Published in final edited form as:

J Immunol Methods. 2016 October ; 437: 1–12. doi:10.1016/j.jim.2016.05.004.

Human Immunophenotyping via Low-Variance, Low-Bias, Interpretive Regression Modeling of Small, Wide Data Sets: Application to Aging and Immune Response to Influenza Vaccination

Tyson H. Holmes^{a,*} and Xiao-Song He^{b,c}

^aStanford University Human Immune Monitoring Center, Institute for Immunity Transplantation and Infection, Stanford University School of Medicine, Stanford, CA, 94305, USA, Tel: 650-725-2925, Fax: 650-498-7495

^bStanford University School of Medicine, 291 Campus Drive, Stanford, CA, USA, 94305

^cVA Palo Alto Healthcare System, 3801 Miranda Ave, Palo Alto, CA, USA, 94304

Abstract

Small, wide data sets are commonplace in human immunophenotyping research. As defined here, a small, wide data set is constructed by sampling a small to modest quantity n , $1 < n < 50$, of human participants for the purpose of estimating many parameters p , such that $n < p < 1,000$. We offer a set of prescriptions that are designed to facilitate low-variance (i.e. stable), low-bias, interpretive regression modeling of small, wide data sets. These prescriptions are distinctive in their especially heavy emphasis on minimizing use of out-of-sample information for conducting statistical inference. That allows the working immunologist to proceed without being encumbered by imposed and often untestable statistical assumptions. Problems of unmeasured confounders, confidence-interval coverage, feature selection, and shrinkage/denoising are defined clearly and treated in detail. We propose an extension of an existing nonparametric technique for improved small-sample confidence-interval tail coverage from the univariate case (single immune feature) to the multivariate (many, possibly correlated immune features). An important role for derived features in the immunological interpretation of regression analyses is stressed. Areas of further research are discussed. Presented principles and methods are illustrated through application to a small, wide data set of adults spanning a wide range in ages and multiple immunophenotypes that were assayed before and after immunization with inactivated influenza vaccine (IIV). Our regression modeling prescriptions identify some potentially important topics for future immunological research. 1) Immunologists may wish to distinguish *age-related* differences in immune features from changes in immune features *caused* by aging. 2) A form of the bootstrap that employs linear extrapolation may prove to be an invaluable analytic tool because it allows the working immunologist to obtain accurate estimates of the stability of immune parameter estimates

*To whom correspondence should be addressed.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

with a bare minimum of imposed assumptions. 3) Liberal inclusion of immune features in phenotyping panels can facilitate accurate separation of biological signal of interest from noise. In addition, through a combination of denoising and potentially improved confidence interval coverage, we identify some candidate immune correlates (frequency of cell subset and concentration of cytokine) with B cell response as measured by quantity of HIV-specific IgA antibody-secreting cells and quantity of HIV-specific IgG antibody-secreting cells.

1. Introduction

1.1 Small, wide data set defined

For purposes here, a small, wide data set (SWDS) is defined as a sampling of a small to modest quantity n , $1 < n < 50$, of human participants for the objective of estimating many parameters p , such that $n < p < 1,000$. We are constraining quantity of parameters, somewhat arbitrarily, to $p < 1,000$ in order to focus on the meso-scale setting and not ultra-high-dimensional phenotyping.

1.2 Immunological motivation

SWDSs are commonplace in immunological research because the technologies generating these data are seeing widening use. Among these technologies are (mass) cytometry by time-of-flight (CyTOF; Watson et al. 2009), comprehensive leukocyte immunophenotyping (CLIP; Biancotto et al. 2011), and cytokine multiplex bead arrays (Harris and Chen 2013). These technologies are allowing investigators to explore more deeply and thoroughly the complex structure and function of the human immune system. Because relatively few observations are required (by definition), SWDSs are, in fact, the immunologists entry point, via “small” “pilot” studies, into rich human immunophenotyping. Of course, this richness is in the quantity of features and not in terms of extensive samplings of human participants. As such, SWDSs pose a number of statistical challenges for the working immunologist, especially in terms the variance or “instability” of parameter estimates from these rich feature sets.

Notably, many of these technologies measure features at the single-cell level. Single-cell data collection can generate thousands of observations per participant, yielding on overall data set with far more observations than features p . Even so, with the possible exception of single-participant “case studies,” studies of human immunophenotypes entail *hierarchical* samplings with a first-stage sampling of human participants followed by a second stage sampling of tissue(s) or individual cells within each participant. The samplings at each stage, participant and within-participant, contribute to the variance (e.g., as quantified by standard errors) of parameter estimates (Thompson 1992, pp. 128-129). Because our scope here is studies of human immunophenotypes and their relationship with factors such as age and vaccine exposure, all of which are whole-person level traits, we will restrict attention to variance of parameter estimates as governed by sample size n . In this framework, variance due to within-participant sampling can be regarded as a contributor to mensuration error (Section 3.1.2).

1.3 Theoretical details

Parameter dimensionality p , which is the quantity of parameters to be estimated from the sample, is distinct from feature dimensionality f , which is the quantity of features (e.g., cytokines) measured in the sample. Given that the purpose of sampling is to estimate parameters (and perhaps test hypotheses regarding those parameters), each feature sampled will have at least one associated parameter of interest; so, as a general rule, $f \geq p$.

Formally, f one-to-one or one-to-many maps to p via a specific regression model

parameterization $\mathbb{p}, f \xrightarrow{P} p$. As such, the cardinality of the small, wide problem is $|np|$ and *not* $|nf|$, a distinction often overlooked in the literature, perhaps due to emphasis on parameter estimation via linear modeling, although the literature on p -value combination methods seems well aware of this distinction (e.g., Phillips and Ghosh 2014). Nonlinear models can be particularly enriched in parameters ($f < p$; Neter et al. 1996, p. 534); and even purely linear models have additional parameters for other moments (e.g., the covariance structure) of interest in immunology, so that in nearly all practical regression settings $f < p$.

1.4 Organization of paper for immunologists

This paper is designed to help the working immunologist understand how statistical technologies can be applied to SWDSs to obtain estimates of immune parameters that are stable, low-bias, and immunologically interpretable. As such, each subsection under Theory (Section 3) provides introductory subsections “General concepts” followed by more in-depth subsections on “Theoretical details.” Similarly, each worked example in the Results (Section 4), contains a middle subsection on “Computational details and findings” that is preceded and followed, respectively, by subsections that provide “Immunological background” and “Immunological implications.” Finally, the Discussion consists of separate subsections that are “Immunological” and “Theoretical.” Those subsections that are labeled “theoretical” or “computational” will be especially of interest to immunologists who wish to understand presented concepts at a more statistical or mathematical level.

2. Materials and Methods

2.1 An illustrative data set

Throughout, concepts are illustrated using a SWDS of immune features in an influenza vaccination study. The human participants were described previously (Sasaki et al. 2011). All analyses reported here were performed on completely de-identified data. These participants were vaccinated with a single dose of 2009 seasonal IIV (Fluzone, from Sanofi Pasteur, Swiftwater, PA, USA) that contained an A/Brisbane/59/2007 (H1N1)-like virus, an A/Brisbane/10/2007 (H3N2)-like virus, and a B/Brisbane/60/2008-like virus. Blood samples were collected from each participant at three time points: day 0 (week = 0) before vaccination, day 7 or 8 (week = 1), and day 28 ± 4 (week = 4) post-vaccination. From the week 1 whole blood samples, B-cells were isolated to measure the frequencies of the IIV-specific IgG and IgA antigen secreting cells (ASCs) using an ELISPOT assay as previously described (Sasaki et al. 2011). Sera were isolated from each blood sample and used for measuring the concentration of cytokines and chemokines. Peripheral mononuclear cells (PBMCs) were isolated from the day 0 blood samples and used for determining the

frequencies of specific immune cell subsets. Details of these immunophenotyping assays are given in Appendix A of the accompanying online supplement.

Altogether, this panel of assays yielded information on 2 features quantifying the B cell response to the vaccination as well as 229 immune features to facilitate immunophenotyping on 35 participants before and after the vaccination. The complete list of features is provided in Appendix B of the online supplement. All cytokine features were detrended for nonspecific binding prior to analysis (Montoya et al., *in preparation*). Volunteers' ages ranged from 19 years to 84 years with a mean age of ~ 53 years, where, by design, sampling was stratified into two groups of a) young adults (19 to 28 years) and b) elderly (70 to 84 years).

3. Theory

3.1 A sampling framework for SWDSs

3.1.1 General concepts—In Section 1.2, we characterized a SWDS as two samplings—of participants and of tissue(s) or individual cells within participants. A SWDS can also be viewed as two samplings in another framework—of participants and immune features within participants. This latter framework is very useful because it allows us to isolate two fundamentally different sources of “noise” in SWDSs that can obscure the biological “signal” of interest to the immunologist. Noise due to the size of the sample of participants, and how that sample is selected, is *sampling error*. In contrast, the sampling of features measured on each participant introduces two components of *mensuration error*. One component of mensuration error is measurement error, either random variation among technical repeats and/or measurements that are systematically above or below the true value (measurement bias). The other component of mensuration error arises when the measured feature set does not include all of the features needed to properly characterize the immunological system being studied. Because SWDSs, by definition, will almost certainly contain substantial noise due to sampling error due to small samples of participants, opportunities for noise reduction can be exploited as the SWDS is widened. In fact, addition of noisy features (e.g., technical error), can actually help to facilitate separation of immunological signal from noise, as detailed in Section 3.6. Much of the modern statistical technology that has been and continues to be developed for wide data sets is focused on feature selection—i.e., separating those features that are associated with immunological signal from those that are not. In Section 3.5, we argue that complete removal of any features by means of statistical selection procedures may be at odds with meaningful immunological interpretation. Indeed, in Section 3.2, we argue that bias reduction can be obtained by discovering “hidden” features in the data.

