_i - u_j$. The ensuing "type-II" error is the unavoidable cost of prioritizing "type-I" or "false-positive" error[1]. As the test threshold increases to control *aggregate* type-I over multiple tests, this underfitting can only increase. By collapsing small differences to the default null hypothesis with increasingly high probability in this way, this conventional approach anticipates more recent hierachical shrinkage methods, as discussed below.

Hierarchical Bayesians address the overfitting problem through probabilistic (distributional) constraints on u, not through deterministic constraints like $u_i = u_j$ above, see "Bayesian and frequentist inference". This rests on the additional assumption that unobserved signal u was the outcome of some hidden random process, denoted U. For example, assuming hidden u_i were randomly sampled from an identical, low variance populatian is akin to a null hypothesis $u_i \approx u_j$ (because it is improbable that they differ from each other by much). Under this "null" prior assumption, Bayes theorem automatically shrinks the posterior estimates $u_i - u_j$ toward zero, which mirrors the bias for retaining the null hypothesis in the conventional approach discussed above. As above, this conveniently suppresses spurious inferences due to noise without needing a hypothesis test to filter them out, but risks underfitting whenever our premise is wrong. Importantly then, hierarchical shrinkage or "regularization" is calibrated according to the data itself, by learning the latent generating process for u or, more coloquically, "learning the prior"[2].

To compare frequentist and Bayesian solutions to overfitting more formally we study the maximum infered signal, assuming the family null hypothesis $u_i = 0$ for all $i \in I$. This family null hypothesis models the scenario when there is no signal anywhere in u, and implies for example $u_i = u_j = 0$ for all $i \neq j$, and $u_{max} = 0$. This means that every time the frequentist maxima exceeds test threshold $h > 0$ we falsely conclude there is one or more positive signal and commit a family-wise error, see "Classical maxima and conventional control". Critically, we note that Bayesian procedures also incur a family-wise error when the *posterior maxima* of a hierarchical Bayesian model exceeds h: this error is equivalent to falsely believing one or more elements of u have positive signal, given data generated under the family null[3]. This insight offers a conceptual

[1] In general, there is always a tradeoff between type-II for type-I, underfitting and overfitting. This relates to the famous bias-variance tradeoff. The type-II probability quantifies the test's conservative bias: accepting the default constraint $u_i = u_j$ introduces possible bias (whenever $u_i \neq u_j$) in return for reducing error variance (equating parameters $u_i = u_j$ means we can pool data across conditions to estimate just one unknown, thereby benefiting from the law of large numbers to reduce variance).

[2] Hierarchical methods thus automatically balance underfitting and overfitting in a principled way, without appealing to the arbitrary 0.05% convention to select a test threshold. Instead, this balance depends on a learned probability model of the latent random process which generated u. Shrinkage increases to counter overfitting whenever the null random process is deemed plausible, otherwise it decreases, to avoid unnecessary underfitting or overshrinking.

[3] The posterior maxima is just the credible magnitude of the maximal effect, which is 0 by definition in family null data. We emphasize that the distribution of the posterior maxima is *not* the familiar maxima of the posterior distribution, i.e. not to the maximum *a posteriori* or MAP estimate. Neither does it report posterior belief about which specific index (indices) take value u_{max}, for example it is not the distribution of the index with the largest marginal

bridge for us to compare Bayesian and frequentist family-wise probabilities, despite their different semantics and assumptions. In particular, we can ask which has a lower family wise error probability at a given threshold h. We have chosen to study the maxima because it is applicable across the very different hierarchical and frequentist models used practice: spatial, cross-sectional, temporal, neither or both. For example, the posterior maxima is defined whether a discretely indexed, exchangeable random process, or a continuously-indexed, non-stationary process. Regularization of the maxima.

3 A simple example

To formalize these concepts 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.

The hierarchical approach starts by specifying a plausible family of data-generating models. Our example assumes the data $V \equiv \{V_{i,j} : i \in I, j \in J_i\}$ was generated as follows. The hidden signal u_i are independent mean-zero Gaussian samples, drawn from a common distribution $U_i|\psi_U$, given parameters ψ_U. Conditional on the unobserved $u_i = U_i$, independent mean-zero Gaussian noise replicates $\epsilon_{i,j}$ are then added, giving the observed $V_{i,j} = u_i + \epsilon_{i,j}$. This notation is our shorthand for $(V_{i,j} = U_i + \epsilon_{i,j})|U_i = u_i$. A parameter ψ_V specifies the scale of this observation noise and is shared over all the $\epsilon_{i,j}$. We can write our data generating model as

$$\begin{aligned} \mathrm{U}_i|\psi_U &\overset{iid}{\sim} N(U_i|0, \psi_U) \\ \mathrm{V}_{i,j}|U_i, \psi_V &\overset{iid}{\sim} N(V_{i,j}|U_i, \psi_V). \end{aligned} \tag{1}$$

We define parameters $\psi_U > 0, \psi_V > 0$, which control the randomness of hidden signal[4] U_i and observation noise $\epsilon_{i,j}$ respectively, as Gaussian precision or inverse-variance parameters. Different settings of these parameters specify different possible models of the data-generating process. We ultimately seek to infer these parameters in order to calibrate under and overfitting, as discussed above. The index j describes within-condition, gene or voxel replications. We will sometimes omit this subscript in order to concisely denote the whole, condition-specific set of measurements $V_i \equiv \{V_{i,j} : j \in J_i\}$. (The subscript i of J_i just permits a different number of observations in each setting i). We call the situation in which our observed data $V_{i,j}$ was generated from the above model with $\psi_U \to \infty$ and $\psi_V < \infty$, a *family null generating process* because it implies $u_i = U_i = 0$ for all i, which is exactly the family null hypothesis familiar from conventional, non-hierarchical hypothesis testing framework. We denote the ensuing "family null" data, and inferences based on it, with an asterix. For

exceedence probability. For clarification on different types of maxima, see Appendix 2.

[4] For example, the sampling distribution of condition-specific "parameters" u_i from the general population.

example the family null data process is V^* and a specific sample v^* from this process.

Given family null data v^*, hierarchical Bayesian analysis automatically shrinks, smooths or "nullifies" inference on target u. Consider inference on latent sample realizations u_i, sampled via the above process U_i whose parameters we estimate to be $\hat{\psi}_V \approx 1, \hat{\psi}_U \approx \infty$ based on the complete data-set. Then posterior inference $U_i|v^*, \hat{\psi}_U, \hat{\psi}_V$ on each unobserved u_i is 0 with probability one, see Appendix 1. This plug-in procedure yields what is known as "empirical" Bayesian posterior inference. We will use the shorter notation $U_i|v^*$ to denote "full" Bayesian posterior inference, which averages $U_i|v, \psi_U, \psi_V$ over a set of different plug-ins, weighted according to posterior belief about their plausiblity $\psi_U, \psi_V|v$. It is important for what follows to note that because all posterior components $U_i|v^*$ shrink to zero with bigger estimates $\hat{\psi}_U$ under family null data, so does the posterior maxima, denoted $U_{max}|v^*$. The posterior maxima $U_{max}|v^*$ reports the credible magnitude of our specific u_{max} having seen the full data set v^*. For clarification on different types of maxima, see Appendix 2.

Bayesian and frequentist hierarchical inference

Conventional models only specify the likelihood function, which describes random error ϵ corrupting our measurement of target u. Conventional analysis can therefore infer the specific measurement errors and learn their general distribution, e.g. OLS regression residuals and their residual variance parameter. A hierarchical model assumes additional randomness in the provenance of target u. Hierarchical analysis can therefore hope to infer the specific u and the general sampling process U. It is based on the insight that these tasks are complementary. The noise-corrupt u tells us something about the general data generating process or population U, for example it's (hyper)parameters. This knowledge about process U in turn gives us information on the credible value of our specific, unobserved outcome u. Hierarchical Bayesian inference therefore aims to learn the latent generating distribution on U, before applying it as a constraint on plausible values of u. Bayesians use the term “population prior” to highlight this duality between the general population distribution and prior belief about an unobserved sample from that population. Assuming a parametric distribution $U|\psi_U$, learning this population prior reduces to learning the unknown parameter ψ_U. *Full* Bayesian inference uses Bayes theorem to learn ψ_U. *Empirical* Bayesian inference uses a point estimate $\hat{\psi}_U$, ignoring inferential uncertainty and therefore failing to propogate this uncertainty throughout all levels of the model.

