A general framework for multiple testing dependence

Jeffrey T. Leek[a] and John D. Storey[b,1]

[a]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21287; and [b]Lewis-Sigler Institute and Department of Molecular Biology, Princeton University, Princeton, NJ 08544

We develop a general framework for performing large-scale significance testing in the presence of arbitrarily strong dependence. We derive a low-dimensional set of random vectors, called a dependence kernel, that fully captures the dependence structure in an observed high-dimensional dataset. This result shows a surprising reversal of the "curse of dimensionality" in the high-dimensional hypothesis testing setting. We show theoretically that conditioning on a dependence kernel is sufficient to render statistical tests independent regardless of the level of dependence in the observed data. This framework for multiple testing dependence has implications in a variety of common multiple testing problems, such as in gene expression studies, brain imaging, and spatial epidemiology.

empirical null | false discovery rate | latent structure | simultaneous inference | surrogate variable analysis

In many areas of science, there has been a rapid increase in the amount of data collected in any given study. This increase is due in part to the ability to computationally handle large datasets and the introduction of various high-throughput technologies. Analyzing data from such high-dimensional studies is often carried out by performing simultaneous hypothesis tests for some behavior of interest, on each of thousands or more measured variables. Large-scale multiple testing has been applied in fields such as genomics (1–3), astrophysics (4, 5), brain imaging (6–8), and spatial epidemiology (9). By their very definition, high-dimensional studies rarely involve the analysis of independent variables, rather, many related variables are analyzed simultaneously. However, most statistical methods for performing multiple testing rely on independence, or some form of weak dependence, among the data corresponding to the variables being tested. Ignoring the dependence among hypothesis tests can result in both highly variable significance measures and bias caused by the confounding of dependent noise and the signal of interest.

Here, we develop an approach for addressing arbitrarily strong multiple testing dependence at the level of the original data collected in a high-dimensional study, before test statistics or P values have been calculated. We derive a low-dimensional set of random vectors that fully captures multiple testing dependence in any fixed dataset. By including this low-dimensional set of vectors in the model-fitting process, one may remove arbitrarily strong dependence resulting in independent parameter estimates, test statistics, and P values. This result represents a surprising reversal of the "curse of dimensionality" (10), because of the relatively small sample size in relation to the large number of tests being performed. Essentially, we show that the manifestation of the dependence cannot be too complex and must exist in a low-dimensional subspace of the data, driven by the sample size rather than by the number of hypothesis tests. This approach provides a sharp contrast to currently available approaches to this problem, such as the estimation of a problematically large covariance matrix, the conservative adjustment of P values, or the empirical warping of the test statistics' null distribution.

The main contributions of this article can be summarized as follows. We provide a precise definition of multiple testing dependence in terms of the original data, rather than in terms of P values or test statistics. We also state and prove a theoretical result showing how to account for arbitrarily strong dependence

among multiple tests; no assumptions about a restricted dependence structure are required. By exploiting the dimensionality of the problem, we are able to account for dependence on each specific dataset, rather than relying on a population-level solution. We introduce a model that, when fit, makes the tests independent for all subsequent inference steps. Utilizing our framework allows all existing multiple testing procedures requiring independence to be extended so that they now provide strong control in the presence of general dependence. Our general characterization of multiple testing dependence directly shows that latent structure in high-dimensional datasets, such as population genetic substructure (11) or expression heterogeneity (12), is a special case of multiple testing dependence. We propose and demonstrate an estimation technique for implementing our framework in practice, which is applicable to a large class of problems considered here.

Notation and Assumptions

We assume that m related hypothesis tests are simultaneously performed, each based on an n-vector of data sampled from a common probability space on $\mathbb{R}^n$. The data corresponding to hypothesis test i are $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, for $i = 1, 2, \ldots, m$. The overall data can be arranged into an $m \times n$ matrix $\mathbf{X}$ where the ith row is composed of $\mathbf{x}_i$. We assume that there are "primary variables" $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ collected, describing the study design or experimental outcomes of interest, and any other covariates that will be employed. Primary variables are those that are both measured and included in the model used to test the hypotheses.