3.1.2 Theoretical details—For fixed quantity of observations n , multidimensional “sampling density” (Hastie et al. 2009, p. 23) declines as f grows, further magnifying sampling variance due to the “curse of dimensionality.” SWDSs are, without question, information limited in that information (signal) is enmeshed within often substantial quantities of “noise.” Extending the exposition of Gavish and Donoho (2014), decompose an observed $n \times f^1$ SWDS matrix w as follows,

$$\mathbf{w}(n, f) = \boldsymbol{\theta}\mathbf{1} + \boldsymbol{\rho} + \sigma_{\hat{\boldsymbol{\mu}}}\mathbf{e}_{\hat{\boldsymbol{\mu}}} + \boldsymbol{\sigma}\mathbf{e}, \quad (\text{E1})$$

where $\boldsymbol{\theta}\mathbf{1}$ is a matrix of the (noiseless) grand mean, matrix $\boldsymbol{\rho}$ contains the systematic (noiseless) deviations of each row (observation) and column (feature) from the grand mean due to regression predictors (e.g., age), $\sigma_{\hat{\boldsymbol{\mu}}}\mathbf{e}_{\hat{\boldsymbol{\mu}}}$ is the sampling-error variance structure

(sampling error in estimated $\boldsymbol{\mu} \stackrel{\text{def}}{=} \boldsymbol{\theta}\mathbf{1} + \boldsymbol{\rho}$) and $\boldsymbol{\sigma}\mathbf{e}$ is mensuration error. Define $\xi \stackrel{\text{def}}{=} \frac{\sum_{i=1}^r s_i^2}{\sum_{i=r+1}^m s_i^2}$

as the unitless ratio from ordered, positive, real-valued, singular values (i.e., $s_1 \dots s_r \dots s_m$) of $\mathbf{w}(n, f)$, $1 \leq r \leq m$, $n < f \leq p$, where r is the rank of matrix $\mathbf{w}(n, f)$ and r is what we define as the “signal rank” (cf. Harville 1997, pp. 553, 556-559). We can approximate $\boldsymbol{\theta}\mathbf{1} + \boldsymbol{\rho}$ from the singular-value decomposition of $\mathbf{w}(n, f)$ with the last $n - r$ singular values set to zero (cf. Harville 1997, pp. 556-559). In the current setting, effectively this is a type of shrinkage estimate of $\boldsymbol{\theta}\mathbf{1} + \boldsymbol{\rho}$ and will contain minimal bias if the last $n - r$ singular values are pure noise and $E[\sigma_{\hat{\boldsymbol{\mu}}}\mathbf{e}_{\hat{\boldsymbol{\mu}}}] \approx \mathbf{0}$ and $E[\boldsymbol{\sigma}\mathbf{e}] \approx \mathbf{0}$. As introduced in Section 3.1.2, noise consists of a component $\sigma_{\hat{\boldsymbol{\mu}}}\mathbf{e}_{\hat{\boldsymbol{\mu}}}$ that is a functional of sampling design, including sample size n , and a mensuration error component $\boldsymbol{\sigma}\mathbf{e}$. Mensuration error arises during the process of quantifying a feature as variation among technical repeats; and in this sense mensuration error can vary (possibly widely) from feature to feature, here represented by $\mathbf{e}$. Mensuration error also arises due to feature sampling (selection) in that the set of features observed may not constitute all features needed to fully characterize the biological system targeted for study—a type of finite sampling error, here represented by $\boldsymbol{\sigma}$. Sampling bias (i.e., $E[\sigma_{\hat{\boldsymbol{\mu}}}\mathbf{e}_{\hat{\boldsymbol{\mu}}}] \neq \mathbf{0}$) and mensuration bias (i.e., $E[\boldsymbol{\sigma}\mathbf{e}] \neq \mathbf{0}$) may be present, as is discussed in Section 3.2.

Suppose we represent a SWDS as a “nested” sampling in the following way. Define set $\mathbb{f}(f)$ as a particular instance of sampling $\mathbb{F}(f) \xrightarrow{\mathcal{S}_I} \mathbb{f}(f)$ of features drawn from all such samples $\mathbb{f}(f) \in \mathbb{F}(f)$ of size f . This is sampling on the *feature axis*. Likewise, define set $\mathbb{o}(n|\mathbb{f}(f))$ as a particular instance of sampling $\mathbb{O}(n|\mathbb{f}(f)) \xrightarrow{\mathcal{D}} \mathbb{o}(n|\mathbb{f}(f))$ of observational units drawn from all such samples $\mathbb{o}(n|\mathbb{f}(f)) \in \mathbb{O}(n|\mathbb{f}(f))$ of size n . This is sampling on the *observation axis* and is conditional on the feature sampling $\mathbb{f}(f)$ because observations can only be drawn once an *initial* feature set has been identified, via $\mathcal{S}_I$. For example, an investigator would not draw a sample of patients and then decide what virus to study. An observed instance of a SWDS $\mathbf{w}(n, f)$ can be represented by the Cartesian product $\mathbf{w}(n, f) \stackrel{\text{def}}{=} \mathbb{o}(n|\mathbb{f}(f)) \times \mathbb{f}(f)$. Ideally, $\mathcal{D}$ (“D” for design) is typically some probability (i.e., random) sampling design (Cochran 1977, p. 9), whereas $\mathcal{S}_I$ (“S” for feature selection) is typically systematic, although at times $\mathcal{S}_I$ is treated as a probability sampling as an expedient for estimation (Section 3.4.3.2 and Diedrichsen et al. 2011). Stratified random (Cochran 1977, p. 99-101), cluster (Cochran 1977, p. 242), double (Cochran 1977, pp. 341-342), and other $\mathcal{D}$ that are structured sampling design algorithms can generate SWDSs that provide more stable estimates of ξ relative to

¹Here and throughout the text, we recognize that $n \times f$ may represent $n \times p$ through parameterization $\mathbf{P}$ per Section 2.

simple random sampling in that $\sigma_{\hat{\mu}} \mathbf{e}_{\hat{\mu}}$ is $\mathcal{D}$ -noise. Careful choice of feature sampling $\mathcal{S}_I$ via selection of features of low mensuration error as well as selection of features that together exhaustively characterize the system being studied ($\sigma \rightarrow 0 \cup E[\mathbf{e}] \approx \mathbf{0}$) can also enhance $\mathbf{g}$. Accordingly, $\sigma \mathbf{e}$ is $\mathcal{S}_I$ -noise. Importantly, noisy features, in $\mathbf{e}$, are not necessarily problematic, and, in fact, may be *desirable*, as they can facilitate isolation of signal (Section 3.6) in SWDSs. Feature sampling can be in two stages, $\mathcal{S}_I$ and $\mathcal{S}_{II}$. For $\mathcal{S}_I$, prior to data collection, investigation begins by selecting a total panel of f_I features to be measured (e.g., an array of 51 cytokines). For $\mathcal{S}_{II}$, once data collection is complete, “hard” and “soft” feature selection techniques (*sensu*Fan and Li 2001) can be applied to obtain a lower-dimensional systematic sampling of f_{II} signal features from the original f -feature space, $f_{II} < f_I$ (but see Sections 3.2 and 3.5), which in turn permits a more parsimonious parameterization ($f_{II} \xrightarrow{p} p_{II}$) for use in, say, a subsequent testing sample of resultant greater multidimensional sampling density.² Some classic vehicles for $\mathcal{S}_{II}$ are the lasso (Tibshirani 1996) and its derivatives.

3.2 Bias and bias reduction

3.2.1 General concepts—The literature on feature selection has burgeoned in recent years. To some extent, this may have been to the neglect of discussions of sampling design for sampling participants, at least in the following sense. While sampling design may be a theoretically “mature” topic, many studies continue to rely on “convenience” samples (*sensu*Hedt and Pagano 2011) in practice. Especially in medical studies, convenience sampling can be unavoidable and, indeed, quite necessary to ensure adequate protections for research participants. Convenience sampling may be more prevalent among SWDSs because attaining width can enhance participant burden in terms of quantity of assessments; and only a (perhaps highly self-selected) subset of otherwise qualified research participants will be willing and able to volunteer for intensive measurement. Because convenience sampling is a departure from pure probability sampling, or pure systematic sampling for that matter, samplings of participants, as well as samplings of features, are *confined* (e.g., to within specific patient referral networks, to biomarkers available on commercially available arrays, etc.). This possibly introduces substantial bias into parameter estimates, which only adds to the existing problem of high variance due to small n , and also inflates mensuration variance due to incomplete sampling of the features needed to fully characterize the immunological system being studied. One possible solution to these problems may be to apply statistical methods that discover hidden (latent) features within the SWDS. A class of particularly important hidden features are unobserved confounders. By definition, a *confounding feature* causes variation in the predictor and in the outcome. Unless these confounders are identified, and we adjust for them, estimates of the true causal relationship between predictor and outcome can be biased. We illustrate in Section 4.1 how the hidden feature discovery methods of Section 3.2.2 can be used to reduce possible bias in the causal effects of aging on immune cells subsets and cytokine levels.

