CROSS-STUDY ANALYSES OF MICROBIAL ABUNDANCE USING GENERALIZED COMMON FACTOR METHODS

by

Mary Gunn Hayes

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
May 2020

Table of Contents

List of Tables

List of Figures

Abstract

Micro-organisms seem to flagellate about wherever they please, in our bodies and in the natural and built environments, but they are more cunning than their meandering behavior would suggest. By creating networks of biochemical pathways, communities of microbes are able to modulate the properties of their environment and even biochemical processes within their hosts. Next-generation high-throughput sequencing has led to a new frontier in microbiology and microbial ecology which promises the ability to leverage the microbiome for good in every facet of our lives, and the stakes are high as global society hurtles toward several apocalyptic ecological crises. However, along with the fascinating complexity of microbial community dynamics comes equally complex data considerations for researchers: genomic data are high-dimensional, sparse, noisy, and refuse to cooperate with authorities. In fact, they will not even cooperate with each other, which prohibits the sorts of consensus-based validation and meta-analysis that we rely on in science. In this thesis we propose an ensemble approach for cross-study exploratory analyses of microbial abundance data, in which we first estimate the variance-covariance matrix from each dataset assuming Poisson sampling, and subsequently model these covariances jointly so as to find a shared low-dimensional subspace of the feature space. By viewing the projection of the latent true abundances onto this common structure, the variation is pared down to that which is shared among all datasets, and is likely to reflect more generalizable biological signal than can be inferred from an individual dataset. We investigate several ways of achieving this, and demonstrate that they work well on simulated and real metagenomic data in terms of signal retention and interpretability.

List of Abbreviations and Symbols Used

PCA ..... principal components analysis

PLNPCA ..... Poisson log-normal PCA

CPCA ..... common principal component analysis

FCPCA ..... Flury's common principal components analysis

SCPCA ..... stepwise common principal components analysis

MSFA ..... multi-study factor analysis

iid ..... independently and identically distributed

$O(n)$ ..... the set of orthogonal $n \times n$ matrices

I_n ..... the $n \times n$ identity matrix

tr ..... the matrix trace function

$|\cdot|$ ..... the matrix determinant function

diag ..... the diagonal function $f : \mathbb{R}^n \to \mathbb{R}^{n \times n}$

diag* ..... the diagonal function $f : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$

$\circ$ ..... the element-wise product of two matrices

Acknowledgements

I am very grateful to my readers Dr. Kenney and Dr. Ho for volunteering their time to read this eldritch text. Of course, I would like to express my deepest gratitude to my supervisors Dr. Gu and Dr. Langille for all their support and insight over the last several years. In fact, I would never have completed this thesis had Dr. Gu not promised to invite me over to make dumplings afterward.

Chapter 1

Introduction

Communities of microbes inhabit all areas of the environment, including the body cavities and exterior surfaces of larger organisms. Each microbiome can be thought of as a community of shared genes and metabolic pathways, which act as complex systems in ways defined in part by their composition, or which microbial taxa are present and in what abundance (Boon et al., 2014). These communities vary widely in composition; in studies of microbial communities that inhabit the human skin or within the human gastrointestinal tract, distributions of bacterial taxa are distinct among individuals. However, certain patterns of colonization seem characteristic across individuals, and analogously across environmental sites, given similar conditions. In fact, mounting evidence suggests that micro-organisms comprising the gut microbiome interact with host systems in myriad ways, and that understanding these associations could have profound implications for our ability to predict, diagnose, or treat pathologies. Similarly, the environmental microbiota, such as the communities characterizing strata of the soil or water column in a particular geographic region, has enormous influence on local physiochemical conditions with far-reaching implications for ecology, agriculture, fisheries, and biotechnology. Moreover, the study of micro-organisms, their community dynamics and emergent properties, and their interactions with the complex cellular systems of animals (such as immunological pathways) continues to help researchers elucidate how the fabric of life on earth was woven by these tiny puddles of protoplasm.

Although microbial communities are rife with all sorts of bundles of genetic material—including archaea, microbial eukaryotes, helminths, and viruses—bacteria tend to be given the most attention, perhaps because of their sheer abundance and diversity. Two

of the most common ways to characterize the composition of the bacterial microbiome involve the use of high-throughput next-generation sequencing technologies, which have drastically widened the scope of genomics research in recent years. One of these methods, shotgun metagenomics, sequences every DNA fragment found in a sample regardless of its origin. Bioinformatic techniques are then used to assemble the reads, infer which genes are present in the sample, and assign taxonomy, resulting in counts of how many instances each taxon was sampled. The other method, called 16S targeted-gene or amplicon sequencing, selectively sequences DNA fragments matching the ubiquitously conserved gene encoding bacterial (16S) ribosomal RNA. Reads are then clustered based on similarity, and finally one is left with counts of how many instances each bacterial "taxon" was sampled. In the 16S case, these "taxa" are called operational taxonomic units (OTUs) if determined by clustering thresholds, or amplicon sequence variants (ASVs) if resolved exactly by correcting sequences for run-specific error patterns (Callahan, 2017); with this understanding, for the sake of simplicity we will henceforth use the terms taxon/taxa to describe micro-organisms of ambiguous taxonomic level regardless of method.

With 16S sequencing, only the counts of bacterial constituents of the sample are available, while metagenomics allows us to infer the full taxonomic profile and also makes possible analysis of functional pathways in the community, since we have inferred every gene in our sample. Of course, just like in macro-ecology, distribution of functions, or niches of particular taxa, can be highly specific to the conditions of an individual community, and things are further complicated in the microbiome by the fact that genes are also laterally transferred between microorganisms. While taxonomy may not be able to tell the whole story, in this thesis we have already implicitly restricted our attention to analyzing the microbiome by taxonomic rather than functional profiling. Let us now make that explicit: comparison of the utility of metagenomic vs. 16S sequencing, especially with respect to taxonomy vs. function, is beyond our scope. Suffice it to say that 1) while 16S (and hence taxonomy-based) studies have some drawbacks, they also have benefits significant enough that they are likely to remain popular for the foreseeable future (Langille, 2018), and 2) abundance data from either 16S and metagenomic sequencing pose similar obstacles to classical statistical analyses, and so the contents of this thesis apply to both.

Regardless of sequencing protocol, the output comprises sequences of base pairs from a random sample of the total collection of genes in the community. This means that we only observe a given count of each taxon, and not its true abundance in the

community. In addition, the "depth" of sequencing, or the average number of reads in a sample that align to a known reference, varies significantly by sample and is thus a source of multiplicative error on the counts. Since many short fragments of a sequence have to be read and aligned with each other in order for that sequence to be recognizable, samples with lower sequencing depth have lower observed counts and more uncertainty. Furthermore, thousands of microbial taxa can be present in a single sample, many of which are present in extremely low numbers while the number of samples is, as in any experiment, limited by cost. As a result, each dataset has far more features than samples (Kurtz et al., 2015), and in a given sample there will be zero instances of many taxa. The high dimension and sparsity of these data invalidate most existing methods of inferring factors associated with large variation among conditions (Sill, Saadati & Benner, 2015). There is also evidence that the "compositional effects" arising from sampling can cause spurious correlations when distributions of taxa are unbalanced. Despite the promise held by microbial abundance data, analysis is so statistically challenging (Tsilimgras & Fodor, 2016) that scope for application is currently limited.

As it happens, further complications arise when genomic samples from different studies are compared with one another. In order to sequence DNA, it first has to be isolated from a sample, fragmented, and potentially amplified. Each of these processes requires a number of laboratory techniques and reagents, and procedures vary substantially between labs. Sequencing platforms also differ, and presumably there are also machine calibration differences between two sequencers of the same model. The result is that even when studies look at similar samples (for example, if two independent European studies each extracted DNA from fecal samples of men and women with or without colorectal cancer, and performed metagenomic sequencing with Illumina HiSeq technology) or even when the exact same samples are sent to two different labs for sequencing, the signal patterns are very different due to dominating "batch effects" that arise from differences in reagents, sequencing platforms, machine calibration, and other sources of technical variation. This noise persists even under highly controlled conditions (Oytam et al., 2016) and can obscure the signal of interest, such that machine learning classifiers enjoying good within-study performance become grossly inaccurate when applied cross-study (Sze & Schloss, 2016). Batch effects impair our ability to determine whether results generalize to other cohorts and preclude meaningful cross-validation and meta-analysis (Buhule et al., 2016; Leek et al., 2010; Miller et al., 2016).

It is only because of this perfect storm of challenges that microbiologists eventually had no choice but to descend into darkness to seek counsel from statisticians, who until this time were only known to emerge from their dank caverns under the cover of night to feed on insects and tree-borne fruit. Microbial abundance has often been modelled as continuous by computing proportions of observed counts to the read depth of the sample (relative abundance), and some workflows involve rarefying counts which sacrifices observed data in order to equalize read depth. Hence, two general recommendations by statisticians for better statistical treatment of abundance data are 1) that an appropriate discrete generating distribution be used to model the sampling of counts, such as Poisson, negative binomial, or multinomial, and 2) that sequencing depth error be treated within a statistical framework (McMurdie & Holmes, 2014). Additionally, it is well supported in both microbial and macro-ecology that the logarithmic scale can be useful when modeling populations of organisms in a community (Oleson et al., 2016). However, the issue of batch effects has not been as thoroughly investigated.

In the RNA microarray literature, several approaches to correcting batch effects have been proposed, the most popular of which performs gene-wise Bayesian location-scale adjustment (Johnson et al., 2007). Several methods that combine regression and singular value decomposition (Leek et al., 2007; Gagnon-Bartsch et al., 2013) have also been proposed, aiming to project away noise, which is determined as such based on gene expression signatures gleaned from regression. However, microarray data is very different from microbial abundance data; critically, we have no equivalent of "housekeeping genes" with which to base inferences about signal source. With the goal of pooling data across case-control microbial abundance studies, Gibbons et al. (2018) proposed a within-study non-parametric normalization technique, in which abundance of taxa in case samples are converted to percentiles of the abundance of equivalent taxa in control samples. However, their results are based on naive relative abundance models, which were run on a subset of taxa chosen in an ad hoc fashion (i.e., those that occurred in at least one third of case or one third of control samples).

This brings us to our purpose, which is to address the broadly meta-analytic difficulties presented by batch effects or technical variation in high-throughput genomic data. We will operate under the assumption that signal shared among different datasets on the same variables is likely to represent biological variability of interest, whereas disparate signal likely reflects variability attributable only to unimportant differences in

experimental conditions. Accordingly, we propose that if the variances are estimated reasonably well from individual datasets, then an existing multi-group method can be applied to these estimates in a simultaneous fashion to remove unshared variation. We assume Poisson sampling, estimate quantities on the logarithmic scale, and account for sequencing depth in our models. We deal with sparsity and the high-dimensional feature space by analyzing abundance at the genus level, which is a trade-off we make in order to obtain high-quality variance estimates. Our overall approach is to 1) estimate the variance-covariance matrices $\Sigma_1, \ldots, \Sigma_S$ from $n_1, \ldots, n_S$ observations of counts on the same p variables from different experiments, accounting for Poisson error and sequencing depth error, and then 2) use $\hat{\Sigma}_1, \ldots, \hat{\Sigma}_S$ to find a common low-dimensional subspace of the original feature space. By projecting the latent vectors of Poisson means, Λ_{is}, $i = 1, \ldots, n_s$, $s = 1, \ldots, S$ into this subspace, we can eliminate most technical noise while retaining shared biological signal, which should facilitate novel exploratory findings as well as improved machine learning prediction.

The rest of the thesis is organized as follows. In Chapter 2 we will review the candidate statistical methods that will be used to perform each part of the two-step ensemble method: Section 2.1 reviews the candidate methods for estimating the variance of the latent Poisson means on the log scale, while Section 2.2 reviews the candidate methods for estimating the common loadings from the variance estimates, and the method we use to estimate the projections of the underlying microbial abundances onto the new coordinate system defined by the estimated loadings. Chapter 3 presents our simulation study designed to test the ability of each candidate ensemble method to recover known signal, and provides the results of these experiments. Chapter 4 presents our analysis of two real metagenomic datasets, comparing the ability of each candidate ensemble method to find a low-dimensional representation of the underlying taxonomic abundances that preserves genuine biological signal. Finally, in Chapter 5, we discuss the implications of our findings and propose future work.

Chapter 2

Review of Methods

2.1 Estimation of Variance

Before we can estimate the common signal among several datasets, we first have to deal with the Poisson error in each dataset individually. In this work we will consider two approaches for which the observations are conditionally Poisson-distributed. Since our application is the analysis of microbiome composition, we are primarily interested in treating the means on the logarithmic scale, which arises naturally in the first method via the canonical link function, and in the second method can be achieved by a transformation.

The first approach is to leverage the Poisson log-normal PCA (PLNPCA; Chiquet et al., 2018), a fully parametric model that extends Tipping & Bishop's (1999) probabilistic PCA such that the emission layer is Poisson (or any natural exponential family distribution) rather than normal. That is, the log of the Poisson mean is a function of a latent variable which follows a multivariate normal distribution of lower dimension than the feature space, and the observed counts are—given the log Poisson mean—independently Poisson-distributed. Using this framework we have the option of introducing row-wise sums as an offset to treat sequencing depth as observed sampling effort, and can obtain an estimate of the variance-covariance matrix.

The second approach is to use Poisson PCA (PoissonPCA; Kenney et al. 2019) to sidestep a complex likelihood-based model in favor of assuming only that the observed counts are—given the latent Poisson means—independently Poisson-distributed, potentially including sequencing depth as a nuisance random variable. The authors derived an unbiased variance estimate for any non-linear transformation of the la-

tent means, as well as a semi-parametric method for estimating the scores of the transformed means that we will use in Section 2.

Poisson Log-Normal PCA

Let $X_i \in \mathbb{N}^p, \;\; i = 1, \ldots, n$ be a random vector of which we have observed n realizations. Tipping & Bishop (1999) showed that PCA can be cast as a parametric model in which the X_i are conditionally independent given the iid latent variables $w_i \in \mathbb{R}^\ell$ as follows:

$$X_i | w_i \sim \mathcal{N}(\beta w_i + \mu, \sigma^2 \mathrm{I}_p), \quad \text{with } w_i \sim \mathcal{N}(0_\ell, \mathrm{I}_\ell), \quad i = 1, \ldots, n \tag{2.1}$$

Clearly this resembles classical factor analysis in that normally distributed latent variables w_i are "loaded" by a $p \times \ell$ matrix β so as to produce the observation X_i along with some noise, although in contrast to factor analysis, the error variance σ^2 is constant. The marginal distribution of X_i is also normal, with mean μ and variance $\beta\beta^T + \sigma^2 \mathrm{I}$, and so it can be shown fairly easily that the global maximum of the likelihood is attained when the columns of β are the principal eigenvectors of the sample covariance matrix, and that for this optimal β the maximum likelihood estimate of σ^2 is given by the average of the $p - \ell$ smallest eigenvalues.

Probabilistic PCA is of interest to us particularly because of its ability to create an elegant covariance model of the data, since $\widehat{\mathrm{Var}(X)}$ (where X is the $n \times p$ matrix comprised of rows X_i^T) is defined entirely by the estimated loading matrix $\hat{\beta}$ and the residuals $\hat{\sigma}^2$. Since we are dealing with count data, we turn to the natural exponential family extension to probabilistic PCA presented by Chiquet et al. (2018). We will review their results and show more detailed derivations for the Poisson case, i.e., PLNPCA.

Now, note that equation (2.1) could be equivalently expressed as conditional on the linear transformation $Z_i = \beta w_i + \mu$, where Z_i is multivariate normal with mean $\mu \in \mathbb{R}^p$ and variance $\Sigma = \beta\beta^T \in \mathbb{R}^{p \times p}$, since w_i has mean vector 0_ℓ and identity variance. So, it is through Z_i that β and w_i characterize the data. PLNPCA inherits this same latent structure, but we no longer have a multivariate normal distribution for $X_i | Z_i$. Instead each independent X_{ij}, given Z_{ij}, is a realization from a Poisson distribution with parameter $Z_{ij}, \quad j = 1, \ldots, p$. If each of these Poisson distributions has mean Λ_{ij}, then using the canonical parameterization for exponential families, we have that $\eta(\Lambda_{ij}) = \log \Lambda_{ij}$. Thus, in the case of PLNPCA, $Z_i = \log \Lambda_i$, where log is

assumed to be applied element-wise whether we speak of $\log \Lambda_i \in \mathbb{R}^p$ or the $n \times p$ matrix $\log \Lambda$. Since we are interested in $\log \Lambda_i$ as opposed to the noisy observed data, we can find an estimator for $\Sigma = \text{Var}(\log \Lambda)$. Of course, like probabilistic PCA, PLNPCA also provides the opportunity to estimate a low-rank covariance matrix depending on the dimension chosen for the latent space, but since we wish only to reduce dimension based on the common variation across all S datasets, we will be using a rank-p estimate. Also, we are not interested in accounting for covariates, as we want to avoid disrupting signal in a supervised fashion, so we replace the main effects μ with $\xi_i \in \mathbb{R}^p$ to denote sample-wise offsets (i.e., $\xi_{i1} = \xi_{i2} = \cdots = \xi_{ip}$). In the microbiome setting, we choose these offsets to correspond to log-total read count, which is the closest we have to an observed value for sequencing depth. Should we wish not to apply a sequencing depth correction, the following results in the rest of this section hold as written for $\xi_i = 0_p$. So, equivalently to probabilistic PCA, we have latent $w_i \in \mathbb{R}^\ell$ iid $\mathcal{N}(0_\ell, I_\ell)$ and $\log \Lambda_i \in \mathbb{R}^p$, where

$$\log \Lambda_i = \xi_i + \beta w_i, \quad i = 1, \ldots, n \tag{2.2}$$

or again equivalently, $\log \Lambda_i$ is multivariate normal with mean ξ_i and variance $\Sigma = \beta\beta^T \in \mathbb{R}^{p \times p}$. No uniqueness constraints are put on β. Observations X_{ij} are generated according to the Poisson distribution with mean Λ_{ij}:

$$p(X_{ij} | \log \Lambda_{ij}) = \frac{1}{X_{ij}!} \exp(X_{ij} \log \Lambda_{ij} - \exp(\log \Lambda_{ij}))$$

such that, for the canonical log link, $\log(\mathbb{E}(X_{ij} | \log \Lambda_{ij})) = \log(\exp(\log \Lambda_{ij})) = \log \Lambda_{ij}$. The complete log-likelihood of the observed data and the unobserved latent w is thus given by

$$\mathcal{L}(X, w; \xi, \beta) = \sum_{i=1}^{n} \big(\log p(X_i | w_i; \xi_i, \beta) + \log p(w_i) \big) \tag{2.3}$$

$$\propto \sum_{i=1}^{n} \left[\sum_{j=1}^{p} \big(X_{ij}(\xi_{ij} + \beta_j^T w_i) - \exp(\xi_{ij} + \beta_j^T w_i) \big) - \sum_{r=1}^{\ell} \frac{1}{2} w_{ir}^2 \right] \tag{2.4}$$

We want to maximize the marginal likelihood $\mathcal{L}(X; \xi, \beta)$, but this isn't necessarily analytic for all choices from the natural exponential family. The authors suggest that most numerical methods, even expectation-maximization, could be computationally challenging, and so they propose integrating out w under a variational approximation

of $p(w|X)$. Variational inference uses the distribution that we have assumed for the latent variable to construct a lower bound for the marginal distribution of the data. We can then maximize this lower bound instead of the likelihood (Blei et al., 2017).

Let $\tilde{p}_\rho(w|X)$ denote an approximation of $p(w|X)$, where the former will be henceforth written with its parameter(s) ρ left implicit. Let

$$\mathrm{KL}[\,\tilde{p}(w|X)\,||\,p(w|X)\,] = \int_w \tilde{p}(w|X) \log \frac{\tilde{p}(w|X)}{p(w|X)} \geq 0$$

denote their Kullback-Leibler divergence. Then we have

$$\begin{aligned}
\mathrm{KL}[\,\tilde{p}(w|X)\,||\,p(w|X)\,] &= \int_w \tilde{p}(w|X) \log \frac{\tilde{p}(w|X)}{p(w|X)} \\
&= \int_w \tilde{p}(w|X) \log \frac{p(X)\tilde{p}(w|X)}{p(w,X)} \\
&= \int_w \tilde{p}(w|X) \left[\log p(X) + \log \frac{\tilde{p}(w|X)}{p(w,X)}\right] \\
&= \int_w \log p(X)\tilde{p}(w|X) + \int_w \tilde{p}(w|X) \log \frac{\tilde{p}(w|X)}{p(w,X)} \\
&= \log p(X) \int_w \tilde{p}(w|X) + \int_w \tilde{p}(w|X) \log \frac{\tilde{p}(w|X)}{p(w,X)} \\
&= \log p(X) + \int_w \tilde{p}(w|X) \log \frac{\tilde{p}(w|X)}{p(w,X)}, \qquad (2.5)
\end{aligned}$$

since $\int_w \tilde{p}(w|X) = 1$. For a good approximation, $\tilde{p}(w|X)$ and $p(w|X)$ should be as close to each other as possible, and so we want to minimize their Kullback-Leibler divergence with respect to the parameters ρ of $\tilde{p}$. To do so, we need only minimize the second term of the RHS of equation (2.5), since $\log p(X)$ is constant with respect to ρ. The second term can be written as

$$\begin{aligned}
\int_w \tilde{p}(w|X) \log \frac{\tilde{p}(w|X)}{p(w,X)} &= \mathbb{E}_{\tilde{p}} \log \frac{\tilde{p}(w|X)}{p(w,X)} \\
&= \mathbb{E}_{\tilde{p}}[\log \tilde{p}(w|X) - \log p(X,w)] \\
&= \mathbb{E}_{\tilde{p}}[\log \tilde{p}(w|X) - \log \big(p(X|w)p(w)\big)] \\
&= \mathbb{E}_{\tilde{p}}[\log \tilde{p}(w|X) - \big(\log p(X|w) + \log p(w)\big)] \\
&= \mathbb{E}_{\tilde{p}}[\log \tilde{p}(w|X) - \log p(X|w) - \log p(w)]
\end{aligned}$$

where $\mathbb{E}_{\tilde{p}}[\,\cdot\,]$ is shorthand for $\mathbb{E}_{\tilde{p}_\rho(w|X)}[\,\cdot\,]$. So equivalently we can maximize

$$J(\tilde{p}, \rho) = \mathbb{E}_{\tilde{p}}[-\log \tilde{p}(w|X) + \log p(X|w) + \log p(w)], \tag{2.6}$$

where the last two terms form the complete log-density that we used to compute the log-likelihood in equation (2.3). Moreover, since the Kullback-Leibler divergence is non-negative, it is clear from equation (2.5) that $J(\tilde{p}, \rho)$ provides a lower bound for $\log p(X)$, the marginal log-density of X, which provides the justification for maximizing $J(\tilde{p}, \rho)$ instead of the marginal log-likelihood.

In the PLNPCA setting, the lower bound we seek is $J(\tilde{p}, \xi, \beta, M, S) \leq \mathcal{L}(X; \xi, \beta)$. Since in our case w_i is multivariate normal, let $\tilde{p}_i = \mathcal{N}(m_i, \operatorname{diag}(s_i \circ s_i))$, and so $\tilde{p} = \prod_{i=1}^n \mathcal{N}\big(m_i, \operatorname{diag}(s_i \circ s_i)\big)$ is the product distribution of the variational parameters M and S (where M and S are $n \times \ell$ matrices with rows m_i^T and s_i^T respectively), which act as the mean and standard deviation of our variational approximation to the conditional density of w given X. Then, using equation (2.6) and equation (2.4), and letting Y be an $n \times \ell$ matrix of independent standard normal variates, the variational lower bound is given by the following, omitting constant terms:

$$\begin{aligned}
J(\tilde{p}, \beta, M, S) &= \mathbb{E}_{\tilde{p}}\big[\log p(w) + \log p(X|w; \beta) - \log \tilde{p}(w)\big] \\
&= \sum_{i=1}^n \mathbb{E}_{\tilde{p}_i}\big[\log p(w_i) + \log p(X_i|w_i; \beta) - \log \tilde{p}_i(w_i)\big] \\
&= \sum_{i=1}^n X_i^T(\xi_i + \beta m_i) - \frac{1}{2}\big(||m_i||_2^2 + ||s_i||_2^2\big) + 1_\ell^T \log s_i \\
&\quad - 1_p^T \mathbb{E}_{\tilde{p}_i}[\exp(\xi_i + \beta w_i)] \\
&= 1_n^T\big(X \circ (\xi + M\beta^T)\big)1_p - \frac{1}{2}1_n^T(M \circ M + S \circ S - 2\log S)1_\ell \\
&\quad - 1_n^T \mathbb{E}_{\tilde{p}}\big[\exp(\xi + w\beta^T)\big]1_p \\
&= 1_n^T\big(X \circ (\xi + M\beta^T)\big)1_p - \frac{1}{2}1_n^T(M \circ M + S \circ S - 2\log S)1_\ell \\
&\quad - 1_n^T \mathbb{E}\big[\exp(\xi + (M + S \circ Y)\beta^T)\big]1_p \\
&= 1_n^T\big(X \circ (\xi + M\beta^T)\big)1_p - \frac{1}{2}1_n^T(M \circ M + S \circ S - 2\log S)1_\ell \\
&\quad - 1_n^T \exp\big[\xi + M\beta^T + \frac{1}{2}(S \circ S)(\beta^T \circ \beta^T)\big]1_p
\end{aligned}$$

since linear combinations of normals are normal, and a log-normal random variable Θ, $\log \Theta_i \sim \mathcal{N}(\mu_i, \Sigma)$, has expectations of the form $\mathbb{E}[\Theta_{ij}] = \exp(\mu_{ij} + \frac{1}{2}\Sigma_{jj})$ (Aitchison

& Ho, 1989). In the above equalities for $J(\tilde{p}, \beta, M, S)$, $\mathbb{E}_{\tilde{p}}$ denotes $\mathbb{E}_{\tilde{p}_{M,S}(w|X)}$. Under this framework, the variance over $\tilde{p}_i$ of $\log \Lambda_i$ is given by

$$\begin{aligned}
\mathrm{Var}_{\tilde{p}_i}(\log \Lambda_i) &= \frac{1}{n}\mathbb{E}_{\tilde{p}_i}\big[(\log \Lambda_i - \mathbb{E}_{\tilde{p}_i}[\log \Lambda_i])(\log \Lambda_i - \mathbb{E}_{\tilde{p}_i}[\log \Lambda_i])^T\big] \\
&= \frac{1}{n}\mathbb{E}_{\tilde{p}_i}\big[(\xi_i + \beta w_i - (\xi_i + \beta m_i))(\xi_i + \beta w_i - (\xi_i + \beta m_i))^T\big] \\
&= \frac{1}{n}\mathbb{E}_{\tilde{p}_i}\big[(\beta w_i - \beta m_i)(w_i^T\beta^T - m_i^T\beta^T)\big] \\
&= \frac{1}{n}\mathbb{E}_{\tilde{p}_i}\big[\beta w_i w_i^T\beta^T - \beta w_i m_i^T\beta^T - \beta m_i w_i^T\beta^T + \beta m_i m_i^T\beta^T\big] \\
&= \beta\mathbb{E}_{\tilde{p}_i}[w_i w_i^T]\beta^T - \beta m_i m_i^T\beta^T \\
&= \beta\big(\mathrm{Var}_{\tilde{p}_i}(w_i) + \mathbb{E}_{\tilde{p}_i}[w_i]\mathbb{E}_{\tilde{p}_i}[w_i]^T\big)\beta^T - \beta m_i m_i^T\beta^T \\
&= \beta\big(\,\mathrm{diag}(s_i \circ s_i) + m_i m_i^T\big)\beta^T - \beta m_i m_i^T\beta^T \\
&= \beta\,\mathrm{diag}(s_i \circ s_i)\beta^T,
\end{aligned}$$

using the fact that $\mathrm{Var}(X_i) = \mathbb{E}(X_i X_i^T) - \mathbb{E}[X_i]\mathbb{E}[X_i]^T$. So the estimator of Σ that we seek is given by

$$\begin{aligned}
\hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\tilde{p}_i}\big[(\log \Lambda_i - \xi_i)(\log \Lambda_i - \xi_i)^T\big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(\mathbb{E}_{\tilde{p}_i}[\log \Lambda_i \log \Lambda_i^T] - \xi_i\mathbb{E}_{\tilde{p}_i}[\log \Lambda_i^T] - \mathbb{E}_{\tilde{p}_i}[\log \Lambda_i]\xi_i^T + \xi_i\xi_i^T\big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(\mathrm{Var}_{\tilde{p}_i}(\log \Lambda_i) + \mathbb{E}_{\tilde{p}_i}[\log \Lambda_i]\mathbb{E}_{\tilde{p}_i}[\log \Lambda_i]^T - \xi_i\mathbb{E}_{\tilde{p}_i}[\log \Lambda_i^T] \\
&\quad - \mathbb{E}_{\tilde{p}_i}[\log \Lambda_i]\xi_i^T + \xi_i\xi_i^T\big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(\beta\,\mathrm{diag}(s_i \circ s_i)\beta^T + (\xi_i + \beta m_i)(\xi_i^T + m_i^T\beta^T) - \xi_i(\xi_i^T + m_i^T\beta^T) \\
&\quad - (\xi_i + \beta m_i)\xi_i^T + \xi_i\xi_i^T\big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(\beta\,\mathrm{diag}(s_i \circ s_i)\beta^T + \beta m_i m_i^T\beta^T\big) \\
&= \beta\left(\frac{1}{n}\big(\,\mathrm{diag}(1_n^T(S \circ S)) + M^T M\big)\right)\beta^T
\end{aligned}$$

again using the fact that $\mathrm{Var}(X_i) = \mathbb{E}(X_i X_i^T) - \mathbb{E}[X_i]\mathbb{E}[X_i]^T$.

PLNPCA is implemented in the R package 'PLNmodels' (Chiquet et al., 2018).

PoissonPCA

Let $X_i \in \mathbb{N}^p, \quad i = 1, \ldots, n$ be a random vector of which we have observed n realizations. PoissonPCA assumes that conditional on the latent Poisson means Λ_{ij}, the X_{ij} are independently distributed as Poisson with parameter Λ_{ij}. In contrast to PLNPCA, the distribution of Λ itself is not parameterized. Using method of moments, we can find variance estimators for any non-linear transformation of Λ by taking advantage of Taylor series (or truncated power series, in the case of the log transformation). Since we are interested in $\log \Lambda$, we will present the main results from Kenney et al. (2019) for estimating $\Sigma = \mathrm{Var}(\log \Lambda)$.