We assume that the goal is to perform a hypothesis test on $E[\mathbf{x}_i|\mathbf{Y}]$. We will also assume that $E[\mathbf{x}_i|\mathbf{Y}]$ can be modeled with a standard basis-function model, which would include linear models, nonparametric smoothers, longitudinal models, and others. To this end, we write $E[\mathbf{x}_i|\mathbf{Y}] = \mathbf{b}_i\mathbf{S}(\mathbf{Y})$, where $\mathbf{b}_i$ is a $1 \times d$-vector and $\mathbf{S}(\mathbf{Y})$ is a $d \times n$ matrix of basis functions evaluated at $\mathbf{Y}$ $d < n$. When there is no ambiguity, we will write $\mathbf{S} = \mathbf{S}(\mathbf{Y})$ to simplify notation. Note that $\mathbf{Y}$ can be composed of variables such as time, a treatment, experimental conditions, and demographic variables. The basis $\mathbf{S}$ can be arbitrarily flexible to incorporate most of the models commonly used in statistics for continuous data.

The residuals of the model are then $\mathbf{e}_i = \mathbf{x}_i - E[\mathbf{x}_i|\mathbf{Y}] = \mathbf{x}_i - \mathbf{b}_i\mathbf{S}$. Analogously, we let $\mathbf{E}$ be the $m \times n$ matrix, where the ith row is $\mathbf{e}_i$. We make no assumptions about what distribution the residuals follow, although by construction $E[\mathbf{e}_i|\mathbf{S}(\mathbf{Y})] = \mathbf{0}$. We allow for arbitrary dependence across the tests, i.e., dependence across the rows of $\mathbf{E}$. We assume that the marginal model for each $\mathbf{e}_i$ is known or approximated sufficiently when performing the hypothesis tests. That is, we assume that the marginal null model for each test is correctly specified.

Fig. 1. A schematic of the general steps of multiple hypothesis testing. We directly account for multiple testing dependence in the model-fitting step, where all the downstream steps in the analysis are not affected by dependence and have the same operating characteristics as independent tests. Our approach differs from current methods, which address dependence indirectly by modifying the test statistics, adaptively modifying the null distribution, or altering significance cutoffs. For these downstream methods the multiple testing dependence is not directly modeled from the data, so distortions of the signal of interest and the null distribution may be present regardless of which correction is implemented.

In matrix form, the model can be written as

$$\mathbf{X} = \mathbf{BS} + \mathbf{E}. \qquad [1]$$

The goal is then to test m hypotheses of the form:

$$H_{0i} : \mathbf{b}_i \in \Omega_0 \quad \text{vs.} \quad H_{1i} : \mathbf{b}_i \in \Omega_1$$

where the null and alternative hypothesis tests are identically defined for each of the tests. This setup encompasses what is typically employed in practice, such as in gene expression studies and other applications of microarrays, brain imaging, spatial epidemiology, astrophysics, and environmental modeling (4, 7, 9, 13, 14).

Two Open Problems

The classical approach to testing multiple hypotheses is to first perform each test individually. This involves calculating a 1-dimensional statistic for each test, usually as some comparison of the model fit under the constraint of the null hypothesis to that under no constraints. By utilizing the observed test statistics and their null distributions, we calculate a P value for each test (15). An algorithm or point estimate is then applied to the set of P values to determine a significance threshold that controls a specific error measure at a specific level (16), such as the false discovery rate (FDR) at 10% (17, 18). Variations on this approach have been suggested, such as estimating a q-value for each test (19) or a posterior error probability (20). Regardless of the approach, the validity and accuracy of these procedures are essentially determined by whether the null distributions are correctly specified (or conservatively specified) and whether the data are independent (or weakly dependent) across tests (21, 22).

Two open problems in multiple testing have received a lot of recent attention. The first is concerned with controlling multiple testing error measures, such as the FDR, in the presence of dependence among the P values (23, 24). This dependence is usually formulated as being present in the "noise" component of the models used to obtain the P values. The second open problem is concerned with the fact that latent structure among the tests can distort what would usually be the correct null distribution of the test statistics (11, 25–27). The approach proposed here shows that

both problems actually stem from sources of variation that are common among tests, which we show is multiple testing dependence, and both problems can be simultaneously resolved through one framework.

The current paradigm for addressing these two problems can be seen in Fig. 1, where the steps taken to get from the original data $\mathbf{X}$ to a set of significant tests are shown. It can be seen that existing approaches are applied far downstream in the process. Specifically, adjustments are performed after 1-dimensional summaries of each test have been formed, either to the test statistics or P values. As we show below, the information about noise dependence and latent structure is found in the original data $\mathbf{X}$ by modeling common sources of variation among tests. Our proposed approach addresses multiple testing dependence (from either noise dependence or latent structure) at the early model-fitting stage (Fig. 1), at which point the tests have been made stochastically independent and the null distribution is no longer distorted.