The latent random quantities U in a hierarchical model are typically more interesting than the measurement noise ϵ. For example, a sample realization u may encode the specific value of interesting random effects, while the scale of U reports how much variance this process explains in general. In another example, u may be the specific hidden path or trajectory of a stochastic spatio-temporal dynamical system, and the parameters of U may encode the system's general tendency for diffusion, or its connectivity.

To many readers, the frequentist approach to hierarchical models may be more familiar in the special case of “repeated measure ANOVA”, popular throughout experimental science. Here random “within-subject” or “first-level” variation ϵ is nested within random “between-subject” or “second-level” variation U. For other readers, it will have arisen in the more general brand of “hierarchical regression” popular in social science or in the linear mixed effects models popular in economics. Yet another common setting is latent “structural equation models” and “structural causal models” of U, with additive measurement error ϵ. Here conditional independence structure between components of U can be represented as a directed acyclic graph.

The “hierarchical frequentist” approach aims to estimate/test the general distributional constraints on U, for example variance component parameters ψ_U, with minumum error variance over repeated draws from the data-generating process. In the social and behavioral sciences, these parameter estimates are just used to adjust standard errors on the “fixed effects” of primary interest. (Adjustments are needed because hierarchical data generation typically violates the strong Gauss-Markov assumptions behind OLS). Elsewhere, hierarchical frequentists have been more ambitious, offering tools for filtering, smoothing and prediction of specific u itself, including “kridging”, Bridge and James-Stein estimation.

Classical maxima and conventional control

Here we discuss the behaviour of conventional estimation and testing in "large u" problems, based on repeated samples from the family null generative process described above. In this setting, conventional approaches can use the frequency distribution of $\hat{U}^*_{max}$ to control over-testing. This is because knowing the distribution of $\hat{U}^*_{max}$ directly permits one to calibrate test thresholds to control FWER under the family null.

Recall that conventional *estimators* of u are chosen to limit the probability of large errors (over repeated data replications). This is done by identifying a minimum variance estimating function for each individual $\hat{U}_i = f_i(V_i)$. Similarly a null hypothesis *test* of component i involves finding a threshold h that errors will rarely exceed, assuming some null equality constraint, such as $u_i = 0$ or $u_i - u_j = 0$, is true. Conventional dogma says rare means $\alpha = 0.05$.

These random estimation/test errors increase whenever we jointly infer all components $\hat{U}_i = f_i(V_i)$ for all $i \in I$. This is because each noisy inference offers a new opportunity for big error to arise by chance. For example, the chance that the *largest* error in our estimator exceeds h is $P(\hat{U}^*_{max} > h) = 1 - P(\cap_{i \in I}\{\hat{U}^*_i \leq h\})$, where $\hat{U}^*_{max}$ denotes the maximum function $\max_{i \in I} \hat{U}_i$. This extreme-error probability *increases* with $|I|$, the number or volume of the index-set I, because in general $P(\cap_{i \in I} A_i) \leq P(\cap_{i \in I'} A_i)$ whenever $I' \subseteq I$: this monotonicity follows directly from Kolmogorov's axioms. (The maxima also increases with statistical independence between different components of our estimator $\hat{U}_i$.) This aggregate error $\hat{U}^*_{max} > h$ only occurs when one or more $\hat{U}^*_i$ erroneously exceed h, and is therefore called the family-wise error, over the "family" indexed by $i \in I$. Solving for h in $P(\hat{U}^*_{max} > h) = \alpha$, gives the threshold beyond which error will rarely arise in *any* component $\hat{U}^*_i$, assuming the family null. In frequentist jargon, this h "weakly controls the family-wise error rate (FWER) of a test to level α" and "solves the multiple testing" problem incurred in "large u" problems.

The relation between FWER and maxima suggests that one can control aggregate error either by "correcting" (increasing) the test threshold h or by correcting (shrinking) the aberrant extrema $\hat{U}^*_{max}$ itself. The former is the approach taken in conventional inference on fixed signal models; the latter can be achieved in modern hierarchical models, for which extreme inferences *decrease* automatically with $|I|$. In neuroimaging, these arise as conventional "statistical parametric mapping" and hierarchical "posterior probability mapping". The relation between these two approaches can be obscured by differences in the terminology and formulation. For example, the classical school "controls" over-testing, while the latter "regularizes" estimation. This paper attempts to help bridge this gap.

4 Posterior maxima and hierarchical control

footnoteSadly, in many common hierachical Bayesian models, including our simple example, the posterior probability that $u_i = U_i = 0$ must itself equal zero for technical reasons, even while the posterior probability mass converges to zero with $|I|$. This is because the prior and posterior are both absolutely continuous with respect to the underlying Lesbesgue measure, e.g. Gaussian. Our solution here is to ask how far posterior belief is from the true $u_i = U_i = 0$: this is just $\mathbb{P}(U_{max} > h|v^*)$, which we contrast with the conventional FWER $\mathbb{P}(\hat{U}^*_{max} > h)$. Alternatively we could move to a class of generative models which can assign positive probability to $u_i = U_i = 0$. This offers a route to control of Bayesian FWER. Here $1-\theta$ encodes signal sparcity which can be learned via hierarchical Bayesian procedures. If this reveals $\theta \approx 1$ *a posteriori*, then we conclude the family null hypotheses is plausible and smoothed inferences on $U_{max}|v^*$ should shrink to exactly zero..

We study the distribution of hierarchical posterior maxima $U_{max}|v^*$ based on a single "family null" data-set v^*, comparing it to the distribution of the frequentist maximum $\hat{U}^*_{max} \equiv \hat{U}_{max}(V^*)$, based on repeated family null realizations. We will see that $\hat{U}^*_{max}$ and $U_{max}|v^*$ behave very differently as the problem dimensionality $|I|$ increases, where $|I|$ is the dimensionality of u. Because the infered constraint $\hat{\psi}_U$ increases with $|I|$ under family null data, the posterior maxima $U_{max}|v^*$ "self-corrects", exceeding even *uncorrected* conventional thresholds with less than 0.05 probability. Depending on the application, this means with enough experimental conditions, voxels or genes i test corrections become unnecessary. Finally, we study the frequentist behavior of hierarchical Bayesian inference, to show that hierarchical Bayesian procedures self-correct, both conditionally and marginally, i.e. over repeated draws from the family null.

The posterior maxima distribution can be evaluated in more general models. For example, the posterior maxima of a continuous, spatial posterior reflects learned constraints on the scale of spatial gradients $\lim_{t \to 0} \frac{U_x - U_{x+t}}{t}$, where $x \in \mathbb{R}$ is the real-valued index of a continuous, latent random field of effect sizes. (Contrast this with a constraint on $U_i - U_j$ in our simple example above). In the case of Gaussian processes with a squared exponential kernel for example, this constraint is found in the "length-scale" parameter. We propose that the posterior maxima complements other overfitting metrics, such as cross-validation error or the expected Euler characteristic from topological statistics.

Ultimately we want to show that, under a general class[5] of hierarchical priors, Bayesian posteriors converge asymptotically with $|I|$, while conventional, unbiased frequentist estimators diverge.

[5] The conditions on this class are that there exists a uniquely identifiable parameter setting ψ^*_U which specifies the family null. In other words, there exists an identifiably setting ψ^*_U which satisfies $P(|U_i| > c|\psi^*_U) = 0$ for all $i \in I$ and $c > 0$ (which is in the support of the prior on ψ_U). In addition, observation noise must have at most "weak" statistical dependence within and across the $i \in I$.

5 Simulations

The distribution of conditional Bayesian posterior maxima $U_{max}|v^*$ and the distribution of frequentist estimator maxima $\hat{U}^*_{max}$ can be derived analytically or by Monte-Carlo.