Finding an estimator for the variance of the log-transformed latent variables is contingent on finding a corresponding (element-wise) transformation f for the data. First, we derive the total variance $\mathrm{Var}(f(X))$, which is

$$\begin{aligned}
\mathrm{Var}(f(X)) &= \mathbb{E}[f(X)f(X)^T] - \mathbb{E}[f(X)]\mathbb{E}[f(X)]^T \\
&= \mathbb{E}\big[\mathbb{E}[f(X)f(X)^T|\Lambda]\big] - \mathbb{E}\big[\mathbb{E}[f(X)|\Lambda]\big]\big(\mathbb{E}\big[\mathbb{E}[f(X)|\Lambda]\big]\big)^T \\
&= \mathbb{E}\big[\mathrm{Var}(f(X)|\Lambda) + \mathbb{E}[f(X)|\Lambda]\big(\mathbb{E}[f(X)|\Lambda]\big)^T\big] \\
&\quad - \mathbb{E}\big[\mathbb{E}[f(X)|\Lambda]\big]\big(\mathbb{E}\big[\mathbb{E}[f(X)|\Lambda]\big]\big)^T \\
&= \mathbb{E}[\mathrm{Var}(f(X)|\Lambda)] + \mathrm{Var}(\mathbb{E}[f(X)|\Lambda]),
\end{aligned} \tag{2.7}$$

The PoissonPCA model specifies that conditionally on Λ, the mean of X is Λ, which implies that we are looking for $f(X)$ such that conditionally on Λ, the mean of $f(X)$ is $\log \Lambda$. Accordingly, substituting $\log \Lambda$ for $\mathbb{E}[f(X)|\Lambda]$ in equation (2.7), we get the formula

$$\Sigma = \mathrm{Var}(\log \Lambda) = \mathrm{Var}(f(X)) - \mathbb{E}[\mathrm{Var}(f(X)|\Lambda)] \tag{2.8}$$

To find $f(X)$, one might hope to make use of the fact that a Maclaurin series multiplied by $e^{-\Lambda_{ij}}$ gives a Poisson conditional mean, since we know that our conditional mean must equal $\log \Lambda$ to be unbiased, but this won't work for us because the logarithmic function is not analytic at zero. Instead PoissonPCA approximates $\log \Lambda_{ij}$ for small values with a Taylor series about a specified point and truncated at a specified value, whereas for larger X it sets $f(X) = \log X$ since $\mathbb{E}[\log X|\Lambda] \approx \log \Lambda$.

Now that we have a way of computing $f(X)$, in order to estimate Σ we just need to estimate the conditional variance from equation (2.8). This is achieved in PoissonPCA by finding a function $k(X)$ such that the average conditional mean of $k(X)$ is

approximately $\text{Var}(\log X|\Lambda)$.

Now, to address sequencing depth correction, we revisit the original PoissonPCA model formulation but additionally consider sequencing depth $\xi_i \in \mathbb{R}^p$ as a random variable such that $X_{ij}|(\Lambda_{ij}, \xi_{ij}) \sim \text{Poisson}(\xi_{ij}\Lambda_{ij})$, where $X_{i1}, \ldots, X_{ip}$ are independent given Λ_i and ξ_i. Unlike in PLNPCA, ξ_i is not considered to be observable, so under this model we end up estimating $\Sigma = \text{Var}(\log \Lambda) = \text{Var}(\log(\xi \circ \Lambda^*))$ when in fact what we want is $\Sigma^* = \text{Var}(\log \Lambda^*)$. Thus, in order to account for sequencing depth error, we have to add constraints to the variance estimator given by equation (2.8) in order to get the desired estimate Σ^* for the true underlying Poisson means. Kenney et al. (2019) suggest that the best method for correcting Σ is to imbue it with the characteristic properties of a variance-covariance matrix of a compositional random vector, which is to say that Σ^* should be symmetric and contained in the orthogonal complement of the vector 1. Obviously we want to preserve the map defined by Σ, and so Σ^* should have the same bilinear form as Σ in the orthogonal complement of the vector 1. With these constraints, it can be proven that the sequencing depth-corrected variance estimate is

$$\Sigma^* = \Sigma - (pI_p + 1_p 1_p^T)^{-1}\Sigma 1_p 1_p^T - 1_p 1_p^T \Sigma^T ((pI_p + 1_p 1_p^T)^{-1})^T \tag{2.9}$$

Note that the PoissonPCA variance estimate will be full rank if there is no sequencing depth correction applied, while equation (2.9) gives $\text{rank}(\hat{\Sigma}^*) = p - 1$. Nonetheless, we will henceforth drop the notation Σ^* and refer to either estimate as Σ, since in the following chapters it should be clear from context whether or not a sequencing depth correction has been applied. In addition, the construction of these estimators makes no guarantee about the definiteness of the estimates, and in practice there can be several negative terminal eigenvalues. Since multi-group analyses typically require a variance-covariance matrix that is positive-definite, if any $\hat{\Sigma}_s$, $s = 1, \ldots, S$ computed in our analyses was indefinite, we eigen-decomposed it, replaced the negative eigenvalues with small decreasing positive values, and then used the eigenvectors to reconstruct the variance.

PoissonPCA is implemented in the R package 'PoissonPCA' (Kenney et al., 2019).

2.2 Multi-Group Analysis

We are ultimately interested in estimating the loadings that are common to all groups, and projecting the estimated latent Poisson means from each group into a common space spanned by the loadings. To achieve this, the natural choice would be to apply multi-group extensions of PCA or of factor analysis to the Poisson-corrected estimates from each group, since by dealing with abundance on the log scale we are able to decompose our variance estimates under Wishart/multivariate normal assumptions. Both PCA-based and factor analysis-based approaches allow us to find a common space of low dimension, but the latter is prescriptive in this sense while the former inherits the exploratory nature of PCA.

Perhaps the most direct multi-group generalization of PCA is common principal components analysis (Flury, 1988), which assumes that there exists an orthogonal matrix that can approximately diagonalize the covariance matrices of all S groups simultaneously. Using a generalized PCA approach, the dimension of the common space can be chosen after estimating the full loadings matrix, based on which q loading vectors are associated with the largest variances for all groups simultaneously. Unlike the usual PCA, Flury's common principal components analysis (FCPCA) assumes that the sample covariance matrices follow a Wishart distribution, and so the common loadings matrix is estimated in a maximum likelihood framework. More recently the model was revisited by Trendafilov, who criticized the fact that FCPCA does not constrain the eigenvalues of each group to be simultaneously decreasing, which in some cases could disallow its use as a dimension reduction strategy. Instead, Trendafilov (2010) suggested the stepwise approach to common principal components analysis (SCPCA), which finds the common loadings sequentially in order of variance explained. In the present study, we compared the performance of FCPCA and SCPCA for simultaneously decomposing the S covariance matrices that have been estimated using either PoissonPCA or PLNPCA.

The second model under consideration is De Vito's (2019) multi-study factor analysis (MSFA), which is an extension of classical factor analysis and likewise assumes that the latent variables and measurement error are multivariate normal and hence that the observations have a multivariate normal marginal distribution. However, in MSFA, there are q latent common factors and $\ell_1, \ldots, \ell_S$ latent unique factors, and so $S + 1$ loadings matrices have to be estimated by maximum likelihood, which is a much more difficult optimization problem than that posed by classical factor analysis. Also, as with any factor analytic approach, dimensions of the latent subspaces must be

considered a hyperparameter of the generative model.

Finally, after estimating the common loadings with either SCPCA, FCPCA, or MSFA, we want to express the underlying abundances in each sample with respect to our new common basis, and we will refer to these quantities as the scores.

Common Principal Components Analysis

Let $\Sigma_s \in \mathbb{R}^{p\times p}$ be symmetric and positive definite and assume that each $(n_s-1)\hat{\Sigma}_s$ is independently distributed as $\mathrm{W}_p(n_s-1,\ \Sigma_s),\ \ s=1,\dots S$, where W_p is the p-variate Wishart distribution. If there is a rotation matrix $V \in O(p)$ for which

$$V^T\Sigma_s V = D_s \tag{2.10}$$

for all s, where $D_s = \mathrm{diag}(d_{s1},\dots,d_{sp})$, then the subspace spanned by the columns of V is common to all groups; the assumption in CPCA is that V exists as such. Hence, CPCA comes down to simultaneous diagonalization of the S sample covariance matrices. Of course, positive definite real symmetric matrices are not necessarily commutative, so we are not guaranteed that a set of covariance matrices will be simultaneously diagonalizable, and thus CPCA requires that we find an optimal approximation. However, for $\hat{\Sigma}_1,\dots,\hat{\Sigma}_S$ positive definite, their quadratic forms are strictly convex, and so the optimization problem posed in CPCA is highly tractable. The objective function can be derived via maximum likelihood. The log-likelihood is given by

$$\begin{aligned}\mathcal{L}(\Sigma_1,\dots,\Sigma_S) &= \log\prod_{s=1}^{S} p(\hat{\Sigma}_s;\Sigma_1,\dots,\Sigma_S)\\ &\propto \sum_{s=1}^{S}\big(-\frac{n_s-1}{2}\log|\Sigma_s| - \frac{n_s-1}{2}\mathrm{tr}(\Sigma_s^{-1}\hat{\Sigma}_s)\big)\end{aligned}$$

Assuming that the desired common rotation matrix $V=(v_1,\dots,v_p)$ exists such that equation (2.10) holds, we can substitute $\Sigma_s = VD_sV^T$ in both the determinant and trace, and then cyclically permute V in the trace to rewrite $\mathrm{tr}(\Sigma_s^{-1}\hat{\Sigma}_s) = \mathrm{tr}(D_s^{-1}V^T\hat{\Sigma}_sV)$. Since V is an orthogonal matrix and D_s is diagonal, $|VD_sV^T| = |D_s| = \prod_{j=1}^{p} d_{sj}$. Taking the log of $\prod_{j=1}^{p} d_{sj}$, and using the definition of the trace and again the fact that D_s^{-1} is diagonal, we can write both terms as summations over j

of v_j and d_{sj}, $j = 1, \ldots, p$, to get

$$\mathcal{L}(\Sigma_1, \ldots, \Sigma_S) \propto \sum_{s=1}^{S} -\frac{n_s - 1}{2} \Big(\sum_{j=1}^{p} \log d_{sj} + \sum_{j=1}^{p} \frac{v_j^T \hat{\Sigma}_s v_j}{d_{sj}} \Big)$$

and finally multiplying by -2 we arrive at an objective function g,

$$g(v_1, \ldots, v_p, d_{11}, \ldots, d_{1p}, d_{21}, \ldots, d_{sp}) = \sum_{s=1}^{S} (n_s - 1) \Big(\sum_{j=1}^{p} (\log d_{sj} + \frac{v_j^T \hat{\Sigma}_s v_j}{d_{sj}}) \Big),$$

with which to construct a minimization problem:

$$\begin{aligned} \hat{V} &= \underset{\substack{v_h^T v_j = 0,\ h \neq j \\ v_h^T v_j = 1,\ h = j}}{\operatorname{argmin}} \; g(v_1, \ldots, v_p, d_{11}, \ldots, d_{1p}, d_{21}, \ldots, d_{sp}) \\ &= \underset{\substack{v_h^T v_j = 0,\ h \neq j \\ v_h^T v_j = 1,\ h = j}}{\operatorname{argmin}} \sum_{s=1}^{S} (n_s - 1) \Big(\sum_{j=1}^{p} (\log d_{sj} + \frac{v_j^T \hat{\Sigma}_s v_j}{d_{sj}}) \Big), \end{aligned} \tag{2.11}$$

Flury (1984) computed derivatives of g with respect to d_{sj} and found that the optimal unique eigenvalues are

$$d_{sj} = v_j^T \hat{\Sigma}_s v_j, \quad \text{for } j = 1, \ldots, p, \quad s = 1, \ldots, S \tag{2.12}$$

which implies that

$$\sum_{j=1}^{p} \frac{v_j^T \hat{\Sigma}_s v_j}{d_{sj}} = p, \quad \text{for } s = 1, \ldots, S. \tag{2.13}$$

So $D_s = \operatorname{diag}^*(V^T \hat{\Sigma}_s V)$ for all $s = 1, \ldots, S$, which is to say that D_s is the diagonal $p \times p$ matrix whose entries $d_{s1}, \ldots, d_{sp}$ are the diagonal entries of $V^T \hat{\Sigma}_s V$. Differentiating g with respect to v_j shows that the solution satisfies

$$v_h^T \Big(\sum_{s=1}^{S} (n_s - 1) \frac{d_{sj} - d_{sh}}{d_{sh} d_{sj}} \hat{\Sigma}_s \Big) v_j = 0 \tag{2.14}$$

for $h, j = 1, \ldots, p$ and $h \neq j$, that is, that $\sum_{s=1}^{S} (n_s - 1)(D_s^{-1} V^T \hat{\Sigma}_s V)$ is symmetric.

FCPCA uses the FG algorithm (Flury & Gautschi, 1986) to compute $\hat{V}$ by approxi-

mate simultaneous diagonalization of a set of S positive definite symmetric matrices $C_1, \ldots, C_S$ given weights $m_1, \ldots, m_S$, as determined by a measure of simultaneous diagonality,

$$\Gamma(C_1, \ldots, C_S \ ; \ m_1, \ldots, m_S) = \prod_{s=1}^{S} \left[\frac{|\operatorname{diag} C_s|}{|C_s|} \right]^{m_s}$$

which leads to the same optimization problem as outlined above. In the FG algorithm, all pairs of vectors of V are rotated in each iteration to satisfy equation (2.14), where each rotation is determined by an inner cycle that computes the solutions to the 2-dimensional analogs of equation (2.14). Using FCPCA we can reduce the basis to $v_1, \ldots, v_q$ only provided that the last $p-q$ eigenvalues are small for all S groups.

On the other hand, Trendafilov's (2010) SCPCA starts with the same objective function, but performs minimization to find the optimal axes sequentially based on the fact that if covariance matrices $C_1, \ldots, C_S$ share a common eigenvector e, then e is also an eigenvector of the average of $C_1, \ldots, C_S$ weighted by their unique eigenvalues associated with e. Looking again to our CPCA minimization problem given by equation (2.11), due to equations (2.12) to (2.14) we can write the same problem compactly as

$$\hat{V} = \underset{\substack{v_h^T v_j = 0, \ h \neq j \\ v_h^T v_j = 1, \ h = j}}{\operatorname{argmin}} \sum_{s=1}^{S} (n_s - 1) \Big(\sum_{j=1}^{p} \log(v_j^T \hat{\Sigma}_s v_j) \Big). \tag{2.15}$$

If the eigenvalues estimated by FCPCA for each estimated covariance matrix are all simultaneously decreasing, then

$$v_1^T \hat{\Sigma}_s v_1 \geq v_2^T \hat{\Sigma}_s v_2 \geq \cdots \geq v_p^T \hat{\Sigma}_s v_p, \quad \text{for } s = 1, \ldots, S,$$

which implies that

$$\sum_{i=1}^{S} (n_s - 1) \log(v_1^T \hat{\Sigma}_s v_1) \geq \cdots \geq \sum_{i=1}^{S} (n_s - 1) \log(v_p^T \hat{\Sigma}_s v_p), \tag{2.16}$$

for $s = 1, \ldots, S$. We can solve the corresponding sequence of minimization prob-

lems,

$$v_j = \underset{v^T v=1}{\operatorname{argmin}} \sum_{s=1}^{S} (n_s - 1)\Big(\sum_{j=1}^{p} \log(v^T \hat{\Sigma}_s v)\Big),$$

which gives the v_j in ascending order of the minima, so we have $v_1, \ldots, v_p \in O(p)$ starting with v_p. Then obtaining the j^{th} eigenvector associated with the j^{th} largest eigenvalue is equivalent to solving

$$\hat{v}_j = \operatorname{argmin} \sum_{s=1}^{S} (n_s - 1)\Big(\sum_{j=1}^{p} \log(v^T \hat{\Sigma}_s v)\Big) \tag{2.17}$$

$$\text{subject to } \; v^T v = 1, \; v^T (v_{j+1}, \ldots, v_p) = 0 \tag{2.18}$$

which can be shown to be equivalent to taking

$$\hat{v}_j = \sum_{s=1}^{S} (n_s - 1)\Big(\sum_{j=1}^{p} \log(v_j^T \hat{\Sigma}_s v_j)\Big) \tag{2.19}$$

which gives us the same result as FCPCA. However, if the FCPCA eigenvalues are not simultaneously decreasing in all S groups, then the stepwise approach will not solve equation (2.15). Trendafilov argues that despite this, the stepwise solution is useful: since equation (2.16) holds for equations (2.17) and (2.19) by construction, applying Jensen's inequality to equation (2.16) and using the pooled sample covariance matrix, he shows that the sum over s of the variances explained by each v_j (weighted by the ratio of each group's degrees of freedom to the total) is greater than or equal to the equivalent sum for v_{j+1}. This means that equations (2.17) and (2.19) can always be used to find a set of $q \leq p$ common principal component vectors forming a basis for $\mathbb{R}^q$, such that their variances $d_{s1}, \ldots, d_{sq}$ are approximately simultaneously decreasing for all $s = 1, \ldots, S$, and all $d_{1j}, \ldots, d_{Sj}$ are as similar as possible for a given j, $j = 1, \ldots, q$.

FCPCA and SCPCA were implemented using code adapted and modified from source code in the R packages 'multigroup' and 'cpca' respectively, as errors were found in both functions.

Multi-Study Factor Analysis

Let $X_{is} \in \mathbb{R}^p$, $i = 1, \ldots, n_s$, $s = 1, \ldots, S$ be a random vector with n_s realizations in the s^{th} group. MSFA assumes that there exist iid latent variables $f_{is} \in \mathbb{R}^q$ and $w_{is} \in \mathbb{R}^{\ell_s}$, which generate X_{is} by

$$X_{is} = \Phi f_{is} + \beta_s w_{is} + \mu_s + \epsilon_{is}, \quad \text{with} \quad f_{is} \sim \mathcal{N}(0_q, \mathrm{I}_q), \; w_{is} \sim \mathcal{N}(0_{\ell_s}, \mathrm{I}_{\ell_s})$$

for $i = 1, \ldots, n_s$, $\; s = 1, \ldots, S$, where $\mu_s \in \mathbb{R}^p$ is the mean vector, ϵ_{is} are iid with $\epsilon_{is} \sim \mathcal{N}(0_p, \Psi_s)$, where $\Psi_s = \mathrm{diag}(\psi_{s1}, \ldots, \psi_{sp})$, and ϵ_{is} is independent from the latent factors f_{is} and w_{is}. Φ is a $p \times q$ matrix of common loadings, and β_s is a $p \times \ell_s$ matrix of group-specific loadings, which provide structure to the latent factors. Alternatively, we can say that each X_{is} is conditionally independent given the latent variables f_{is} and w_{is} as follows:

$$X_{is} | f_{is}, w_{is} \sim \mathcal{N}(\Phi f_{is} + \beta_s w_{is} + \mu_s, \Psi_s) \;\; \text{with} \;\; f_{is} \sim \mathcal{N}(0_q, \mathrm{I}_q), \; w_{is} \sim \mathcal{N}(0_{\ell_s}, \mathrm{I}_{\ell_s})$$

Since we have a multivariate normal distribution for X_{is} conditional on the multivariate normal latent variables, X_{is} has a multivariate normal marginal distribution with $\mathbb{E}[X_{is}] = \mu_s$, so the covariance matrix Σ_s of X_s (the $n_s \times p$ matrix of stacked rows X_{is}^T) is given by

$$\begin{aligned}
\Sigma_s &= \frac{1}{n} \mathbb{E}\big[(X_s - \mu_s)(X_s - \mu_s)^T\big] \\
&= \frac{1}{n} \mathbb{E}\big[(\Phi f_s + \beta_s w_s + \mu_s + \epsilon_s - \mu_s)(\Phi f_s + \beta_s w_s + \mu_s + \epsilon_s - \mu_s)^T\big] \\
&= \frac{1}{n} \big(\Phi \mathbb{E}[f_s f_s^T] \Phi^T + \beta_s \mathbb{E}[w_s w_s^T] \beta_s^T + \mathbb{E}[\epsilon_s \epsilon_s^T]\big) \\
&= \frac{1}{n} \big(n \Phi \mathrm{I}_q \Phi^T + n \beta_s \mathrm{I}_{\ell_s} \beta_s^T + n \Psi_s\big) \\
&= \Phi \Phi^T + \beta_s \beta_s^T + \Psi_s
\end{aligned}$$

So the marginal distribution is $X_{is} \sim \mathcal{N}(\mu_s, \Sigma_s = \Phi\Phi^T + \beta_s \beta_s^T + \Psi_s)$. Hence, the log-likelihood is given by

$$\begin{aligned}
\mathcal{L}(\Phi, \beta_s, \Psi_s) &= \log \prod_{s=1}^{S} \prod_{i=1}^{n_s} p(X_{is} | \Phi, \beta_s, \Psi_s) \\
&\propto \sum_{s=1}^{S} \big(- \frac{n_s}{2} \log |\Sigma_s| - \frac{1}{2} \sum_{i=1}^{n_s} (X_i - \mu_{is})^T \Sigma_s^{-1} (X_i - \mu_{is}) \big)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=1}^{S} \big(-\frac{n_s}{2} \log|\Sigma_s| - \frac{1}{2} \sum_{i=1}^{n_s} \mathrm{tr}\big(\Sigma_s^{-1}(X_i - \mu_{is})^T(X_i - \mu_{is})\big)\big) \\
&= \sum_{s=1}^{S} \big(-\frac{n_s}{2} \log|\Sigma_s| - \frac{1}{2} \mathrm{tr}\big(\Sigma_s^{-1} \sum_{i=1}^{n_s}(X_i - \mu_{is})^T(X_i - \mu_{is})\big)\big) \\
&= \sum_{s=1}^{S} \big(-\frac{n_s}{2} \log|\Sigma_s| - \frac{n_s}{2} \mathrm{tr}(\Sigma_s^{-1}\hat{\Sigma}_s)\big)
\end{aligned}
$$

As identifiability constraints, the authors impose that Φ, $\beta_1, \ldots, \beta_S$ each be lower triangular matrices, which is a typical of classical factor analysis and forces the first loading to correspond to only the first axis of the factor space, the second loading to the first and second axes, so on and so forth. An additional constraint that $\mathrm{rank}\big([\Phi\ \beta_1 \ldots \beta_S]\big) = q + \sum_{s=1}^{S} \ell_s$ is needed to ensure uniqueness of the solution, since we have to estimate S group-specific loadings plus the common loadings from the S covariance matrices. $\hat{\Phi}$, $\hat{\beta}_1, \ldots, \hat{\beta}_S$, and $\hat{\Psi}_s$ are estimated by expectation-conditional maximization (ECM).

Of course, we are interested only in $\hat{\Phi}$, the matrix which describes how the latent factors characterizing the common signal are weighted to generate the observed data. By projecting the unobserved Poisson means into a common subspace with dimension $q < p$, we hope to remove much of the unique signal that may be obscuring signal of interest. However, we are interested in the subspace spanned by a set of orthonormal vectors in order of the variance on the common factors, and so what we seek are the rotated loadings $v_1, \ldots, v_q$, the first q columns of $V \in O(p)$ computed from $\hat{\Phi}\hat{\Phi}^T$ by the following spectral decomposition,

$$
\hat{\Phi}\hat{\Phi}^T = VAV^T, \tag{2.20}
$$

where A is the diagonal matrix of eigenvalues.

MSFA was implemented using the source code from the R package 'msfa' (DeVito et al., 2019), modified so as to obtain estimates from $\hat{\Sigma}_1, \ldots, \hat{\Sigma}_S$ instead of from the standard sample covariance matrices $\frac{1}{n_1 - 1}(X_1 - \mu_1)(X_1 - \mu_1)^T, \ldots, \frac{1}{n_S - 1}(X_S - \mu_S)(X_S - \mu_S)^T$.

Computing Scores

Since we are interested in the scores of the unobserved $\log \Lambda_{is}$, rather than the scores of the observed X_{is}, we adopt the following procedure from Kenney et al. (2019). To apply the classical PCA criterion of minimizing the squared reconstruction error between the original points and their projections onto principal component space, we would like to minimize the squared reconstruction error between $\log \Lambda_{is}$ and its projection P_{is}^q onto the q-dimensional common subspace spanned by the orthonormal vectors $v_1, \ldots, v_q$ that we estimated using CPCA or MSFA. Since Σ_s is the variance of $\log \Lambda_s$, this error is given by the squared distance

$$D^2 = (\log \Lambda_{is} - \mu_s - P_{is}^q)^T \Sigma_s^{-1} (\log \Lambda_{is} - \mu_s - P_{is}^q), \tag{2.21}$$

where, because the v_j are orthonormal, the projection is $P_{is}^q = \sum_{j=1}^q v_j v_j^T (\log \Lambda_{is} - \mu_s)$. At the same time, since $\log \Lambda_{is}$ is unobserved, we need to maximize the likelihood of the observed X_{is}. Whether we use PoissonPCA or PLNPCA to estimate $\Sigma_s = \operatorname{Varlog} \Lambda$, we assume that X_{ijs} was generated by a Poisson distribution with mean Λ_{ijs}, and so the log-likelihood is

$$\mathcal{L}(\Lambda_{is}) \propto \log \prod_{i=1}^{n_s} \Lambda_{is}^{X_{is}} \exp \Lambda_{is} = \sum_{i=1}^{n_s} X_{is}^T \log \Lambda_{is} + 1^T \Lambda_{is} \tag{2.22}$$

So combining equations (2.21) and (2.22) we arrive at an objective function,

$L(\Lambda_{is}) = \sum_{i=1}^{n_s} \big(X_{is}^T \log \Lambda_{is} + 1^T \Lambda_{is} - (\log \Lambda_{is} - \mu_s - P_{is}^q)^T \Sigma_s^{-1} (\log \Lambda_{is} - \mu_s - P_{is}^q) \big)$, which is optimized by Newton-Raphson iteration.

2.3 Ensemble Method

We have reviewed two very different ways of estimating the full- (or near full-) rank variance-covariance matrix from a set of conditionally independent realizations of Poisson sampling in which the Poisson means are subject to additional multiplicative noise: PoissonPCA and PLNPCA, each of which can either be performed with a sequencing depth correction or without. These methods are applied to each data set $X_1, \ldots, X_S$ individually. We went on to explore two distinct methods that can take a set of S estimated variance-covariance matrices and estimate a set of $q < p$ common vectors forming a shared orthogonal basis for $\mathbb{R}^q$: CPCA (for which we have a choice of two algorithms, SCPCA and FCPCA) and MSFA. Chapters 4 and 5 will

	Ensemble Methods	
No SDC	PoissonPCA + SCPCA	PLNPCA + SCPCA
	PoissonPCA + FCPCA	PLNPCA + FCPCA
	PoissonPCA + MSFA	PLNPCA + MSFA
SDC	PoissonPCA + SCPCA	PLNPCA + SCPCA
	PoissonPCA + FCPCA	PLNPCA + FCPCA
	PoissonPCA + MSFA	PLNPCA + MSFA

Table 2.1: The twelve candidate ensemble methods.

therefore demonstrate simulation and real data analysis results for all twelve possible combinations of methods, as given in Table 2.1, to compare their performance as ensemble methods. 'SDC' in tables and figures henceforth refers to 'sequencing depth correction'.

Chapter 3

Simulation Study

We performed simulation studies for two groups across several scenarios. These scenarios differed on the true signal (whether the eigenvalues used to simulate the variance-covariance matrices were simultaneously decreasing), the sample sizes n_1 and n_2, whether or not the sample sizes were unbalanced, and whether or not sequencing depth correction was performed in the Poisson modeling stage. For each simulation experiment, the process of simulating Poisson log-normal data and applying each method was performed 100 times. For each of the 100 replicates, synthetic data for two "groups" were simulated as follows. Synthetic eigenvectors E_1 were constructed for Group 1 by spectral decomposition of a synthetic covariance matrix FF^T where each column F_j, $j = 1, \ldots, p$ was a normalized length-p vector of standard normal deviates. We then constructed the eigenvectors E_2 for Group 2 so as to share the first q columns of E_1 for several values of q, while the remaining columns $e_{2,q+1}, \ldots, e_{2,p}$ were replaced by vectors of standard normal deviates regressed on the preceding q columns and normalized. These eigenvectors, along with a pre-determined set of eigenvalues for each group, were used to construct variance-covariance matrices Σ_1 and Σ_2. Note that with decreasing eigenvalues, for $q << p$, by this construction each shared eigenvector will be a principal eigenvector of the two covariance matrices, whereas with non-decreasing eigenvalues some of the large variances will not be associated with the shared axes. We performed simulation experiments for both scenarios. Next, these covariance matrices Σ_1 and Σ_2 in turn were used to simulate the transformed latent Poisson means $\log \Lambda_1$ and $\log \Lambda_2$ using the multivariate normal, with mean vectors for each group consisting of p normal random variates (see Table 3.1). We then performed scalar multiplication of Λ_1 and Λ_2 respectively by length-n_1 and length-n_2 vectors of gamma random variates to simulate sequencing depth error, and finally these means

	Group 1	**Group 2**
$\log \Lambda_{is} \sim \mathcal{N}(\mu_{is}, \Sigma_s)$	$\mu_{ij1} \sim \mathcal{N}(4, 3)$	$\mu_{ij2} \sim \mathcal{N}(3, 2)$
$X_{ijs} \sim \text{Poisson}(\gamma_{is}\Lambda_{ijs})$	$\gamma_{i1} \sim \text{Gamma}(7, 1)$	$\gamma_{i1} \sim \text{Gamma}(10, 1)$

Table 3.1: Distributions used to simulate Poisson log-normal data.

were used to generate $n_1 \times p$ and $n_2 \times p$ synthetic data matrices of Poisson random variates for each group respectively (see Table 3.1). Then, the candidate ensemble methods listed in Table 2.1, and some single-group alternatives (PCA on the concatenated datasets, PCA on the log-transformed concatenated datasets, PoissonPCA on the concatenated datasets, or PLNPCA on the concatenated datasets) were performed on the synthetic data. Before log-transforming count or relative abundance data for PCA, zero count were first imputed with 0.001.

In the case of SCPCA and FCPCA, the explained variances (eigenvalues) for each estimated orthogonal loading vector were computed by

$$\hat{d}_{sj} = \hat{v_j}^T \Sigma_s \hat{v_j}, \;\; j = 1, \ldots, p, \;\; s = 1, 2 \tag{3.1}$$

where Σ_s is the true variance-covariance matrix of Λ_s. Cumulative sums of $\hat{d}_{11}, \ldots, \hat{d}_{1p}$ and $\hat{d}_{21}, \ldots, \hat{d}_{2p}$ were divided by the true eigenvalues $\sum_{j=1}^{p} d_{1j}$ and $\sum_{j=1}^{p} d_{2j}$ respectively to find the proportion of the true variance explained by each method.

In the case of MSFA, as the common loadings are not constrained to orthogonality, the variance $\hat{\Phi}\hat{\Phi}^T$ of the common factors was computed and then eigen-decomposed as described in equation (2.20). The resultant $\hat{v}_1, \ldots, \hat{v}_p$ (of which only the first q contain signal, but all p are retained in this case for ease of visibility in plots) were used to compute the estimated eigenvalues and the proportion of true variance explained as according to equation (3.1). Representative results of the simulations are given by Figures 3.1 and 3.12, where in each plot the true eigenvalues are given in black.

First, Figures 3.1 and 3.2 show results of the candidate ensemble methods in the decreasing eigenvalues case, where sequencing depth corrections were applied for PoissonPCA and PLNPCA in the former but not the latter, and where the variance-covariance matrices of Group 1 and Group 2 shared one eigenvector, and the sample sizes $n_1 = 200 > n_2 = 100$ were unbalanced. Figure 3.3 shows the corresponding results using each of the four naive PCA methods on the concatenated data, as well as PoissonPCA or PLNPCA alone on the concatenated data with their respective

sequencing depth corrections. Figures 3.5 to 3.6 show the analogous results for when Σ_1 and Σ_2 shared five principal eigenvectors.

Then, Figures 3.8 to 3.9 again show results for the candidate ensemble methods with sequencing depth correction, the candidate ensemble methods without sequencing depth correction, and the single-group methods (where sequencing depth correction was applied for PoissonPCA and PLNPCA), this time in the non-decreasing eigenvalues case. Thus, in these plots, principal eigenvectors of Σ_1 and Σ_2 were not necessarily shared. Group 1 and Group 2 shared one eigenvector, and the sample sizes $n_1 = 200 > n_2 = 100$ were again unbalanced. Figures 3.11 to 3.12 show the analogous results for when Σ_1 and Σ_2 shared five common eigenvectors. In each of these figures, the true and estimated eigenvalues corresponding to the q shared eigenvectors were plotted in descending order (indicated by solid points for the true eigenvalues), and subsequently the estimated eigenvalues corresponding to the unique eigenvectors were plotted in descending order (indicated by outline-only points for the true eigenvalues).