Proposed Framework

Definition of Multiple Testing Dependence. Multiple testing dependence has typically been defined in terms of P values or test statistics resulting from multiple tests (21, 24, 26, 28, 29). Here, we form population-level and estimation-level definitions that apply directly to the full dataset, $\mathbf{X}$. The estimation-level definition also explicitly involves the model assumption and fit utilized in the significance analysis. When fitting model **1**, we denote the estimate of $\mathbf{B}$ by $\widehat{\mathbf{B}}$.

Definition: We say that population-level multiple testing dependence exists when it is the case that:

$$\Pr(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m | \mathbf{Y}) \neq \Pr(\mathbf{x}_1 | \mathbf{Y}) \times \Pr(\mathbf{x}_2 | \mathbf{Y}) \times \cdots \times \Pr(\mathbf{x}_m | \mathbf{Y}).$$

We say that estimation-level multiple testing dependence exists when it is the case that:

$$\Pr(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m | \widehat{\mathbf{B}}, \mathbf{S}(\mathbf{Y})) \neq \Pr(\mathbf{x}_1 | \widehat{\mathbf{B}}, \mathbf{S}(\mathbf{Y})) \times \cdots \times \Pr(\mathbf{x}_m | \widehat{\mathbf{B}}, \mathbf{S}(\mathbf{Y})).$$

Multiple testing dependence at the population level is therefore any probabilistic dependence among the $\mathbf{x}_i$, after conditioning on $\mathbf{Y}$. In terms of model **1**, this is equivalent to the existence of

dependence across the rows of $\mathbf{E}$; i.e., dependence among the $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$. Estimation-level dependence is equivalent to dependence among the rows of the residual matrix $\mathbf{R} = \mathbf{X} - \widehat{\mathbf{B}}\mathbf{S}$. It will usually be the case that if population-level multiple testing dependence exists, then this will lead to estimation-level multiple testing dependence. The framework we introduce in this article is aimed at addressing both types of multiple testing dependence.

A General Decomposition of Dependence. Dependence among the rows of $\mathbf{E}$ and among the rows of $\mathbf{R} = \mathbf{X} - \widehat{\mathbf{B}}\mathbf{S}$ are types of multivariate dependence among vectors. The standard approach for modeling multivariate dependence is to estimate a population-level parameterization of the dependence and then include estimates of these parameters when performing inference (30). For example, if the $\mathbf{e}_i$ are assumed to be Normally distributed with the columns of $\mathbf{E}$ being independently and identically distributed random m-vectors, then one would estimate the $m \times m$ covariance matrix which parameterizes dependence across the rows of $\mathbf{E}$. One immediate problem is that because $n \ll m$, it is computationally and statistically problematic to estimate the covariance matrix (31).

A key feature is that, in the multiple testing scenario, the dimension along which the sampling occurs is different than the dimension along which the multivariate inference occurs. In terms of our notation, the sampling occurs with respect to the columns of $\mathbf{X}$, whereas the multiple tests occur across the rows of $\mathbf{X}$. This sampling-to-inference structure requires one to develop a specialized approach to multivariate dependence that is different from the classical scenarios. For example, the classical construction and interpretation of a P value threshold is such that a true null test is called significant with P value $\leq \alpha$ at a rate of α over many independent replications of the study. However, in the multiple testing scenario, the P values that we utilize are not P values corresponding to a single hypothesis test over m independent replications of the study. Rather, the P values result from m related variables that have all been observed in a single study from a single sample of size n. The "sampling variation" that forms the backbone of most statistical thinking is different in our case: we observe one instance of sampling variation among the variables being tested. Therefore, even if each hypothesis test's P value behaves as expected over repeated studies, the set of P values from multiple tests in a single study will not necessarily exhibit the same behavior. Whereas this phenomenon prevents us from invoking well-established statistical principles, such as the classical interpretation of a P value, the fact that we have measured thousands of related variables from this single instance of sampling variation allows us to capture and model the common sources of variation across all tests. Multiple testing dependence is variation that is common among hypothesis tests.