We also examine the frequentist behavior of hierarchical Bayesian inference, by deriving the *marginal* distribution of the Bayesian posterior maxima which we simple denote U^*_{max} (to contrast with hatted, frequentist $\hat{U}^*_{max}$). This marginal entails averaging the conditional posterior maxima $U_{max}|v^*$ over repeated draws from the family null data-generating process V^*, in other words $U^*_{max} \equiv \mathbb{E}_{V^*}[U_{max}|V^*] = \int_{v^*} p(u_{max}|v^*)p(v^*)dv^*$.

In our simple linear-Gaussian example, the family null data-generating process is parameterized with $\psi_U = \infty, \psi_V < \infty$, so the marginal distribution on U^*_{max} may be approximated, for example, via Monte Carlo

$$\int_{v^*} p(u_{max}|v^*)p(v^*|\psi_U = \infty, \psi_V)dv^* \approx \sum_l^L p(u_{max}|v_l^*).$$

This approximation just averages L posterior maxima distributions, each being conditional on a single, independent realization v_l^* from the family null $p(^*_v|\psi_U = \infty, \psi_V)$. (This marginalization can be applied both to full Bayesian and "Empirical" Bayesian plugin solutions to hierarchical Bayesian inference, the latter is equivalent to just studying $U_{max}|V^*, \hat{\psi}_U(V^*), \hat{\psi}_V(V^*)$, over the distribution of repeated draws from V^*.) We compare the ensuing marginal Bayesian maxima distribution over U^*_{max} with frequentist distribution over $\hat{U}^*_{max}$. Let h_α denote the conventional, uncorrected threshold, which only guarantees to control the frequentist false-positive error at a single index i to a level of α, say 0.001, but is famously uncorrected/wrong for high-dimensional $u \equiv \{u_i : i \in I\}$. We say that a hierarchical Bayesian procedure automatically offers *corrected*, marginal, weak FWER control at level α, if $\mathbb{P}(U^*_{max} > h_\alpha) \leq \alpha$.

5.1 Spatial Gaussian process

We could have used a different posterior summary, for example the an $|I|$-dimensional set of marginal exceedence probabilities. The index i with the largest marginal posterior exceedence probability tells us which

for example the an $|I|$-dimensional set $\{T_i : i \in I\}$ of binary indicators, where T_i equals 1 if and only if the marginal exceedence at i is above say $1 - \alpha$, i.e. $\mathbb{P}_{U_i|v^*}(U_i > h|v^*) > 1 - \alpha$. Or we could use 1-dimensional $\mathbb{P}(\sum_{i \in I} T_i > 0)$, the family-wise error. We could have used the $|I|$-dimensional set of marginal *maximum a posteriori* estimates reported by the James-Stein estimator. While decision theory[6] could formally justify our choice of this posterior summary, we appeal directly to the analogy with the frequentist FWER. Accordingly, $\mathbb{P}(U^*_{max})$ describes the long-run frequentist behavior of this posterior summary over repeat sampling from the data-generating process.

[6] Given a loss function, a prior process, and a specific data-set v, the generalized Bayes risk tells us which posterior summary minimizes the expected loss.

What is Posterior exceedance $\mathbb{P}_{V^*}(\rho^*_{max})$, where $\rho^*_{max} \equiv \max_{i \in I} \mathbb{P}_{U_i|v^*}(U_i > h|v^*)$

Frequentist behavior of the subjective maxima

Frequentist behavior of different posterior summaries. The frequentist properties of *maximum a posteriori* are well documented in our simple example, see James stein estimation for example.

This non-frequentist "test" is therefore simply a summary of our posterior belief.

Alternatively consider the two-sided case.

Earlier work on the frequentist properties of Bayesian inference has focused on the behavior of posterior summaries, such as the posterior mode. For the case of our simple example, James-Stein showed that the posterior mode has lower aggregate MSE than conventional estimation. From a FWER testing perspective, more relevant is the maximum posterior mode, the $maxMAP$, under the family null. We could contrast the frequentist distribution of the maximum MAP estimator and the conventional, unconstrained ML estimator. If the former places less than α mass beyond uncorrected thresholds, we could say that it self-corrects. Supra-threshold MAPs are automatically significant at a corrected FWER. This is perhaps more appropriate for the summary statistic approach used by SPM?

The *Bayes risk* is defined as scalar functional of the measure on the prior process $\mathbb{P}_U$ and the decision rule $\delta(V)$, namely $r(\mathbb{P}_U, \delta)) = \mathbb{E}_{\mathbb{P}_U}[E_{\mathbb{P}_{V|U}}[L(U, \delta(V)]]$. For a given u, the inner expectation is the *frequentist risk* function, over repeated realizations V. The outer expectation is over the random generating process U, or a Bayesian prior. Bayes risk is therefore the expectation of the expected loss. Alternatively, we can write Bayes risk as $r(\mathbb{P}_U, \delta(V))) = \mathbb{E}_{\mathbb{P}_V}[E_{\mathbb{P}_{U|V}}[L(U, \delta(V)]]$, where the inner expectation is the Bayesian posterior expected loss (sometimes called the *generalized Bayes risk*), and the outer expectation is the marginal distribution of the data V. The Bayesian posterior expected loss is well defined if we observed the data v, but the Frequentist expected loss (or frequentist risk) is still a function of unknown u. So need further principles - beyond the expected loss principle - are needed to guide the choice of δ. These include unbiasedness, admissability, minimax, invariance, etc.

A Bayes rule minimizes the Bayes risk. The generalized Bayes rule $\delta(v)$ minimizes generalized Bayes loss.

The *generalized Bayes risk* or *Bayesian expected loss* is more recognizably Bayesian, being the expected posterior loss given fixed dataset v, namely $r(\mathbb{P}_U, \delta, v) = \mathbb{E}_{\mathbb{P}_{U|v}}[L(U, \delta(v))]$. For each v we aim to find a decision rule $\delta(v)$ that minimizes the expected loss (known as a generalized Bayes rule with respect to the prior process $\mathbb{P}_U$).

The basic disagreement between Bayesians and frequentists: Bayesians say we don't know V we know v. Frequentists say we don't know u is the outcome of a random variable (in non-hierarchical applications u may be a constant such as the speed of light). This objection is less keen when frequentist acknowledges that U is a population or sampling process.

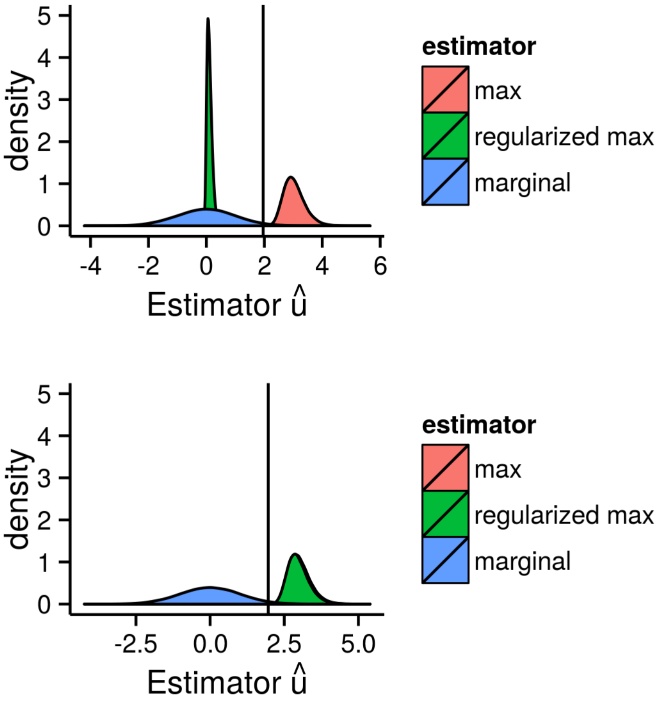

Figure 1: The marginal distribution of one element $\hat{U}_i|u^*$ of a classical signal estimator $\hat{U}|u^*$ in blue, the distribution of the maximum of that classical estimator $\hat{U}_{max}|u^*$ in red, and the distribution of the maximum hierarchical Bayesian posterior $U_{max}|v, u^*$ in green (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.