Results for balanced sample sizes and large sample sizes are provided in Supplementary Figures A.1 and A.24.

The simulations showed generally good performance by the ensemble methods and by PoissonPCA or PLNPCA alone. This is especially clear in comparison to the naive PCA methods on relative abundance or raw counts which were not able to find good rotations for the data, and as a result showed very slow increases in their cumulative explained variance and no improvement as a result of increasing the number of shared eigenvectors. In contrast, PCA on the log-counts performed decently. While both variance estimation methods seem to do fairly well, PoissonPCA is seen here to have consistently outperformed PLNPCA in terms of the reconstruction of the dominant signal in each group. Interestingly, the difference in performance between PoissonPCA and PLNPCA on the first few common axes is negligible without sequencing depth correction, which suggests that the PoissonPCA compositional correction to the variance estimate is superior to PLNPCA's treatment of observed read count as an offset. Incidentally PoissonPCA is also an order of magnitude faster to run. However, regardless of which method was used to estimate the variance, in the cases where no sequencing depth correction was applied, the first CPC explained very little variance because that axis instead captured the unique variation from sequencing depth. The second CPC then usually explains a large proportion of common variance, and thereafter the points track the true signal.

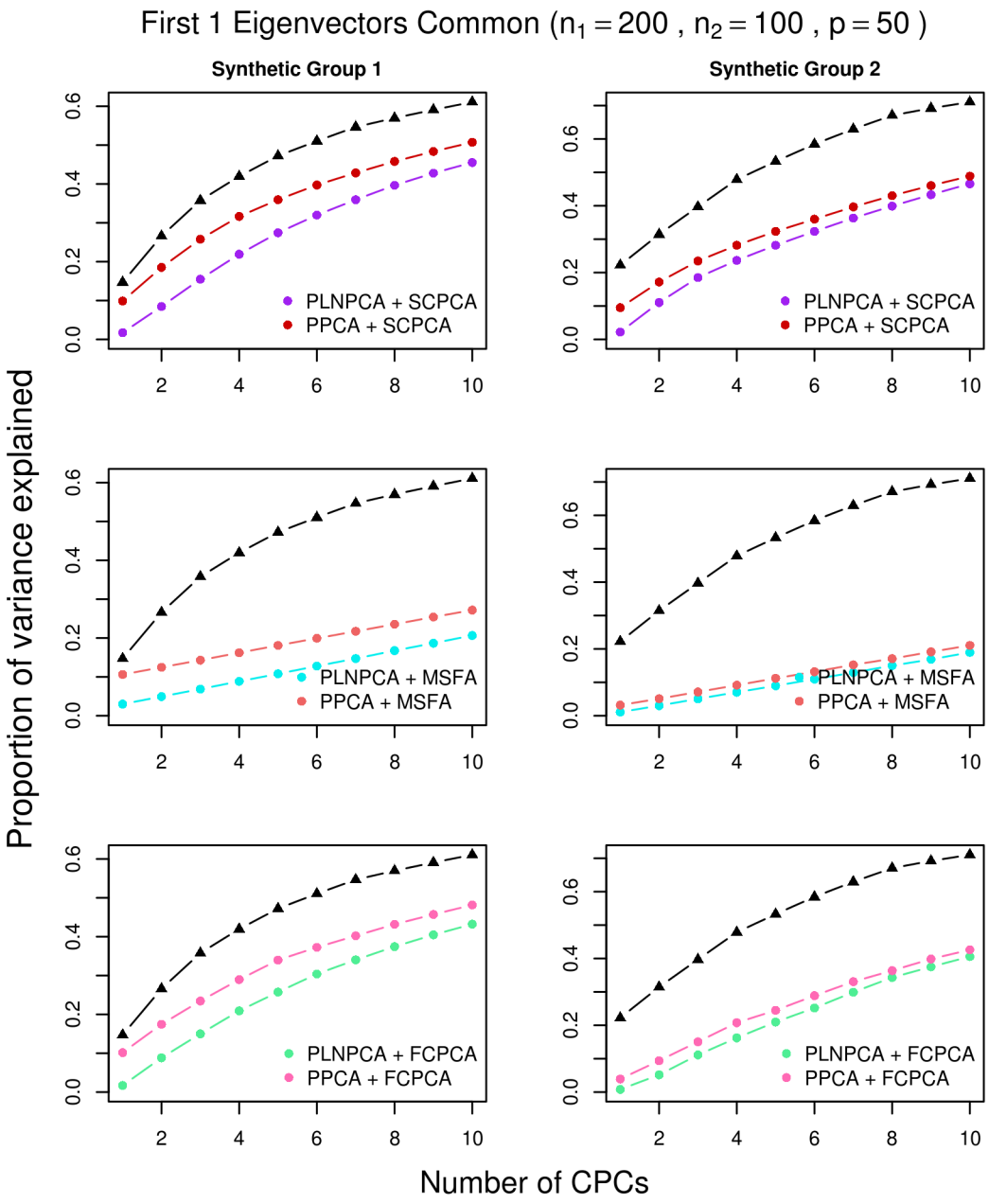

Figure 3.1: Simulation results for decreasing eigenvalues and one common eigenvector, with SDC; p=50, $n_1 = 200$, $n_2 = 100$.

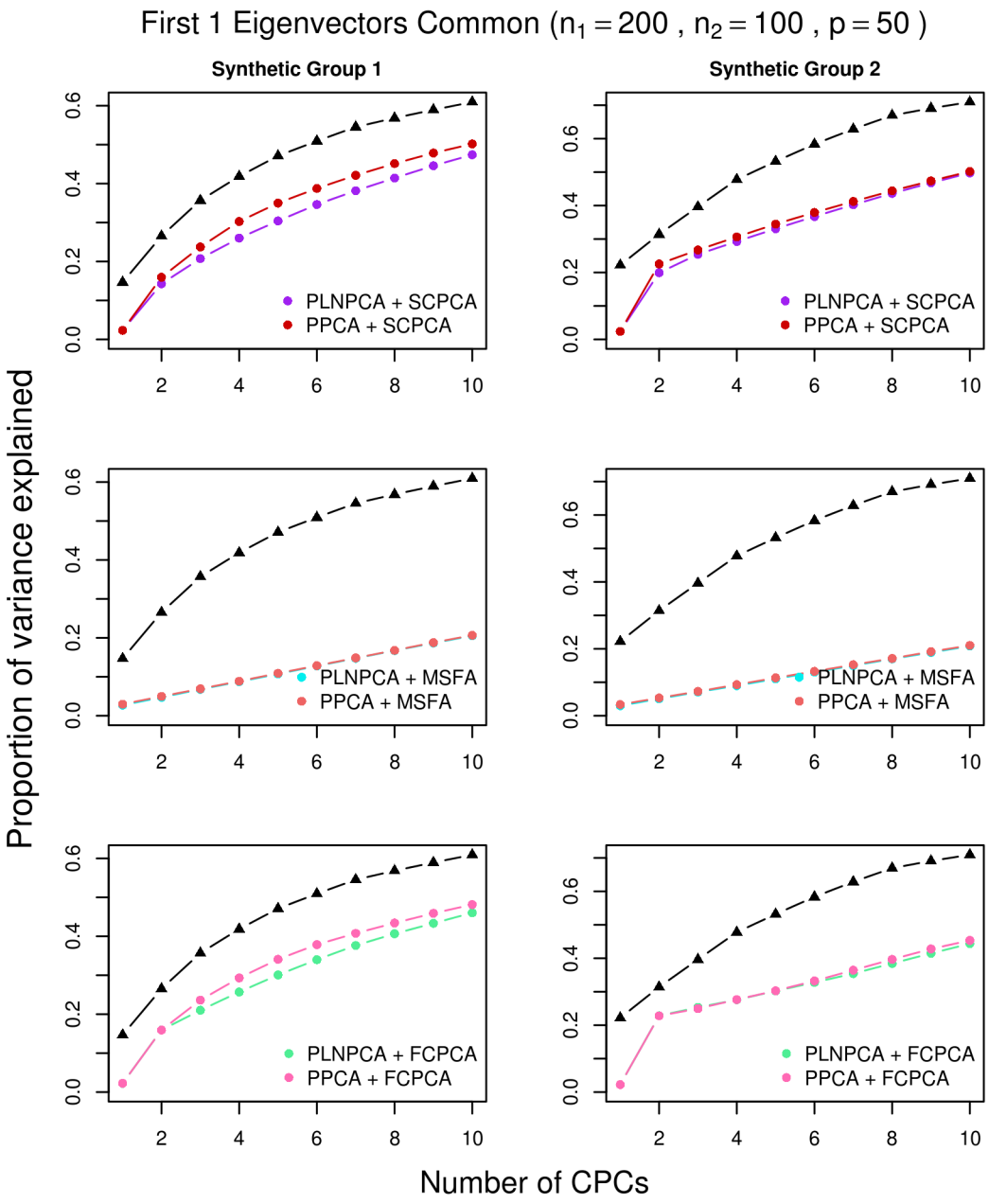

Figure 3.2: Simulation results for decreasing eigenvalues and one common eigenvector, with no SDC; p=50, $n_1 = 200$, $n_2 = 100$.

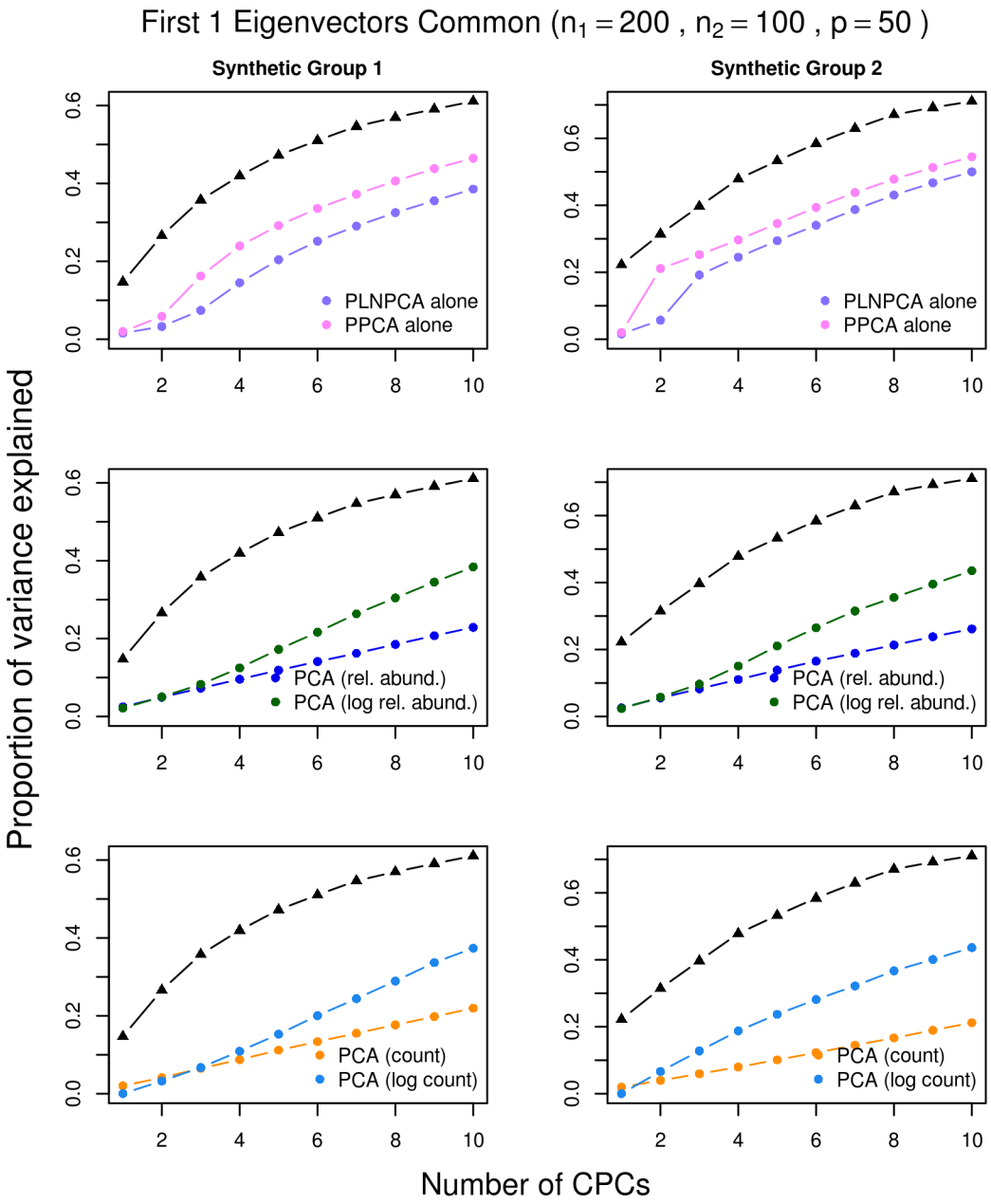

Figure 3.3: Simulation results for decreasing eigenvalues and one common eigenvector, using single-group methods (with SDC for P/PLN) on the concatenated data; p=50, $n_1 = 200$, $n_2 = 100$.

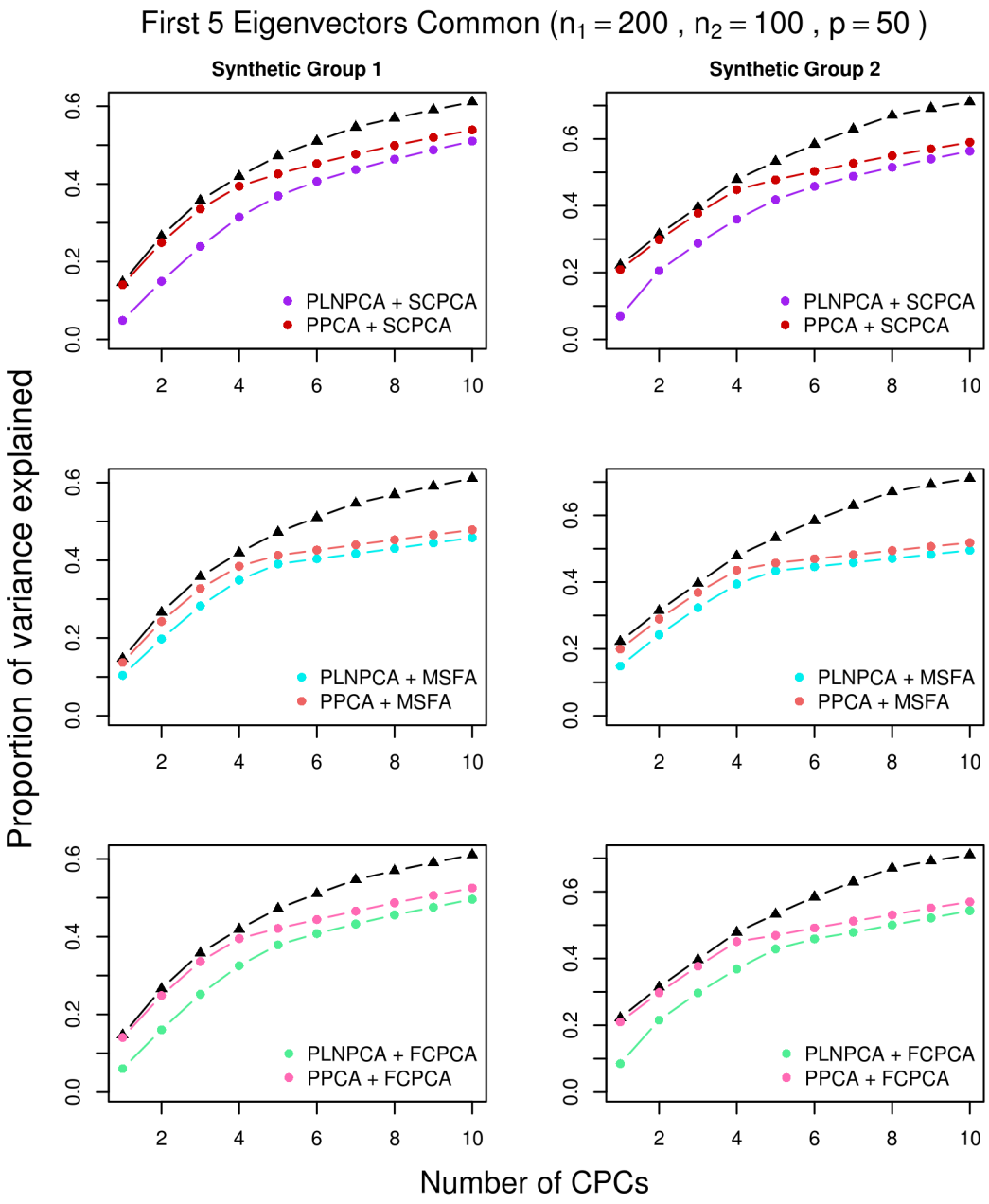

Figure 3.4: Simulation results for decreasing eigenvalues and five common eigenvectors, with SDC; p=50, $n_1 = 200$, $n_2 = 100$.

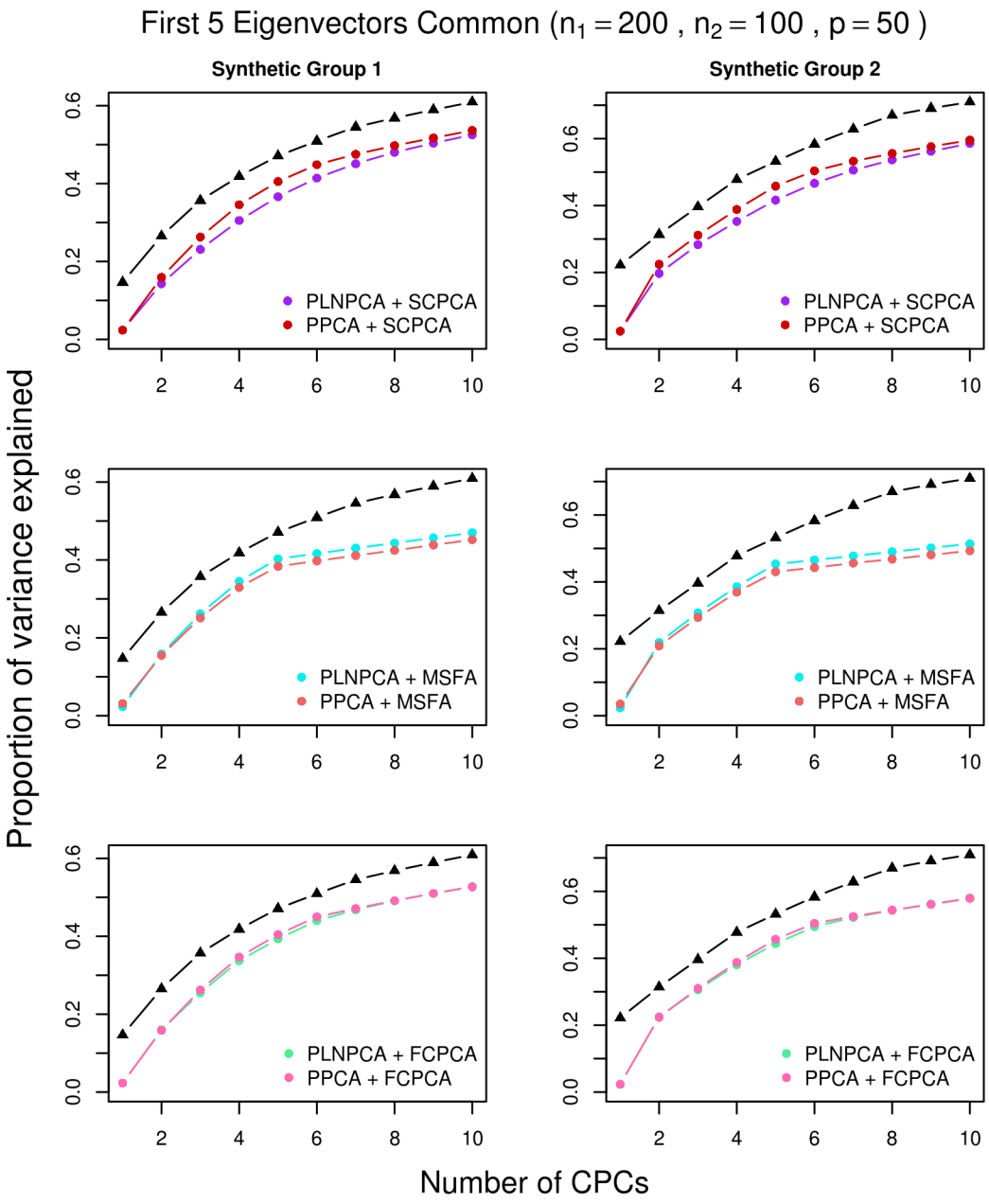

Figure 3.5: Simulation results for decreasing eigenvalues and five common eigenvectors, with no SDC; p=50, $n_1 = 200$, $n_2 = 100$.

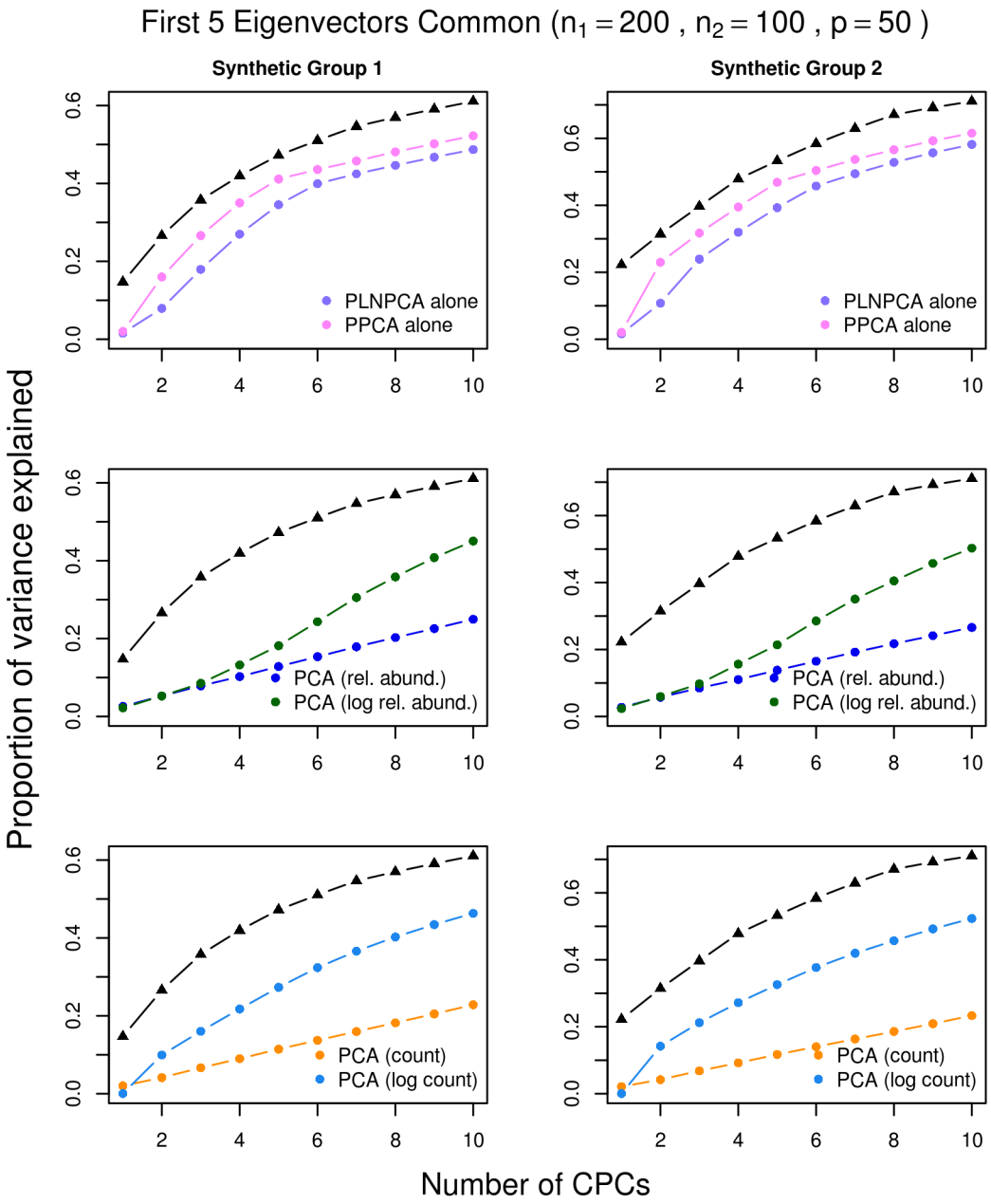

Figure 3.6: Simulation results for decreasing eigenvalues and five common eigenvectors, using single-group methods (with SDC for P/PLN) on the concatenated data; p=50, $n_1 = 200$, $n_2 = 100$.

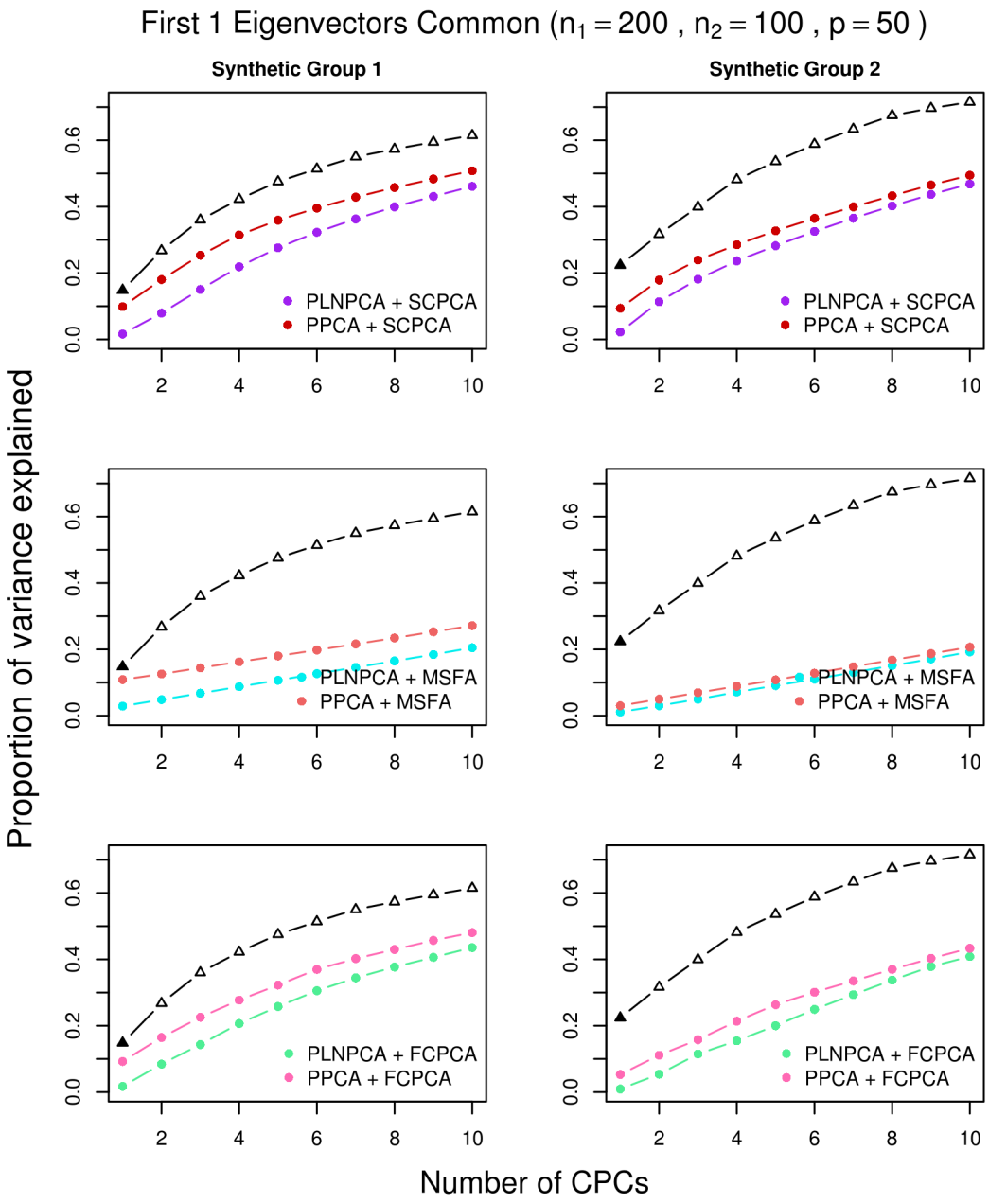

Figure 3.7: Simulation results for non-decreasing eigenvalues and one common eigenvector, with SDC; p=50, $n_1 = 200$, $n_2 = 100$.

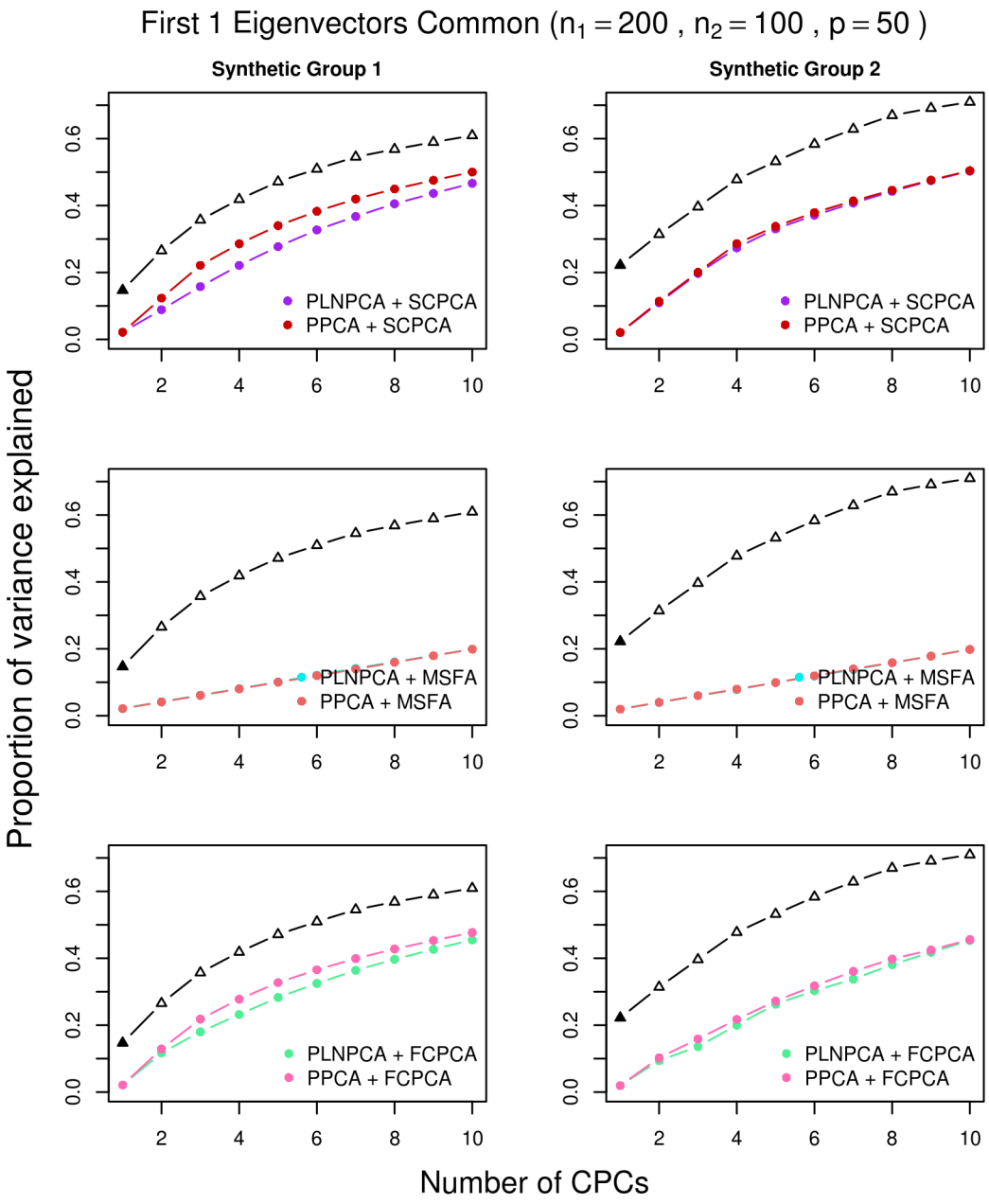

Figure 3.8: Simulation results for non-decreasing eigenvalues and one common eigenvector, with no SDC; p=50, $n_1 = 200$, $n_2 = 100$.