Thus, rather than proposing a population-level approach to this problem (which includes the population of all hypothetical studies that could take place in terms of sampling of the columns of $\mathbf{X}$), we directly model the random manifestation of dependence in the observed data from a given study, by aggregating the common sampling variation across all tests' data. Including this information in the model during subsequent significance analyses removes the dependence within the study. Therefore dependence is removed across all studies, providing study-specific and population-level solutions. To directly model the random manifestation of dependence in the observed data, we do the following: (*i*) additively partition $\mathbf{E}$ into dependent and independent components, (*ii*) take the singular value decomposition of the dependent component, and (*iii*) treat the right singular values as covariates in the model fitting and subsequent hypothesis testing. To this end, we provide the following result, which shows that any dependence can be additively decomposed into a dependent component and an independent component. It is important to note that this is both for an arbitrary distribution for $\mathbf{E}$ and an arbitrary (up to degeneracy) level of dependence across the rows of $\mathbf{E}$.

Proposition 1. *Let the data corresponding to multiple hypothesis tests be modeled according to Eq. 1. Suppose that for each $\mathbf{e}_i$, there is no Borel measurable function g such that $\mathbf{e}_i = g(\mathbf{e}_1, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m)$ almost surely. Then, there exist matrices $\mathbf{\Gamma}_{m \times r}$, $\mathbf{G}_{r \times n}$ ($r \leq n$), and $\mathbf{U}_{m \times n}$ such that*

$$\mathbf{X} = \mathbf{B}\mathbf{S} + \mathbf{\Gamma}\mathbf{G} + \mathbf{U}, \qquad [2]$$

where the rows of $\mathbf{U}$ are jointly independent random vectors so that

$$\Pr(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m) = \Pr(\mathbf{u}_1) \times \Pr(\mathbf{u}_2) \times \cdots \times \Pr(\mathbf{u}_m).$$

Also, for all $i = 1, 2, \ldots, m$, $\mathbf{u}_i \neq \mathbf{0}$ and $\mathbf{u}_i = h_i(\mathbf{e}_i)$ for a non-random Borel measurable function h_i.

A formal proof of *Proposition 1* and all subsequent theoretical results can be found in the supporting information (SI) *Appendix*. Note that if we let $r = n$ and then set $\mathbf{U} = \mathbf{0}$ or set $\mathbf{U}$ equal to an arbitrary $m \times n$ matrix of independently distributed random variables, then the independence of the rows of $\mathbf{U}$ is trivially satisfied. However, our added assumption regarding $\mathbf{e}_i$ allows us to show that a nontrivial $\mathbf{U}$ exists where $\mathbf{u}_i \neq \mathbf{0}$ and $\mathbf{u}_i = h_i(\mathbf{e}_i)$ for a deterministic function h_i. In other words, $\mathbf{u}_i$ is a function of $\mathbf{e}_i$ in a nondegenerate fashion, which means that $\mathbf{U}$ truly represents a row-independent component of $\mathbf{E}$. The intuition behind these properties is that our assumption guarantees that $\mathbf{e}_i$ does indeed contain some variation that is independent from the other tests. For hypothesis tests where there *does* exist a Borel measurable g such that $\mathbf{e}_i = g(\mathbf{e}_1, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m)$, then the variation of $\mathbf{e}_i$ is completely dependent with that of the other tests' data. In this case, one can set $\mathbf{u}_i = \mathbf{0}$ and the above decomposition is still meaningful.

The decomposition of *Proposition 1* immediately indicates one direction to take in solving the multiple testing dependence problem, namely to account for the $\mathbf{\Gamma}\mathbf{G}$ component, thereby removing dependence. To this end, we now define a "dependence kernel" for the data $\mathbf{X}$.

Definition: An $r \times n$ matrix $\mathbf{G}$ forms a **dependence kernel** for the high-dimensional data $\mathbf{X}$, if the following equality holds:

$$\mathbf{X} = \mathbf{B}\mathbf{S} + \mathbf{E}$$
$$= \mathbf{B}\mathbf{S} + \mathbf{\Gamma}\mathbf{G} + \mathbf{U}$$

where the rows of $\mathbf{U}$ are jointly independent as in *Proposition 1*.

In practice, one would be interested in minimal dependence kernels, which are those satisfying the above definition and having the smallest number of rows, r. *Proposition 1* shows that at least one such $\mathbf{G}$ exists with $r \leq n$ rows. As we discuss below in *Scientific Applications*, the manner in which one incorporates additional information beyond the original observations to estimate and utilize $\mathbf{\Gamma}$ and $\mathbf{G}$ is context specific. In the *SI Appendix*, we provide explicit descriptions for two scientific applications, latent structure as encountered in genomics and spatial dependence as encountered brain imaging. We propose a new algorithm for estimating $\mathbf{G}$ in the genomics application and demonstrate that it has favorable operating characteristics.