²In fact, selection is typically applied to the parameter set which causes a reduction in the cardinality of the feature set per Section 2.

3.2.2 Theoretical details—In theory, consistent estimates of parameters can be obtained from a convenience sample through inverse-probability weighting techniques (e.g., Haneuse et al. 2009), if information on prevalences is available in the referent population, or via regression adjustment. Weighting is unattractive, even stabilized weights (e.g., Robins et al. 2000), because oftentimes weights are estimates from a sampling of the referent population, which introduces yet another source of sampling variance into an estimation process that is already highly enriched in sampling variance due to low sampling density (i.e., SWDS). A potential drawback to regression adjustment is that sampling bias through $\mathcal{D}$ may arise along dimensions (features) that were not observed/selected (Groenwald et al. 2009) during $(\mathcal{S}_1)$. Here is where the special structure of the SWDS may offer a solution. Namely, rather than being *subtractive*, make selection $\mathcal{S}_1$ *generative* to encompass *latent* feature discovery. To illustrate, suppose $\mathcal{S}_1$ identified a set of f_1 features. Optionally, basis expansions or truly nonlinear parameterizations may be performed for some or all of these features, generating parameter *sets*, where the respective quantities of parameters are $p_1, \dots, p_A$, $f_1 < p \stackrel{\text{def}}{=} \sum p_i$. In the first stage, we fit f_1 (possibly nonlinear) regression models (one for each of the features and corresponding parameter sets in $\mathcal{S}_1$), specified as $\vec{Y}_i = h_i(\mathbf{X}_i | \vec{\beta}_i) + \vec{E}_i$, with $n \times 1$ outcome vector $\vec{Y}_i$, $n \times p_i$ design matrix of candidate predictors $\mathbf{X}_i$, $b_i \times 1$ parameter vector $\vec{\beta}_i$, and some (possibly nonlinear) function $h_i(\cdot)$. Signal separation methods such as independent component analysis (Teschendorff et al. 2011) can be applied to the sampling of the *residual space*, represented by the observed $n \times f_1$ matrix of column-concatenated residual vectors $\mathbf{E} \stackrel{\text{def}}{=} [\vec{E}_1 \dots \vec{E}_{f_1}]$, for the purpose of extracting an $n \times f_R$ residual *signal*/latent feature matrix $\mathbf{R}$, $0 < f_R < f_1$. Note that the closed interval for f_R includes zero because no latent features may exist (i.e., residuals are pure noise). By definition, a *confounding feature causes variation in predictor and outcome*. In the second stage of the analysis, that set of original outcomes $\vec{Y}_i$, $i \in \{1, \dots, f_1\}$ that are associated with the k^{th} extracted residual feature vector $\vec{R}_k$, $k \in \{1, \dots, f_R\}$ are subjected to signal separation methods to form a matrix $\mathbf{L}_k$ of $f_{L(k)}$ latent features $f_{L(k)} < f_1$. These extracted features, unlike the residual latent features $\mathbf{R}$, can be, by construction, associated with candidate predictors $\mathbf{X}_i$ and with outcome—a necessary condition for *confounding*.³ That column vector $\vec{L}_k^*$ of $\mathbf{L}_k$ that is found to be most strongly associated with the $\vec{R}_k$ is designated as a *surrogate variable* that can be used subsequently to adjust for confounding. When this process has been completed for all f_R residual features, adjustment for confounding is performed via $\vec{Y}_i = h_i(\mathbf{X}_i | \vec{\beta}_i) + \mathbf{L}^* \vec{\gamma}_i + \vec{E}_i$, where $\mathbf{L}^* \stackrel{\text{def}}{=} [\vec{L}_1^* \dots \vec{L}_{f_R}^*]$ with associated parameter vector $\vec{\gamma}_i$. The extracted set $\mathbf{L}^*$ may not (and probably will not) contain the unobserved confounding features themselves but, more specifically (and ideally), some basis that spans the same

³This association is a necessary but, alone, not sufficient condition for confounding. By definition, a confounding feature causes variation in predictor and outcome. Some latent features discovered by this method may not cause variation in the predictor but may instead be another outcome with its variation caused by the predictor. For this reason, this feature discovery method for adjusting for possible confounding should be coupled with mechanistic studies of immune system function and/or medium-term longitudinal studies (see Section 4.3.3 and 5.1).

space as these features (Leek and Storey 2007). Function $h_f(\cdot)$ should be a flexible function that captures all aspects of the true relationship between the candidate predictors and the outcome. That way, $\mathbf{E}$ should only contain truly unobserved factors and not *avoidable* misspecifications of the relationship between $\mathbf{X}_i$ and $\vec{\mathbf{Y}}_i$ (e.g., curvilinear terms). Available methods of surrogate variable analysis include Leek and Storey (2007), Leek and Storey (2008) and Teschendorff et al. (2011).

3.3 The problem of out-of-sample information

3.3.1 General concepts—The potential discoverability of hidden (latent) features raises a broader question. Just how information-limited are SWDSs? One of the temptations under small n in general, not necessarily only when $1 < n < 50$ and $n < p < 1,000$, is to flood estimation and hypothesis testing with *out-of-sample* information (i.e., information external the SWDS) on the presumption that a sample of small scale on the observation axis *must* be highly information limited. However, heavy reliance on out-of-sample information does place the immunological investigator in a defensive position, especially when data sets are small and evidence is limited as regards the applicability of any invoked out-of-sample assumptions. The remainder of Section 3.3 describes some additional means of making fuller use of what information is contained within a SWDS. This frees the immunologist to proceed with analysis without being concerned about imposing (and defending) strong out-of-sample assumptions.

3.3.2 Theoretical details—A classic example is Bayesian estimation, which has often been criticized when informative (out-of-sample) priors are applied that then dominate posterior distributions derived from small samples—a property that extends to larger samples as f grows (Diaconis and Freedman 1986). “Semiparametric” generalized maximum (cross) entropy estimation has been characterized as highly sensitive to the particular choice of the prior support (Léon et al. 1999), which may be selected based on out-of-sample information. In the frequentist setting, parametric estimation and testing are often applied in samples that may be too small to permit assessment of whether or not those out-of-sample parametric assumptions apply to the population being sampled (e.g., Rasch et al. 2011). All such out-of-sample information arises in at least one of three components of a complete formulation of the regression model: 1) parameterization $\mathbb{p}$; 2) imposition of constraints on model parameters β_j via priors and inequality constraints; and 3) higher-moment assumptions, if any, about the distribution of the $\mathbf{Y}_i$ conditional upon the $E[\mathbf{Y}_i] = h(\mathbf{X}_i|\beta_j)$ (e.g., $\mathbf{Y}_i$ are drawn from a lognormal distribution). In our view, progress in the empirical sciences is largely data-driven. As such, maximally efficient exploitation of available *in-sample* information in the SWDS setting is essential with minimal imposition of out-of-sample structure other than that due to certain irrefutable constraints (e.g., non-negative analyte concentrations). Mining $\mathbf{E}$ for latent feature structure, as outlined above, is one means of making fuller use of in-sample information for purposes of hypothesis testing, specifically to address misspecification of $h_f(\mathbf{X}_i|\beta_j)$ and especially if that mining imposes few additional out-of-sample assumptions (e.g., non-Gaussianity and independent component analysis). The next section discusses two more means of more fully exploiting in-sample information.

3.4 Tail extrapolation and bi-directional shrinkage

3.4.1 Tail extrapolation

3.4.1.1 General concepts: Imagine drawing a random sample of participants and estimating some parameter of interest, such as the mean of a particular immune parameter. Further imagine repeating this process 1,000 times, each time with a new random sample of participants. Together these 1,000 means provide an estimate of the *sampling distribution* of the parameter estimate. Estimates of sampling distributions are very useful because they allow immunologists to make approximate probability statements, such as confidence intervals, about a parameter estimate of interest.

Bootstrap resampling (i.e., repeatedly sampling with replacement from the observed sample) is a popular and highly flexible tool for characterization of the sampling distribution of a parameter estimate via minimal out-of-sample assumptions. With that flexibility comes a limitation. Unlike methods of confidence interval construction that rely on out-of-sample information (e.g., assuming that sampling is from a normal distribution), in small samples of participants, bootstrap procedures can underestimate the frequency of the more extreme values of the sampling distribution (“tails” of distribution) and that can cause confidence intervals to be too narrow and hypothesis testing based on confidence intervals to reject the null hypothesis too often (too liberal). In the remainder of Section 3.4.1, a technique is described due to Hutson (2002) that can reduce the bootstrap's tendency to underestimate the tails of sampling distribution; and Hutson's technique achieves this with a minimum of out-of-sample assumptions, relying only on linear extrapolation. Here we call this the $Q(n)$ -bootstrap for reasons detailed in Section 3.4.1.2.

3.4.1.2 Theoretical details: The current setting is a small sample $\phi(n|f(f))$, $1 < n < 50$, of observations. The observed distribution of a parameter estimate under nonparametric bootstrap resampling is confined to that *lattice* within that convex hull C_p defined by sample $\phi(n|f(f))$ within the parameter space generated by $\mathbb{p}$. Perturbation resampling (e.g., Jin et al. 2001) can provide smoother (“between lattice vertices”) interrogation of the region defined by C_p but nevertheless remains confined to within C_p . As a result, when n is small, tail regions of the resampling distribution of the estimate may be especially under-represented (e.g., confidence bounds that are too narrow, “ p -values” are underestimated).