5.2 Details of Simulation

Using real data from X, we compared frequentist and Bayesian inference on a parameter that we knew to be 0, by ground truth, over the whole brain.

1) We first estimated the GLM parameter associated with a single random regressor (independent, standard Gaussian realizations), using both frequentist SPM and Bayesian PPM schemes. The frequentist estimation scheme provides us with uncorrected h_u and the corrected h_c (from the standard approximate SPMM survival function at high h). We used Monte-Carlo simulation to estimate the corresponding PPMM survival function. We simply sampled from the joint posterior PPM and choose the maxima (this gives a PPM maxima distribution, comparable to the SPM maxima distribution given by RFT etc). Specifically, we sampled from the PPM maxima distribution by first drawing 1000 samples from the N-dimensional joint PPM, then extracting the observed maxima over the entire brain. The proportion of these realized maxima which exceed h is called a Monte-Carlo estimate of BEP_h.

2) We used permutation Monte-Carlo to assess the SPM maxima distribution non-parametrically. This has the advantage of giving us (an estimate of) the entire SPM maxima distribution (where the RFT approximation is only valid for high thresholds). We can then see unambiguously that SPM maxima dominate PPM maxima.

The data can be found at:

http://www.fil.ion.ucl.ac.uk/spm/data/face$_rfx/$

http://www.fil.ion.ucl.ac.uk/spm/data/face$_rfx/face_rfx_spm12_batch.m$

Henson, R.N.A, Shallice, T., Gorno-Tempini, M.-L. Dolan, R.J (2002). Face repetition effects in implicit and explicit memory tests as measured by fMRI. Cerebral Cortex, 12, 178-186.

6 Hierarchical Bayesian stochastic process models

Moving beyond the specific examples above, we now consider hierarchical modeling more generally. “Hierarchical” modeling is based on the fact that the joint distribution of a set of random variables decomposes into a series of smaller, conditional models. All hierarchical models - spatial, temporal, neither or both - are formulated in three levels (possibly with sub-levels).

Parameter model: [noise-parameters ψ_V and signal-parameters ψ_U]

Hidden signal model: [signal U given signal-parameters ψ_U]

Observed data model: [outcome V given signal and noise-parameters ψ_V].

This hierarchy may be written

$$\begin{aligned} \psi = (\psi_U, \psi_V) &| S \\ U &| \psi_U, S \\ V &| U, \psi_V, S \end{aligned} \tag{2}$$

This notation can be interpreted generally, so that ψ_U (ψ_V) are the signal (noise) parameters of any arbitrary hierarchical model, and not just the inverse-precision parameters of our simple linear-Gaussian example model above. S denotes the (often implicit) model-specific, parametric and conditional independence constraints within and across levels. Here U is a general stochastic process[7].

We have emphasized that knowledge of the generating process or distribution on U constrains which samples are probable and credible: infering and applying this constraint therefore limits over-fitting/testing under the family null. To understand distributional constraints on a general stochastic process U it helps to ask whether two arbitrary process subsets $U_A \equiv \{U_i : i \in A\}$ and $U_B \equiv \{U_i : i \in B\}$, indexed by possibly overlapping index subsets A and B, are statistically and/or parametrically dependent[8]. Statistical dependence implies the restriction that U_A and U_B are not free to vary independently, $p(u_A, u_B|\psi_U) \neq p(u_A|\psi_U)p(u_B|\psi_U)$. Parametric dependence means they are not free to differ arbitrarily in *marginal* distribution, but must share common signal parameters, i.e. $p(u_A|\psi_U), p(u_B|\psi_U)$. Just as in our simple example, inference on these general constraints can help adaptively control posterior error variation.

We have said that U may be constrained by qualitative (in)dependence constraints, not just quantitative hyper-parameters. Permissible statistical dependence may be represented by the edges between nodes of a graph (all random variables in the hierarchical model get a node): a missing edge between two subsets of the process implies a conditional independence. For example, U_A and U_B may be conditionally independent, conditional on (hyper)parameters ψ_U alone (e.g. global shrinkage priors and our simple example), or independent given parameters ψ_U and some other set of signal U_N (e.g. the neighbourhood N separates them, as assumed in Markov random field spatial prior) or independent asymptotically as A and B become separated by increasing, continuous distance (Gaussian diffusion priors).

Parametric dependence - e.g. identical parameters - refers to relationships between parameters of the marginal distributions on U_A and U_B. Such dependence arises when the joint distribution over finite subsets of the signal process remain identical under some transformations of the index set I. For example, “exchangeability”, “stationarity” or “isotropy” describe invariance of all finite subsets of the process to permutation, translation or rotation, respectively. (These invariants of the process imply that process subsets share *all* parameters,

[7] Recall that a *random variable* $U_i(\omega)$ is any measurable, real-valued function of the outcome $\omega \in \Omega$ of a random experiment: the random ω means $U_i(\omega)$ is also random. Each experiment generates a specific outcome ω and therefore a specific real number $u_i = U_i(\omega)$. A *stochastic process* is simply a set of random variables $U(\omega) \equiv \{U_i(\omega) : i \in I\}$, indexed by some index set I. Depending on the model, this index set may be discrete or continuous, countable or uncountable. Here an outcome ω implies a whole set of real numbers, $u \equiv \{u_i : i \in I\} \equiv \{U_i(\omega) : i \in I\}$. For clarity we omit the argument ω throughout.

[8] It also helps to ask whether the process index-set is a finite subset of natural numbers, a graph-valued index set, a (uncountably infinite) subset of real euclidean space or some non-euclidean topologies.

because two finite subsets share identical joint marginal distributions! For this reason we refer to "parametric dependence").

Another very general restriction on a process is the assumption that it follows some known parametric form. The most common spatial priors in neuroimaging are a convolution of Gaussian processes: Either conditionally white processes (global priors in PPM), Markov-conditionally white (local shrinkage "spatial" priors) or asymptotically white priors (badly named "Gaussian process" priors, parametrized by some kernel K). Relaxing these parametric restrictions leads to non-parametric stochastic processes (e.g. spatial Dirichlet process) which go beyond our current scope.

The qualitative issues which we have discussed dictate whether, given family null data V^*, $U_{max}|v^*$, and conditional Bayesian FWER, can be controlled by hierarchical inference. Precisely quantifying such control is a larger job which we only begin in this work[9].

7 Spatio-temporal neuroimaging

We now illustrate how these concepts relate to neuroimaging. Functional imaging measurements V_{st} are typically indexed by space and time, $V = \{V_{st} : s \in S, t \in T\}$. The "continuous field" perspective on such data recognizes that the sets of time T and space S are continuous: technology alone limits our measurements to a finite subset of points. In contrast, the "discrete field" perspective restricts itself from the outset to considering *aggregates* over small continuous regions of space and/or time, e.g. indexed by discrete scan and voxel number.

Consider the general model $\{V_{st} : V_{st} = f(x_{st}, u_{st}, e_{st}), s \in S, t \in T\}$ where the observed data is related to three other quantities - a known vector of predictors at each space-time $\{x_{st}\}$, an unobserved target vector $\{u_{st}\}$, and unobserved noise $\{e_{st}\}$ which corrupts each measurement (for notational clarity, we omit the condition $s \in S, t \in T$ on each of these sets). Note that $\{u_{st}\}$ is a large set of unknowns, e.g. number of scan-times x number of spatial voxels x dimensionality C of our vector u_{st}. Naive attempts to fit our model to data (or test) will suffer from a serious "over-fitting", "multiple estimation" or "multiple testing" problem, arising from the aggregation of small estimation errors over a large signal.

[9] The most extreme value - and the topological complexity - of a process increases with the size (or Lesbegue measure) of index set $|I|$, the point-wise variance at each $i \in I$ and the independence across $i \in I$. Shannon's information entropy also increases with size, marginal variance and joint independence but is more general, and applies to nominal random variables, such as class labels, for which extrema are meaningless. Because entropy increases with the dimension $|I|$ it is used to regularize the dimensionality of probabilistic PCA/ICA decompositions. Relative entropy - or KL divergence - of prior and posterior process offer a general way to understand Bayesian control of over-estimating. For example the celebrated "model evidence" or "marginal likelihood" of our signal parameter ψ_U can be written as $p(v|\psi_U) = exp(E_{u|v,\psi_U} lnp(v|u, \psi_U) - KL(p(u|v, \psi_U)||p(u|\psi_U)))$. This says that the posterior deviations from the prior distribution $KL(p(u|v, \psi_U)||p(u|\psi_U))$ incur an exponential penalty (again KL is relative entropy).