Dependence Kernel Accounts for Dependence. An important question arises from *Proposition 1*. Is including $\mathbf{G}$, in addition to $\mathbf{S(Y)}$, in the model used to perform the hypothesis tests sufficient to remove the dependence from the tests? If this is the case, then only an $r \times n$ matrix must be known to fully capture the dependence. This is in contrast to the $m(m-1)/2$ parameters that must be known for a covariance matrix among tests, for example. To put this into context, consider a microarray experiment with 1,000 genes and 20 arrays. In this case, the covariance has ~500,000 unknown parameters, whereas $\mathbf{G}$ has, at most, 400 unknown values. The following two results show that including $\mathbf{G}$ in addition to

$S(Y)$ in the modeling is sufficient to remove all multiple hypothesis testing dependence.

Corollary 1. *Under the assumptions of* Proposition 1, *all population-level multiple testing dependence is removed when conditioning on both* **Y** *and a dependence kernel* **G***. That is,*

$$\Pr(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m | \mathbf{Y}, \mathbf{G})$$
$$= \Pr(\mathbf{x}_1 | \mathbf{Y}, \mathbf{G}) \times \Pr(\mathbf{x}_2 | \mathbf{Y}, \mathbf{G}) \times \cdots \times \Pr(\mathbf{x}_m | \mathbf{Y}, \mathbf{G}).$$

If instead of fitting model **1**, suppose that we instead fit the decomposition from *Proposition 1*, where we assume that **S** and **G** are known:

$$\mathbf{X} = \mathbf{BS} + \boldsymbol{\Gamma}\mathbf{G} + \mathbf{U}. \qquad [3]$$

It follows that estimation-level multiple testing independence may then be achieved.

Proposition 2. *Assume the data for multiple tests follow model* **1***, and let* **G** *be any valid dependence kernel. Suppose that model* **3** *is fit by least squares, resulting in residuals* $\mathbf{r}_i = \mathbf{x}_i - \widehat{\mathbf{b}}_i\mathbf{S} - \widehat{\boldsymbol{\gamma}}_i\mathbf{G}$*. When the row space jointly spanned by* **S** *and* **G** *has dimension less than n, the residuals* $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m$ *are jointly independent given* **S** *and* **G***, the* $\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \ldots, \widehat{\mathbf{b}}_m$ *are jointly independent given* **S** *and* **G***, and*

$$\Pr(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m | \widehat{\mathbf{B}}, \mathbf{S}, \widehat{\boldsymbol{\Gamma}}, \mathbf{G})$$
$$= \Pr(\mathbf{x}_1 | \widehat{\mathbf{B}}, \mathbf{S}, \widehat{\boldsymbol{\Gamma}}, \mathbf{G}) \times \cdots \times \Pr(\mathbf{x}_m | \widehat{\mathbf{B}}, \mathbf{S}, \widehat{\boldsymbol{\Gamma}}, \mathbf{G}).$$

The analogous results hold for the residuals and parameter estimates when fitting the model under the constraints of the null hypothesis.

Since **G** will be unknown in practice, the practical implication of this proposition is that we have to estimate only the relatively small $r \times n$ matrix **G** well in order to account for all of the dependence, while the simple least-squares solution to $\boldsymbol{\Gamma}$ suffices. When the row space jointly spanned by **S** and **G** has dimension equal to n, then the above proposition becomes trivially true. However, if we assume that **S**, **G**, and $\boldsymbol{\Gamma}$ are known, then the analogous estimation-level independence holds. In this case, we have to estimate $\boldsymbol{\Gamma}$ and **G** well in order to account for dependence. These $(m + r)n$ parameters are still far smaller than the unknown $m(m - 1)/2$ parameters of a covariance matrix, for example.

Strong Control of Multiple Testing Error Rates. Many methods exist for strongly controlling the family-wise error rate (FWER) or FDR (16, 18, 19, 21, 24, 32). These methods are applied to the P values calculated from multiple hypothesis tests. Most of these methods require the P values corresponding to true null hypotheses to be independent in order for the procedure to provide strong control. For example, finite-sample strong control of several FDR procedures (21, 24) and the conservative point estimation of the FDR (19) all require the true null P values to be independent. Several methods exist for controlling FWER or FDR when dependence is present. However, these either tend to be quite conservative or require special restrictions on the dependence structure (21, 24).