Certainly, various correctives do exist, such as Gaussian bias correctives (Davison and Hinkley 1997, p. 204) or peak-over-threshold tail extrapolation, which relies on local parametric approximation via, for example, a generalized Pareto distribution (El-Aroui and Diebolt 2002). Though flexible, these parametric correctives rely upon out-of-sample information. Instead, perhaps the closest example to a purely in-sample approach to tail estimation is a method such as Hutson (2002). In brief, he proposed a method of estimating the quantile function Q that is linearly interpolative within the closed interval defined by sample $\phi(n|f(f))$ in combination with linear tail extrapolation. For independent and identically distributed observations $X_i = x_i$ from a sample of size n , $i \in \{1, \dots, n\}$, he defines

$$\hat{Q}(n) \stackrel{\text{def}}{=} \begin{cases} x_1 + (x_2 - x_1) \log((n+1)U), & 0 < U \leq \frac{1}{n+1} \\ (1-\varepsilon)x_{\lfloor (n+1)U \rfloor} + \varepsilon x_{\lfloor (n+1)U \rfloor + 1}, & \frac{1}{n+1} < U < \frac{n}{n+1} \\ x_n + (x_n - x_{n-1}) \log((n+1)(1-U)), & \frac{n}{n+1} \leq U < 1 \end{cases}$$

for sample order statistics $x_1 \leq x_2 \leq \dots \leq x_n$, $e \stackrel{\text{def}}{=} (n+1)U - \lfloor (n+1)U \rfloor$, and floor function $\lfloor \cdot \rfloor$. Application of the inverse transform over the support of the uniform distribution, $U \in (0, 1)$, permits, as with perturbation resampling, a smoothed form of resampling. Computationally, this is straightforward and simple to implement. We refer to this as the $Q(n)$ -bootstrap.

Discussion thus far has been in univariate setting (i.e., examining each feature separately). In Section 4.3.2, we extend the $Q(n)$ -bootstrap to resampling from joint (multivariate) distributions.

3.4.2 Alternatives for (largely) in-sample SWDS inference

3.4.2.1 General concepts: Hutson's linear extrapolation technique (the $Q(n)$ -bootstrap) is one way that immunologists can make inferences about immune parameters of interest in SWDSs with a minimum of out-of-sample assumptions. Other possibilities exist as well, such as Bayesian methods (Holmes and Lewis 2014), permutation testing (Good 1994), and assuming that distributions are symmetric (Modarres 2002); but all of these approaches do include some strong assumptions that limit ease-of-use for the working immunologist.

3.4.2.2 Theoretical details: Given the goal of minimizing reliance on out-of-sample information (Section 3.3.1), ideally Bayesian inference would be “nonparametric.” Semiparametric methods exist such as specification of a mixture of Polya trees for the distribution of residual error; however, these methods may entail parametric distributions for the hyperparameters (Hanson and Johnson 2002). Discrete Bayes is more fully nonparametric but does so by boosting the quantity of parameters to be estimated (i.e., via parameterization $\mathbb{p}$ of $f \xrightarrow{\mathbb{p}} p$) and in that sense exacerbates the “width” of the regression problem (and lowers multidimensional sampling density). Like Bayesian inference, permutation testing is exact but, in most applications, inference is conditional on the observed sample, which may limit statistical power (Mehrotra et al. 2003), exacerbating the problem of high variance caused by n small; although this conditioning might be relaxed (cf. Mehrotra et al. 2003). Concerns have also been raised about proper error control under permutation testing in wide data sets (Kaizar et al. 2011).

Suppose one is willing to make the out-of-sample assumption that the observations within $\mathfrak{o}(n|f)$ are drawn from a symmetric distribution or a distribution that can be transformed to symmetry. This permits extension outside the region defined by C_p due to the introduction (really a form of imputation) of values that “mirror” observed values in $\mathfrak{o}(n|f)$. For an introduction to this topic, see Section 1 of Modarres (2002). This too is the basic concept underlying the *wild bootstrap*, although here the symmetry assumption is imposed conditionally, to address heteroscedasticity, upon fitted regression model residuals

(Davidson and Flachaire 2008). In our view, symmetry assumptions, which are out-of-sample, must be strongly motivated on scientific grounds.

3.4.3 Bidirectional shrinkage

3.4.3.1 General concepts: At times, an immunologist may be interested in *analogous* parameters based on the same feature(s) but each observed within a distinct subpopulation represented within a SWDS. For instance, the linear slopes of a particular set of immune features on age are to be estimated within patients receiving different immunosuppressive regimens following lung transplants.

One means of reducing variance (instability) is to apply *shrinkage* methods. Shrinkage methods shift the original (unconstrained) estimate of the parameter to an *optimal degree* toward some target value. This introduces some bias into the resultant shrunken (constrained) estimate but with a concomitant reduction in variance of the estimate such that the shrunken estimate has *lower total error* (bias + variance) than the original estimate. In the present setting, for the same immune feature, parameters may be estimated separately within each subpopulation represented within the SWDS; and then these original estimates could be shrunk toward each other to improve their stability (reduce variance) and reduce total error.

3.4.3.2 Theoretical details: Hierarchical estimation frameworks, notably including fully Bayesian and empirical Bayes, permit shrinkage of analogous parameters' estimates, typically toward each other (i.e., the average subpopulation), with possibly attendant reduction in risk (expected loss); although, in theory, other shrinkage targets may be employed (e.g., constants such 0 or 1). Shrinkage toward an in-sample mean may have more desirable risk properties (Weigel et al. 1991). Here, shrinkage across subpopulations for the same parameter(s) is denoted as “*o*-shrinkage.”

Shrinkage may also be possible *across* different non-analogous parameters (features) within observational units (e.g., Diedrichsen et al. 2011), denoted here as “*f*-shrinkage.” For example, estimates of a linear slope over time could be shrunk toward each other (e.g., an average) across different cytokines. To mitigate unnecessary introduction of bias, *f*-shrinkage should be confined within groups of functionally similar features (e.g., pro-inflammatory) so that the shrinkage target (e.g., average across these features) is sufficiently close—within the “effective interval” (Mehta and Srinivasan 1971)—to each unconstrained estimate to permit gains in reduction of estimation risk. In-sample *f*-shrinkage can require accounting for covariance between the original estimate and the shrinkage target.

As a simple example, suppose we wished to estimate means of immune features by age group (young vs. elderly). Here *o*-shrinkage would be of the two different age groups' means toward each other; *f*-shrinkage might be within cell types (e.g., within B cell features, within T-cell features, etc.).

Though attractive in theory for estimation risk reduction, bi-directional shrinkage offers some important challenges to implementation. Among these are the specific choice of shrinkage method and whether or not the same method is used for *o*-shrinkage and *f*-shrinkage. Also, how to partition features into groups may often not be clear. Within the

influenza data set, for example, cytokines could be grouped as pro-inflammatory, anti-inflammatory, etc.; but many cytokines have different functions depending upon concentration and/or cell type. One solution may be to allow a given feature to shrink toward multiple targets (George 1986). Bidirectional shrinkage merits further research.

3.5 Immunological interpretation and an appeal for parameter retention

3.5.1 General concepts—Generally speaking, regression models are fit to a SWDS for two distinct purposes, *prediction* and *hypothesis testing*. Prediction models are used for forecasting, prognostics, classification, or calibration, wherein performance is measured in terms of how well predicted values of the outcome variable agree, in the statistical sense, with *observed* outcome.⁴ Prediction models are not of further interest here. Rather, focus is on hypothesis testing that is performed for the specific purpose of understanding the degree to which each individual immunological feature contributes to variation in outcome. *Removal or “selecting out” of individual features (Section 3.1) is completely at cross-purposes with this interpretive goal.*⁵ Instead, each feature and its parameter(s) should be retained along with the hypothesis test(s), parameter estimate(s), confidence region(s), etc. that together constitute a *complete body of strength of evidence* regarding that feature's contribution, including the case of no contribution at all, to variation in the outcome variable.

The argument can also be made that measured features are just that—*only what immunologists and other scientists can measure*, oftentimes quite imperfectly—of true underlying biological constructs (e.g., networks, cascades). Even though we do not observe these constructs directly, we may be able to derive *latent* features from observed features that more closely approximate these true biological constructs. That is, the problem becomes one of segregating pure noise from dimensions or *axes* of biological signal.

3.5.2 Theoretical details—Discussion thus far has been of $\mathcal{S}_{\Pi}$ that treat each of the p parameters as a separate “sampling unit,” such that each parameter can be included or excluded separately to generate an instance $\mathbf{p}(p)$, although certainly methods such as the group lasso can perform selection on multi-parameter sets. More to the point, two approaches that are aligned with a hypothesis-testing, interpretive framework are a) to shrink parameter estimates smoothly toward a pertinent target and/or b) to sample some orthogonal representation of the feature space and not remove observed features but instead remove dimensions of pure noise from that feature space (and corresponding parameter space generated by $\mathbf{p}$). Denoising (Section 3.6) combines these approaches.