A conventional solution would be to assume some equality constraints on the spatio-temporal expression of signal u. For example, neuroimaging analyses typically assume the unknown vector $u_{st} = u_s$ and $x_{st} = x_t$. Thus u_s at each spatial position s is a time-invariant "fixed effect" with respect to time, and the predictor x_t is space-invariant. These help limit temporal over-fitting at any point in space[10]. This reduces the inference problem to $C \times V$, which nonetheless remains huge.

I change indices from st to ij!

Conventional "region-of-interest" analysis attempts to rescue this situation through another appeal to equality constraints, namely $u_s = u_{s'} = u$ for all $s \neq s'$, i.e. unknown parameters are invariant over some spatial region. This reduces estimated signal down to one C dimensional vector. To summarize, the region-of-interest constraint $u_{st} = u_{s't'} = u$ restricts how much independent signal there is to over-fit down to just the C components of u. This comes at the expense of introducing under-fitting bias whenever the signal u is not really time or space invariant. How can we assess whether such constraints are appropriate? In principle, one could use some "omnibus" test of the family null hypothesis, e.g. $u_i = u_j$ for all i, j. This is akin to the familiar conventional F-test, which is parametrized by the number of constraints imposed (possibly adjusted for non-sphericity). Generally such testing requires the conditional frequency distribution of a scalar model comparison index, example maximum likelihood ratio, given the null setting/model. In reality, practitioners seldom perform this test. This amounts to infinitely precise priors on the constraint. In principle this constraint can be rejected by model comparison with either a more general conventional or hierarchical alternative. Don't test prior on the family null.

Moving beyond region-of-interest analyses, in general there are untested equality constraints implicit in all conventional forms of pre-specified spatial smoothing. These go from the simple identity $u_i = u_j$ assumption above to assumptions about stationarity such as $u_i = u_{i-k}$ or isotropy such as $u_i - u_j = f(|i-j|)$. The hierachical approach replaces these with *statistical* constraints of exchangeability, stationarity or isotropy, see "Hierarchical Bayesian stochastic process models". In practice this means that fixed equality constraints such as $u_i = u_j$ or equivalently $u_i - u_j = 0$, are replaced with random equality constraint $p(u_i - u_j|\psi_U) \approx 0$, sometimes expressed in terms of a discrete slope between neighbours $p((u_i - u_j)/\delta|\psi_U) \approx 0$ (or continuous derivative), the covariance $cov(u_i, u_j|\psi_U)$, etc. Importantly, these restrictions are not pre-specified, can be imposed - or relaxed - by different, inferable settings of ψ_U.

[10] A typical analysis additionally assumes linearity in the unknowns $\{u_s, e_{st}\}$, i.e. $V_{st} = f(x_{st}, u_s, e_{st}) = x_t^T u_s + e_{st} = \sum_{c \in C} x_{tc} u_{sc} + e_{st}$, conveniently permitting us to write the observed data in matrix notation as $V = XU + E$. Here the visible data V_{st} is arranged into a $N \times V$ matrix (number of measured time-points × number of measured space points), X is a space-invariant experimental design matrix (N × number predictors C), unobservable parameters U form a $C \times V$ matrix, and E is a $N \times V$ matrix of stochastic errors.

Future research: Use simulation under other existing spatial imaging models, to confirm that they incidentally control FWER.
Use global shrinkage with a spatial structural/neuromodulatory predictor. This would provide a way to functionally validate structural models, and to test whether spatial distribution of neuromodulator predicts fMRI.
Bayesian model selection over space, that incidentally controls FWER.
Validate connections between hierarchical control and FDR.
Hierarchical regularization of noisy single trial ERP.
Nested PPMs: scale-space PPMS! Easy!

8 Conclusion

Maximum $\max_{i \in I} U_i$ is a non-linear function on a process. Contrast with other functions on U, say $\sum_{i \in I} U_i$, which may be linear.

Our confidence in a conventional interval comes from the fact that the inference method we used works on average over repeat sampling from the DGP. Confidence in a Bayesian interval summarizes uncertain belief about the unknown true value in our specific sample.

Bias and variance in estimators, e.g. causal or non-causal parameters of population. For causal inference: bias - simpson confounding, berkson selection - is bad. Unbiasedness - or at least known shrinkage bias - is important in casual inference. Pearl has alot to say about bias in estimation of population causal parameters and individual causal states. He defines it clearly, and enumerates. Not much on variance. Bias wrt the true population parameter. Bias with respect to a causal population *parameter*. Bias in individual level causal *inference* (Pearl). Bias: Ecological/Simpson's bias is an underfitting bias. Selection bias is not. diagnose by DAG, prevent by randomization. Measurement bias. Big data problems (don't average out in big data). Variance: Random sampling error of estimator = random deviation of estimator due to random sampling from population. Frequentist bias-variance tradeoff: bias useful (especially if we *know* the bias, e.g. shrinkage bias given by $\hat{\psi}_U$).

Avoid biased population estimators by random selection, random intervention (assignment of units to treatments). Sampling selection bias (berksons nonrandom sample selection) versus simpsons modeling (underfitting) bias. Nonrandom intervention (confounding such as paradox). Experiment. Blocked randomized replicated experiment for bias-variance reduction: Block for variance-reduction, then randomize within block for bias-reduction. This randomization increases variance - randomization variation cf sampling variation just named after the source - in causal estimators. So use replication.

Re-randomization/pemutation test (units randomized to treatments, re-randomize), Re-sampling estimator.

FWER ignores different costs of error. Incorrect/correct decisions have different costs for different stakeholds: a game.

Many treatments of multiple testing focus on $p-values$. In general a decision on one $p - value$ is not *seperable* but can depend on all $p - values$: this pays off when test statistics, and therefore $p - values$ are dependent.

FDR control via BH rests on ordering the p-values. Under the homogenous variance assumption in our simple example, the only thing that differs between $i \in I$ are the components of u itself. So an ordering of p-values reduces to ordering on estimated effect sizes $\hat{U}$: BH accepts $\hat{U}_{max}$ if its p-value is beneath α/m.

Often it is unsafe to assume a parametric form for the population process, as we have throughout.

Multilevel models support multilevel generalization: in a parametric, nested design for example this means generalizing to new data from a subject ψ_V versus to new subjects ψ_U. The parameters ψ_U, ψ_V respectively govern generatization to new indices i', j' in $V_{i',j'}$ respectively (of course generalization may be local to one dimension of the index set).

Our work develops on Gelman et al. (2012). The FWER of 0.05 or even 0.001 is arguably far too low.

Conventional control of *test* error incidentally reduces *estimation* error because complete data/parameter pooling is the default, unless the null can be rejected. We have emphasized that, conversely, hierarchical control of estimation error incidentally reduces hierachical test error. Thus, using either quantitive estimation or qualitative testing, the result is comparable: one correction is enough. In our simple example, the hierarchical control parameter ψ_U mimicks the conventional test control parameter h_α. (h_{alpha} is a parameter of the decision rule, ψ_U is a parameter of the data-generating process: that they both control over-fitting means decision rules can but need not be explicitly be designed to reduce aggregate error). Both shrink or nullify model complexity under the family null hypothesis, implementing an Occam's razor.

To relate these two schools we considered the multiplicity problem. We emphasized that scale of this problem increases with signal dimensionality in "large-u" problems, incurring a severe "multiple test problem". This error may be seen in the distribution of maximum estimator under the family null, for which $\mathbb{P}(\hat{U}^*_{max} > h) \leq \mathbb{P}(\hat{U}^*_{max} > h)$, for any $I \subseteq I'$ and some null u^*. Thus, the probability that a big error ($> h$) arises somewhere in $\hat{U}$ increases with the size/volume of the signal vector (all else constant). We showed that in hierarchical Bayesian analysis conditional on family null data this condition is reversed i.e. $\mathbb{P}(\hat{U}^*_{max} > h|v^*) \geq \mathbb{P}(\hat{U}^*_{max} > h|v, u^*)$. Thus multiplicity can be a blessing, not a curse for hierarchical models. The maxima provides one way to illustrate how hierarchical inference can automatically solve the over-estimating/testing problem.