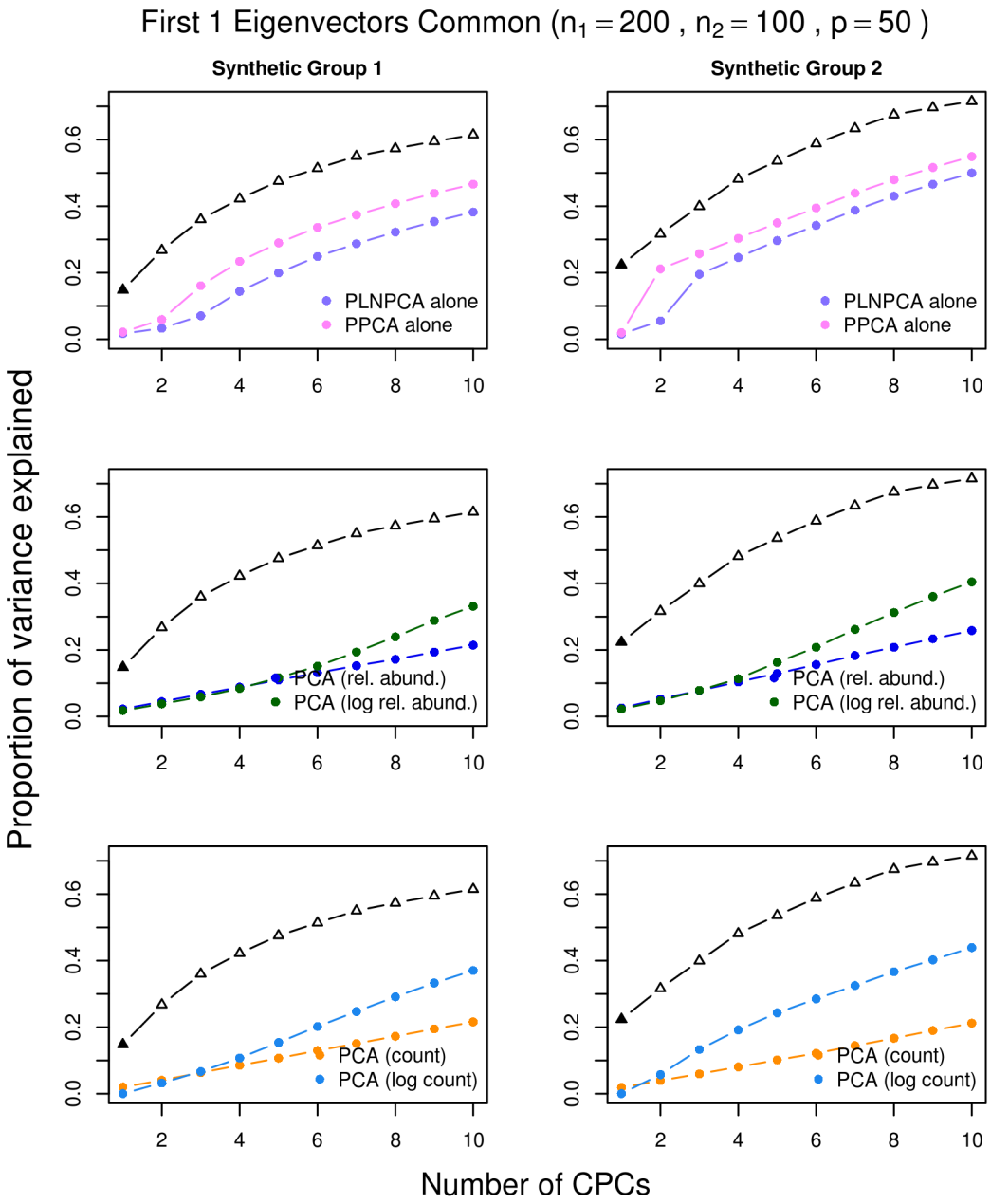

Figure 3.9: Simulation results for non-decreasing eigenvalues and one common eigenvector, using single-group methods (with SDC for P/PLN) on the concatenated data; p=50, $n_1 = 200$, $n_2 = 100$.

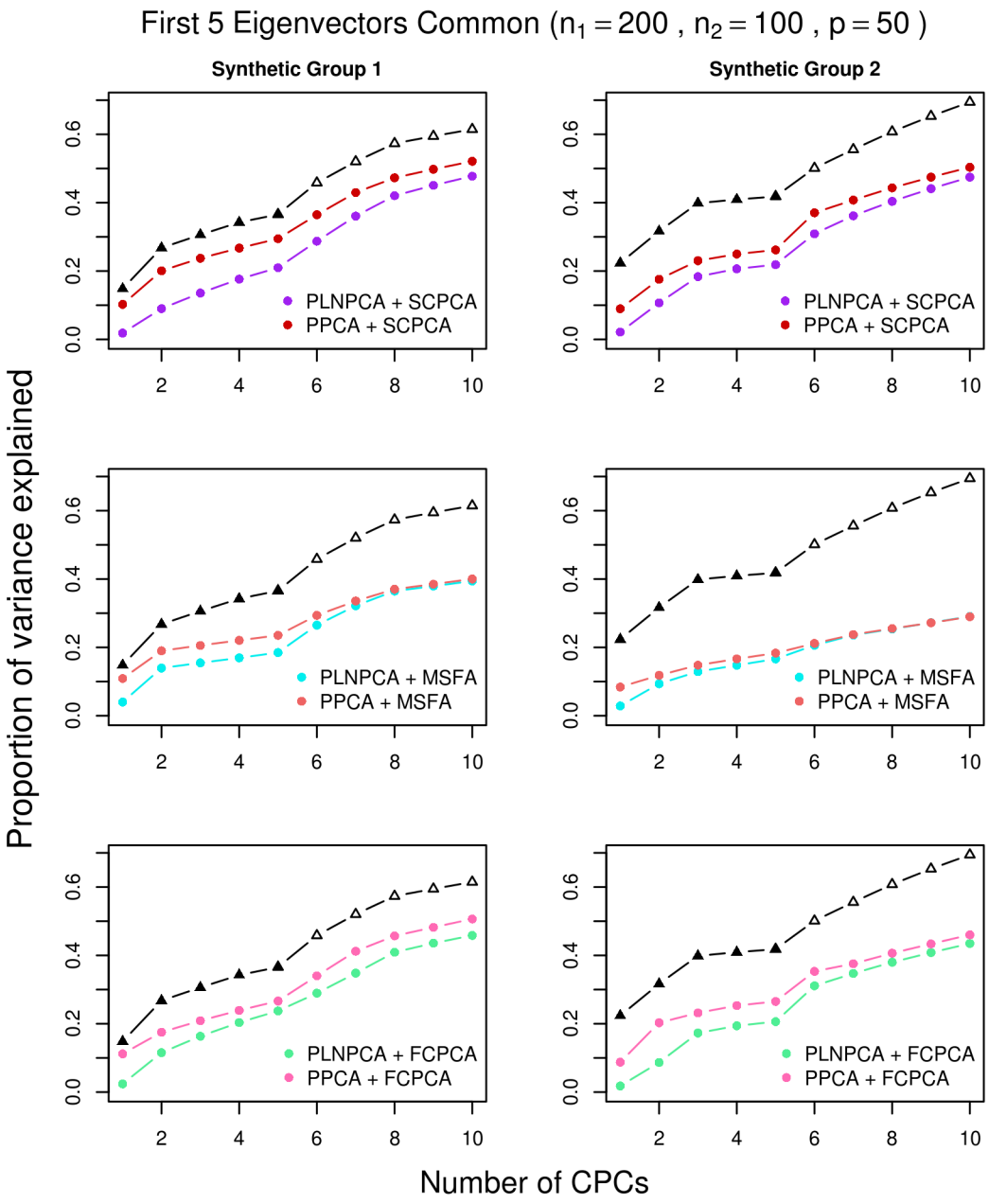

Figure 3.10: Simulation results for non-decreasing eigenvalues and five common eigenvectors, with SDC; p=50, $n_1 = 200$, $n_2 = 100$.

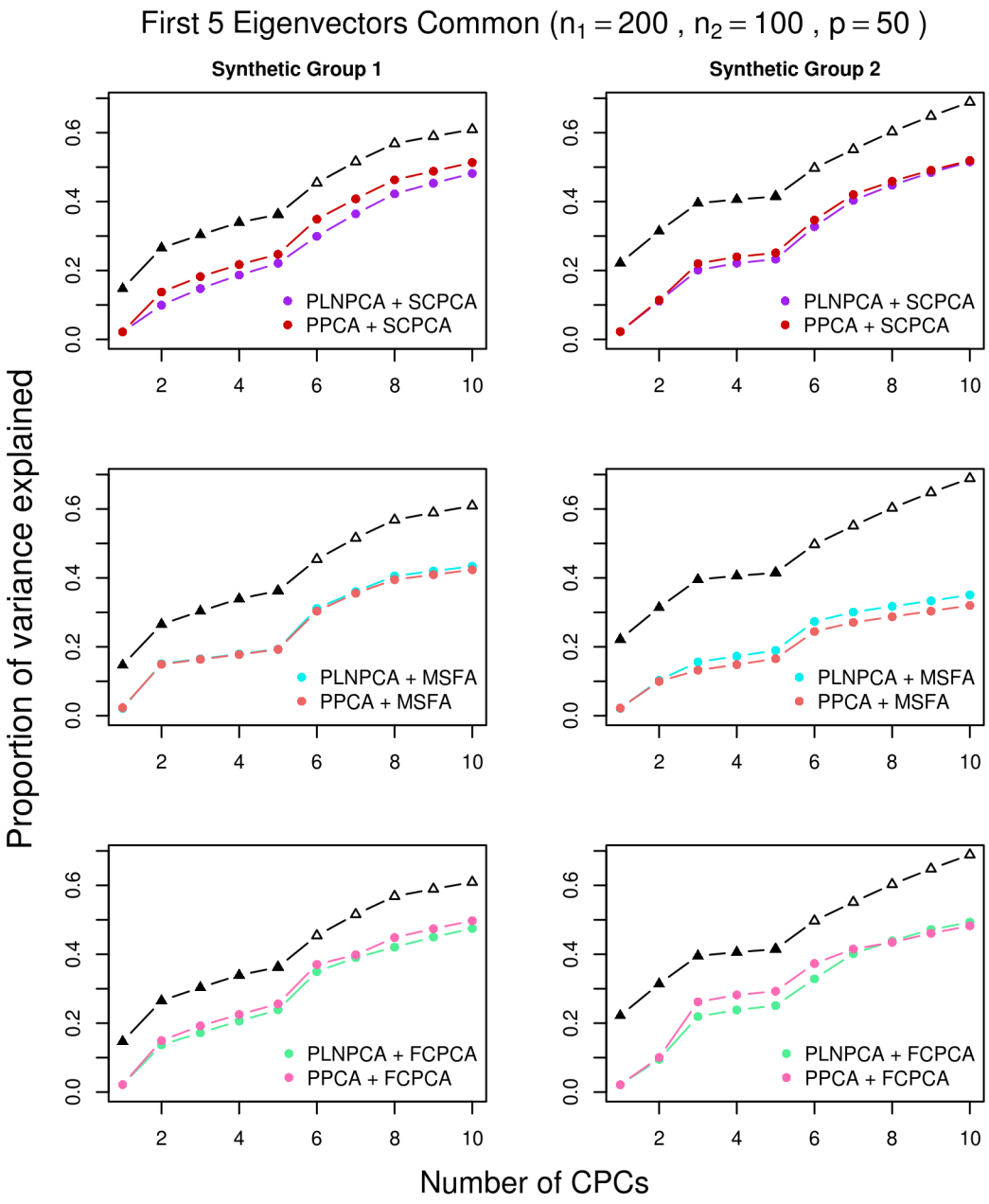

Figure 3.11: Simulation results for non-decreasing eigenvalues and five common eigenvectors, with no SDC; p=50, $n_1 = 200$, $n_2 = 100$.

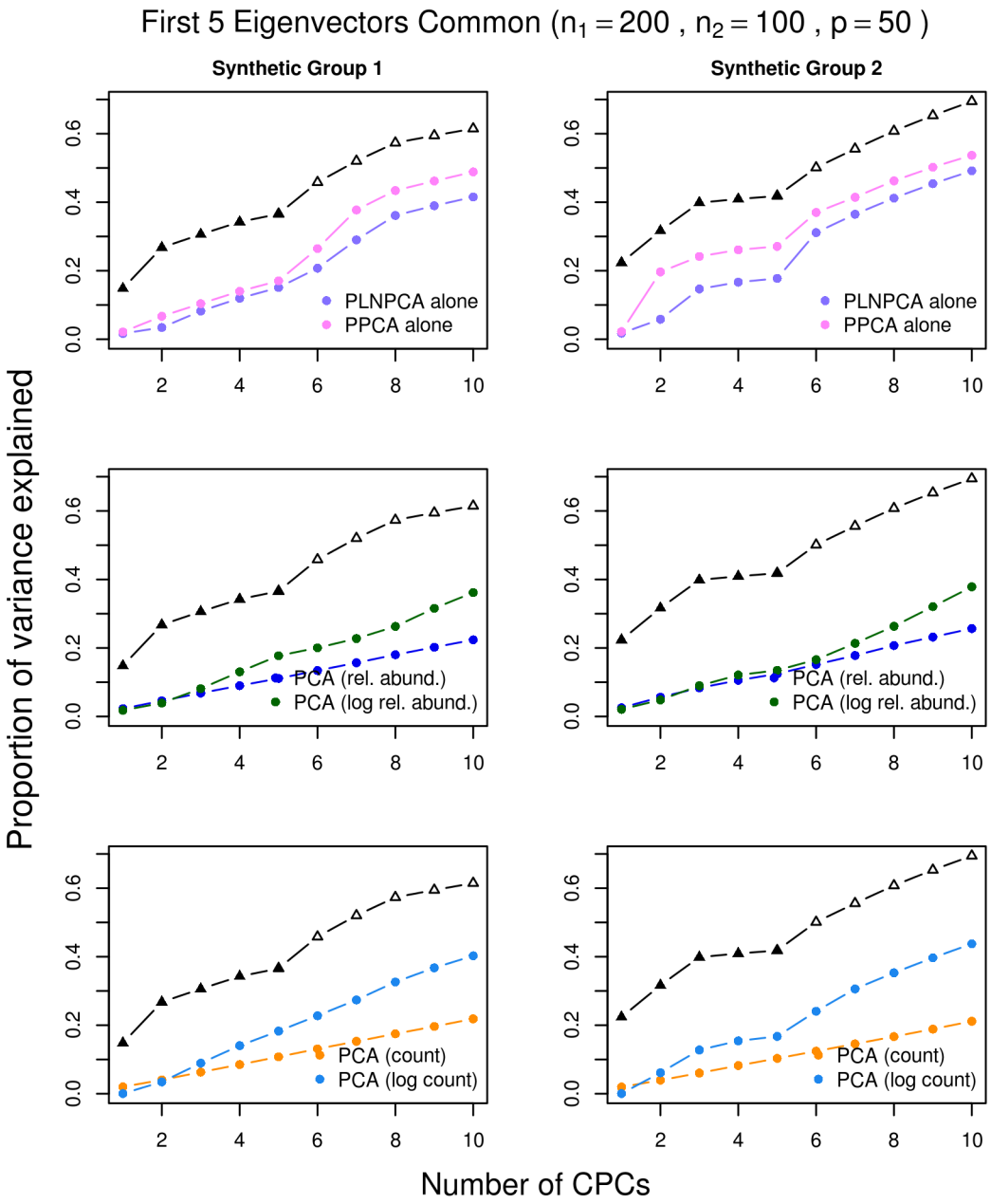

Figure 3.12: Simulation results for non-decreasing eigenvalues and five common eigenvectors, using single-group methods (with SDC for P/PLN) on the concatenated data; p=50, $n_1 = 200$, $n_2 = 100$.

As for the estimation of $v_1, \ldots, v_q$, the two CPCA methods are very similar and both did well. SCPCA appeared to outperform FCPCA when the signal was more difficult to resolve, such as when there were very few shared eigenvectors and the true cumulative variance increased slowly (e.g., Group 2 with one shared eigenvector). While technically SCPCA, FCPCA, and MSFA are all designed for positive-definite symmetric matrices, in practice SCPCA performed well even when the covariance matrix was indefinite, which was often true in the PoissonPCA case. SCPCA performs almost indistinguishably from FCPCA when the eigenvalues are non-decreasing. MSFA performed well, which is unsurprising, since MSFA assumes only q common axes whereas CPCA assumes that all p axes are common, and furthermore we were able to specify the true q. Interestingly, the difference in performance between MSFA applied to PoissonPCA vs. PLNPCA variance estimates was often not large, whereas SCPCA and FCPCA both performed much better on the PoissonPCA variance estimate. However, when common signal was very low (q=1), MSFA struggled to capture any variation on the first axis for Group 2 compared to SCPCA. Moreover, the MSFA optimization routine failed so frequently that the simulation results for MSFA had to be averaged over only the successful replicates (typically 80%-90%). In addition, MSFA was by far the slowest method, and so from these simulations it would appear that the CPCA approaches have more practical utility.

Finally, PoissonPCA or PLNPCA alone on the concatenated data performed similarly to either method in combination with a common factor extraction method, with the first axis capturing unique signal even with a sequencing depth correction, and the second axis showing very good recovery. The fact that these methods alone can adequately reconstruct the signal of the two groups while naive PCA methods fail to do so suggests that misspecification of models with respect to sampling may pose a larger obstacle to cross-study microbial abundance analyses than the heterogeneous study-specific sources of variation in and of themselves. However, the CPCA methods outperform PoissonPCA or PLNPCA alone when some of the large eigenvalues of Σ_s are not associated with the shared eigenvectors. When the true eigenvalues are non-decreasing, the cumulative variance explained by PoissonPCA or PLNPCA alone tracks the same trajectory of the true cumulative variance, instead of loading as much variance as possible on the first few axes like SCPCA, FCPCA, or MSFA.

Finally, Figure 3.13 contains scree plots for the estimated eigenvalues from PoissonPCA followed by SCPCA for each group, where Σ_1 and Σ_2 share either 5 or 10 common eigenvectors and the true eigenvalues are simultaneously decreasing. These

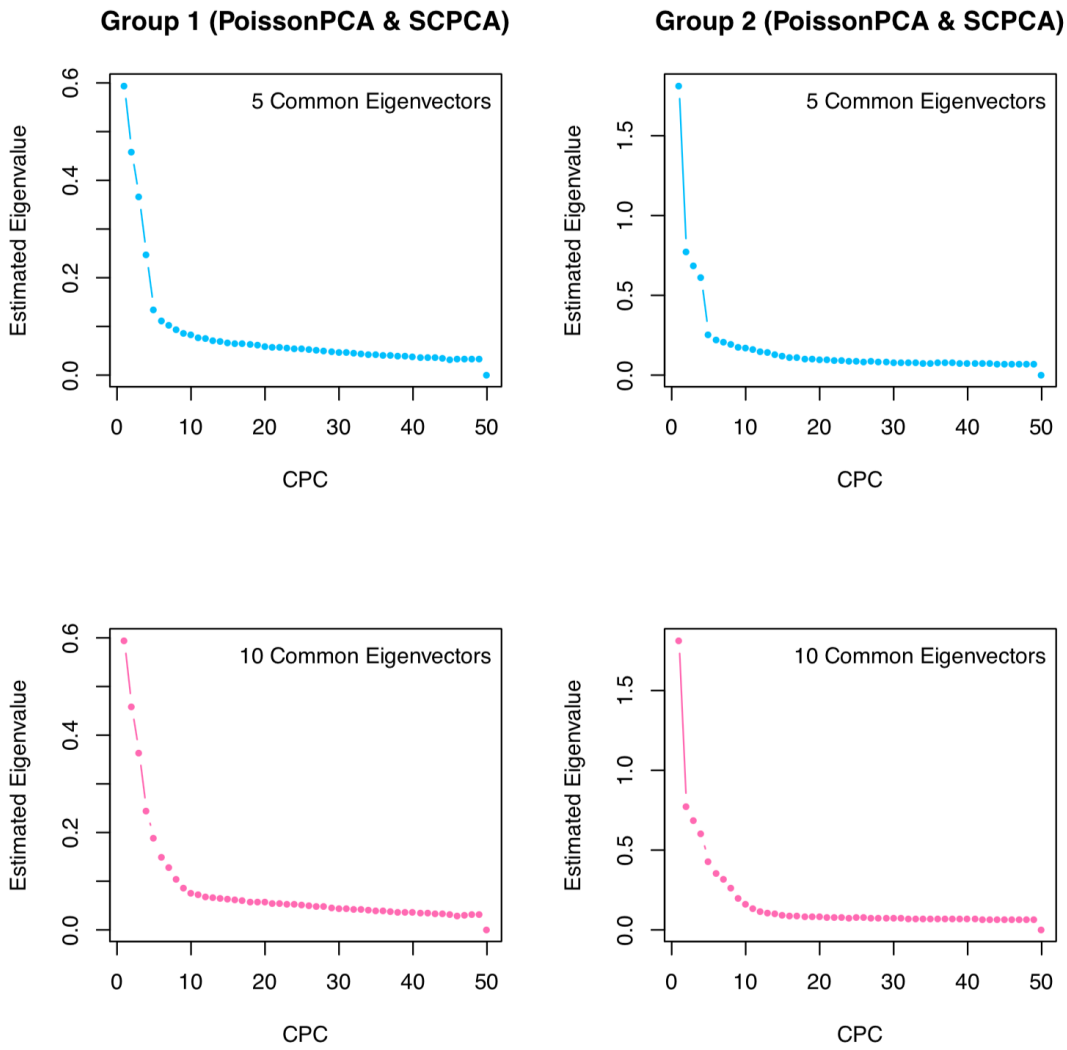

Figure 3.13: Scree plots of estimated eigenvalues from PoissonPCA & SCPCA for each group with 5 or 10 shared eigenvectors.

plots show that the differences between successive estimated eigenvalues drop to near zero when the number of CPCs is larger than the true number of shared eigenvectors, which provides evidence that we will be able to make a good choice of q when applying the method to real data for which we cannot know the true number of shared eigenvectors.

Chapter 4

Data Analysis

The communities comprising the human gut microbiome have been implicated in processes such as drug metabolism (Blaser et al., 2013), immune system function (Thais et al., 2016), energy balance (Nieuwdorp et al., 2014), and in numerous disease states (Blaser et al., 2013). It is not surprising, then, that many studies have found links between the gut microbiota and cancer of the colon. To investigate the performance of our candidate ensemble methods on real data, we re-analyzed the metagenomic datasets from Feng et al. (2015) and Zeller et al. (2014), who each collected fecal samples from participants diagnosed with colorectal carcinoma (CRC) or non-malignant colorectal adenoma, and from controls (study and participant characteristics are summarized in Table 4.1). Fecal samples were processed according to each research team's workflow, and samples underwent shotgun metagenomic sequencing on the Illumina HiSeq platform. Although each team had their own bioinformatic pipeline to process the raw reads, the taxonomy tables used for the present data analysis were obtained from the R package 'curatedMetagenomicData' (Pasolli et al., 2017), whose authors applied a standard pipeline for assembly, gene prediction, and

	Zeller et al. (2014)	**Feng et al. (2015)**
Number of samples	199	154
Country of origin	France, Germany	Austria
Sequencing technology	Illumina HiSeq	Illumina HiSeq
Number of CRC samples	91	46
Number of adenoma samples	42	47

Table 4.1: Comparison of Zeller et al. (2014) and Feng et al. (2015).

taxonomic assignment to the raw files from each study.

The taxonomic abundance tables for each dataset included strain-level taxa from all three domains of life, as well as viruses. Since our candidate methods do not involve any regularization, we separately collapsed the data to the genus level and removed features with near zero variance to reduce the number of taxa to $p = 104 < \min(n_Z = 199, n_F = 154)$. We then subsetted the features to include only those common to the two datasets, and ran the twelve candidate combinations of methods listed in Table 2.1 just as we did for the simulations in Chapter 3. We then computed the scores by projecting the latent means into the common space.

Figure 4.1 depicts selected score plots for each candidate ensemble method with sequencing depth corrections applied (see Supplemental Figure A.25 for the equivalent plots with no sequencing depth correction) which show participants with CRC clustering distinctly from participants without CRC on the common axes. From Figure 4.2, which shows several axes on each dataset separately for PoissonPCA & SCPCA with sequencing depth correction, we can see that the Feng et al. (2015) data seem to be better behaved than Zeller et al. (2014) in terms of clustering. In terms of lab-specific signal, Supplemental Figure A.26 through Supplemental Figure A.36 show score plots for each ensemble method colored by study of origin, with Supplemental Figure A.37 through Supplemental Figure A.41 containing those for the single-group and naive methods. While the sequencing depth correction does appear to create some clustering by study on some axes, the CPCs that best discriminate disease state do not show this. For example, for PoissonPCA with sequencing depth correction followed by SCPCA, the best clustering of control vs. CRC samples is shown by CPC 4 and CPC 3, and Figure 4.3 shows that lab-driven signal only appears in plots of CPC 1 or CPC 6. This suggests that these axes indeed correspond to common, generalizable CRC-related biological signal. In contrast, although PoissonPCA with sequencing depth correction does show clustering by disease state on the selected PCs in Figure 4.5, the score plots in Figure 4.4 show clustering by study on these and most other PCs. Of the naive PCA methods, each one shows some clustering by disease state (Figure 4.5). Supplemental Figure A.37 through Supplemental Figure A.41 do not suggest pervasive clustering by study of origin, although this is expected since the data were individually mean-centered by study.

We then investigated the predictive ability of these scores. Since we, like Zeller et al. (2014), found that samples positive for colorectal adenoma tended to cluster with control samples, we collapsed the two groups together. As a benchmark of the dis-

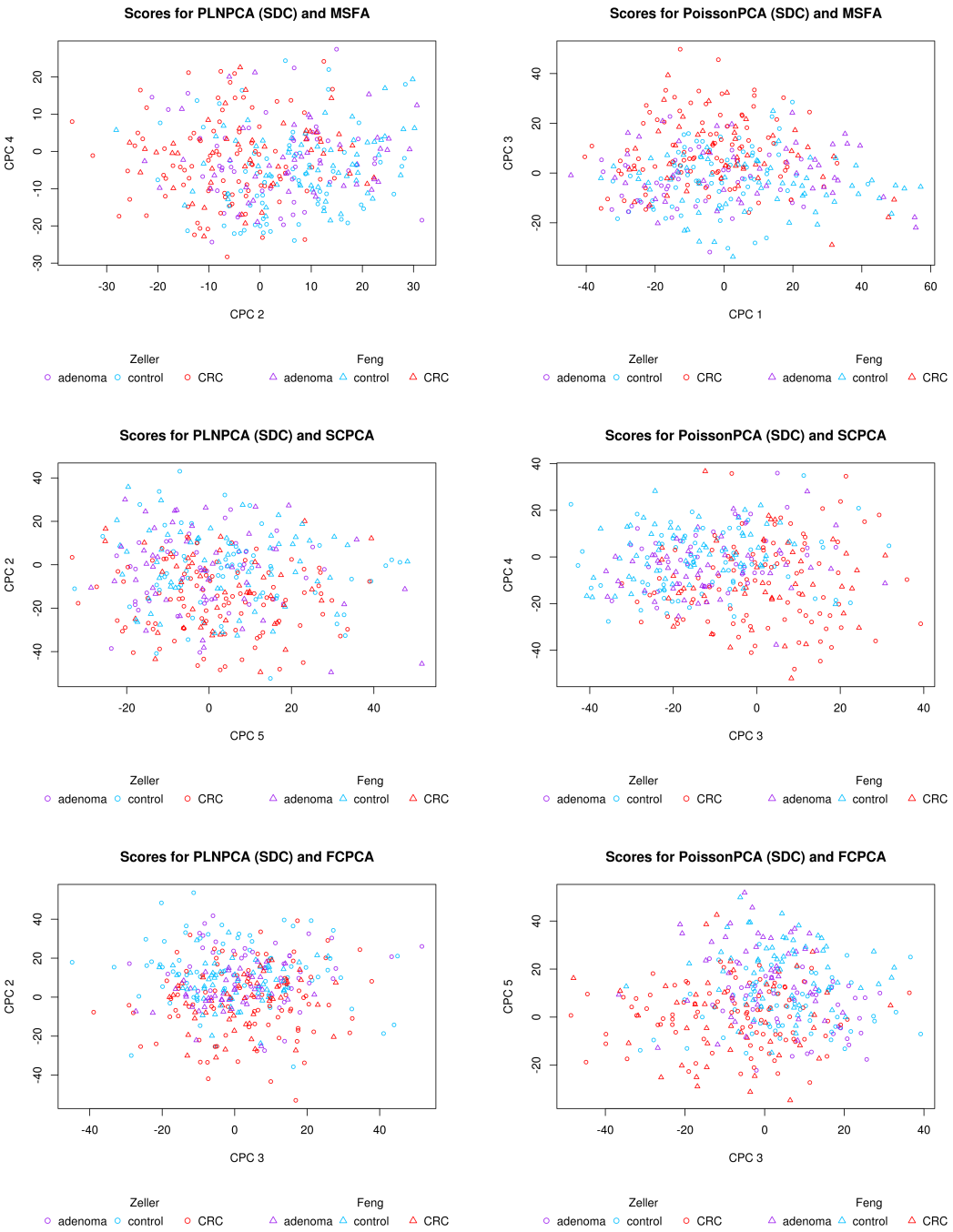

Figure 4.1: Scores from ensemble methods with SDC by disease state.