When utilizing model **3**, the statistics formed for testing the hypothesis should be based on a function of the model fits and residuals. When this is the case, we achieve the desired independence of P values.

Corollary 2. *Suppose that the assumptions of* Proposition 2 *hold, model* **3** *is utilized to perform multiple hypothesis tests, and* **G** *is a known dependence kernel. If P values are calculated from test statistics based on a function of the model fits and residuals, then the resulting P values and test statistics are independent across tests.*

In other words, *Corollary 2* extends all existing multiple testing procedures that have been shown to provide strong control

when the null P values are independent to the general dependence case. Instead of deriving new multiple testing procedures for dependence at the level of P values, we can use the existing ones by including **G** into the model fitting and inference carried out to get the P values themselves.

Scientific Applications

Two causes for multiple testing dependence can be directly derived from scientific problems of interest. In each case, the dependence kernel **G** has a practical scientific intepretation.

Spatial Dependence. Spatial dependence usually arises as dependence in the noise because of a structural relationship among the tests. In this case, we will consider the $\mathbf{e}_i$ of model **1** to simply represent "noise," an example being the spatial dependence for noise that is typically assumed for brain-imaging data (6–8). In this setting, the activity levels of thousands of points in the brain are simultaneously measured, where the goal is to identify regions of the brain that are active. A common model for the measured intensities is a Gaussian random field (6). It is assumed that the Gaussian noise among neighboring points in the brain are dependent, where the covariance between two points in the brain is usually a function of their distance.

In Fig. 2 *A* and *B*, we show two datasets generated from a simplified 2-dimensional version of this model. It can be seen that the manifestation of dependence changes notably between the two studies, even though they come from the same data generating distribution. Using model **3** for each dataset, we removed the $\boldsymbol{\Gamma}\mathbf{G}$ term. In both cases, the noise among points in the 2-dimensional space becomes independent and the P value distributions of points corresponding to true null hypotheses follow the Uniform distribution. It has been shown that null P values following the $Uniform(0, 1)$ distribution is the property that confirms that the assumed null distribution is correct (22). Additionally, it can be seen that the null P values from the unadjusted data fluctuate substantially between the two studies, and neither follows the $Uniform(0, 1)$ null distribution. This is due to varying levels of correlation between **S** and **G** from model **3**. In one case, **S** and **G** are correlated producing spurious signal among the true null hypotheses; this would lead to a major inflation of significance. In the other case, they are uncorrelated leading to a major loss of power. By accounting for the $\boldsymbol{\Gamma}\mathbf{G}$ term, we have resolved these issues.

Latent Structure. A second source of multiple testing dependence is a latent structure due to relevant factors not being included in the model. It is possible for there to be unmodeled factors that are common among the multiple tests but that are not included in **S**. Suppose there exists unmodeled factors **Z** such that $\mathrm{E}(\mathbf{x}_i | \mathbf{Y}) \neq \mathrm{E}(\mathbf{x}_i | \mathbf{Y}, \mathbf{Z})$ for more than one test. If we utilize model **1** when performing the significance analysis, there will be dependence across the rows of **E** induced by the common factor **Z**, causing population-level multiple testing dependence. Likewise, there will be dependence across the rows of **R** causing estimation-level multiple testing dependence. A similar case can arise when the model for $\mathbf{x}_i$ in terms of **Y** is incorrect. For example, it could be the case that $\mathrm{E}[\mathbf{x}_i | \mathbf{Y}] = \mathbf{b}_i\mathbf{S}^*(\mathbf{Y})$, where the differences between **S** and $\mathbf{S}^*$ are nontrivial among multiple tests. Here, there will be dependence across the rows of **R** induced by the variation common to multiple tests due to $\mathbf{S}^*$ but not captured by **S**, which would cause estimation-level multiple testing dependence. Failing to include all relevant factors is a common issue in genomics leading to latent structure (11, 12). The adverse effects of latent structure due to unmodeled factors on differential expression significance analyses has only recently been recognized (12).