Weak control of FWER? When the family null is false.

The general generating process for U gives information on the credible values of u. Conversely, the (noise-corrupt) u gives information on the generating process U.

We have shown here that both hierarchical and non-hierarchical procedures control false-positive error rate. This does not make them equivalent. Control of false-negative error is vastly worse in conventional procedures. The problem with conventional procedures in principle is underfitting not overfitting.

Conventional procedures take great care to control false-positive error, strictly penalizing complex models with too many unconstrained parameters. They do this by assuming a simple null model by default, unless the null can be rejected by a test. In large-scale conventional inference, post-hoc test corrections which weakly control FWER assume the family null by default. This framing completely neglects false *negatives* which arise if the family null is false. This means underfitting, or oversmoothing true effects. Hierarchical procedures avoid this trap by infering the plausibility of the family null first. Shrinkage is relaxed when the family null is implausible: this reduces false negative errors. Otherwise, shrinkage penalizes false-positives.

”The classical approach to the multiple testing problem relies on constructing procedures maximizing the power while controlling FWER”

There are different ways to quantify aggregate inferential error, for example, the family-wise error over tests, the mean squared error over estimators and the maxima error over estimators. Conventional procedures prioritize the first, hierarchical procedures the second (James-Stein). We propose the third metric, the estimator maxima, bridges these approaches. In particular, the maxima distribution shows why conventional test correction is necessary under the family null, and why hierarchical test correction is not. Non-uniform priors are analogous to ridge regression, whose out-of-sample predictive performance exceeds least squares because it penalizes the sum of squared estimators.

We have shown that hierarchical Bayesian inference weakly controls frequentist FWER at uncorrected test thresholds, both conditionally and marginally. We do not recommend uncorrected test thresholds as standard practice when reporting Bayesian posterior inferences.

Conventional and Hierarchical methods differ in their treatment of unobserved heterogeneity across members of some index-set I. The conventional approach can only reject equality constraints, while hierarchical approach learns the complete latent sampling process.

Non-hierarchical inference, both Bayesian and frequentist, uses fixed constraints to limit overfitting: a fixed Bayesian prior or a fixed null hypothesis. To avoid overfitting large models, these constraints increase with the problem dimensionality or ”multiple comparisons” (DCM prior precision depends on number of params, same for MCP). Hierarchical procedures - Bayesian or frequentist shrinkage or FDR - do not use fix constraints, blindly increasing constraints with the problem dimensionality. Instead constraints are derived from the data itself. This means overfitting is penalized more when the family null is plausible a postiori (not when it is simply higher dimensional). In our simple example, hierarchical learning means infering $\psi_U \approx \infty$ from data.

Conversely, when the family null is implausible, false-positive error control is relaxed and false-negative error is reduced. Herer inference on component u_i is freed from hierarchical constraint. For hierarchical Bayesians the family null is a (degenerate) latent generating process, in other words a point mass on the 0 vector: the choice of whether to police false-positive versus false-negative test error - equivalently over-fitting versus under-fitting error - is made automatically, by infering the family null plausibility. For non-hierarchical procedures, the family null false-positive error is *assumed*, even if it is not plausible given the data. These procedures therefore blindly penalize false positives, parameter overfitting or model over-complexity, even when false-negatives, underfitting or model over-simplicity are the real problem.

Pedogogy: 1. show target realization 2. show likelihoods (with dispersion, each only centered on target in expectation) 3. now remove the target realizations and show generating process instead, explain that if you had to choose amoung likely arguments of the (dispersed) likelihood function, you'd probably favour the shrunk side. This is bayesian analysis. 4. now hide the generating process, can it be learned by hierarchical learning (MLII, full Bayesian or whatever).

decision theory: control! i.e. min EL(s,a) causal theory: control!

9 Appendix 1: A simple example

Given ψ_U, ψ_V, it can be shown by Bayes theorem that uncertain posterior inference on the latent sample realization u_i, denoted $U_i|v, \psi_U, \psi_V$, 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 $\hat{u}_i = \frac{1}{J_i}\sum_{j=1}^{J} v_{i,j}$ is the observed sample mean. These equations show that a large prior constraint - large ψ_U - on the sampling process U_i causes the posterior mean inference on realization u_i to shrink towards zero and away from the observed sample mean $\hat{u}_i$. Posterior uncertainty about u_i, given by variance $1/\psi_U$, also shrinks to zero, effectively trapping all posterior mass close to zero. These facts are useful because the *estimated* distributional constraint $\hat{\psi}_U$ should be large when the family null is true: plugging this estimate into the formulas for the posterior mean and precision above will naturally shrink all posterior components $U_i|v^*, \hat{\psi}_U, \hat{\psi}_V$ towards zero. It is by this mechanism that hierarchical methods suppress noise or counter over-fitting under the family null hypothesis.

10 Appendix 2: Different types of maxima

We should clearly distinguish different maxima. Given a specific set of, typically unobserved, real numbers of interest $u \equiv \{u_i : i \in I\}$, u_{max} is just the largest value in that set. Given a conventional noise-corrupt estimator of $\hat{U}(V)$ of u, $\hat{U}_{max}(V)$ is the distribution of the maximal estimate over repeated, noise-corrupt samples V *with* u *fixed*. Assuming the family null means assuming $u_i = u_j = u_{max} = 0$ for all $i, j \in I$. We denote ensuing family null samples V^* and note $\hat{U}^*_{max} \equiv \hat{U}_{max}(V^*)$ can be used to quantify and weakly control FWER, see "Classical maxima and conventional control".

Assume a general latent generating process $U|\psi_U \equiv \{U_i : i \in I\}|\psi_U$, then $U_{max}|\psi_U$ is the distribution of u_{max} over repeat draws from the latent generating process. Bayesians might use this general sampling distribution as *prior* belief about credible value of a specific unobserved u_{max}, i.e. the largest value in our u, see "Bayesian and frequentist hierarchical inference". One might call it the prior maxima distribution. Given Bayesian *posterior* belief $U|v^*$ about our specific unobserved target u, $U_{max}|v^*$ is the credible posterior magnitude of our specific u_{max}.

By Bayes theorem, $p(u_{max}|v) = p(v|u_{max})p(u_{max})/p(v)$, where $p(u_{max}) = \int_{\psi_U} p(u_{max}|\psi_U)$. The distribution of the prior maxima encodes maximum overfitting, i.e. over fitting that occurs without having data. $U_{max}|v^*$ encodes actual overfitting, conditional on family null data. ($U_{max}|v$, for non-null v, is not generally of interest.)

11 Appendix on Bayesian versus frequentist inference

Bayesians and frequentists differ in whether or not you can use probability to represent uncertainty about an unobserved constant, be it a fixed parameter or the unobserved realisation of random variable. Imagine a fair coin flipping process. The prospective relative frequency of heads is one half: this is the frequent notion of probability. If you believe a single, unobserved retrospective outcome was heads with probably one half, then you are Bayesian: a frequentist would say this is not random so has no probability distribution.

Hierarchical predictions generalise to new latent realisations - via ϕ_U (and maybe all other data? c.f. Gaussian process) and new x.

Different components of the prior have different semantics. Prior to ignorance about what? What do we know about the unknown number, logical constraint $0 < u < 1$, sampling constraints, i.e. is it the realisation from a latent frequency distribution? For the latter, hierarchical, case prospective frequency - i.e. the population prior - dictates retrospective credibility of a sample. The outcome distribution is both the prediction for new data realisation, and subjective belief about an unobserved realization (which is no longer random, so frequentist cannot use probabilities). The prior predictive outcome distribution. The posterior predictive outcome distribution.