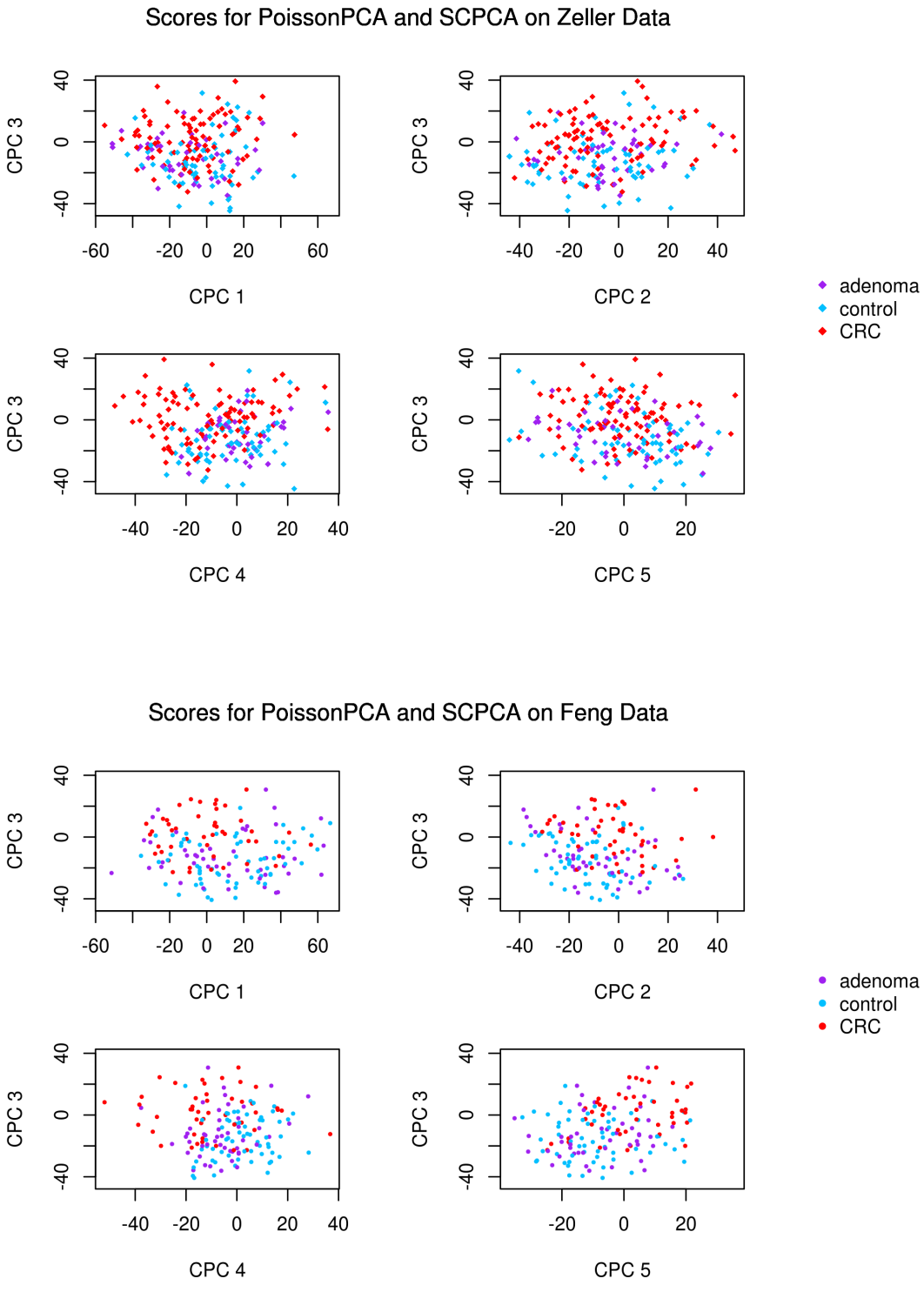

Figure 4.2: Score plots by disease state (PoissonPCA with SDC & SCPCA).

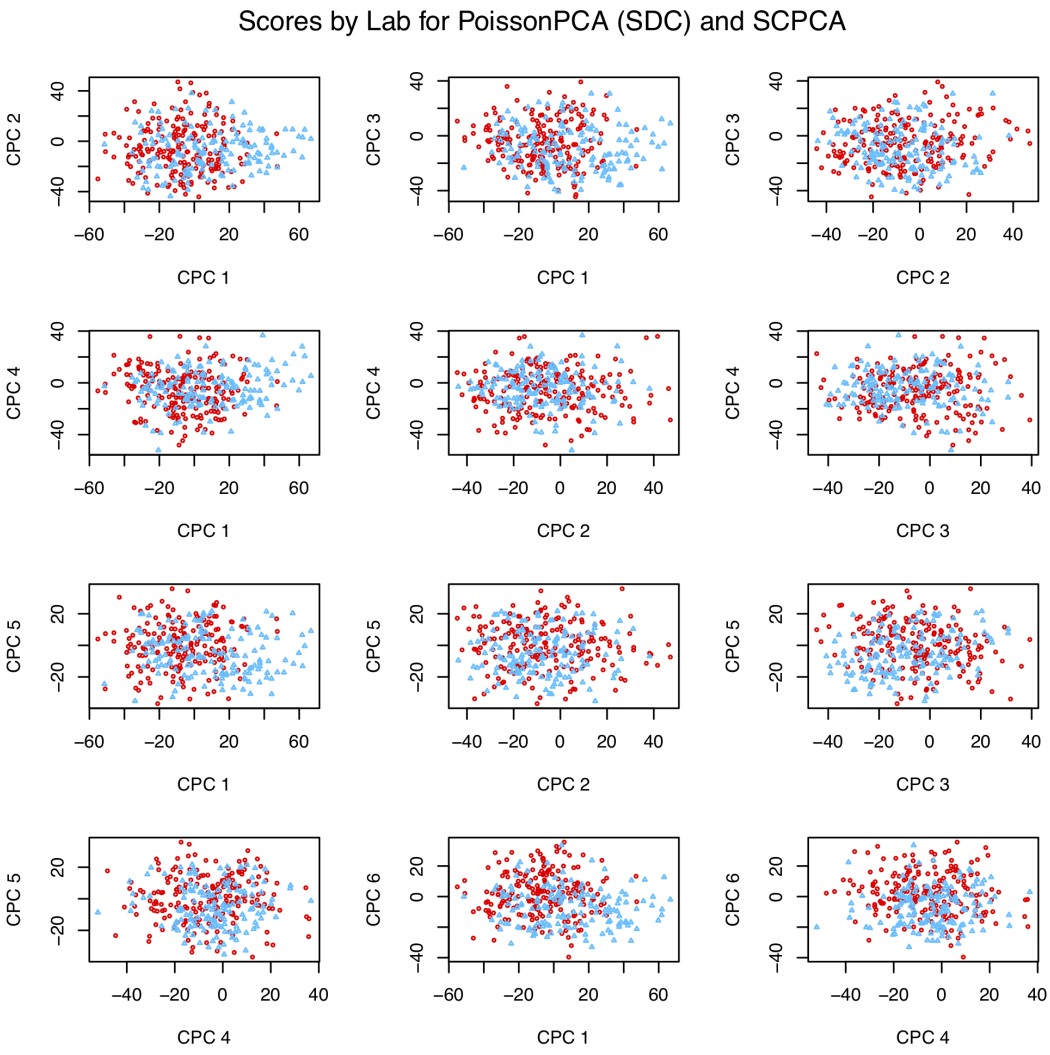

Figure 4.3: Scores by study of origin (PoissonPCA with SDC and SCPCA).

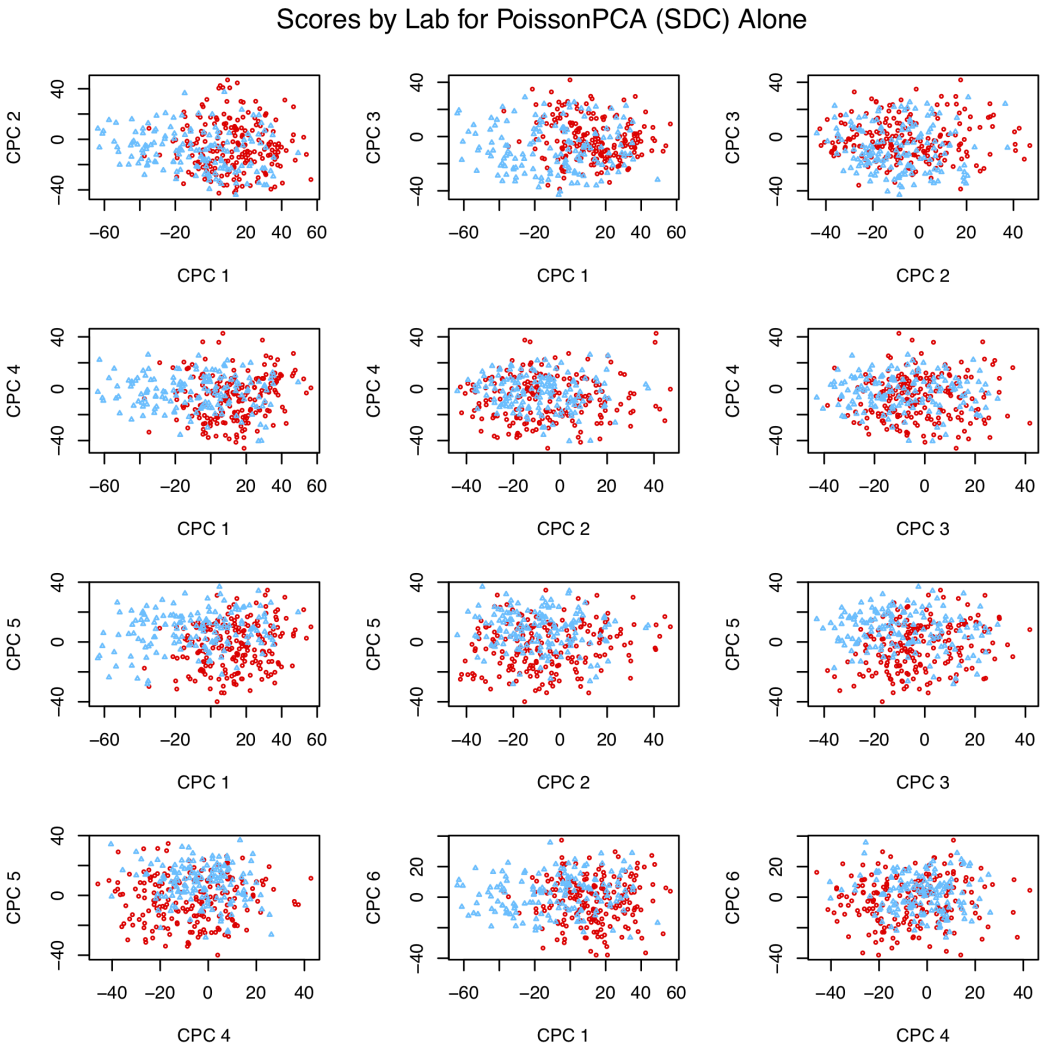

Figure 4.4: Scores from PoissonPCA alone with SDC by study of origin.

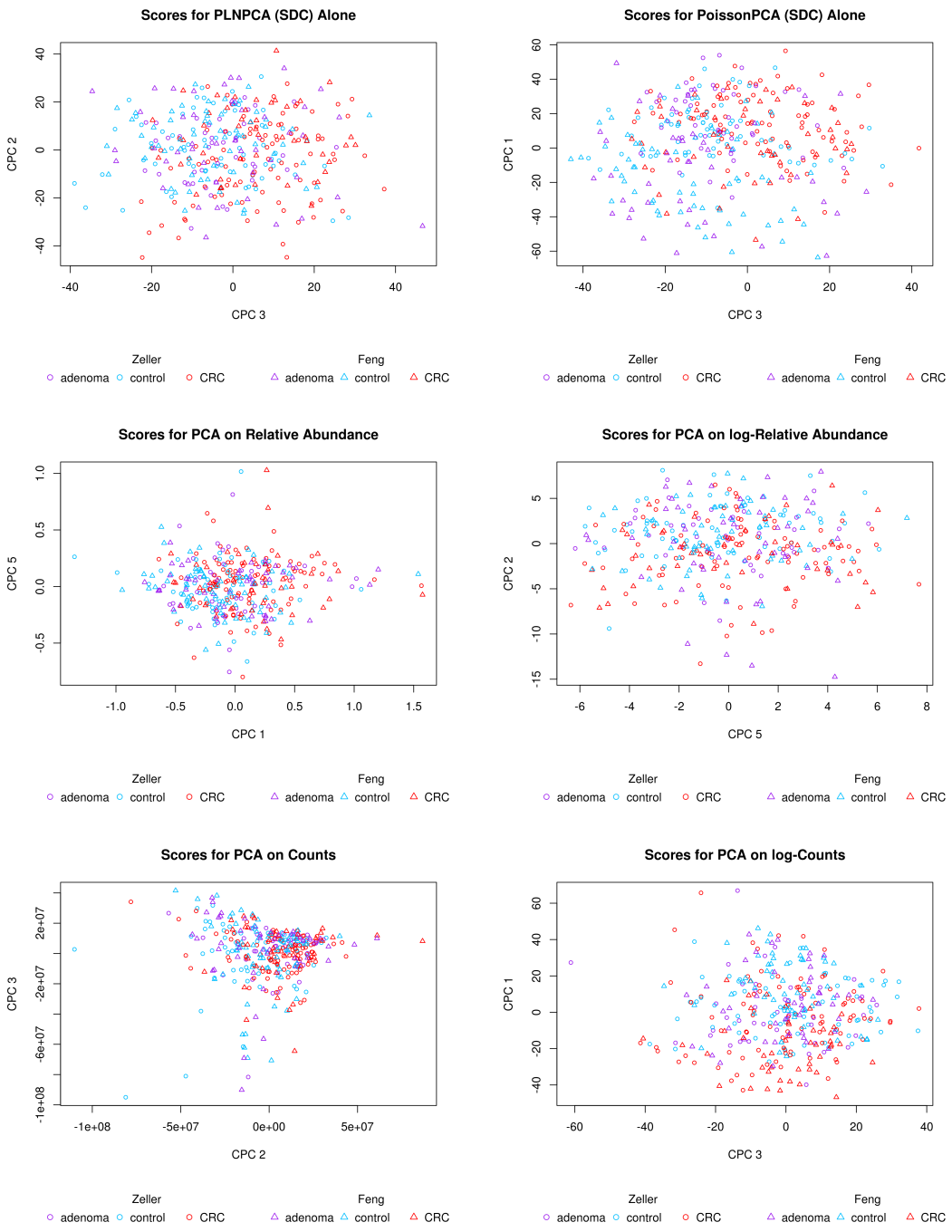

Figure 4.5: Scores from single-group and naive methods (with SDC for PoissonPCA/ PLNPCA) by disease state.

		First 5 CPCs		First 10 CPCs	
	Scores	**AUC**	**Acc.**	**AUC**	**Acc.**
No SDC	PoissonPCA + SCPCA	0.835	0.757	0.839	0.786
	PLNPCA + SCPCA	**0.877**	0.800	**0.885**	**0.814**
	PoissonPCA + FCPCA	0.826	0.771	0.832	0.786
	PLNPCA + FCPCA	**0.868**	0.786	**0.860**	0.800
	PoissonPCA + MSFA, $q = 4$	0.804	0.729	**0.860**	0.771
	PLNPCA + MSFA, $q = 3$	0.811	0.786	**0.858**	0.757
SDC	PoissonPCA + SCPCA	**0.860**	**0.814**	0.848	0.771
	PLNPCA + SCPCA	0.766	0.671	0.710	0.700
	PoissonPCA + FCPCA	**0.863**	**0.814**	0.849	0.800
	PLNPCA + FCPCA	0.780	0.771	0.779	0.743
	PoissonPCA + MSFA, $q = 4$	0.843	0.771	**0.860**	0.786
	PLNPCA + MSFA, $q = 5$	0.752	0.743	0.826	0.800
	PoissonPCA alone	0.818	0.771	**0.853**	0.771
	PLNPCA alone	0.744	0.629	0.756	0.714
Naive	PCA of Proportion	0.669	0.657	0.645	0.614
	PCA of Log-Proportion	0.677	0.671	0.738	0.743
	PCA of Count	0.637	0.643	0.628	0.614
	PCA of Log-Count	0.822	0.743	0.816	0.771
		AUC	**Acc.**		
	All Genus-Level Features	0.828	0.771		

Table 4.2: AUC and test set accuracy for classification of CRC samples.

criminating signal in the data, we trained random forest models on 80% of the full genus-level concatenated data (in which training and testing sets were built using stratified sampling by disease state) using five-times-repeated five-fold cross validation. We fit 5000 trees and selected the number of variables available for splitting at each node based on classification accuracy. We used the final model to predict disease state on the test set, and consider the performance of this nonlinear classifier to represent the extent to which the signal in these data can be used to discriminate CRC samples. We then trained logistic regressions on the first five or ten common scores (for all samples concatenated across the two studies) estimated by each method. The results are summarized in Table 4.2; area under the receiver operating curve (AUC) entries greater than 0.85 and test accuracy entries greater than 0.800 are in bold.

In general, the ensemble methods performed well with many obtaining high accuracy using only five CPCs, insinuating that the biological signal characterizing CRC sam-

ples across the two groups was captured in very few CPCs. In fact, the linear classifier using the scores as predictors often performs better than the nonlinear classifier using all genus-level taxa, suggesting that uninformative noise was reduced. Sequencing depth correction does not appear to increase predictive ability of scores estimated by CPCA, which may suggest that the sequencing depth noise is not complex enough to throw the classifier off the scent of the CRC-related signal, which provides further evidence that correcting estimates for Poisson error is the critical factor in these analyses.

PoissonPCA on the concatenated data performed well with no common factor estimation, as did PCA of the individually mean-centered log-count data. This is not entirely surprising given their performances in simulation. Also, Figure 4.2 suggest that the ability to discriminate disease state in Zeller samples may be the main determinant of classification performance, and so the success of PoissonPCA and log-count PCA on the concatenated data may be owed to some unique biological signal that is relevant to CRC status in the Zeller data but missing or undetectable in Feng, which would naturally give the single-group procedures an edge over the ensemble methods for prediction. In general, unique signal could be helpful or unhelpful for machine learning prediction of a given response, but in either case it could potentially obstruct meaningful interpretation of the predictors. The results in Table 4.2 support our hypothesis that the proposed ensemble methods are able to find a very low-dimensional representation of the data that retains virtually all discriminating biological signal that is shared among groups, which is our main interest.

To choose the optimal number of common principal components to describe the common signal on CRC status, we can choose q using scree plots of the estimated eigenvalues or by prediction such that the test set misclassification error is lowest. For example, using PoissonPCA followed by SCPCA, scree plots in Figure 4.6 suggest that the Feng data have a pronounced elbow, while the eigenvalues level off more slowly in the Zeller data. However, in both datasets, the last large difference between successive eigenvalues occurs by about $q = 11$. In Figure 4.7, which shows the cross-validation and test classification accuracies, it is apparent that after 6 CPCs the misclassification error tends to increase with additional CPCs and so we choose $q = 6$. The prediction results in Table 4.2 corroborate that in general, such parsimonious models perform well.

Finally, we turn to biological interpretation of the common loadings, wherein lies the strength of the ensemble method. Even assuming that single-group methods

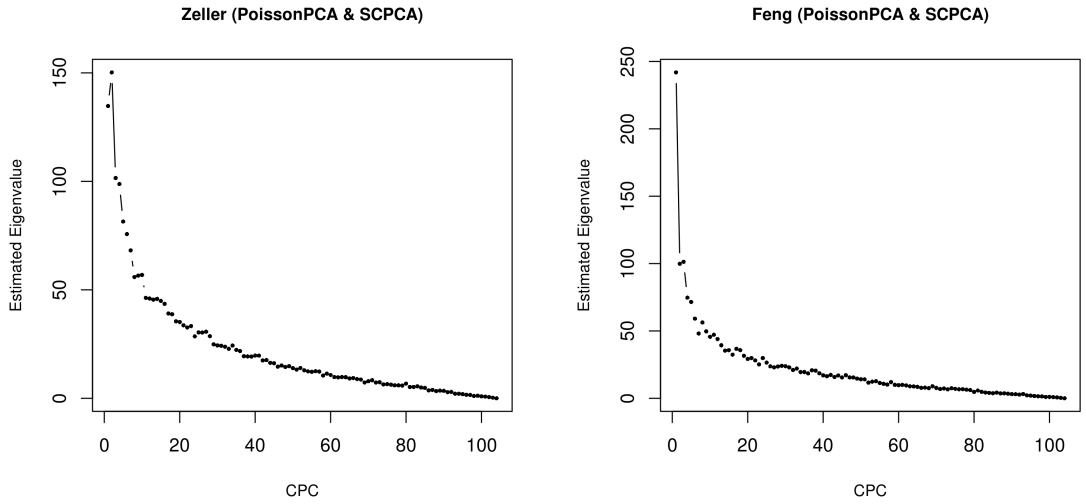

Figure 4.6: Scree plots of Zeller and Feng data.

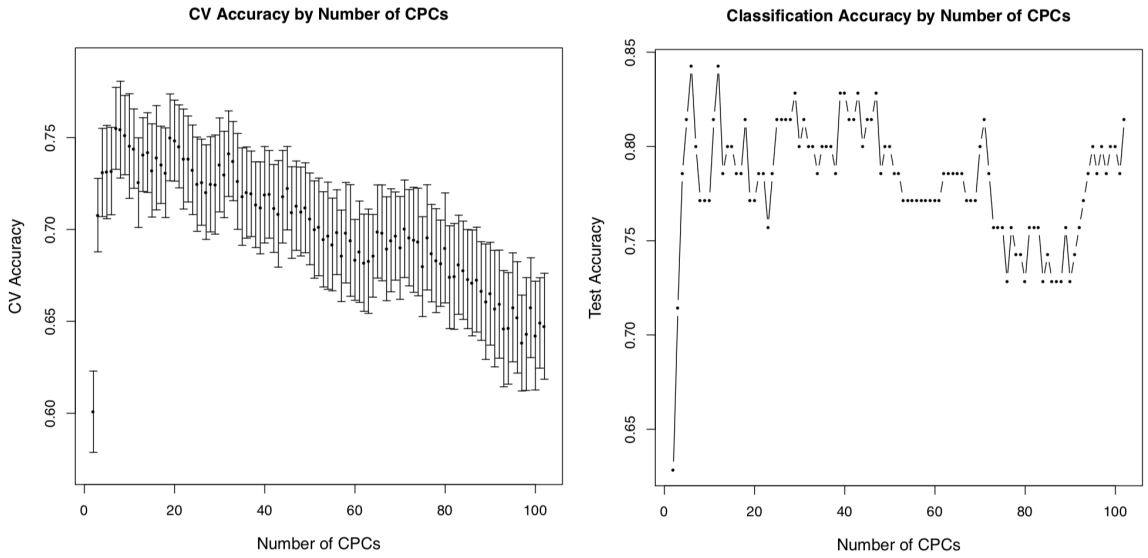

Figure 4.7: Cross-validation and test accuracy for Zeller and Feng.

were able to resolve some common underlying variation patterns, the resultant PCs reflecting this would be indistinguishable from those capturing within-study variation or contrasts between studies. However, we have been assured by our simulation results and assessment of the CPC scores that the ensemble methods can yield $q << p$ common axes that load on fully shared factors with strong signal. Furthermore, since our estimated loadings are related to the decomposition of variance on the log-scale, we can interpret the values given by these loadings as log-ratios of geometric means between groups of taxa with large magnitude values in opposite directions. The following was observed from the common loadings estimated from SCPCA on the PoissonPCA sequencing depth-corrected variance estimate.

Within our two CRC datasets, we found that on the two estimated common loading vectors that appeared to differentiate between samples with and without CRC, taxa with loading coefficient magnitudes in the 90^{th} or greater percentile were often described in other studies of CRC-associated microbiota. Peptostreptococcus, Butyrivibrio, Phascolarctobacterium, Fusobacterium, and Porphorymonas genera were identified by our common loadings as well as by the analyses of Zeller et al. (2014), while Acidaminococcus, Parvimonas, Gemella, and Peptostreptococcus were found in common with Feng et al. (2015). Genera we identified that have been previously associated with CRC include Fusobacterium (Purcell et al., 2017; Chen et al., 2012), Parvimonas (Feng et al., 2015), Porphorymonas (Purcell et al., 2017; Chen et al., 2012), Acidaminococcus (Feng et al., 2015; Azcarate-Peril et al., 2011), Phascolarctobacterium (Weir et al., 2013), Enterococcus (Kong et al., 2019), Gemella (Wong et al., 2017; Kong et al., 2019), Klebsiella (Chen et al., 2012), Prevotella (Purcell et al., 2017) and the Siphoviridae viruses (Handley & Devkota, 2019). Of those highlighted by our analyses, only Dialister (Weir et al., 2013), Butyrivbrio (Zeller et al., 2014) and Flavonifractor (Ai et al., 2019) have been reported to be protective. Solobacterium has been observed to be enriched in CRC samples in some studies (Wong et al., 2017) and decreased in others (Yang et al., 2019). Lastly, our loadings suggest that Adlercreutzia contributes to distinguishing CRC samples, but this genus has not yet been implicated in the literature, although it has been observed to participate in the metabolism of flavonoids (Li et al., 2018), which are antioxidants generally known to be anti-inflammatory and anti-tumorigenic. Moreover, on the loading in which Butyrivibrio (a genus of butyrate producers) and Adlercreutzia have large negative coefficients, all other large magnitude coefficients are positive. Flavonoids, along with dietary fibre, are linked to butyrate production, all of which have been linked to metabolic health and decreased risk of cancer (Jakobsdottir et al., 2013; Wu

et al., 2018; Ohkawara et al., 2006), although little is known about how flavonoids are metabolized. This loading vector may then represent a contrast between these genera—which may be involved in some protective pathway related to the metabolic intermediates of flavonoids—and the other genera dominating the loading. This is an insight that neither Zeller et al. (2014) nor Feng et al. (2015) were able to resolve from their single-group analyses.

Chapter 5

Conclusion

We have demonstrated that by appropriate modeling of microbial abundance count datasets so as to estimate their respective variances, followed by estimation of a basis for a common low-dimensional subspace for the true underlying abundances, we can remove unwanted variation while retaining shared variation patterns. Our ensemble method performs well in simulation and appears to have utility for real data analysis.

While we have demonstrated the method on data from two studies, in theory it can be applied to any number of datasets. PoissonPCA and SCPCA, the constituent methods showing the best performance, computation speed, and flexibility, can scale up to handle estimation of a common basis for many groups. Hence, our next step is to find a larger pool of similar datasets to determine whether strong common signal can still be resolved, since the relative amount of common signal will presumably drop with each additional dataset. We can also validate our findings by classification prediction or graphically by applying leave-one-out cross-validation across the groups, and comparing the projections of the estimated abundances for the held-out data and the rest of the data onto the subspace common to all except the held-out set. If we find that the ensemble method can indeed find common generalizable signal across many groups, this will lead us closer to impactful meta-analysis and the translation of research to application.

Similarly, we have shown the performance of our ensemble method on independent genomic samples from different experiments with very similar characteristics, but it remains to be investigated what range of common-signal-to-unique-noise ratios the method can handle. In this thesis we assumed that the variation of interest is shared

across samples, which is true at least in the sense that the common metadata collected in all samples reflects what is of interest to us in a given exploratory analysis. However, the common signal may be harder to differentiate from strong unique biological signatures than it is from technical noise, and in our analyses in Chapter 4 we avoided this issue by choosing datasets from Feng et al. (2015) and Zeller et al. (2014), two clinical studies with very similar designs. That is, certain types of samples may be presumed to have more constraints on their similarity than other types: two independent fecal samples from different individuals are more similar than would be, for instance, two independent water samples from different lakes, because we would expect that the conditions within the human intestinal lumen are more constrained across individuals (by physiology) than the conditions of lake water are constrained across lakes. Because of the abundance of biomedical and clinical data available from studies of human and animal gut microbiota which tend to investigate similar themes, our method already has a solid basis of applicability based on our results. The next step in our research is to investigate whether our proposed method can resolve the common axes among different microbial abundance data even when there is presumably large unique biological variation among groups, in addition to unique technical noise and common biological variation. If so, the scope of our method could be extended to exploring commonalities in community structure that are conserved across a diverse range of communities.

In other future work, we may develop a multi-group extension of PLNPCA, which could make direct use of the dimension reduction capabilities of modeling counts as conditional on linear combinations of common and unique factors lying in separate low-dimensional subspaces, potentially removing the need for an ensemble method. However, we would have to consider carefully how best to fit this model, as PLNPCA and especially MSFA are already slow to optimize parameter estimates. We would also need to consider whether the restrictive parameterization that these models share might be inappropriate in general for exploratory analyses of metagenomic and 16S data, and that in fact an ensemble method using semi-parametric approaches is preferable, as was found in our analyses. Yet another option would be to build an equivalent hierarchical Bayesian factor analysis model, although MCMC run times scale poorly with dimension compared to likelihood methods. Even so, given the difficulty that MSFA's ECM algorithm had in achieving convergence in our simulations and data analysis, it is possible that a Bayesian approach would be easier to implement.

In conclusion, the study of microbiota poses a number of challenges, and signal-

obstructing noise is heterogeneous and will forever evade capture. We can only do our best to treat these data in the most appropriate way possible for a given aim. In this thesis, we addressed the lack of cross-study generalization in microbial abundance data, and proposed a framework to "remove" some of the observed within-study variation by seeking the latent structure that provides a scaffold for the common variance.

References

Ai, D., Pan, H., Li, X., Gao, Y., Liu, G., & Xia, L. C. (2019). Identifying gut microbiota associated with colorectal cancer using a zero-inflated lognormal model. Frontiers in Microbiology, 10, 826.

Aitchison, J. & Ho, C. H. (1989). The multivariate Poisson log-normal distribution. Biometrika, 76(4), 643–653.

Blaser, M., Bork, P., Fraser, C., Knight, R. and Wang, J. (2013). The microbiome explored: recent insights and future challenges. Nature Reviews Microbiology 11, 213-217. doi:10.1038/nrmicro2973

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518), 859-877.

Boon, E., Meehan, C. J., Whidden, C., Wong, D. H.-J., Langille, M. G. I. and Beiko, R. G. (2014). Interactions in the microbiome: communities of organisms and communities of genes. FEMS Microbiology Review, 38, 90-118. doi:10.1111/1574-6976.12035

Buhule, O. D., Minster, R. L., Hawley, N. L., Medvedovic, M., Sun, G., Viali, S., ... Weeks, D. E. (2014). Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale. Frontiers in Genetics, 5, 354. doi:10.3389/fgene.2014.00354

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME Journal, 11(12), 2639.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS ONE, 6(2).

Chen, W., Liu, F., Ling, Z., Tong, X., & Xiang, C. (2012). Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. PLoS ONE, 7(6).

Chiquet, J., Mariadassou, M., & Robin, S. (2018). Variational inference for probabilistic Poisson PCA. The Annals of Applied Statistics, 12(4), 2674-2698.

De Vito, R., Bellio, R., Trippa, L., & Parmigiani, G. (2019). Multiâstudy factor analysis. Biometrics, 75(1), 337-346.

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., ... & Su, L. (2015). Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nature Communications, 6, 6528.

Flury, B. N. (1984). Common principal components in k groups. Journal of the American Statistical Association, 79(388), 892-898.

Flury, B. N., & Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. SIAM Journal on Scientific and Statistical Computing, 7(1), 169-184.

Gagnon-Bartsch, J. A., Jacob, L., & Speed, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. Berkeley: Technical Reports from the Department of Statistics at the University of California, 1-112.

Gibbons, S. M., Duvallet, C., & Alm, E. J. (2018). Correcting for batch effects in case-control microbiome studies. PLoS Computational Biology, 14(4), e1006102.

Handley, S. A. & Devkota, S. (2019). Going viral: A novel role for bacteriophage in colorectal Cancer. mBio, 10(1), e02626-18. doi:10.1128/mBio.02626-18

Jakobsdottir, G., Blanco, N., Xu, J., Ahrne, S., Molin, G., Sterner, O., & Nyman, M. (2013). Formation of short-chain fatty acids, excretion of anthocyanins, and microbial diversity in rats fed blackcurrants, blackberries, and raspberries. Journal of Nutrition and Metabolism, 2013. doi:10.1155/2013/202534

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1), 118-127.