Fig. 2 *C* and *D* shows independently simulated microarray studies in this scenario, where we have simulated a treatment effect
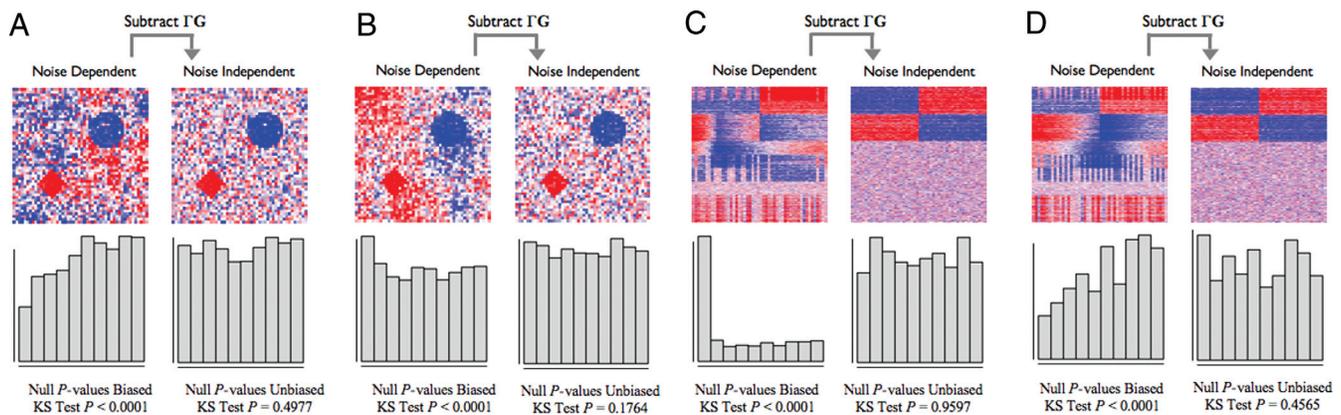
Fig. 2. Simulated examples of multiple testing dependence. *A* and *B* consist of spatial dependence examples as simplified versions of that encountered in brain imaging, and *C* and *D* consist of latent structure examples as encountered in gene expression studies. In all examples, the data and the null *P* values are plotted both before and after subtracting the dependence kernel. The data are plotted in the form of a heat map (red, high numerical value; white, middle; blue, low). The signal is clearer and the true null tests' *P* values are unbiased after the dependence kernel is subtracted. (*A* and *B*) Each point in the heat map represents the data for one spatial variable. The two true signals are in the diamond and circle shapes, and there is autoregressive spatial dependence between the pixels. (*A*) An example where the spatial dependence confounds the true signal, and the null *P* values are anticonservatively biased. (*B*) An example where the spatial dependence is nearly orthogonal to the true signal, and the null *P* values are conservatively biased. (*C* and *D*) Each row of the heat map corresponds to a gene's expression values, where the first 400 rows are genes simulated to be truly associated with the dichotomous primary variable. Dependence across tests is induced by common unmodeled variables that also influence expression, as described in the text. (*C*) An example where dependence due to latent structure confounds the true signal, and the null *P* values are anticonservatively biased. (*D*) An example where dependence due to latent structure is nearly orthogonal to the true signal, and the null *P* values are conservatively biased.

plus effects from several unmodeled variables. The unmodeled factors were simulated as being independently distributed with respect to the treatment, which is equivalent to a study in which the treatment is randomized. As in Fig. 2 *A* and *B*, it can be seen that the *P* values corresponding to true null hypotheses (i.e., genes not differentially expressed with respect to the treatment) are not Uniformly distributed. When utilizing model **3** for these data and subtracting the term $\mathbf{\Gamma G}$, the residuals are now made independent and the null *P* values are *Uniform*(0, 1) distributed.

Estimating G in Practice

There are a number of scenarios where estimating **G** is feasible in practice. One scenario is when nothing is known about the dependence structure, but it is also the case that $d + r < n$, where d and r are the number of rows of the model **S** and dependence kernel **G**, respectively. This is likely when the dependence is driven by latent variables, such as in gene expression heterogeneity (12). In the *SI Appendix*, we present an algorithm for estimating **G** in this scenario. It is shown that the proposed algorithm, called iteratively reweighted surrogate variable analysis (IRW-SVA), exhibits favorable operating characteristics. We provide evidence for this over a broad range of simulations. Another scenario is when the dependence structure is well characterized at the population level. Here, it may even be the case that $d + r \approx n$. This scenario is common in brain imaging (6, 7) and other spatial dependence problems (9), as discussed above. The fact that $\mathbf{\Gamma}$ is largely determined by the known spatial structure allows us to overcome the fact that $d + r \approx n$ (*SI Appendix*).