12 note

Frequentists attempt to interpret the estimates of the model, which is difficult except when the model is linear, has no inverse link function, and contains no interaction terms. Bayesians can avoid this difficulty simply by inspecting the posterior predictive distribution at different levels of the predictors.

13 Strong versus weak control

There are two popular hierarchical model classes for data with sparse and exchangeable signal (Carvalho horseshoe). The first models true sparsity, where

$u_i = 0$ for many of $i \in I$, by assuming that each u_i is a mixture of a discrete point mass at $u_i = 0$ and an absolutely continuous alternative. The second class models a weaker form of sparsity, where $u_i \approx 0$ for many $i \in I$, by giving the u_i an absolutely continuous "shrinkage" prior, centered at zero. This latter includes the LASSO, relevance vector machine, horseshoe and standard "global shrinkage" PPM.

Numerical problems are common when estimating models from data with extremely sparce signal, such as under the FWN.

14 Critique of global shrinkage PPM implementation

The ReML estimates of the (hyper)parameters have precision linear in the number of voxels (Friston and Penny 2003). Global shrinkage models observation error twice over. Once in Equ 4 of (Friston and Penny 2003) which is voxel-specific, once in Equ 7 of (Friston and Penny 2003) which "can be treated as an estimate of the average [scale of observation error] over voxels". These are combined into the approximation in Equ 10 of (Friston and Penny 2003). It is not stated whether Equ 10 shrinks or regularizes voxel-specific noise parameter estimation.

The default arguments of a call to $spm_PEB.m$ from $spm_{s}pm_{B}ayes.m$ for the purposes of second level global shrinkage PPM, permits negative error variance h. PPM seems to overfit the observation noise parameters, there is at least one per voxel. $spm_{s}pm_{B}ayes.m$ then checks for them and "corrects" with $h = abs(h)$, which is dubious. All "random effect" models make use of both fixed and random constraints. For example, a random intercept model typically assumes the error variance across component nested regressions is fixed, but regularizes the intercept. Model unequal variance parameters as random effects.

15 Critique of global shrinkage PPM model

Independent population prior on β_0^i, β_1^i means voxel-specific intercept and slope are unrelated. If false,

Traditional regression. Called between-voxel homoskedasticity in regression (different error variance) or homogeneity of variance assumption in ANOVA.

Heterogeneous noise and signal (i.e. both the scale of predictor coefficients and noise variance vary across voxels).

Three problems with PPM. Endogeneity from misspecification (residual dependence after modeling exchangeable voxel effects). Despite the model assumptions, the random noise errors are not conditionally independent across voxels, given random effects. Rather there are residual spatial correlations. Despite model assumptions, random effects may be correlated a priori. Nested model (subjects within voxels), but not nested data. The commonly tauted advantage is that submodels (1st level, nested regressions) with low-data benefit from sharing strength. In fMRI, data is balanced, with no voxel having more or less data. This means

Decision theory

Optimal responses to an uncertain world rest on first defining our model of unknowns in the world - i.e. out-of-sample or fundamentally unobservable parameters/states - and then defining a model of the cost of taking each available action, given these unknowns. Under these assumptions, an optimal *policy* is typically defined to be one which minimises the expected utility loss, a.k.a. the "risk". In learning/engineering tasks, for example, we often want to *track* some unknown quantity, i.e. want decisions to match or fit an unknown quantity. The cost of mismatch is formalized by a "decision-utility function", e.g. expected squared prediction error loss (ESPEL) as below, a function which quantifies the (conditional) cost of each specific response given the fixed or random unknown quantity. (These two cases are called estimation and prediction respectively in statistics).

Thus a policy is a mapping from observable data to actions, known variously as a decision-rule, estimator, test, predictor, fitting-procedure/algorithm, statistic, etc. Frequentist statisticians try to devise good functions to identify and quantify hidden quantities of the world, assuming some model. These may be real-valued functions of the data (estimators), set-valued functions (confidence intervals or regions), binary-valued ("test") functions, constructed to meet some desiderata. Statistically optimal policies invariably impose some constraints to adjust, penalize or correct for over-estimating, i.e. to reduce the probability of being "far" from the truth (estimators with small bias/variance) or "false" (tests small false-positive/negative).

Expected squared prediction error loss (ESPEL) exposes a general, famous trade-off between two sources of risk in any policy toward unknown parameters, states or out-of-sample predictions: "statistical bias" and "statistical variance", the first two central moments of the squared signal-prediction error. We ideally want zero statistical bias and variance. The trade-off says that, for any given risk, one can choose either but not both. "Statistical bias" captures the regular or predictable error over imagined replications of the data which arises due to a modeling restriction. As explained in the text, this "statistical bias" relates intimately to "inductive bias" which helps extreme over-estimating by adaptive parameter shrinking. In contrast, statistical variance quantifies unpredictable, over-estimating error due to noise. The conventional school of inference typically takes a different view on this trade-off. Given the trade-off they prize unbiasedness and only then attempt to derive the minimum variance estimator. By pretending to circumvent inductive bias in inference they forgo the full power of hierarchical control of over-estimating.

"Analytic" risk minimisation above differs from the "Empirical risk minimization" (in supervised learning models). This philosophy here is to approximate true risk by "empirical risk", i.e. the average loss function over some training set. The validity of this approach rests on correctly understanding the statistical properties of these cross-validation estimators of the true risk. We do not expand on these here.

16 Notes

Testing or regularization of random and/or fixed parameters to minimizing noise in multivariate signal prediction/estimation. The non-hierarchical fixed-effects approach is to compare nested models, with versus without some functional restriction on signal. This comparison is possible for RFX models too (compare nested models with versus without some restriction on the variance component parameters). In addition,

Identifiability problems whenever one random factor has zero variance (i.e. the null).

Optimal estimation and prediction entails stopping, shrinking, smoothing or filtering out noise variability. We perform an omnibus inference on the global null, by inferring the covariance component parameter which governs between-voxel random effects, e.g. in PMM with global-shrinkage prior. This substitutes a single comparison for multiple comparisons (akin to using a single F-statistic hypothesis test in linear fixed-effects models like ANOVA).

In PPM, voxels are indexed by the level of a categorical variable. PPM is a 2-level hierarchical model, with observations/subjects nested within voxels.

Despite PPM, subjects are crossed with - and not nested within - voxels. This model is not a "hierarchical" in Douglas Bates' terminology (but it is a multilevel).

A common example of wide linear models is when there are many exchangeable groups of observations, therefore parameters.

Regularized causal effect estimates. How to formulate and justify the regularization constraint.

Continuous or grouping (discrete, categorical) variable .

The design shapes the cost function/posterior over unknowns, i.e. the existence and curvature of a uniquely identifiable optima in the cost function over unknowns. Deterministic (linear) or statistical dependence (co-linear or correlated features) and the ratio of unknowns to observations. Co-linear or wide ($P > N$) designs imply there is no single unique optima, only a "ridge" of solutions. Correlated or sparsely observed features (e.g. $P/N \approx 1$) imply the unique optima is broad (large $\mathbb{V}(\hat{\beta})$). An inductive bias on the loss function - i.e. prior or "regularization" constraints on P unknowns, such as some pre-specified unknowns are zero or equivalently omit variables so $N < P$, all unknowns should shrink towards zero (or some constant, c.f. Bridge regression), solutions must be orthogonal to non-principle components of the design (c.f. principle components regression), all unknowns are correlated *a posteriori*, etc - ensures a well-posed optimization problem and reduces variance because there are more observations per parameter, but increases risk of under-fitting mis-specification bias. I.E. regularization bias reduces the number and breadth of optima to one precise optima, at cost of underfitting misspecification bias. With $N = P$ non-colinear features, the parameters fit the data perfectly, e.g. the data *are* the parameters. As P/N shrinks by inductive bias, $under-fitting\ mis-specification$ bias increases but variance. Regularize an ill-posed and/or imprecise optimization problem. The omission inductive bias/constraint (unknown is prespecified to be zero) on the loss function only creates well-posed optimization if there is no remaining deterministic linear independence amoung included features. Pre-specified omission risks omitted feature bias in the right graph. Feature omission (zero weight/effect variable) constraints can induce bias, but so can feature inclusion (Pearl's selection bias). Omitted confounds is not the only source of bias. If loss is a function of unknown *causal* parameters, can use Pearl's backdoor graphical criterion etc, to guide the variable/effect omission (pre-specified zeros) that do not incur bias, e.g. Simpson's causal paradox. For Pearl SEM-based "identifiablity" is not just due to the ratio of observations to unknowns or deterministic/statistical dependence among predictive features, but due to the open back-door paths with respect to the data generating process (omitted confounds, included colliders and/or cyclic causation). Also, increasing number of features/weights can reduce unexplained error variance, hence estimator variance (ceteris paribus, a.k.a. blocking).