Kenney, T., Huang, T. & Gu, H. (2019). Poisson measurement error corrected PCA with application to microbiome data. arXiv:1904.11745v1 [stat.ME]

Kong, F., & Cai, Y. (2019). Study insights into gastrointestinal cancer through the gut microbiota. BioMed Research International, 2019. doi:10.1155/2019/8721503

Kurtz, Z. D., Muller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J. and Bonneau, R. A.(2015). Sparse and compositionally robust inference of microbial ecological networks. PLoS Computational Biology, 11(5). doi:10.1371/journal.pcbi.1004226

Langille, M. G. (2018). Exploring linkages between taxonomic and functional profiles of the human microbiome. MSystems, 3(2), e00163-17.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11(10), 733-739.

Li, Y., Zhang, T., & Chen, G. Y. (2018). Flavonoids and colorectal cancer prevention. Antioxidants, 7(12), 187.

McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Computational Biology, 10(4).

Miller, R. R., Uyaguari-Diaz, M., McCabe, M. N., Montoya, V., Gardy, J. L., Parker, S., ... Group, T. C. (2016). Metagenomic investigation of plasma in individuals with ME/CFS highlights the importance of technical controls to elucidate contamination and batch effects. PLoS One, 11(11). doi:10.1371/journal.pone.0165691

Nieuwdorp, M. Gilijamse, P. W., Pai, N. and Kaplan, L. M. (2014). Role of the microbiome in energy regulation and metabolism. Gastroenterology, 146(6), 1525-1533. doi:10.1053/j.gastro.2014.02.008

Ohkawara, S., Furuya, H., Nagashima, K., Asanuma, N., & Hino, T. (2006). Effect of oral administration of Butyrivibrio fibrisolvens MDT-1 on experimental enterocolitis in mice. Clinical and Vaccine Immunology, 13(11), 1231-1236.

Olesen, S. W., Vora, S., Techtmann, S. M., Fortney, J. L., Bastidas-Oyanedel, J. R., Rodriguez, J., ..., Alm, E. J. (2016). A novel analysis method for paired-sample microbial ecology experiments. PLoS ONE, 11(5). doi:10.1371/journal.pone.0154804

Oytam, Y., Sobhanmanesh, F., Duesing, K., Bowden, J. C., Osmond-McLeod, M., & Ross, J. (2016). Risk-conscious correction of batch effects: Maximising information extraction from high-throughput genomic datasets. BMC Bioinformatics, 17(1), 332. doi:10.1186/s12859-016-1212-5

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., ... & Huttenhower, C. (2017). Accessible, curated metagenomic data through ExperimentHub. Nature Methods, 14(11), 1023.

Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S., & Frizelle, F. A. (2017). Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. Nature Scientific Reports, 7(1), 1-12.

Sill, M., Saadati, M. and Benner, A. (2015). Applying stability selection to consistently estimate sparse principal components in high-dimensional molecular data. Bioinformatics, 31(16), 2683-2690. doi:10.1093/bioinformatics/btv197

Sze, M. A., Schloss, P. D. (2016). Looking for a signal in the noise: revisiting obesity and the microbiome. MBio, 7(4), e01018-16.

Thaiss, C. A., Zmora, N., Levy, M. and Elinav, E. (2016). The microbiome and innate immunity. Nature, 535(7610), 65–74. doi:10.1038/nature18847

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3), 611-622.

Tralau, T., Sowada, J. and Luch, A. (2015). Insights on the human microbiome and its xenobiotic metabolism: what is known about its effects on human physiology? Expert Opinion on Drug Metabolism & Toxicology, 11(3), 411-425. doi:10.1517/17425255.2015.990437

Trendafilov, N. T. (2010). Stepwise estimation of common principal components. Computational Statistics & Data Analysis, 54(12), 3446-3457.

Tsilimigras, M. C. B. and Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Annals of Epidemiology, 26(5), 330â335. doi: 10.1016/j.annepidem.2016.03.002

Weir, T. L., Manter, D. K., Sheflin, A. M., Barnett, B. A., Heuberger, A. L., & Ryan, E. P. (2013). Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. PLoS ONE, 8(8).

Wu, X., Wu, Y., He, L., Wu, L., Wang, X., Liu, Z. (2018). Effects of the intestinal microbial metabolite butyrate on the development of colorectal cancer. Journal of Cancer, 9(14), 2510-2517.

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., ... & Hercog, R. (2014). Potential of fecal microbiota for earlyâstage detection of colorectal cancer. Molecular Systems Biology, 10(11).

Appendix A

Supplementary Results

A.1 Additional Simulation Results

A.2 Additional Score Plots

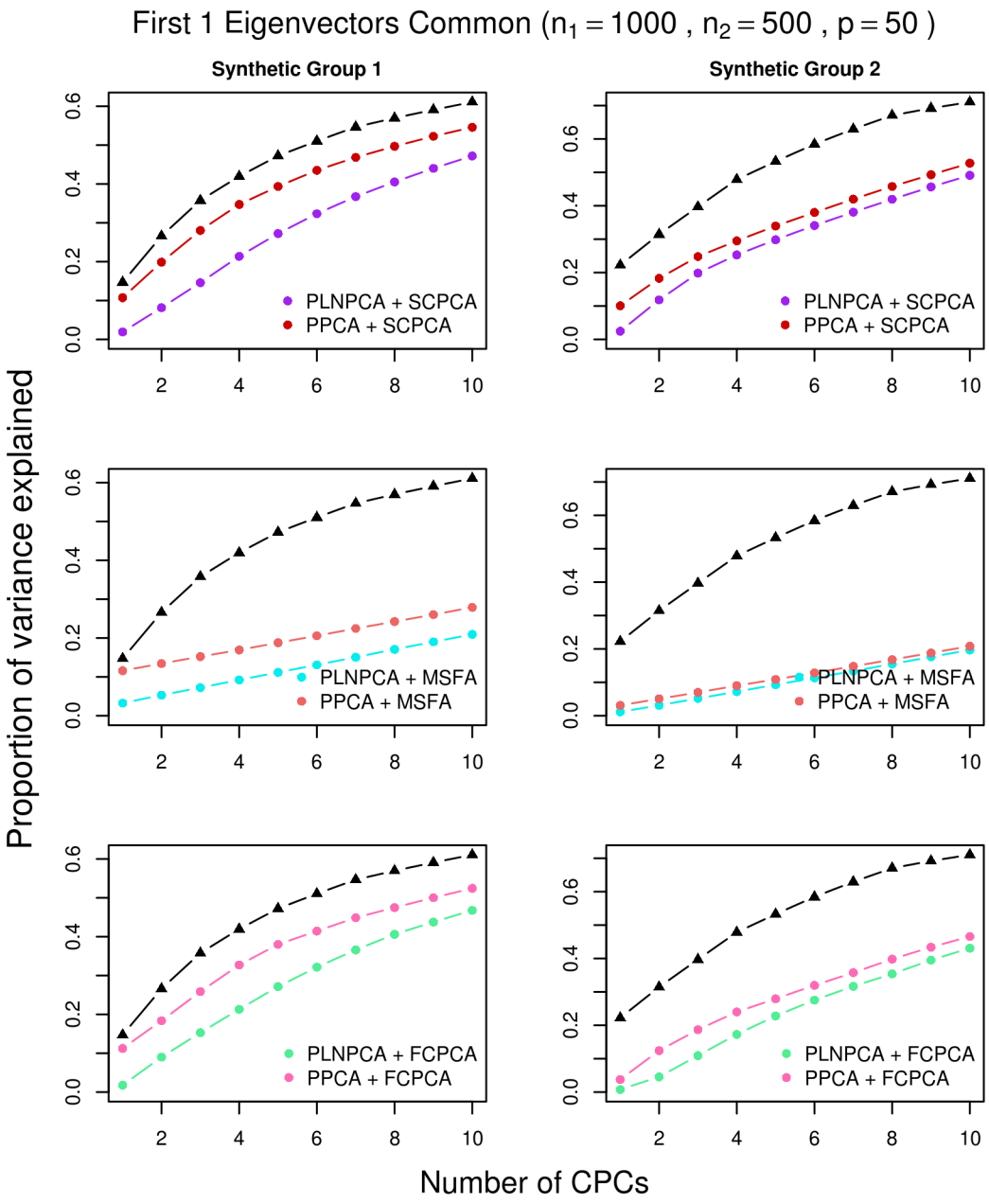

Figure A.1: Simulation results for decreasing eigenvalues and one common eigenvector, with SDC; p=50, $n_1 = 1000$, $n_2 = 500$.

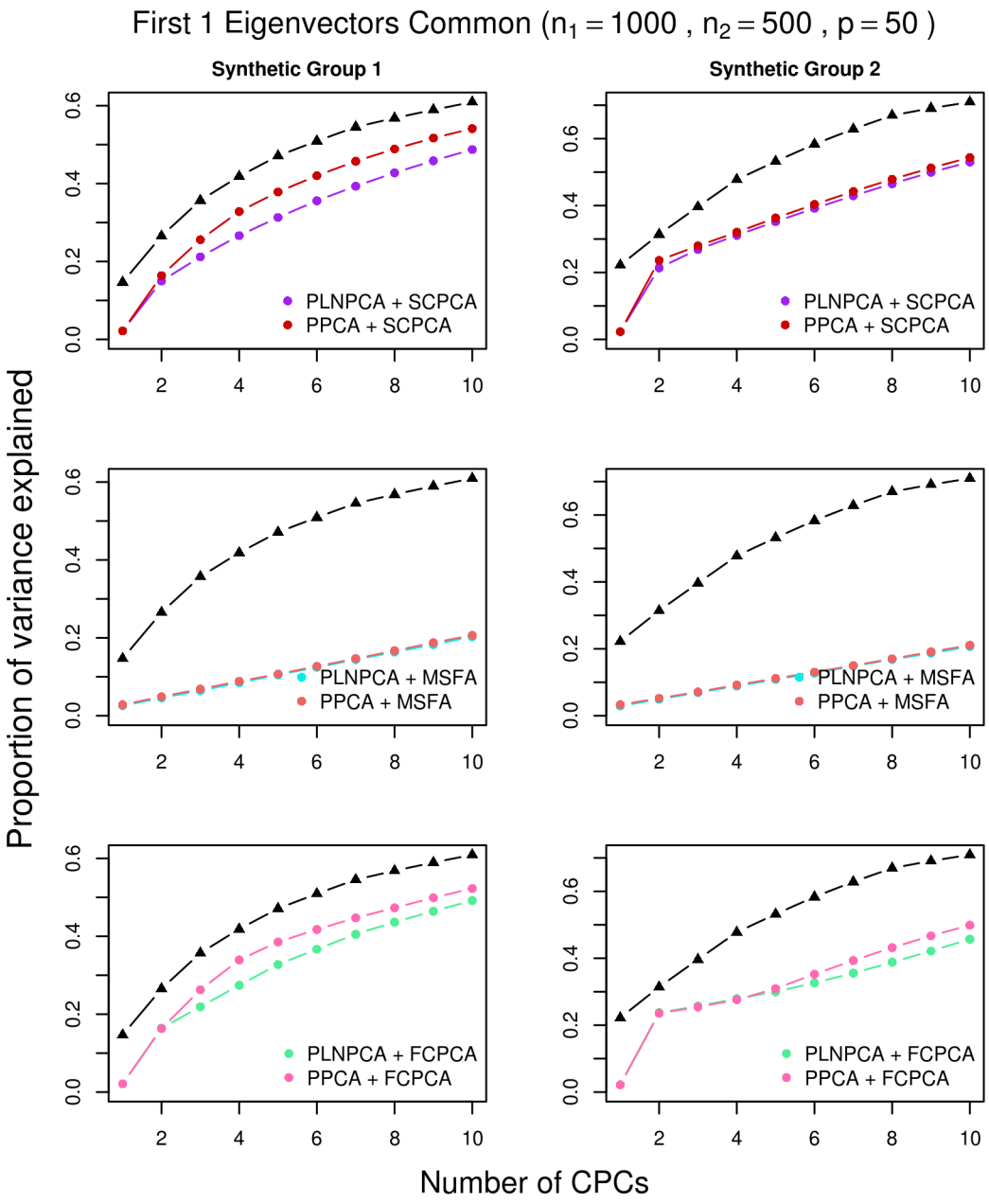

Figure A.2: Simulation results for decreasing eigenvalues and one common eigenvector, with no SDC; p=50, $n_1 = 1000$, $n_2 = 500$.

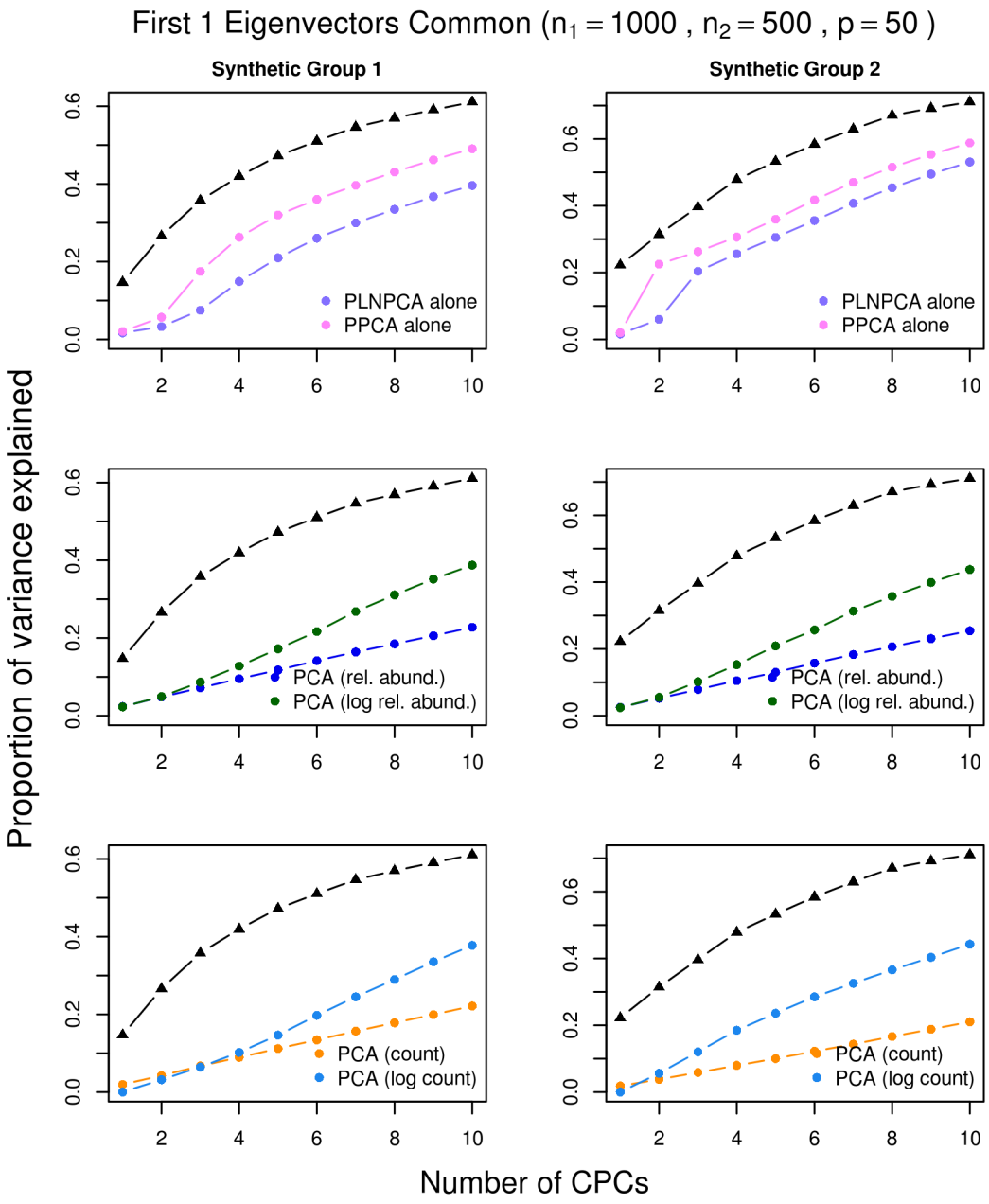

Figure A.3: Simulation results for decreasing eigenvalues and one common eigenvector, with SDC; p=50, $n_1 = 1000$, $n_2 = 500$.

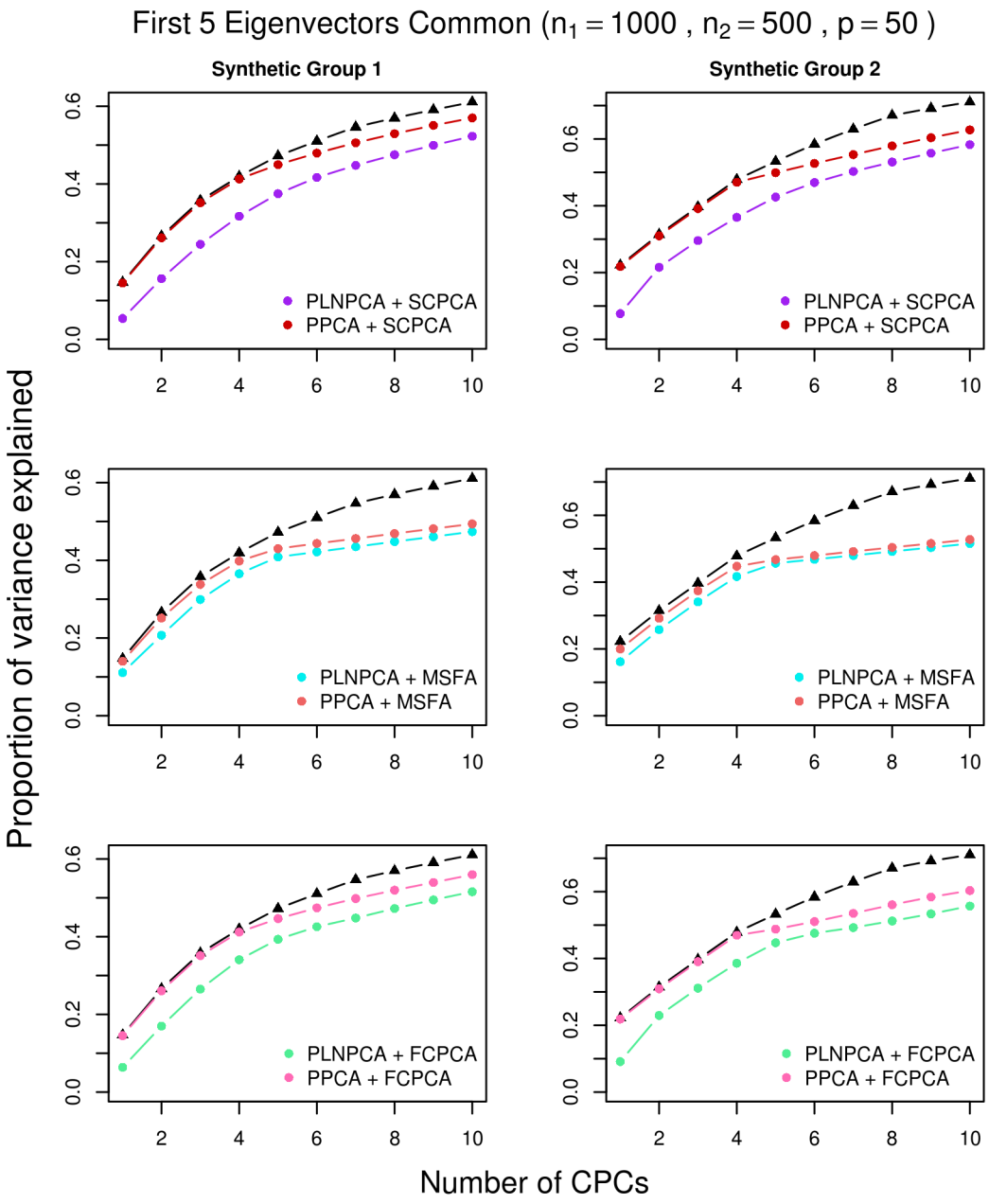

Figure A.4: Simulation results for decreasing eigenvalues and five common eigenvectors, with SDC; p=50, $n_1 = 1000$, $n_2 = 500$.

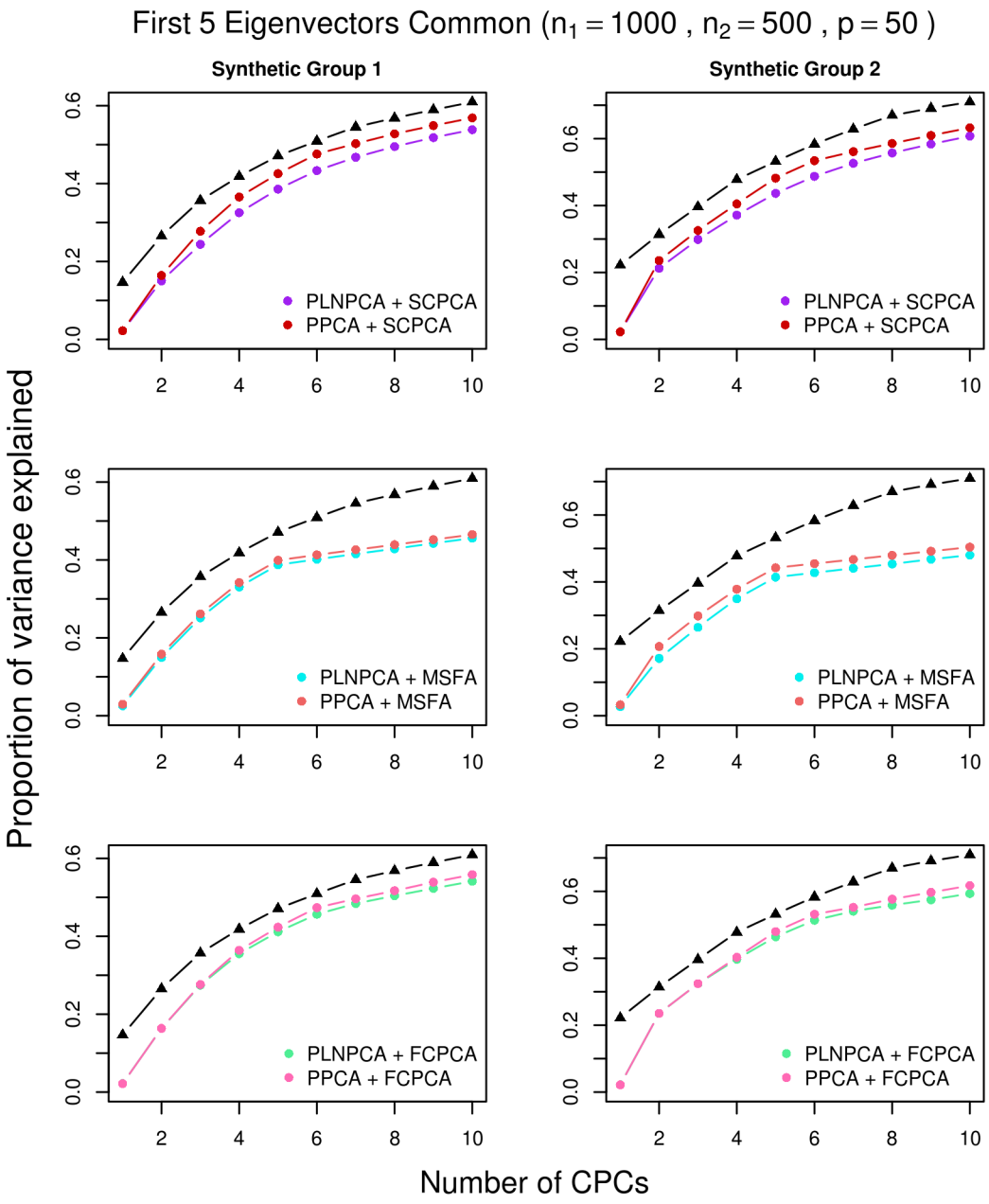

Figure A.5: Simulation results for decreasing eigenvalues and five common eigenvectors, with no SDC; p=50, $n_1 = 1000$, $n_2 = 500$.

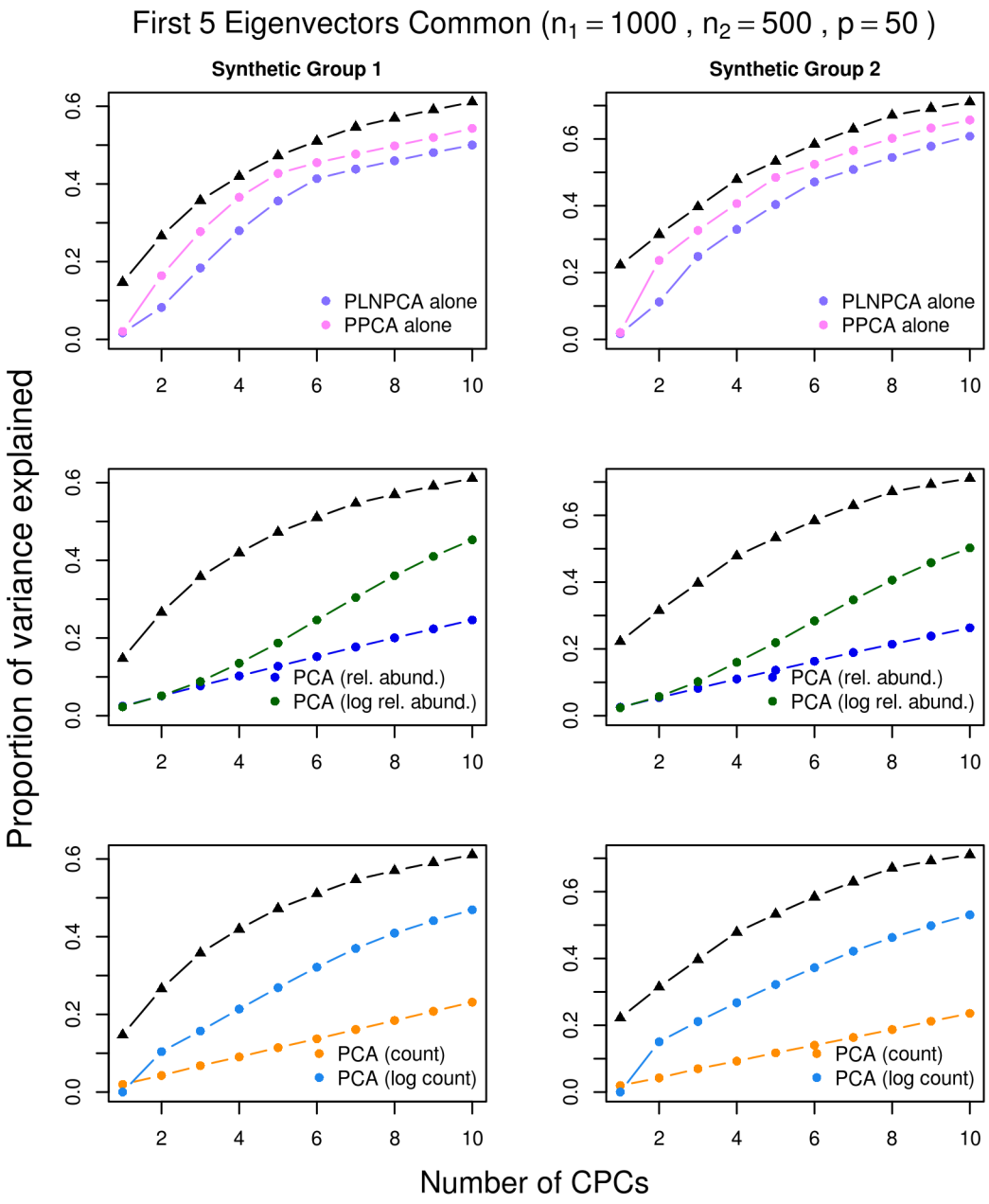

Figure A.6: Simulation results for decreasing eigenvalues and five common eigenvectors, with SDC; p=50, $n_1 = 1000$, $n_2 = 500$.

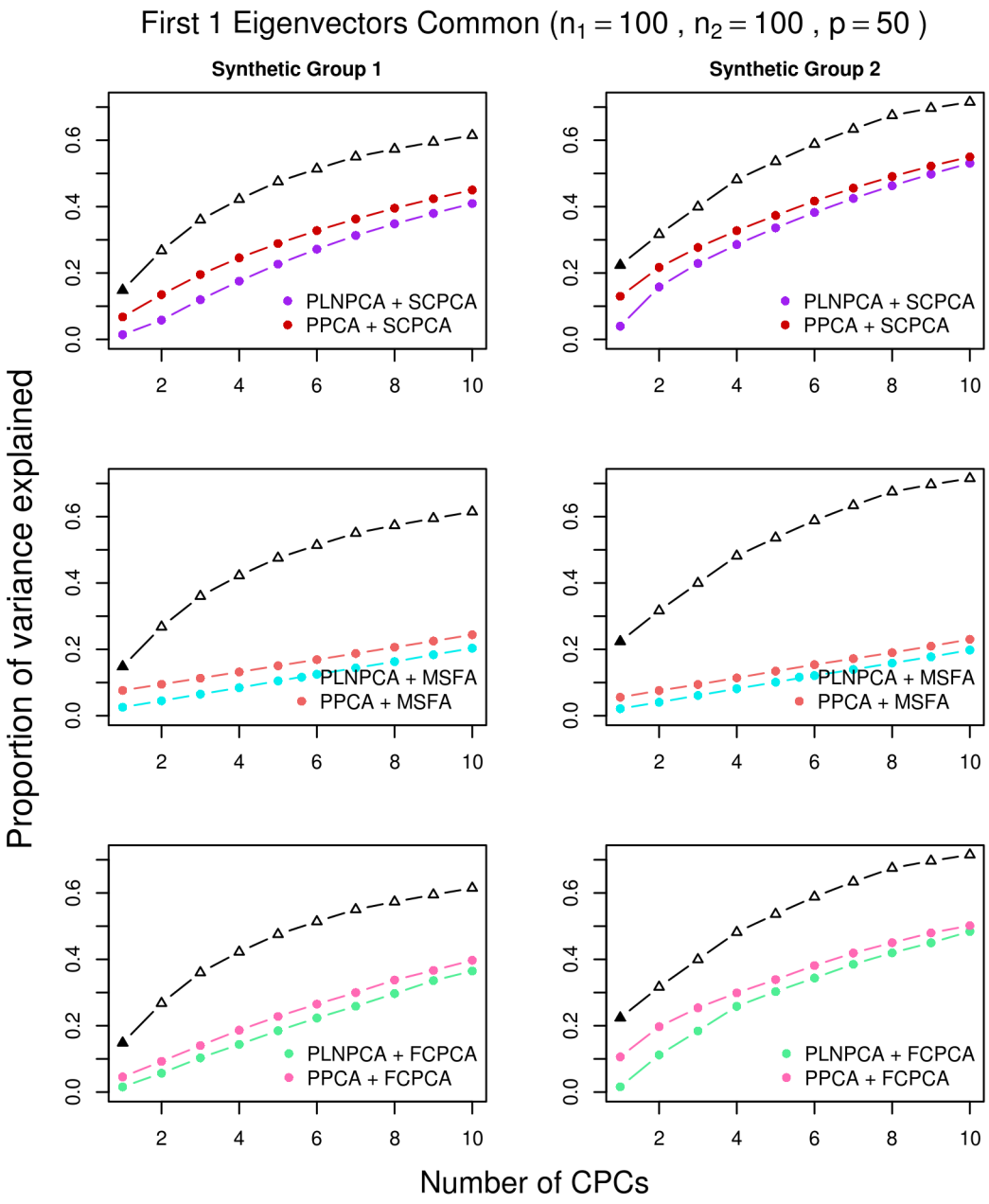

Figure A.7: Simulation results for non-decreasing eigenvalues and one common eigenvector, with SDC; p=50, $n_1 = 100$, $n_2 = 100$.