Discussion

We have described a general framework for multiple testing dependence in high-dimensional studies. Our framework defines multiple testing dependence as stochastic dependence among tests that remains when conditioning on the model is used in the significance analysis. We presented an approach for addressing the problem of multiple testing dependence based on estimating the dependence kernel, a low-dimensional set of vectors that completely defines the dependence in any high-throughput dataset. We have shown that if the dependence kernel is known and included in the model, then the hypothesis tests can be made stochastically independent. This work extends existing results regarding error rate control under independence to the case of general dependence. An additional advantage of our approach is that we can not only estimate dependence at the level of the data, which is intuitively more appealing than estimating dependence at the level of *P* values or test statistics, but we can also directly adjust for that dependence in each specific study. We presented an algorithm with favorable operating characteristics for estimating the dependence kernel for one of the main two scientific areas of interest that we discussed. We anticipate that well behaved estimates of the dependence kernel in other scientific areas are feasible.

One important implication of this work is that multiple testing dependence is tractable at the level of the original data. Downstream approaches to dealing with multiple testing dependence are not able to directly capture general dependence structure (Fig. 1). Another implication of this work is that, for a fixed complexity, the stronger the dependence is among tests, the more feasible it is to properly estimate and model it. It has also been shown that the weaker multiple testing dependence is, the more appropriate it is to utilize methods that are designed for the independence case (21). Therefore, there is promise that the full range of multiple testing dependence levels is tractable for a large class of relevant scientific problems.

1. Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. *Proc Natl Acad Sci USA* 100:9440–9445.
2. Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. *Stat Sci* 18(1):71–103.
3. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci, USA* 98:5116–5121.
4. Miller CJ, *et al.* (2001) Controlling the false-discovery rate in astrophysical data analysis. *Astron J* 122:3492–3505.
5. Starck JL, Pires S, Refregier A (2006) Weak lensing mass reconstruction using wavelets. *Astron Astrophys* 451:1139–1150.
6. Worsley KJ, *et al.* (1996) A unified statistical approach for determining significant signals in images of cerebral activation. *Hum Brain Mapp* 4:58–73.

7. Geneovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* 15:870–878.
8. Schwartzman A, Dougherty RF, Taylor J (2008) False discovery rate analysis of brain diffusion direction maps. *Ann Appl Stat* 2:153–175.
9. Elliott P, Wakefield J, Best N, Briggs D (2001) *Spatial Epidemiology: Methods and Applications* (Oxford Univ Press, New York).
10. Bellman R (1961) *Adaptive Control Processes: A Guided Tour* (Princeton Univ Press, Princeton, NJ).
11. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 65:220–228.
12. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3:e161.
13. Green PJ, Silverman BW (2000) *Nonparametric Regression and Generalized Linear Models* (Chapman & Hall, San Francisco).
14. Worsley KJ (2003) Detecting activation in fmri data. *Stat Methods Med Res* 12:401–418.
15. Rice JA (1995) *Mathematical Statistics and Data Analysis* (Duxbury Press, Pacific Grove, CA), 2nd Ed.
16. Shaffer JP (1995) Multiple hypothesis testing. *Annu Rev Psychol* 46:561–584.
17. Soric B (1989) Statistical discoveries and effect-size estimation. *J Am Stat Assoc* 84:608–610.
18. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc B* 57:289–300.
19. Storey JD (2002) A direct approach to false discovery rates. *J Roy Stat Soc B* 64:479–498.
20. Newton MA, Noueiry A, Sarkar D, Ahlquist P (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5:155–176.
21. Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J Roy Stat Soc B* 66:187–205.
22. Dabney AR, Storey JD (2006) A reanalysis of a published affymetrix genechip control dataset. *Genome Biol* 7:401.
23. Yekutieli D, Benjamini Y (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Stat Plan Inf* 82:171–196.
24. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29:1165–1188.
25. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55:997–1004.
26. Efron B (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Am Stat Assoc* 99:96–104.
27. Price AL, *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909.
28. Klebanov L, Jordan C, Yakovlev (2006) A new type of stochastic dependence revealed in gene expression data. *Stat Appl Genet Mol Biol* 5:7.
29. Qui X, Klebanov L, Yakovlev AY (2005) Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Stat Appl Genet Mol Biol* 4:34.
30. Mardia KV, Kent JT, Bibby JM (1997) *Multivariate Analysis* (Academic, New York).
31. Owen A (2005) Variance of the number of false discoveries. *J Roy Stat Soc B* 67:411–426.
32. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scan J Stat* 6:65–70.

STATISTICS