Both frequentists/Bayesians aim to make predictions and to estimate parameters (where "parameters" are those aspects of a model that generalize to new data) Gelman's blog. JC: in my terms they are general or shared by all current and future indices.

Gelman: lasso (by itself, and as an inspiration for Bayesian regularization) is a great idea.

identifiability and shrinkage via constraints (priors or penalties, which in turn can be learned by full Bayesian, empirical Bayesian or cross-validation methods)

regression coefficients getting out of control

I could go around smugly thinking that lasso was a trivial implementation of a prior distribution, coupled with a silly posterior-mode summary.

Discrete/continuous-smooth shrinkage of random parameters/states (alternatively, Lagrange constrained on regression). PCR and kernel PCR have a discrete shrinkage effect on the eigenvectors of. Contrast with Ridge regression.

Beware: correlated factors and/or failure to block on important factors both increase estimator variance. "Extreme distance": Hoeffding bounds, permutation p-values, parametric tests. Adaptive tests - e.g. Benjamini-Hochberg, and FWER analogies - adaptive estimation. Different thresholds + different interpretations for uncorrected, FDR, FWER.! Resolve endogeneity then resolve multiplicity. If assumptions are strong, uncertainty wrong, null is wrong (hence tests/p-values).

Rule of thumb: simple OLS ignoring/marginalizing over voxel-index (i.e. concatenate all voxel-specific data) is significant for a null effect, we have a problem!

Non-identifiable (e.g. infinite solutions because wide or co-linear design) or identified but noisy (non-orthogonal/correlated regressors).

This "prior", introduced before any access to the data, reflects our prior knowledge on u: general qualitative symmetries (e.g. exchangeability, which may modeled graphically) and quantitative parametric class assumptions such Gaussian (together with uncertainty about the "hyper"-parameters). Posterior *distribution* approximations can also rest on qualitative (e.g. mean-field) and quantitative (e.g. Gaussian) assumptions. Posterior *point* approximation via MAP.

Lasso $\hat{u}(\lambda) = \underset{u}{argmin}\ \frac{1}{N}\sum_{i=1}^{N}(y_i - (u_0 + x_i u))^2 + \lambda\|u\|_1$

Ridge $\hat{u}(\lambda) = \underset{u}{argmin}\ \frac{1}{N}\sum_{i=1}^{N}(y_i - (u_0 + x_i u))^2 + \lambda\|u\|_2$

Because the log prior is a summation, the $u_1, ..., u_I$ are mutually independent and identically distributed (white Laplacian or Gaussian process). We thus omit the subscript i and study the prior distribution of u (c.f. the bayesian bridge with R package *BayesBridge*). In spatial models, the $u_1, ..., u_I$ are not independent, but spatially dependent. Parsimony of lasso: only a small subset of features selected increases the interpretability (especially conjugate with parcels). If n ¿ m, i.e. more features than samples, Lasso will select at most m variables. More difficult to implement than Ridge Regression.

Elasticnet (i.e. laplace * Gaussian prior, i.e. first and second bridge (or powered exponential) takes care of groups of correlated variables (and therefore effects) by shrinking together, but also selects them out. Selection or shrinkage?

R package *EBEN* provides Empirical Bayesian Elastic Net (for wide, collinear and correlated designs).

"Hyperprior" is a prior on the "hyperparameters", in Cressie this simplifies to "prior on parameters", all other effects are random.

R package *EBglmnet* provides empirical Bayesian lasso and elastic net algorithms for variable selection and effect estimation (see *glmnet* to learn hyperparameters by cross-validation instead).

Almost nobody calls it "lasso." Lasso say (equivalently in terms of estimator) that they're doing MAP with double-exponential (Laplace) priors. The electrical engineers are all over L1 and call it "compressed sensing" for its motivating application in their world. Even the people fitting SVMs do it, and they rarely even talk about probability.

Ridge regression was invented - and named - as a solution to non-identifiability (ill-conditioned - e.g. "wide" or co-linear - designs). Its impact on $\mathbb{V}(\hat{\beta}^R)$ is what concerns us. [11]

standard errors for lasso and elastic net coefficients.

Figure: Binary nested model comparison by likelihood-ratio, or deviance difference $dd(l_1, l_2) = dd(l_1/l_2) = -2log(l_1/l_2)$ whose distribution in known $(1, 2$ index nested constraints). If deviance $d(l_1, l_g) = -2log(l_1/l_g)$ then $dd(l_1, l_2) = d(l_1, l_g) - d(l_2, l_g)$. Fraction of deviance explained: as this approaches 1, we are perfectly overfitting.

Elasticnet provides identifiability via the L2 (i.e., "ridge") component and shrinkage to 0 in point estimates via the L1 (i.e., "lasso") component. Regarding identifiability, if you have two predictors x1 and x2 with perfect correlation, L2 regularization identifies a unique result, whereas with L1 regularization any

[11] With multicollinearity, you get a "ridge" in the likelihood function (i.e. a trough in the Loss function such as RSS). Ridge regression "fixes" the ridge. The term "ridge" also relates to the main diagonal λI in $\hat{\beta}^R = (X^T X + \lambda I)^{-1} X^T Y$ which permits a solution to this inverse matrix. $\hat{\beta}^R$ is also the Bayesian MAP or posterior mean.

interpolation between b * x1 and b * x2 (e.g. 0.3 * b + x1 + 0.7 * b * x2) has the same likelihood and the same L1 penalty (namely b). But the L2 penalty is lowest when the two coefficients are equal (lambda * b * x1 + lambda * b * x2). So you get identification.

Coefficients shrunk to exactly zero vastly reduces both the memory needed to store the model, and the time taken to compute a prediction. Whenever we set up a model with a finite number of variables we are already setting a lot of other parameters to zero! So don't critique lasso in this way. We all make binary distinctions all of the time (in/out of model, connected/not connected), so the criticism that DAGs take a binary approach to causality (e.g. either two variables are connected or not) is neither here nor there. We all do this all of the time. A DAG is about laying out explicitly the asumptions behind the joint probability model. Other than that they are completely non-parametric which is to say, compatible with Bayesian estimation and partial pooling.

We are always making approximations and shortcuts. I find that, the cruder my inferential tools, the quicker I am to reach for the shortcut. That's why, in my everyday regression work, I want to move toward routine use of partial pooling, which I think will allow me to more easily include more predictors. That's the problem with the whole "noninformative priors" thing. If you don't have prior information in your prior distribution, you often have to put it in somewhere else, and in a cruder way, such as by preselecting predictors as in or out.

I don't think any of the true coefficients ("true" in the sense of being the values I would get if I could interview the entire population) would be exactly zero. But lots of them would be close to zero, which is the point in practice. In the sorts of problems I work on, there's nothing really special about zero, but there is a need to regularize our estimates. Lasso has proved very useful for this.

Gelman: L2 isn't so great because it shrinks by a uniform percentage, thus doing too much shrinkage of big values and not enough shrinkage of small values. To put it another way, I think the true distribution of coefficients is long-tailed, that's why I think the Cauchy shrinkage worked so well in my paper with Aleks.

predict, explain, or intervene. Gelman: Oooooh, I hate that "spike and slab." Especially the slab. The key to lasso is that it partially pools in a continuous manner (see graph at top of post, which was taken from Tibshirani's slides). It's no slab.

Bridge regression: OLS (= 0), lasso (= 1) and ridge (= 2),

Sampling process constrains the covariance pattern. Domain constraints are as good as domain knowledge.