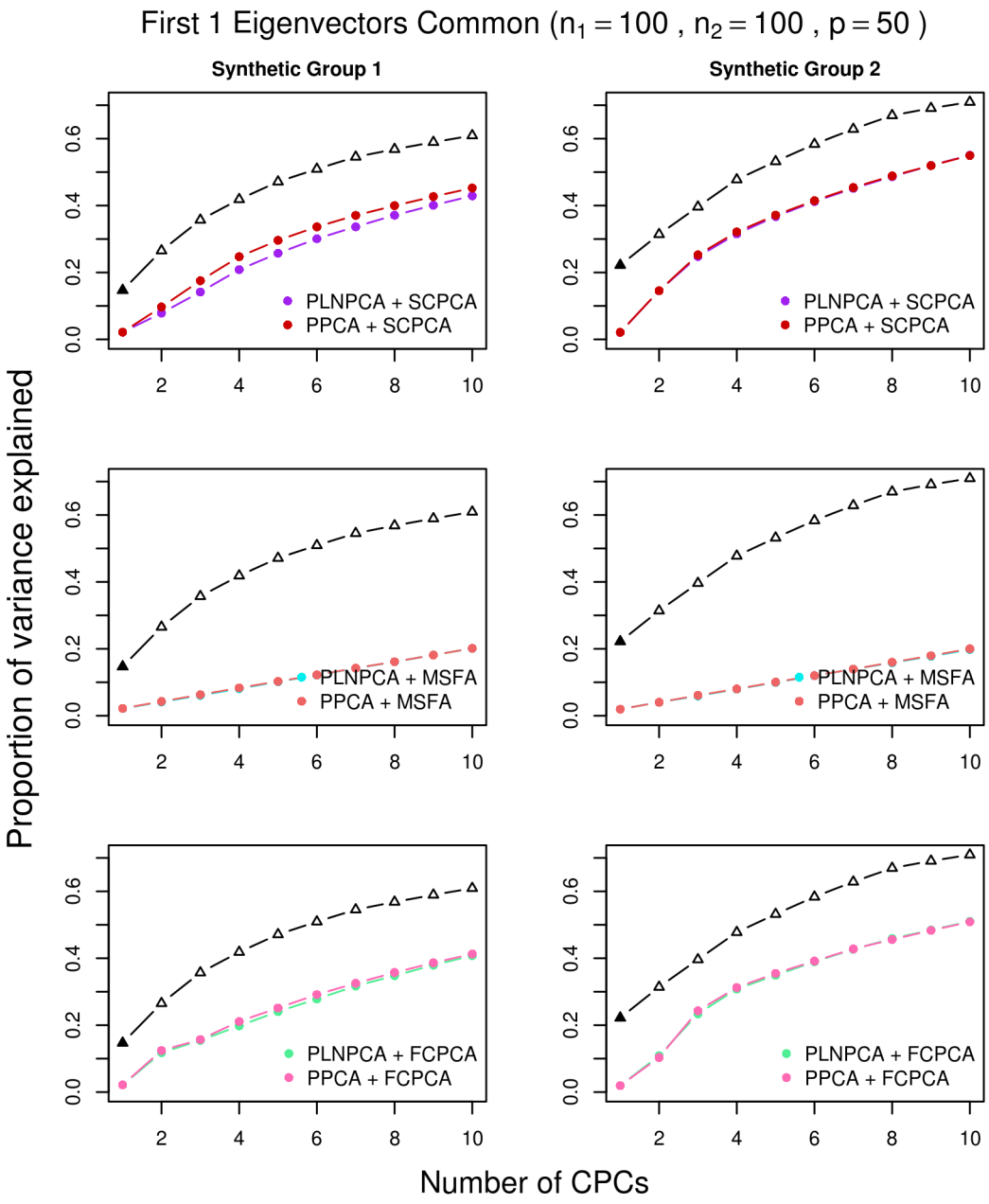

Figure A.8: Simulation results for non-decreasing eigenvalues and one common eigenvector, without SDC; p=50, $n_1 = 100$, $n_2 = 100$.

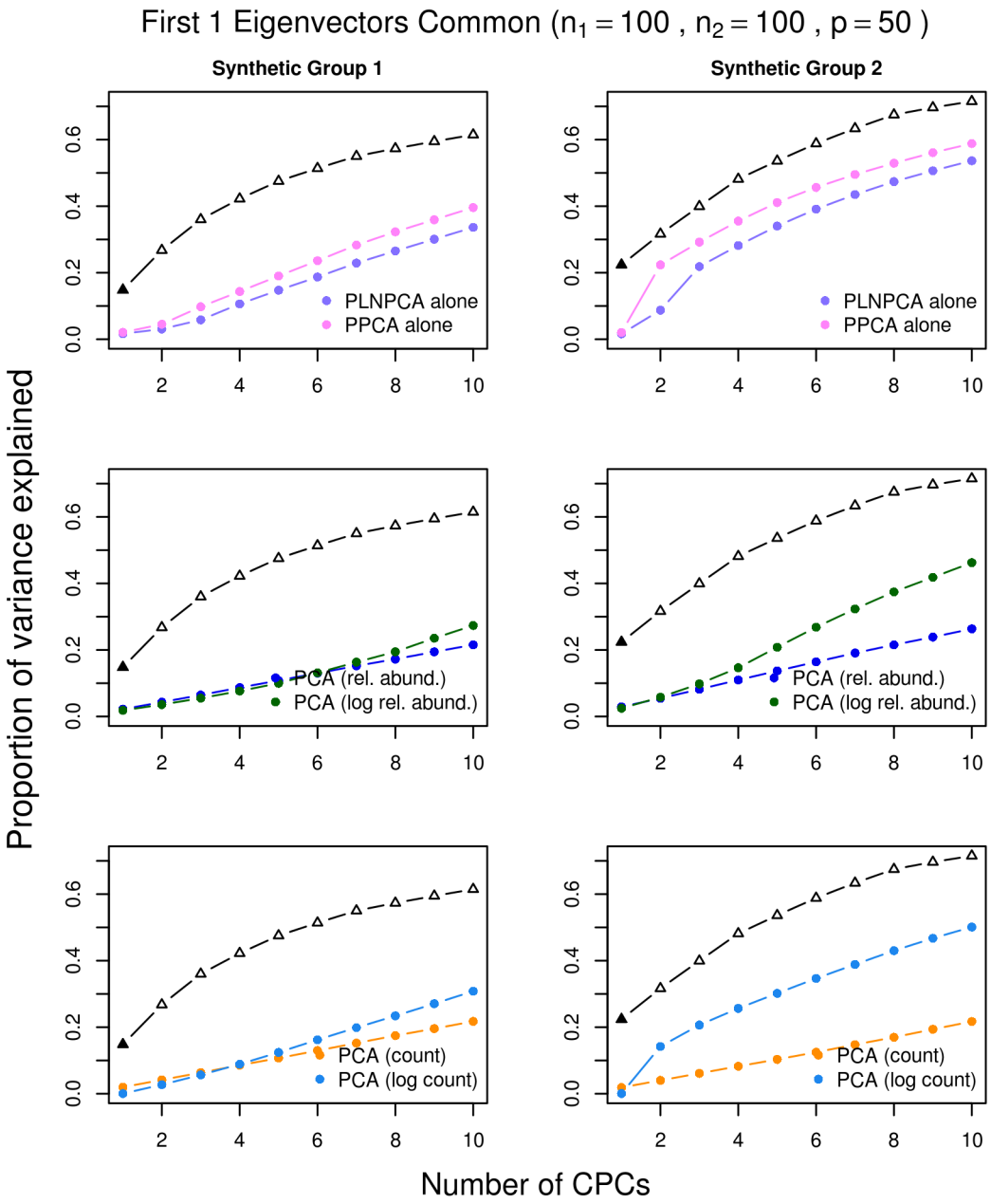

Figure A.9: Simulation results single-group methods for non-decreasing eigenvalues and one common eigenvector, with SDC for PoissonPCA/PLNPCA; p=50, $n_1 = 100$, $n_2 = 100$.

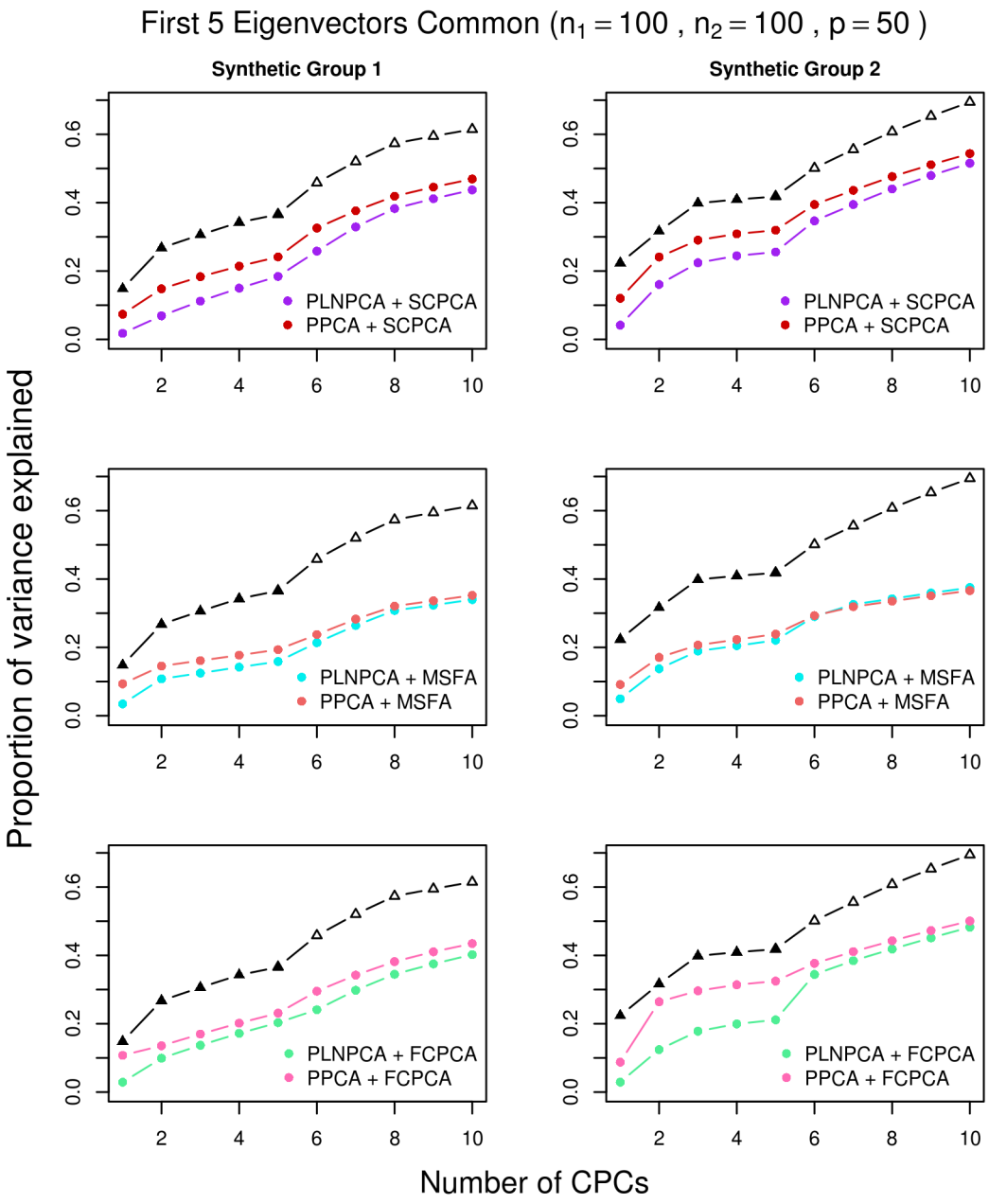

Figure A.10: Simulation results for non-decreasing eigenvalues and five common eigenvectors, with SDC; p=50, $n_1 = 100$, $n_2 = 100$.

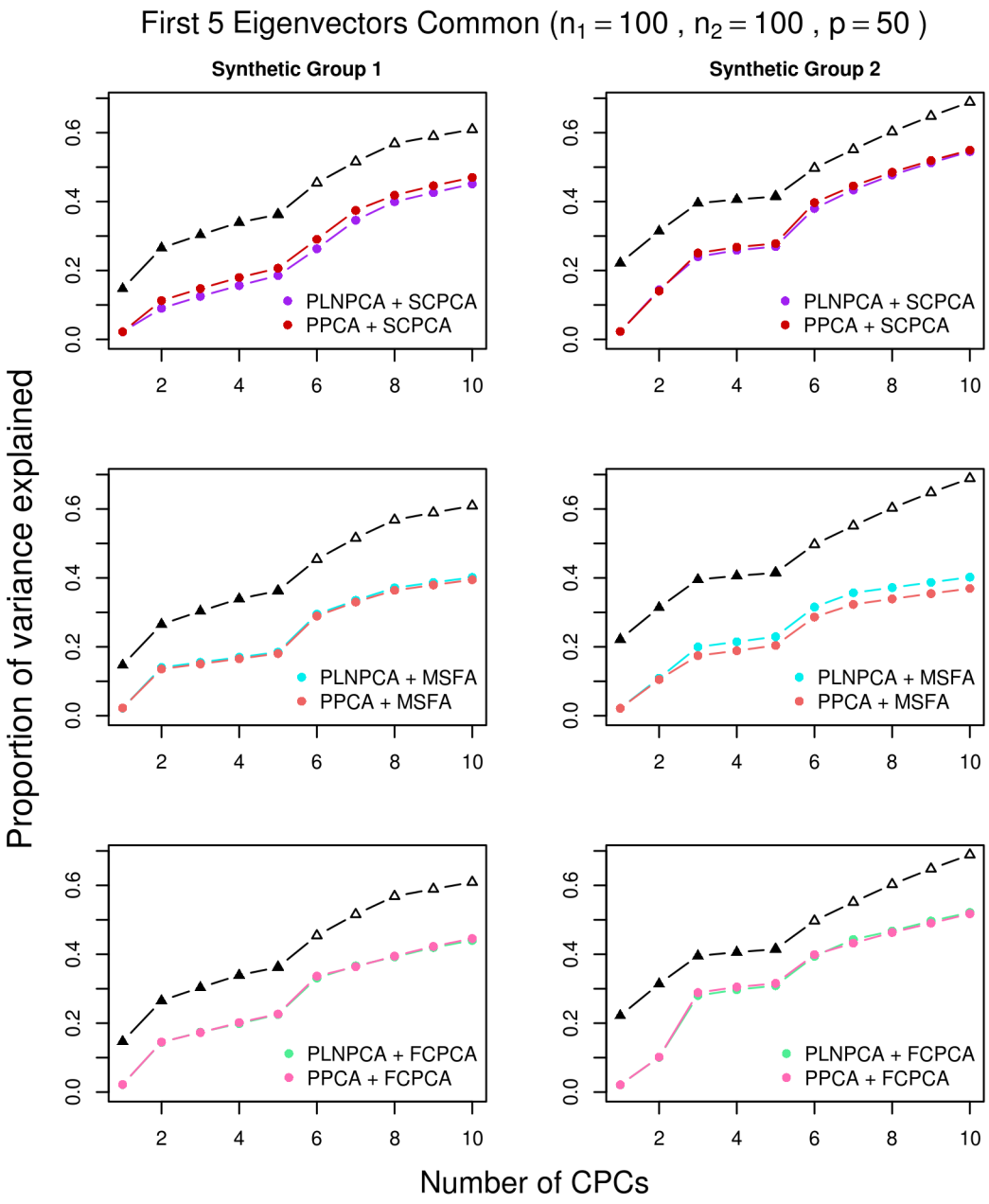

Figure A.11: Simulation results for non-decreasing eigenvalues and five common eigenvectors, without SDC; p=50, $n_1 = 100$, $n_2 = 100$.

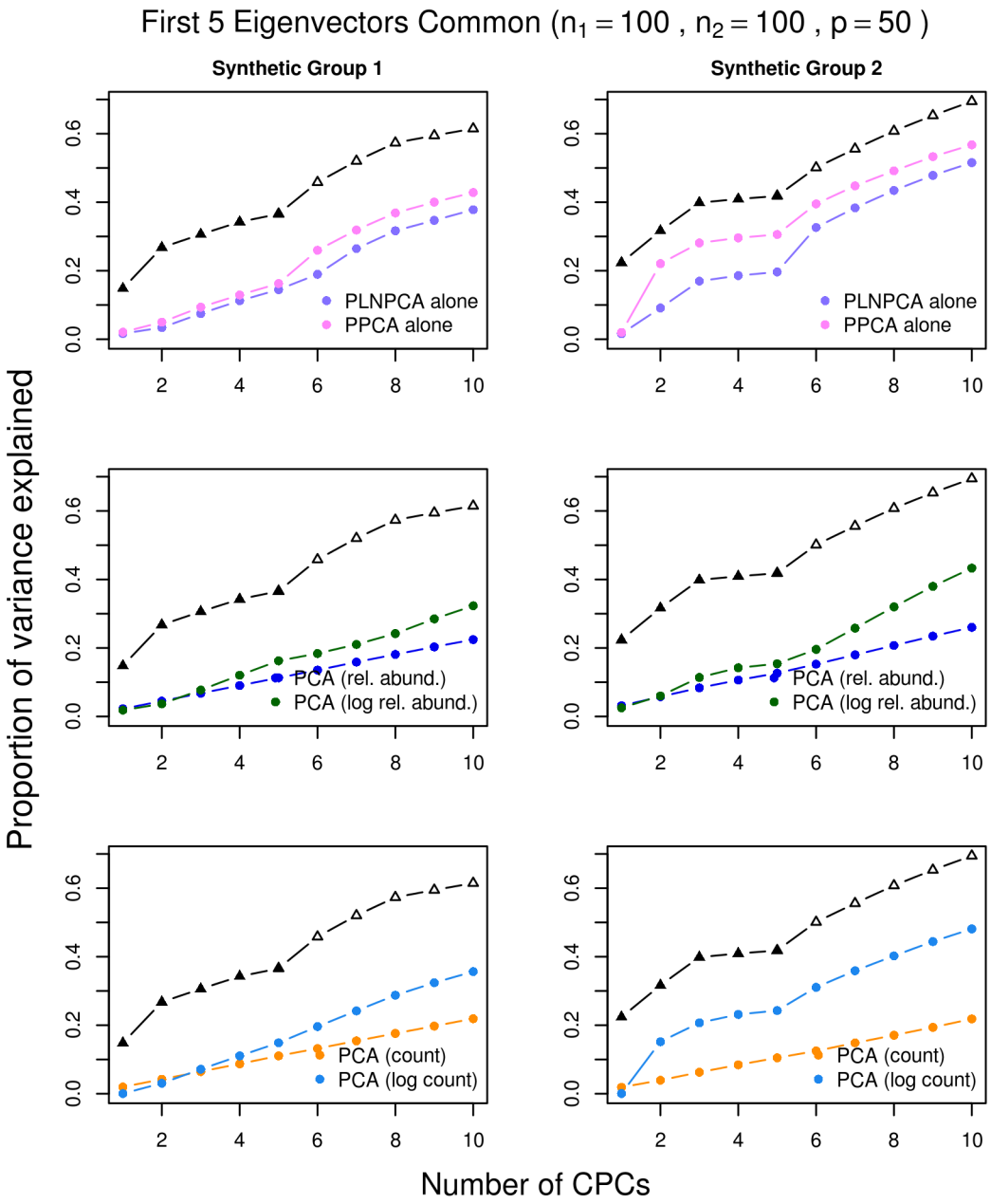

Figure A.12: Simulation results single-group methods for non-decreasing eigenvalues and five common eigenvectors, with SDC for PoissonPCA/PLNPCA; p=50, $n_1 = 100$, $n_2 = 100$.

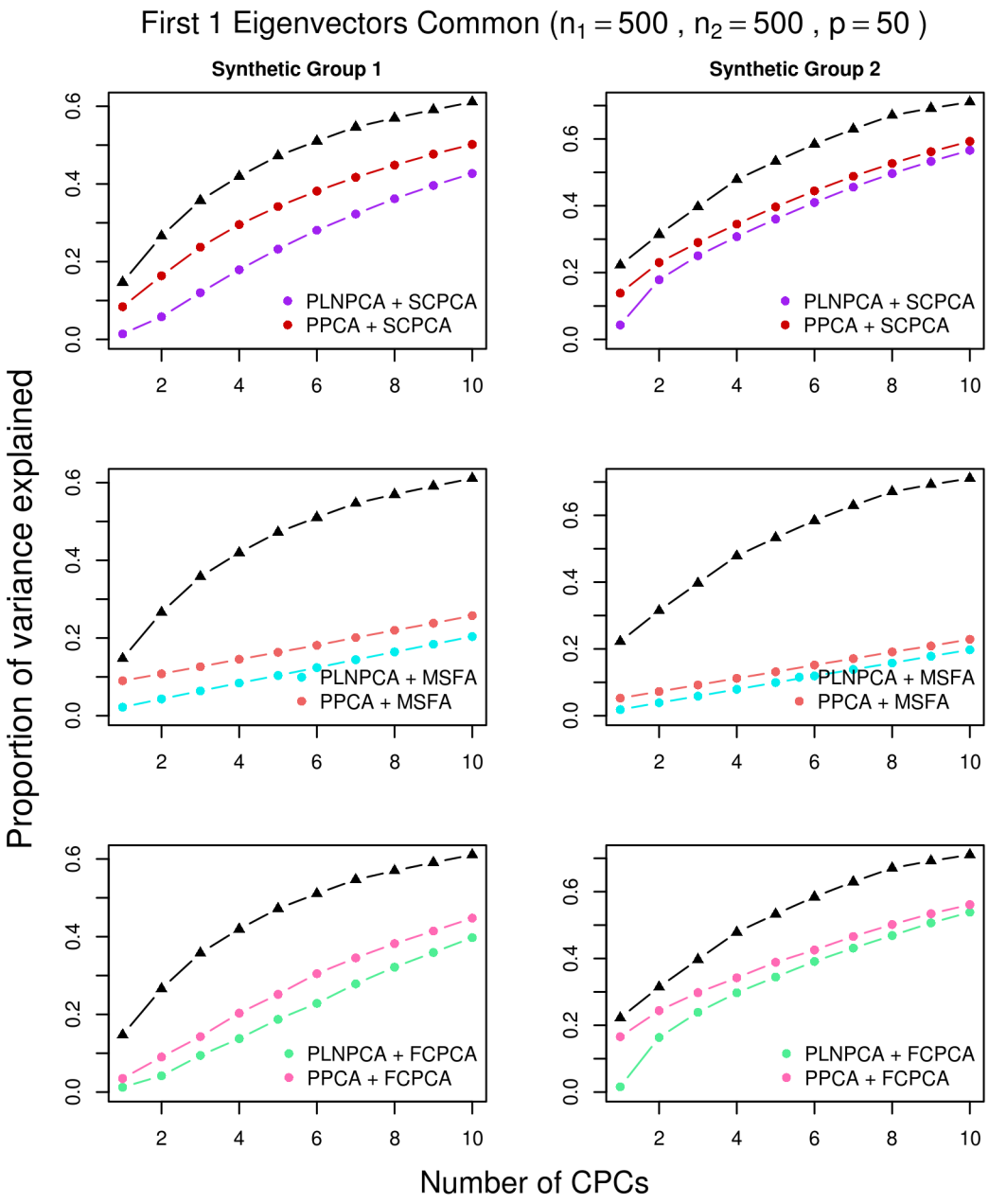

Figure A.13: Simulation results for decreasing eigenvalues and one common eigenvector, with SDC; p=50, $n_1 = 500$, $n_2 = 500$.

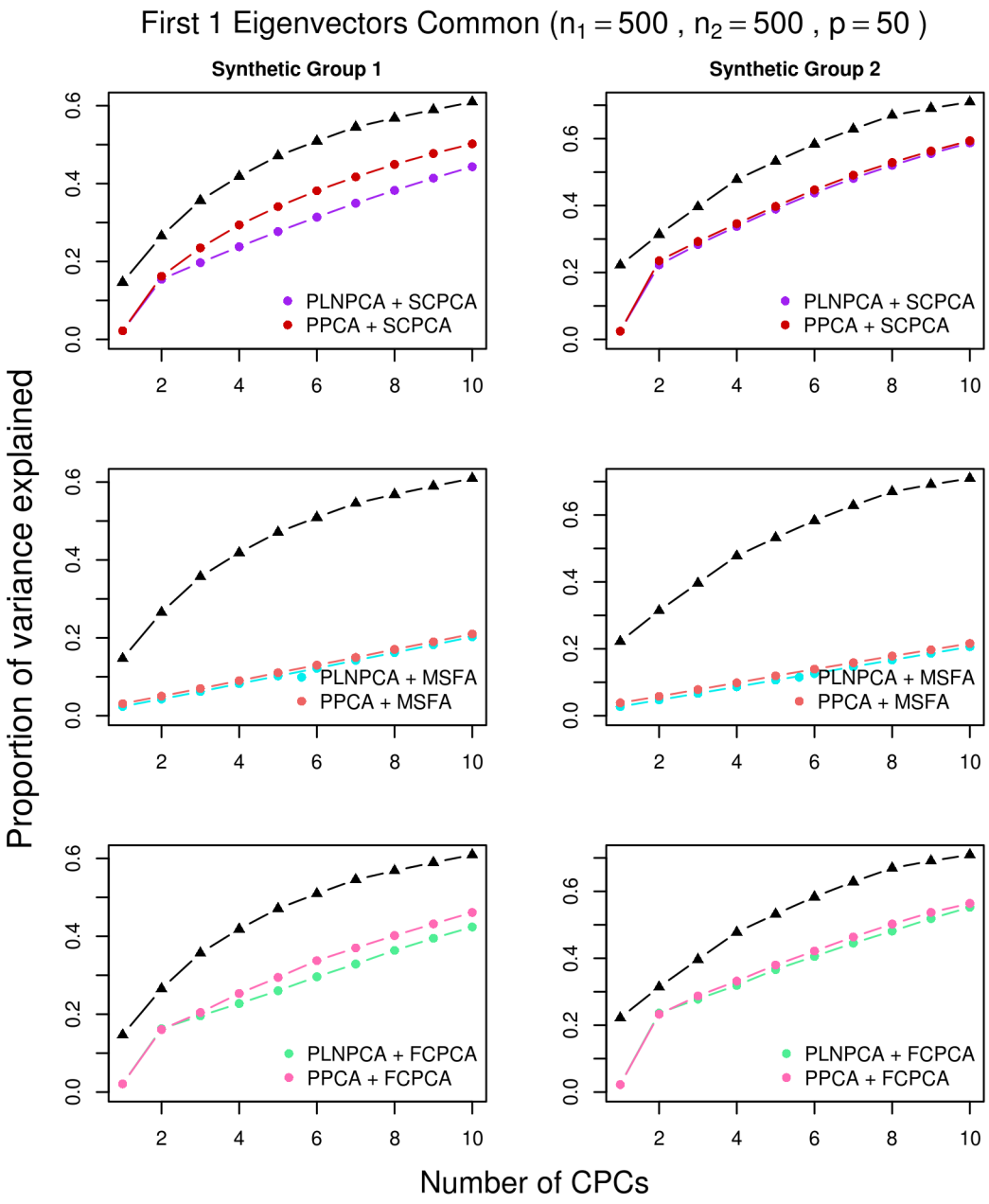

Figure A.14: Simulation results for decreasing eigenvalues and one common eigenvector, with no SDC; p=50, $n_1 = 500$, $n_2 = 500$.

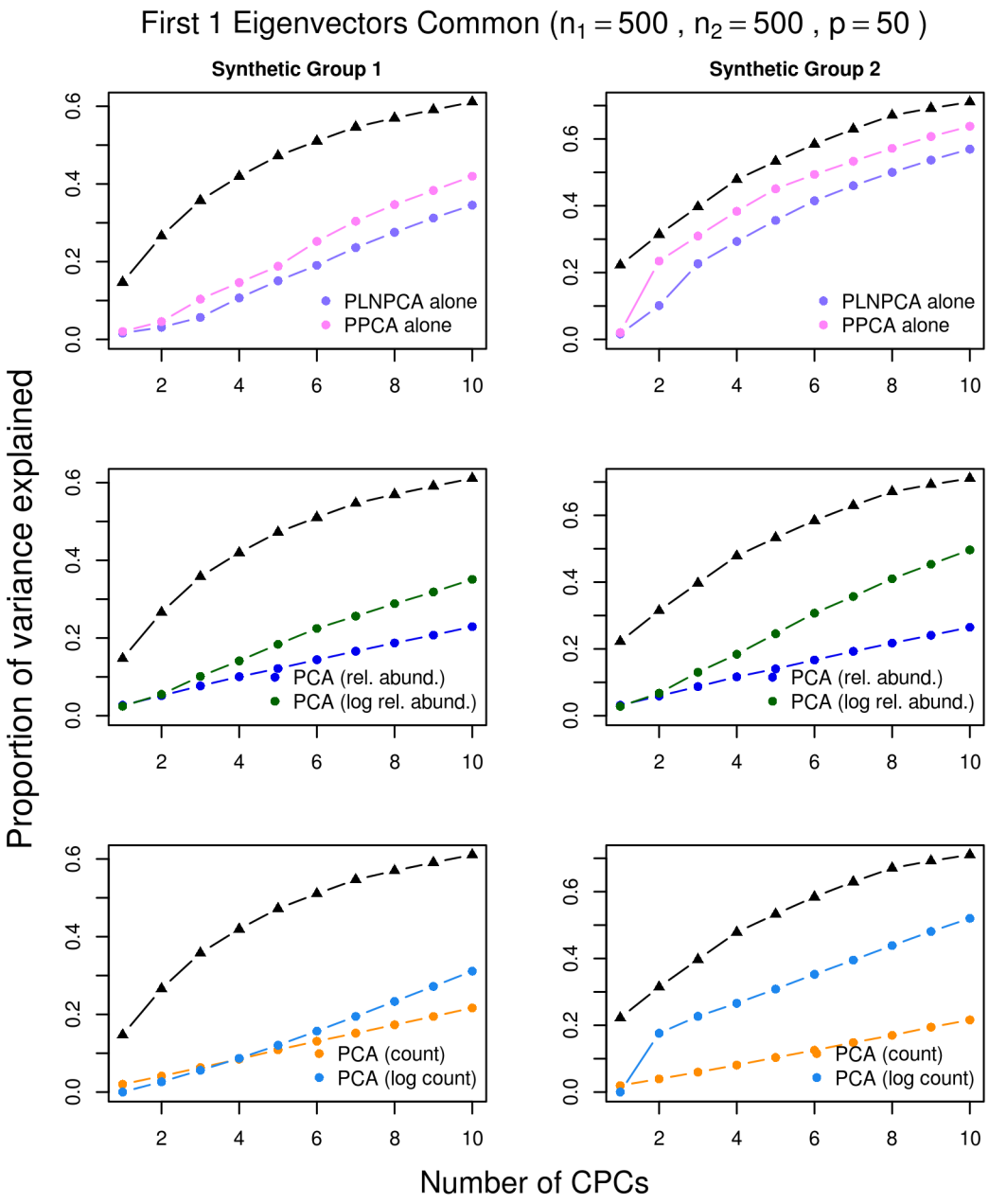

Figure A.15: Simulation results for decreasing eigenvalues and one common eigenvector, with SDC; p=50, $n_1 = 500$, $n_2 = 500$.