A classic approach to dimension selection and derived latent features, via a linear approximation $h_{\lambda}(\mathbf{X}_i\boldsymbol{\beta}_i) \approx h_{\lambda}(\mathbf{X}_i\hat{\boldsymbol{\beta}}_i)$, is (incomplete) principal components regression (IPCR; Rencher 1998, pp. 364-370). The approximation $h_{\lambda}(\mathbf{X}_i\boldsymbol{\beta}_i) \approx h_{\lambda}(\mathbf{X}_i\hat{\boldsymbol{\beta}}_i)$ is not particularly restrictive because many models are constructed of linear predictors to start and nonlinear models oftentimes can be approximated to desired degree through appropriate basis

⁴Measures of statistical agreement in this setting include prediction mean square error and intra-class correlation.

⁵Also an additional statistical problem is introduced—consistency of variable retention (i.e., are the “true” set of explanatory features retained).

expansions in the linear predictor. Exceptions are those cases where the parameters of a nonlinear model have particularly valuable scientific meaning that is lost upon linear re-parameterization.

3.6 Denoising

3.6.1 General concepts—A SWDS is typically represented by a rectangular matrix with a separate column for each feature that contains data from all participants and a separate row for each participant that contains data on all features. (For present purposes, we will assume that no data are missing within the SWDS). Within this matrix are some biological signal components with the remainder being noise (Section 3.1) and this signal may be separable into distinct biological dimensions or axes (Section 3.5.1). A matrix can be *decomposed* in different ways. Roughly, a decomposition partitions a matrix into different axes, each of which has a *direction* and, importantly, an associated *scale*. An axis of large scale represents structure, presumably biological structure,⁶ whereas an axis of small, near-zero scale represents noise. Matrix denoising technologies are designed to segregate structural axes from noise axes thereby permitting regression analyses to operate on biological signal.

3.6.2 Theoretical details—For SWDS $\mathbf{w}(n, f) \stackrel{\text{def}}{=} \mathbf{X}$, briefly, a hard-threshold, reduced-rank approximation to the linear least-squares normal equations is given by $\tilde{\boldsymbol{\beta}} \stackrel{\text{def}}{=} (\mathbf{C}_r \mathbf{D}_r^{-1} \mathbf{C}_r') \mathbf{X}' \mathbf{Y}$, where $\mathbf{D}_r$ is the $r \times r$ diagonal matrix of the eigendecomposition of $\mathbf{X}' \mathbf{X}$ ($\mathbf{X}$ of rank m) of the first r , $1 \leq r \leq m \leq n < f$, eigenvalues and $\mathbf{C}_r$ is the matrix of the corresponding first r eigenvectors (see Section 3.1.2 and cf. Rencher 1998, pp. 364-370). This denoising mitigates loss in estimation efficiency (and attendant statistical power) through elimination of those eigencomponents of greatest contribution to collinearity among regressors (Rencher 1998, p. 364). The immediate question is how to select r . To avoid an out-of-sample restriction, such as multivariate normality for the Marcenko-Pastur distribution (Kyriakopoulos et al. 2009, p. 528) or the asymptomatic ($f \rightarrow \infty$) and strictly low rank ($r \ll f$) framework of Gavish and Donoho (2014), that may not apply in a SWDS, the noise distribution might be obtained instead via iterated random permutation of the vector of observations, independently for each feature (e.g., Peterson and Ford 2013). This approach has three drawbacks. Permutation can be computationally burdensome. Also, permutation procedures do make strong out-of-sample assumptions—notably forms of exchangeability (e.g., Xu and Hsu 2007). Some form of Studentization of the permutation procedure (*sensu* Janssen 1997) may enhance exchangeability. Finally, permutation is *conditional* on the observed SWDS $\mathbf{w}(n, f)$ and thereby completely ignores $\sigma_{\hat{\mu}} \mathbf{e}_{\hat{\mu}}$. This may be an especially important limitation because n is small and thus $\sigma_{\hat{\mu}} \mathbf{e}_{\hat{\mu}}$ can be large. We circumvent the latter two problems by obtaining a null distribution of eigenvalues by *independently* $Q(n)$ -bootstrapping each of the f features (f univariate $Q(n)$ -bootstraps), column-concatenating the resultant f resamples, and performing an eigendecomposition of that null matrix. We only retain those eigenvalues above the 90th percentile τ from the null matrix, which is a form of eigenvalue shrinkage (see, for instance, Chi and Lange 2014) that

⁶As a word of caution, systematic structure may be introduced into a SWDS that is pure experimental artifact, which is why “data normalization” procedures such as correction for nonspecific binding are so important in the analysis of immunological SWDSs.

may result in smooth shrinkage in the parameter space induced by $\mathbf{p}$. We employ the 90th percentile to allow for the possibility that the null eigenvalue distribution may extend into the distribution of signal eigenvalues (Peterson and Ford 2013). Obviously, estimates $\hat{r}$ of signal rank r are design-constrained such that $\hat{r} \leq n$, another limitation of SWDSs. Because n is small and $n < p$, one could attempt to introduce additional stability through initially smoothly shrinking eigenvalue estimates (e.g., directly toward 0 or toward each other or indirectly through shrinkage of the sample correlation matrix (Schafer and Strimmer 2005)) prior to hard eigenvalue thresholding.

Once noise eigenvalues are zeroed (e.g., via the τ threshold), a linear regression model can be fit using IPCR. IPCR is a two-stage, *unsupervised* process of denoising followed by regression estimation. (To contrast, denoised principal components analysis is a one-step alternative in which the denoising step is *supervised* via inclusion of outcome. Namely, define the column-concatenated matrix $\mathbf{Z}^{\text{ref}} = [\mathbf{Y}|\mathbf{X}]$ and perform eigenstructure denoising on $\mathbf{Z}$.)

4. Results

4.1 Bias reduction example

4.1.1 Immunological background—We illustrate this method through application to the immunophenotyping data set of the influenza vaccine recipients. We wished to estimate the relationship between immune features (as outcomes) and age (as *causal* predictor). Age is particularly useful for illustrating methods of confounder adjustment because studies of immune response over the adult lifespan can, practically, only be conducted via cross-sectional observational studies. Cross-sectional studies are potentially highly problematic because many factors can vary across *different* people of different ages that drive variation in immune features but which are distinct from the effect of aging *per se*. For example, older adults have been exposed to a different ensemble of influenza viral strains compared to younger adults.

4.1.2 Computational details and findings—Prior to statistical analysis, each immune feature Y_i , $i \in \{1, \dots, f_1\}$, $f_1 = 229$, was centered and scaled as $(Y_i - m_i)/s_i$ for sample mean m_i and sample standard deviation s_i . We applied the surrogate variable procedure of Leek and Storey (2008) to these data. With surrogate variable adjustment, attained significance levels (p -values) for the effects of age were calculated using the optimal discovery procedure of Woo et al. (2011). Leek and Storey (2008) argue that their surrogate variable analysis can eliminate dependence among hypothesis tests across features. For this reason, we controlled the false discovery rate at 5% across these 229 tests using the data-adaptive, two-stage procedure of Benjamani et al. (2006) —a procedure that was designed for independent tests but has also been shown to provide superior error control for possibly dependent hypothesis tests (Kim and van de Weil 2008), if for example, surrogate variable adjustment imperfectly orthogonalizes hypothesis tests.

4.1.3 Immunological implications—Table 1 lists those features that are detected to vary with age without adjustment for possible unobserved confounders. In contrast, the list of

immune features that are found to vary with age with adjustment for possibly biasing unobserved confounders is longer and includes some major additional innate and adaptive immune cell subsets (B cells, NK cells, and monocytes). Estimates of slopes on age are provided with and without bias adjustment. Another potentially important finding is that estimated slopes on age are steeper (stronger) with surrogate variable adjustment than without, indicating that the presence of potentially biasing unobserved confounders causes the impact of age on these features to be underestimated in standard analyses that fail to correct for this bias. Moreover, if this correction has removed the effects of all unobserved confounders (and unobserved confounders only), we can properly state that these adjusted slopes represent the *attributable* (causal) effect of aging on each of these sixteen immune features (Groenwold et al. 2009). For example, estimates of positive slopes in the presence of bias adjustment suggest that aging is *causing* levels of those immune features to increase. Two examples (monocytes and NK cells) of these effects of age are shown in Figure A of the online supplement.

This distinction between changes *associated* with aging versus changes *caused* by aging should be of interest to immunologists. Cross-sectional studies of age-related changes in immune features (e.g., Hearps et al. 2012, Whiting et al. 2015) do adjust for a few other factors (e.g., gender); but only those factors that they have observed which may not represent a comprehensive correction. When these few factors do exhaust the other age-related but non-aging processes, published results should coincide with those of Table 2. Table 2 indicates that aging is estimated to cause NK cells to increase with age, which is commensurate with the comprehensive review of Gayoso et al. (2011) that this subset is “preserved or increased” with aging. Also, Whiting et al. (2015) noted an increase in serum levels of the cytokine MIG (CXCL9) with age, which agrees with the bias-adjusted positive slope of MIG on age in Table 2. Notably, here in a sample of only $n = 35$ adults we have reproduced the finding of Whiting et al. (2015) in $n = 243$ adults. In contrast, other immune features may be governed by many other strong age-related but non-aging processes. According to Table 2, monocytes and $CD4^+$ T cells are caused to increase due to aging. The majority class of monocytes (classical, $CD14^{++}CD16^-$) has been documented to be *lower* at older ages; however, the aging body apparently compensates with increased proportions of nonclassical ($CD14^+CD16^{++}$) and, especially, intermediate ($CD14^{++}CD16^+$) monocyte subsets (Hearps et al. 2012). Thymic output of naïve $CD4$ T cells decreases with age (Holmes and Lewis 2014); but with age there may also be compensatory increased proliferation of naïve $CD4$ T cells in the periphery (Kilpatrick et al. 2008) and increased longevity of naïve $CD4$ T cells (Dowling and Hodgkin 2009). With monocytes and $CD4$ T cells as examples, older individuals may have less of each these subsets due to the many assaults on the body with the passage of time (e.g., cumulative exposure to pathogens); but the effect of aging per se may be homeostatic, driving compensatory mechanisms to attempt to offset these losses.

4.2 Tail extrapolation example

4.1.2 Immunological background—We illustrate application of the $Q(n)$ -bootstrap using the influenza data. Recall that the purpose of this procedure is twofold. We want to obtain confidence intervals on parameter estimates of interest to immunologists with a

minimum of imposed out-of-sample assumptions while at the same time making sure that confidence interval widths are accurate and not biased toward being too narrow. When confidence intervals are biased toward being too narrow, the immunologist can be misled into thinking that parameter estimates are more stable than they truly are—and that can lead to false and irreproducible discoveries.

Here we illustrate this procedure in application to estimating age-related changes in immune features with age. This $Q(n)$ -bootstrap procedure could be combined with bias reduction (Sections 3.2 and 4.1) and in Section 4.3 we illustrate a multivariate version of the $Q(n)$ -bootstrap to facilitate combination of these two methods. Together these would provide estimates of the causal effects of age with accurate confidence interval coverage with a bare minimum of out-of-sample assumptions.

4.2.2 Computational details and findings—A simple linear regression on age was performed separately for each centered-and-scaled feature, with estimates of slope on age and the intercept estimated via ordinary least squares. The quantile function $Q(n)$ was then estimated from the sample order statistics of the regression residuals from the ordinary-least-squares fit for the i^{th} feature. Resampling was from these residuals to reflect that fact that ages were largely fixed by the stratified sampling study design. A total of 2,500 resamples were drawn from $\hat{Q}_i(n)$ and the linear regression model was re-fit to each resample via ordinary least squares.

Figure 1 summarizes the results for a stratified random sampling of the 229 features. Presented are 95% confidence bounds using 1) percentiles (2.5%, 97.5%) of slope estimates from the $Q(n)$ -bootstrap (Q), 2) percentiles (2.5%, 97.5%) of a basic percentile bootstrap (S; Davison and Hinkley 1997, pp. 202-203) of ordinary-least-squares regression residuals, and 3) the out-of-sample normal approximation based on Student's t distribution (N; Neter et al. 1996). Any confidence band entirely above or below the zero line (gray) suggests the presence of an age-related effect on that outcome.⁷ For example, regardless of which method of confidence interval construction is employed, levels of IL8 at week 0 appear to be increased with age.

Confidence bounds based on the basic percentile bootstrap were narrower than those from the $Q(n)$ -bootstrap for over two-thirds (~68%) of the 229 features (data not shown), which is consistent with the hypothesis that tail coverage tends to be insufficient without tail extrapolation. In a sample this small ($n = 35$), the normal approximation may not always be satisfactory (Ratcliffe 1968), which may account for some of the observed differences between the $Q(n)$ -bootstrap and the normal approximation (e.g., CXCR3+ CD8+T cells, and PAI1 at week 0).

Importantly, through various numerical experiments, Hutson (2002) found the $Q(n)$ -bootstrap to have improved coverage probabilities compared to the simple percentile

⁷This is without correction for unobserved confounders. That would require sampling from the joint distribution of the features in order to permit identification of surrogate variables. We provide extension of the $Q(n)$ -bootstrap to sampling from joint distributions in Section 4.3.

bootstrap, bootstrap- t , and “smoothed kernel quantile estimator” (Sheather and Marron 1990). In particular, he found the simple percentile bootstrap to provide undercoverage relative to the $Q(n)$ -bootstrap.

4.2.3 Immunological implications—Results here suggest that immunological studies conducted to date in small samples, particularly those relying upon strong and unverifiable out-of-sample assumptions (e.g., normal distribution of sampled population) or standard bootstrap procedures, may require re-evaluation where conclusions depended upon statistical inference employing tail regions of sampling distributions (Section 3.4.1.1) of parameter estimates (confidence intervals, p -values).

4.3 Denoising examples

4.3.1 Background—A question of scientific interest from the influenza sample is as follows. Which collection of measured immune features measured at baseline (week 0) are associated with subsequent vaccine-specific B cell response at week 1? Here, B cell response is measured in terms of 1) quantity of IIV-specific IgA ASCs and 2) quantity of IIV-specific IgG ASCs. Given the design of the study (Section 2.1), this question is addressed with adjustment for age.

4.3.2 Computational details and findings—We performed two IPCRs, one for each vaccine-specific B-cell outcome—IgA ASCs and IgG ASCs. Candidate predictors in each IPCR were the remaining features measured at baseline by flow cytometry and Luminex assays (online supplement Appendix B) *plus age*. Our overall procedure included *extending the $Q(n)$ -bootstrap to the multivariate setting*.

1. $w(n, f) \rightarrow w$: Unmix $w(n, f)$ via independent component analysis. Extract $\min\{n, f\} - 1$ independent components, forming w .
2. $w \rightarrow w(n, f)^i$: Draw $b, i \in \{1, \dots, b\}$, $Q(n)$ -bootstrap replicates of each of the independent components of w prior to mixing back to $w(n, f)^i$.⁸ The quantile function is estimated via linear interpolation and n values are drawn randomly from this estimated quantile function (Section 3.4.1.2) separately for each column of w .
3. Perform IPCR on each $w(n, f)^i$ with signal rank r estimated by resampling $w(n, f)$ per the procedure described in Section 3.6.2, giving estimates $\hat{r}^i, i \in \{1, \dots, b\}$.

For application here, $b = 1,250$, $f = p = 130$ baseline features. (Restricted to baseline, sample size was slightly larger at $n = 37$.) For the bootstrap, we drew simple random samples with replacement rather than stratifying on age because age was included as another baseline candidate predictor for purposes of this analysis. The $Q(n)$ -bootstrap estimate of the sampling distribution of estimate $\hat{r}$ was $\hat{r} < 8$ (0% frequency), $\hat{r} = 8$ (0.6 % frequency), $\hat{r} = 9$

⁸The idea to transform to independence was suggested to us by Dr. Alan Huston (personal communication).

(1.5 % frequency), $\hat{r} = 10$ (7.5 % frequency), $\hat{r} = 11$ (26.6 % frequency), $\hat{r} = 12$ (40.0 % frequency), $\hat{r} = 13$ (20.6 % frequency), $\hat{r} = 14$ (3.1 % frequency), $\hat{r} = 15$ (0.1 % frequency), and $15 < \hat{r}$ (0% frequency). At the modal estimate of $\hat{r} = 12$, signal-to-noise was estimated to be $\hat{\sigma} \approx 2.9$. As such, signal is estimated to be concentrated within a small quantity of dimensions in this otherwise noisy wide feature space.

In addition to confidence intervals, the $Q(n)$ -bootstrap resamplings could be used to construct equal-tail bootstrap p -values (Davidson and MacKinnon 2007). Using this technique, for example, yields $p = 0.024$ for the association between quantity of IIV-specific IgG ASCs and baseline TGFB (Figure 3).

4.3.3 Immunological implications—For quantity of IIV-specific IgA ASCs, one of the 130 baseline immune features' 95% confidence bounds fell entirely below zero (CD314+ CD94- NK cells, Figure C of online supplement), which indicates possible negative association with quantity of IIV-specific IgA ASCs. Quantity of IIV-specific IgG ASCs may demonstrate age-related decline (Figure 2), which has been reported elsewhere in this same dataset (Sasaki et al. 2011), although this association with age, *as a causal estimate of the effect of aging*, could be biased by unobserved confounders, which surrogate variable analysis could address (Section 3.2). Indeed, the longitudinal study of Mosterín Höpping et al. (2016) suggests that this age-related decline in IIV-specific IgG ASCs may be due to prior exposure and not a true aging effect. Baseline TGFB (Figure 3) and baseline IFNG may be positively associated with quantity of IIV-specific IgG ASCs. A positive association between TGFB and IgG is mechanistically possible given that IgG can bind TGFB (e.g., Harada et al. 2002). As with any SWDS, these potential associations with cytokines need to be interpreted cautiously, and require verification through additional studies, especially since they do not reflect any form of correction for performing multiple hypothesis testing.⁹

5. Discussion and Conclusions

5.1 Immunological

Presented principles, concepts and methods have been illustrated through application to a SWDS of adults spanning a wide range in ages that were assayed for B cell response to inactivated influenza vaccination and multiple immunophenotypes before and after vaccination. Some potentially important findings for future research are as follows.

1. Immunologists may wish to distinguish age-related differences in immune features from changes in immune features caused by aging. The type of statistical bias adjustment described in Section 3.2 and illustrated in Section 4.1, perhaps coupled with mechanistic studies of immune function (e.g., assessments of the function of the thymus, lifespans of individual cells of particular immune cell types, immune cell differentiation, cytokine binding,

⁹To be fair, strictly speaking, such correction is not without an ongoing history of controversy (Carmer and Walker 1982, Saville 1990, Hurlbert and Lombardi 2012).

etc.) and highly-controlled medium duration longitudinal studies (e.g., Mosterín Höpping et al. 2016), can help the working immunologist to identify those changes in immune features that are truly the homeostatic, compensatory mechanisms caused directly by aging per se.

2. Because immune response is complex, wide panels of features must be captured and the cost of those wide panels limit sample size. The $Q(n)$ -bootstrap may prove to be an invaluable analytic tool in these settings because it allows the working immunologist to obtain accurate estimates of the stability of immune parameter estimates with a bare minimum of imposed assumptions. Code for performing the $Q(n)$ -bootstrap is available upon request.
3. The immunologist should fully exploit technological capabilities to measure many immune features on each participant in a study. Perhaps somewhat obviously, an exhaustive panel of features means that all or most features that truly contribute to the system being studied (i.e., the signal component) will be available for analysis. Perhaps less obvious is that liberal inclusion of noise features (sensu Section 3.1) can facilitate accurate distinction of biological signal of interest from noise, perhaps using the methods of Section 3.6, because noise is more fully characterized, although “curse of dimensionality” applies.

5.2 Theoretical

To help the analyst focus on the biological mechanisms generating data, particular emphasis has been placed on methods that rely upon a minimum of out-of-sample assumptions while, at the same time, maintaining low variance, minimal bias, and, importantly, *interpretability*. This said, no general strategy can be completely assumption-free. Without question, for instance, the $Q(n)$ -bootstrap is not completely free of out-of-sample assumptions as it does presume linearity for extrapolation; however, it remains attractive because that extrapolation does not impose the stronger assumptions associated with any particular parametric distribution and instead relies heavily on information contained within $w(n, f)$; and linear extrapolation should be accurate to first-order. We introduced an extension of the $Q(n)$ -bootstrap to the multivariate setting. Further research would be useful that examines the accuracy of confidence interval coverage under this extension. Extensions to hypothesis testing are also possible. In particular, data sharpening (Braun and Hall 2001) could permit construction of $Q(n)$ -bootstrap null distributions with, by construction, a bare minimum of imposed out-of-sample constraints.

Inescapably, a small sample size n on the observation axis is a limitation. For example, as observed in Section 3.6.2, estimates $\hat{r}$ of signal rank r are design-constrained such that $\hat{r} \leq n$. An analysis that yields a modal $\hat{r} \leq n$ is an immediate “red flag” that the sample on hand is

inadequate to describe the structure within the sampled population. Sampling on the feature axis can be just as limiting. Another warning sign is when a surrogate variable analysis, properly performed, results in extensive changes in estimates of regression model coefficients, especially when no clues exist as to the identity of the unobserved confounders. Further investigation would clearly be needed and straightaway to determine if these confounders are experimental artifact (e.g., correction for instrumentation background), known biological structure that simply was not measured, or, the most intriguing possibility, previously unknown biological structure.

Shrinkage, including that induced by denoising (Section 3.6), can stabilize parameter estimates in SWDSs albeit with some introduction of bias. Many methods exist that provide shrinkage that is optimal on some combined risk measure, such as mean square error; however, to immunological investigators, the assurance that estimation variance plus squared bias is minimal may be too nebulous. Of greater practical utility and transparency, perhaps, would be *bias-constrained* shrinkage estimators wherein efficiency gains are permissible but only to the extent that bias does not exceed some *a priori* explicitly-defined maximum bias threshold.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by NIH/NIAID grant awards 5U19AI057229 to Dr. Mark Davis, Director of the Institute of Immunity Transplantation and Infection, Stanford University. We are grateful to Dr. Holden Maecker, Director of the Human Immune Monitoring Center, Stanford University, for providing the immunophenotyping data sets and for his review of and comments on draft versions of this paper, and to Dr. Harry Greenberg for his encouragement.

References

- Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*. 2006; 93:491–507.
- Biancotto A, Fuchs JC, Williams A, Dagur PK, McCoy JP. High dimensional flow cytometry for comprehensive leukocyte immunophenotyping (CLIP) in translational research. *Journal of Immunological Methods*. 2011; 363(2):245–261. [PubMed: 20600079]
- Braun WJ, Hall P. Data sharpening for nonparametric inference subject to constraints. *Journal of Computational and Graphical Statistics*. 2001; 10(4):786–806.
- Carmer SG, Walker WM. Baby Bear's dilemma: a statistical tale. *Agronomy Journal*. 1982; 74:122–124.
- Chi EC, Lange K. Stable estimation of a covariance matrix guided by nuclear norm penalties. *Computational Statistics and Data Analysis*. 2014; 80:117–128. [PubMed: 25143662]
- Cochran, WG. Sampling techniques. third. New York: John Wiley & Sons; 1977.
- Davison, AC.; Hinkley, DV. Bootstrap methods and their application. Cambridge: Cambridge University Press; 1997.
- Davidson R, Flachaire E. The wild bootstrap, tamed at last. *Journal of Econometrics*. 2008; 146(1): 162–169.
- Davidson R, MacKinnon JG. Improving the reliability of bootstrap tests with the fast double bootstrap. *Computational Statistics and Data Analysis*. 2007; 51(7):3259–3281.
- Diaconis P, Freedman DA. On the consistency of Bayes estimates (with discussion). *Annals of Statistics*. 1986; 14(1):68.

- Diedrichsen J, Ridgway GR, Friston KJ, Wiestler T. Comparing the similarity and spatial structure of neural representations: a pattern-component model. *NeuroImage*. 2011; 55(4):1665–1678. [PubMed: 21256225]
- Dowling MR, Hodgkin PD. Modelling naive T-cell homeostasis: Consequences of heritable cellular lifespan during ageing. *Immunology and Cell Biology*. 2009; 87(6):445–456. [PubMed: 19290017]
- El-Aroui M, Diebolt J. On the use of the peaks over thresholds method for estimating out-of-sample quantiles. *Computational Statistics and Data Analysis*. 2002; 39(4):453–475.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96(456):1348–1360.
- Gavish M, Donoho DL. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*. 2014; 60(8):5040–5053.
- Gayoso I, Sanchez-Correa B, Campos C, Alonso C, Pera A, Casado JG, Morgado S, Tarazona R, Solana R. Immunosenescence of human natural killer cells. *Journal of Innate Immunity*. 2011; 3(4):337–343. [PubMed: 21576928]
- George EI. Minimax multiple shrinkage estimation. *The Annals of Statistics*. 1986; 14(1):188–205.
- Good, P. *Permutation tests*. New York: Springer-Verlag; 1994.
- Groenwold RHH, Hak E, Hoes AW. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. *Journal of Clinical Epidemiology*. 2009; 62(1):22–28. [PubMed: 18619797]
- Haneuse S, Schildcrout J, Crane P, Sonnen J, Breitner J, Larson E. Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology*. 2009; 32(3):229–243. [PubMed: 19176974]
- Hanson T, Johnson WO. Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*. 2002; 97(460):1020–1033.
- Harada M, Tatsugami K, Nomoto M, Nomoto K. Circulating immunoglobulin-bound transforming growth factor β at a late tumour-bearing stage impairs antigen-specific responses of CD4⁺ T cells. *Clinical and Experimental Immunology*. 2002; 128(2):204–212. [PubMed: 12041507]
- Harris G, Chen W. Profiling of cytokine and chemokine responses using multiplex bead array technology. *Methods in Molecular Biology*. 2013; 1061:265–278. [PubMed: 23963943]
- Harville, DA. *Matrix algebra from a statistician's perspective*. New York: Springer; 1997.
- Hastie, T.; Tibshirani, R.; Friedman, JH. *The elements of statistical learning: Data mining, inference, and prediction*. second. New York: Springer; 2009.
- Hearps AC, Martin GE, Angelovich TA, Cheng W, Maisa A, Landay AL, Jaworowski A, Crowe SM. Aging is associated with chronic innate immune activation and dysregulation of monocyte phenotype and function. *Aging Cell*. 2012; 11(5):867–875. [PubMed: 22708967]
- Hedt BL, Pagano M. Health indicators: Eliminating bias from convenience sampling estimators. *Statistics in Medicine*. 2011; 30(5):560–568. [PubMed: 21290401]
- Holmes TH, Lewis DB. Bayesian immunological model development from the literature: example investigation of recent thymic emigrants. *Journal of Immunological Methods*. 2014; 414:32–50. [PubMed: 25179832]
- Hurlbert SH, Lombardi CM. Lopsided reasoning on lopsided tests and multiple comparisons. *Australian and New Zealand Journal of Statistics*. 2012; 54:23–42.
- Hutson AD. A semiparametric quantile function estimator for use in bootstrap estimation procedures. *Statistics and Computing*. 2002; 12:331–338.
- Janssen A. Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statistics and Probability Letters*. 1997; 36(1):9–21.
- Jin Z, Ying Z, Wei LJ. A simple resampling method by perturbing the minimand. *Biometrika*. 2001; 88(2):381–390.
- Kaizar EE, Li Y, Hsu JC. Permutation multiple tests of binary features do not uniformly control error rates. *Journal of the American Statistical Association*. 2011; 106(495):1067–1074.

- Kilpatrick RD, Rickabaugh T, Hultin LE, Hultin P, Hausner MA, Detels R, Phair J, Jamieson BD. Homeostasis of the naive CD4⁺ T cell compartment during aging. *Journal of Immunology*. 2008; 180(3):1499–1507.
- Kim KI, van de Wiel MA. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*. 2008; 9
- Kyriakopoulos F, Thurner S, Puhr C, Schmitz SW. Network and eigenvalue analysis of financial transaction networks. *European Physical Journal B*. 2009; 71(4):523–531.
- Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*. 2007; 3(9):1724–1735. [PubMed: 17907809]
- Leek JT, Storey JD. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(48):18718–18723. [PubMed: 19033188]
- Léon Y, Peeters L, Quinqu M, Surry Y. The use of maximum entropy to estimate input-output coefficients from regional farm accounting data. *Journal of Agricultural Economics*. 1999; 50(3): 425–439.
- Mehrotra DV, Chan ISF, Berger RL. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics*. 2003; 59:441–450. [PubMed: 12926729]
- Mehta JS, Srinivasan R. Estimation of the mean by shrinkage to a point. *Journal of the American Statistical Association*. 1971; 66(333):86–90.
- Modarres R. Efficient nonparametric estimation of a distribution function. *Computational Statistics and Data Analysis*. 2002; 38(5):75–95.
- Montoya JG, Holmes TH, Anderson JN, Maecker HT, Rosenberg-Hasson Y, Valencia I, Norris J, Chu L, Younger JW, Tato CM, Davis MM. A novel profile of circulating cytokines is associated with disease severity in chronic fatigue syndrome patients. 2016 In preparation.
- Mosterín Höpping A, McElhaney J, Fonville JM, Powers DC, Beyer WEP, Smith DJ. The confounded effects of age and exposure history in response to influenza vaccination. *Vaccine*. 2016; 34(4): 540–546. [PubMed: 26667611]
- Neter, J.; Kutner, MH.; Nachtsheim, CJ.; Wasserman, W. *Applied linear statistical models*. fourth. Boston: WCB McGraw-Hill; 1996.
- Peterson, LE.; Ford, CE. Random matrix theory and covariance matrix filtering for cancer gene expression. In: Peterson, LE., Masulli, F., Russo, G., editors. *Computational Intelligence Methods for Bioinformatics and Biostatistics Lecture Notes in Computer Science*. Vol. 7845. Berlin: Springer-Verlag; 2013. p. 173–184.
- Phillips D, Ghosh D. Testing the disjunction hypothesis using Voronoi diagrams with applications to genetics. *The Annals of Applied Statistics*. 2014; 8(2):801–823.
- Rasch D, Kubinger KD, Moder K. The two-sample t test: Pre-testing its assumptions does not pay off. *Statistical Papers*. 2011; 52(1):219–231.
- Ratcliffe JF. The effect on the *t* distribution of non-normality in the sampled population. *Applied Statistics*. 1968; 17(1):42–48.
- Rencher, AC. *Multivariate statistical inference and applications*. New York: John Wiley & Sons, Inc; 1998.
- Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000; 11(5):550–560. [PubMed: 10955408]
- Sasaki S, Sullivan M, Narvaez CF, Holmes TH, Furman D, Zheng N, Nishtala M, Wrammert J, Smith K, James JA, Dekker CL, Davis MM, Wilson PC, Greenberg HB, He X. Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. *Journal of Clinical Investigation*. 2011; 121(8):3109–3119. [PubMed: 21785218]
- Saville DJ. Multiple comparison procedures: The practical solution. *The American Statistician*. 1990; 44(2):174–180.
- Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*. 2005; 4

- Sheather SJ, Marron JS. Kernel quantile estimators. *Journal of the American Statistical Association*. 1990; 85(410):410–416.
- Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*. 2011; 27(11): 1496–1505. [PubMed: 21471010]
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58(1):267–288.
- Thompson, SK. Sampling. New York: John Wiley & Sons, Inc; 1992.
- Watson DA, Gaskill DF, Brown LO, Doorn SK, Nolan JP. Spectral measurements of large particles by flow cytometry. *Cytometry Part A*. 2009; 75(5):460–464.
- Weigel KA, Gianola D, Tempelman RJ, Matos CA, Chen IHC, Wang T, Bunge R, Lo LL. Improving estimates of fixed effects in a linear model. *Journal of Dairy Science*. 1991; 74:3174–3182.
- Whiting CC, Siebert J, Newman AM, Du H, Alizadeh AA, Goronzy J, Krishnan E, Fathman GC, Maecker HT. Large-scale and comprehensive immune profiling and functional analysis of normal human aging. *PLoS ONE*. 2015; 10(7)
- Woo S, Leek JT, Storey JD. A computationally efficient modular optimal discovery procedure. *Bioinformatics*. 2011; 27(4):509–515. [PubMed: 21186247]
- Xu H, Hsu JC. Applying the generalized partitioning principle to control the generalized familywise error rate. *Biometrical Journal*. 2007; 49(1):52–67. [PubMed: 17342949]

Figure 1.

Slope estimates for the regression on age of 49 randomly selected immune features. Values for each feature were centered and scaled as $(Y_i - m_i)/s_i$ for sample mean m_i and sample standard deviation s_i prior to regression analysis (vertical axis). Black lines are 95% confidence bands based on asymptotic normal (N), $Q(n)$ -bootstrap (Q) and basic percentile bootstrap (S) approximations, where $Q(n)$ -bootstrap and basic percentile bootstrap were performed on residuals from ordinary least squares fit to the original data. Horizontal gray line demarcates a slope of zero so that any confidence band entirely above or entirely below the gray line suggests the presence of an age-related effect on that outcome. W0 = week 0 (day 0), W1 = week 1 (days 7 and 8), and W4 = week 4 (days 28 ± 4).

Figure 2.

Caterpillar plot of 95% percent $Q(n)$ -bootstrap confidence bands on the IPCR regression coefficient estimates for the association between each of the 130 candidate baseline predictors, including age, and count of IIV-specific IgG ASC (as outcome). Confidence bands that include zero are in gray while those that do not include zero (suggesting a predictor association) are in black and annotated with the name of that immune feature. Units for vertical axis are change in standard deviations of IIV-specific IgG ASC count per change in standard deviations of the candidate predictor.

Figure 3. Relationship between centered and scaled IgG IIV ASC count and centered and scaled baseline TGFB after ordinary least squares regression adjustment for the relationship between centered and scaled IgG IIV ASC count and age.

Table 1

Estimates of immune features' associations with age (as predictor) *without* adjustment for unobserved confounders (surrogate variables). Attained significance levels (p -values) are given without and with adjustment for multiple comparisons to control the false discovery rate at 5% across the 229 immune features using the data-adaptive two-stage procedure of Benjamini et al (2006). First-stage estimate of percentage of true null hypotheses is ~95.2%. Only those features with FDR-adjusted p -values less than 5% (0.05) are listed here. Highlighted in gray are all features also listed in Table 2.

Feature	Unadjusted Slope	p-value	Adjusted p-value
CD314- CD94- CD8- T cells	0.5440	0.0010	0.0378
CD314+ CD94- CD8- T cells	-0.5143	0.0009	0.0378
CD4+ T cells	0.5163	0.0021	0.0421
HLADR- CD38+ CD8+ T cells	-0.5850	0.0002	0.0171
Central memory CD8+ T cells	0.5780	0.0004	0.0298
Live CD8+ T cells	-0.4951	0.0014	0.0421
IL15 W0	-0.4826	0.0018	0.0421
IL17 W0	-0.4772	0.0020	0.0421
IL2 W0	-0.4816	0.0019	0.0421
IL2 W4	-0.5866	0.0001	0.0171
IL8 W0	0.5504	0.0009	0.0378

Table 2

Estimates of immune features' associations with age (as predictor) with and without adjustment for unobserved confounders (surrogate variables). Attained significance levels (p -values) are given without and with adjustment for multiple comparisons to control the false discovery rate at 5% across the 229 immune features using the data-adaptive two-stage procedure of Benjamini et al (2006). First-stage estimate of percentage of true null hypotheses is ~93.0%. Only those features with FDR-adjusted p -values less than 5% (0.05) after adjustment for unobserved confounders are listed here. Highlighted in gray are all features also listed in Table 1.

Feature	Unadjusted Slope	Adjusted Slope	p-value	Adjusted p-value
CD20- CD3- cells	0.2423	0.4822	0.0026	0.0342
CD314- CD94- CD8- T cells	0.5440	0.5856	0.0011	0.0175
CD314+ CD94- CD8- T cells	-0.5143	-0.6116	0.0009	0.0175
CD4+ T cells	0.5163	0.6034	0.0007	0.0166
CD8- T cells	0.2901	0.5937	0.0002	0.0157
CD8+ T cells	-0.4296	-0.6137	0.0013	0.0182
CXCR3+ B cells	0.1829	0.6482	<.0001	0.0028
NK cells	0.4512	0.6304	0.0005	0.0157
Central memory CD8+ T cells	0.5780	0.5765	0.0010	0.0175
Monocytes	0.3892	0.6379	0.0007	0.0166
IL2 W4	-0.5866	-0.4646	0.0004	0.0157
IL5 W1	0.0866	0.4589	0.0012	0.0179
IL8 W0	0.5504	0.6191	0.0005	0.0157
MIG W1	0.3855	0.4934	0.0005	0.0157
MIG W4	0.4033	0.5137	0.0004	0.0157
VEGF W1	0.1300	0.4736	0.0011	0.0175