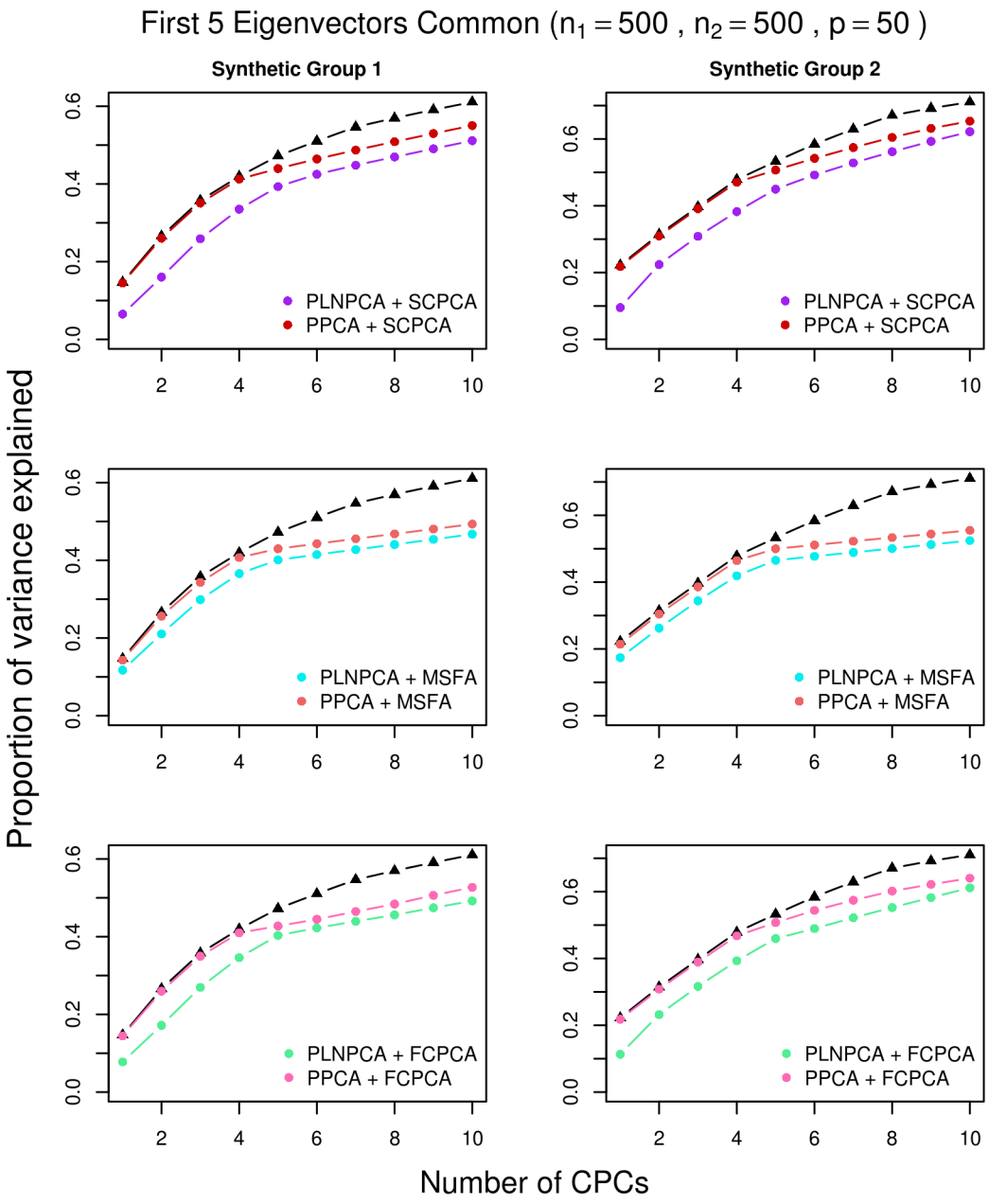

Figure A.16: Simulation results for decreasing eigenvalues and five common eigenvectors, with SDC; p=50, $n_1 = 200$, $n_2 = 500$.

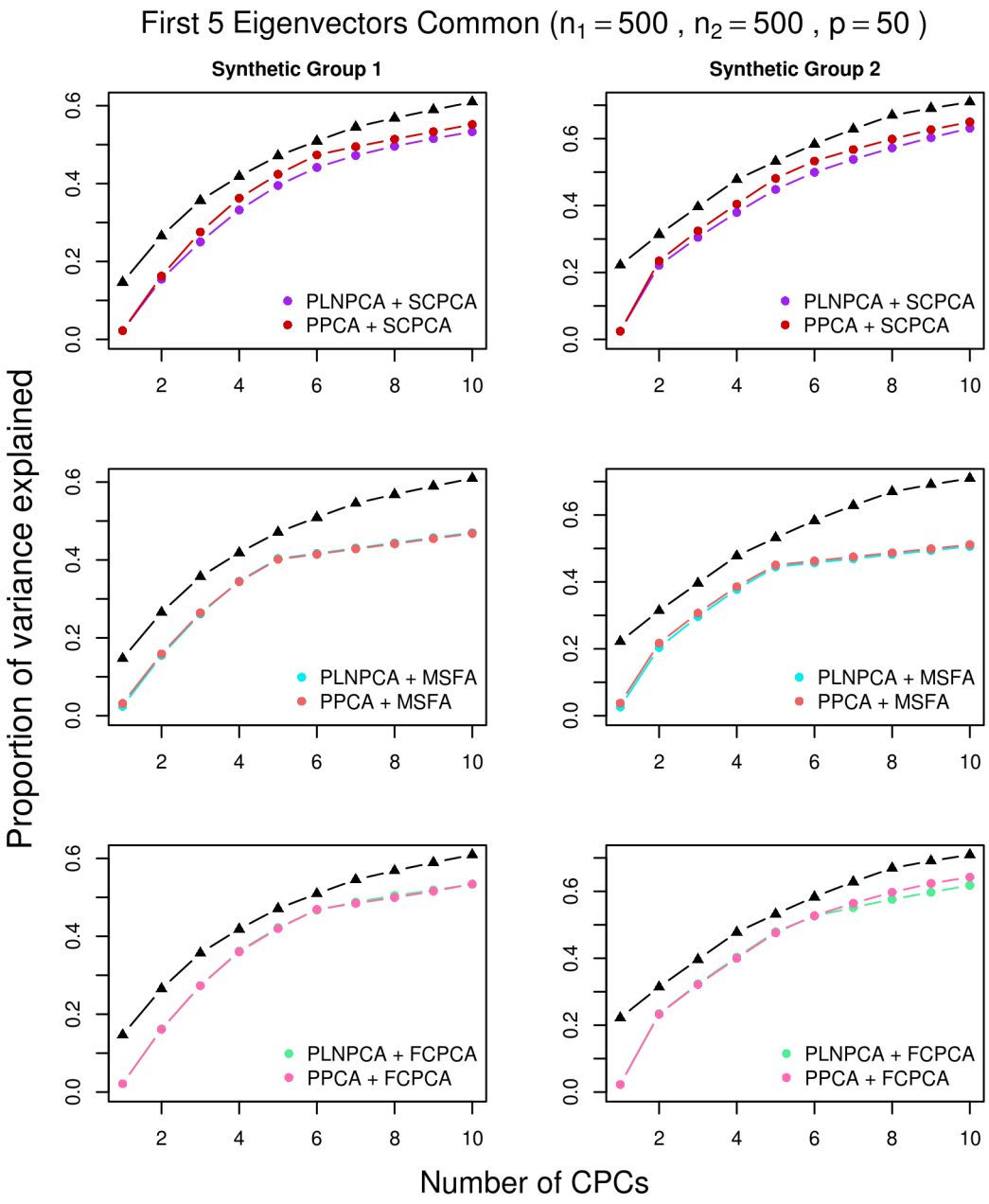

Figure A.17: Simulation results for decreasing eigenvalues and five common eigenvectors, with no SDC; p=50, $n_1 = 500$, $n_2 = 500$.

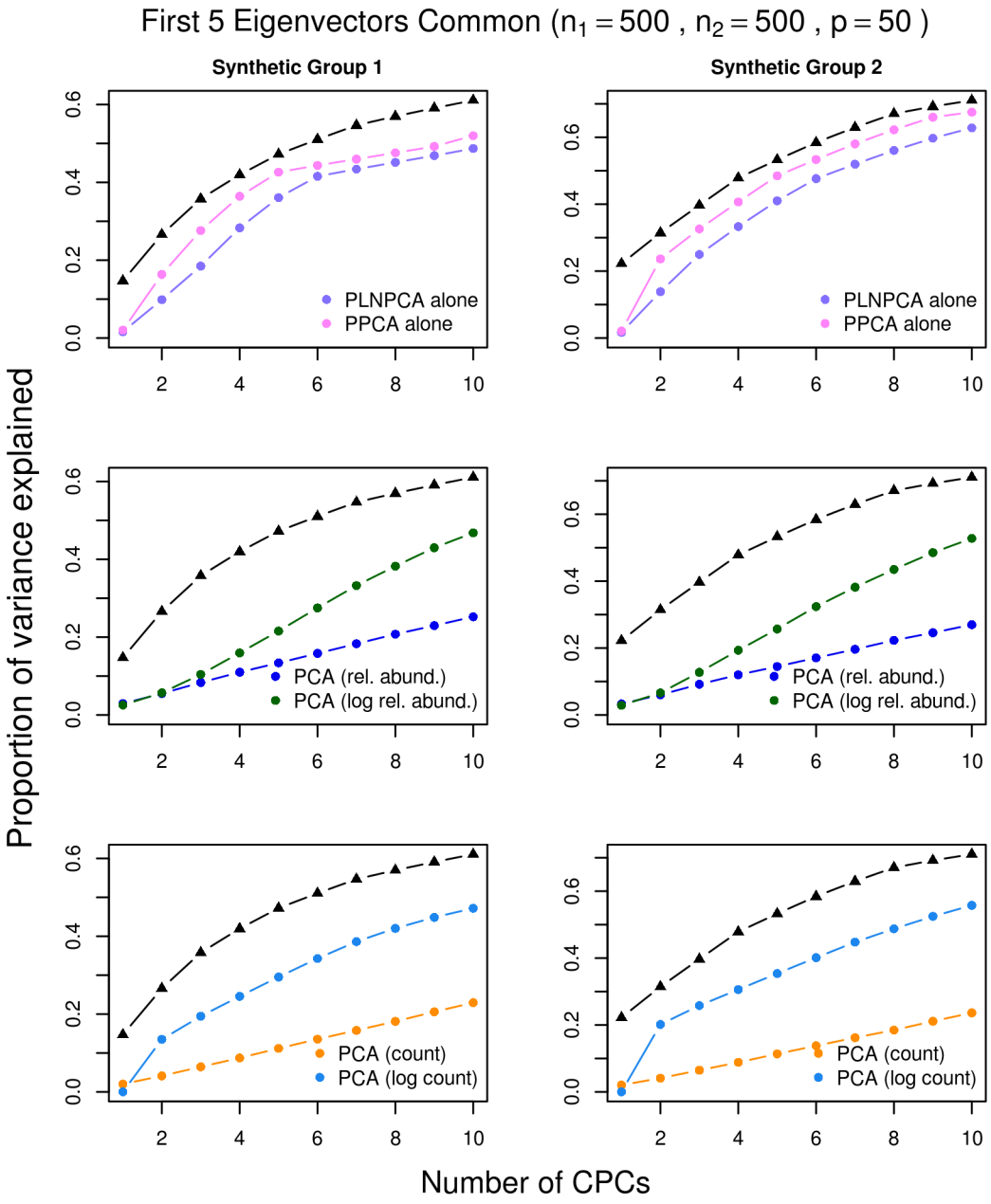

Figure A.18: Simulation results for decreasing eigenvalues and five common eigenvectors, with SDC; p=50, $n_1 = 500$, $n_2 = 500$.

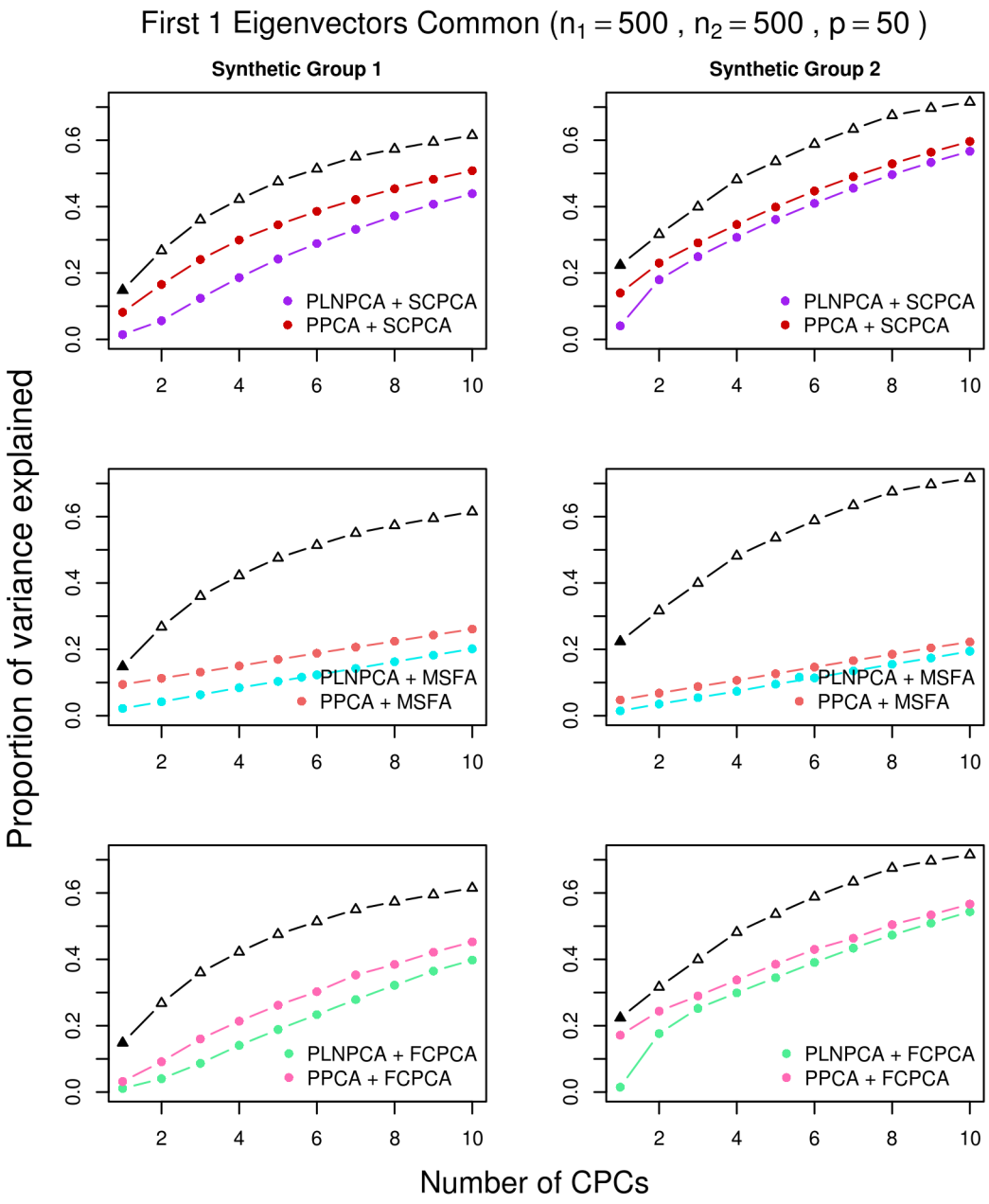

Figure A.19: Simulation results for non-decreasing eigenvalues and one common eigenvector, with SDC; p=50, $n_1 = 500$, $n_2 = 500$.

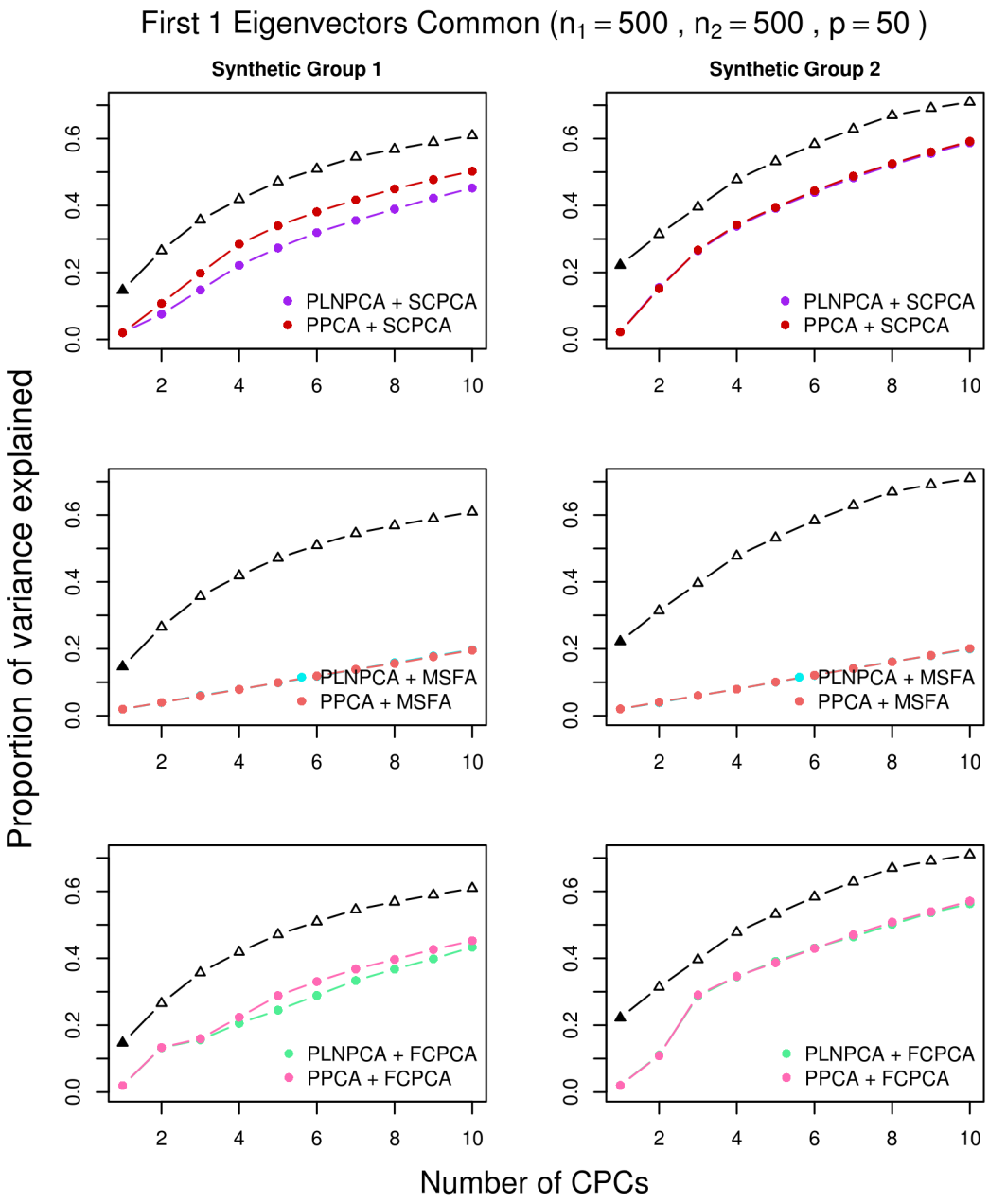

Figure A.20: Simulation results for non-decreasing eigenvalues and one common eigenvector, without SDC; p=50, $n_1 = 500$, $n_2 = 500$.

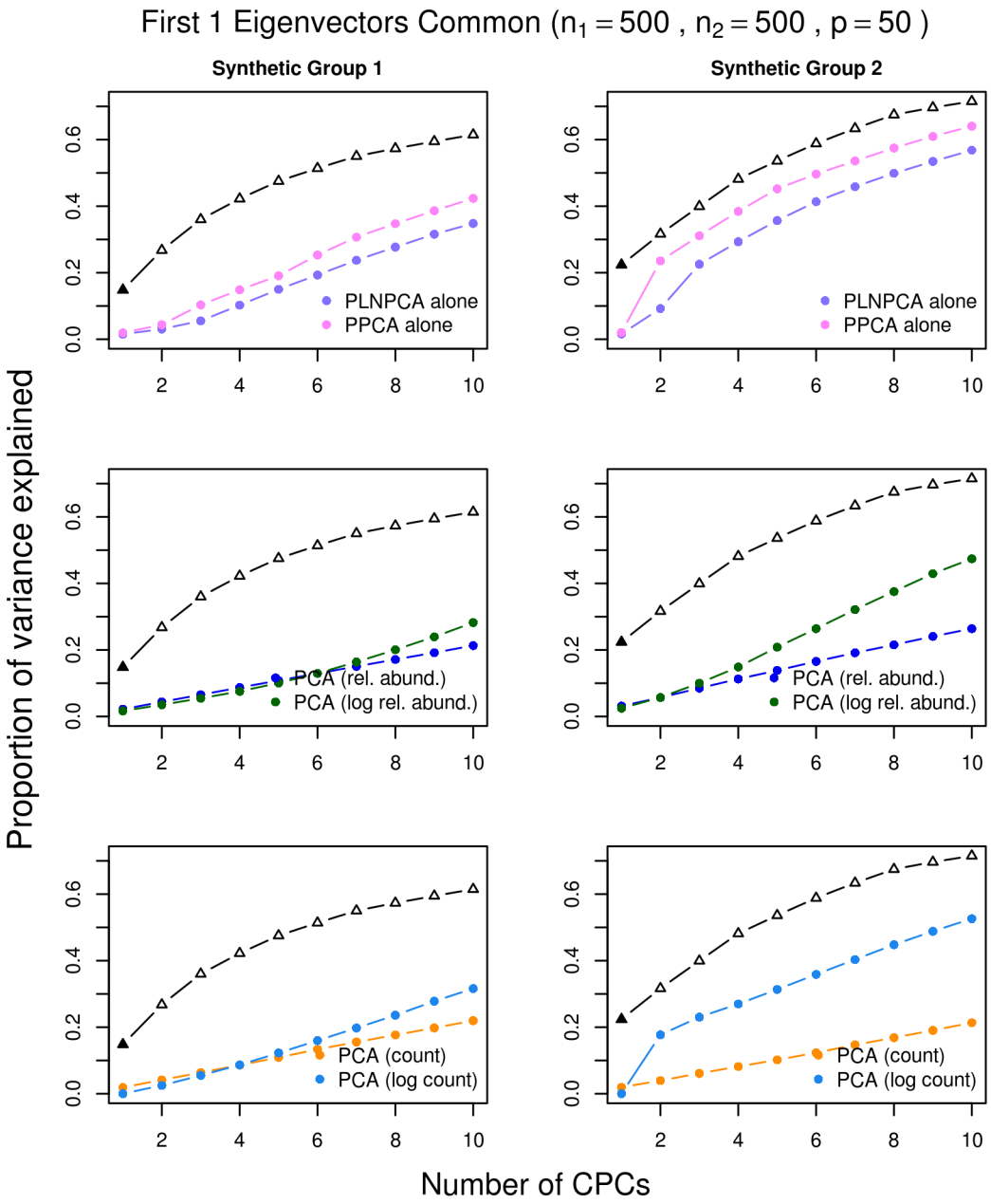

Figure A.21: Simulation results single-group methods for non-decreasing eigenvalues and one common eigenvector, with SDC for PoissonPCA/PLNPCA; p=50, $n_1 = 500$, $n_2 = 500$.

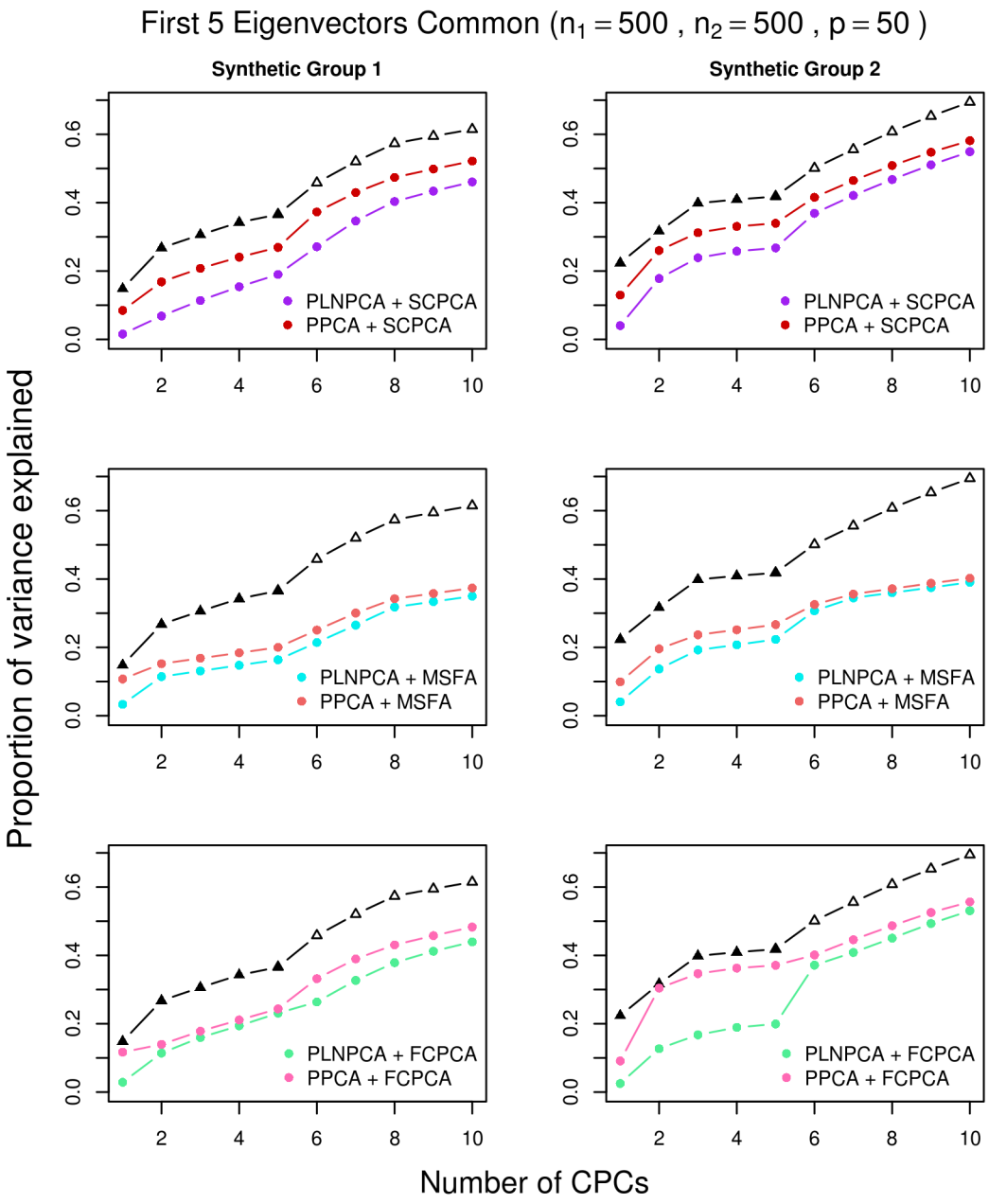

Figure A.22: Simulation results for non-decreasing eigenvalues and five common eigenvectors, with SDC; p=50, $n_1 = 500$, $n_2 = 500$.

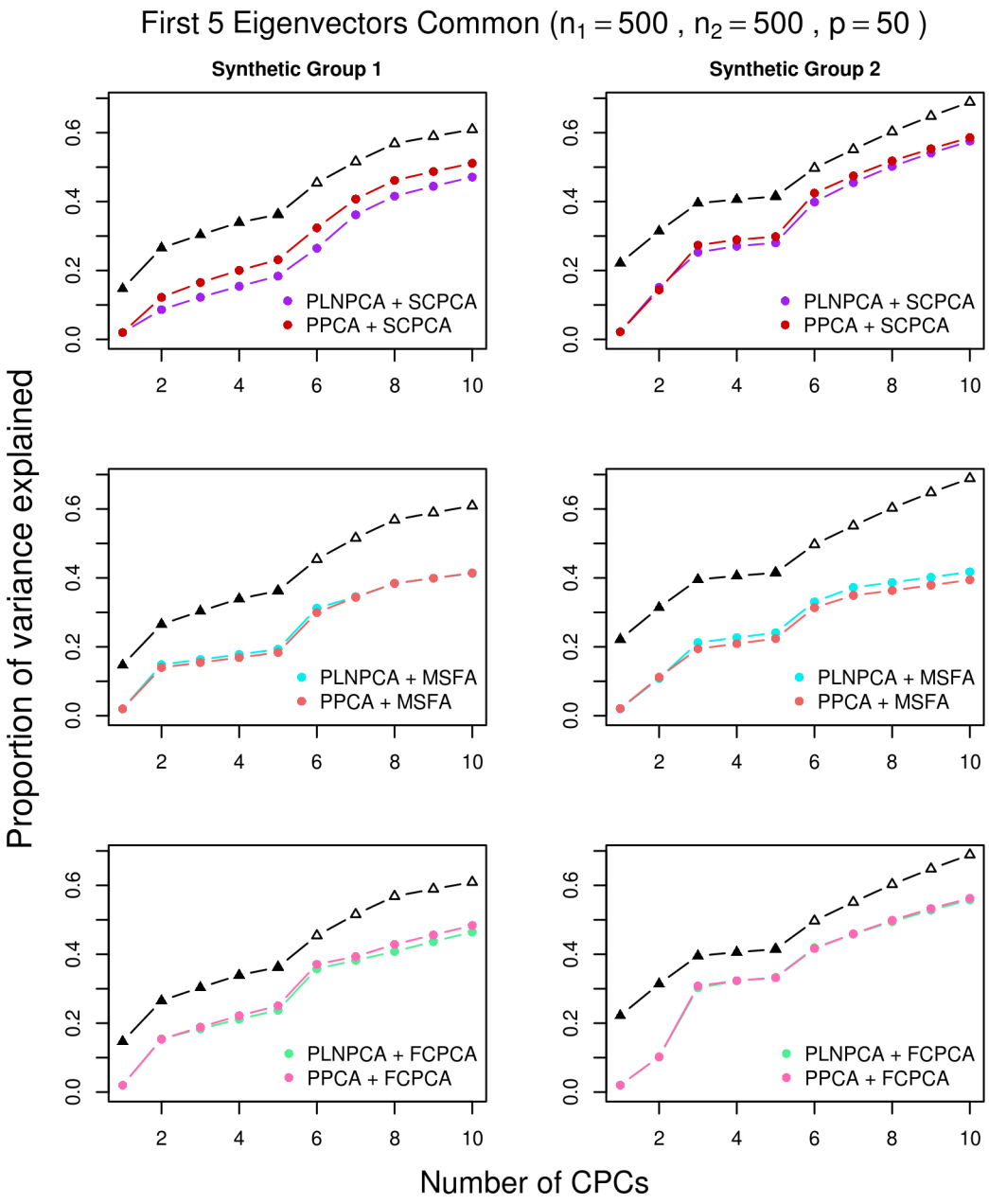

Figure A.23: Simulation results for non-decreasing eigenvalues and five common eigenvectors, without SDC; p=50, $n_1 = 500$, $n_2 = 500$.

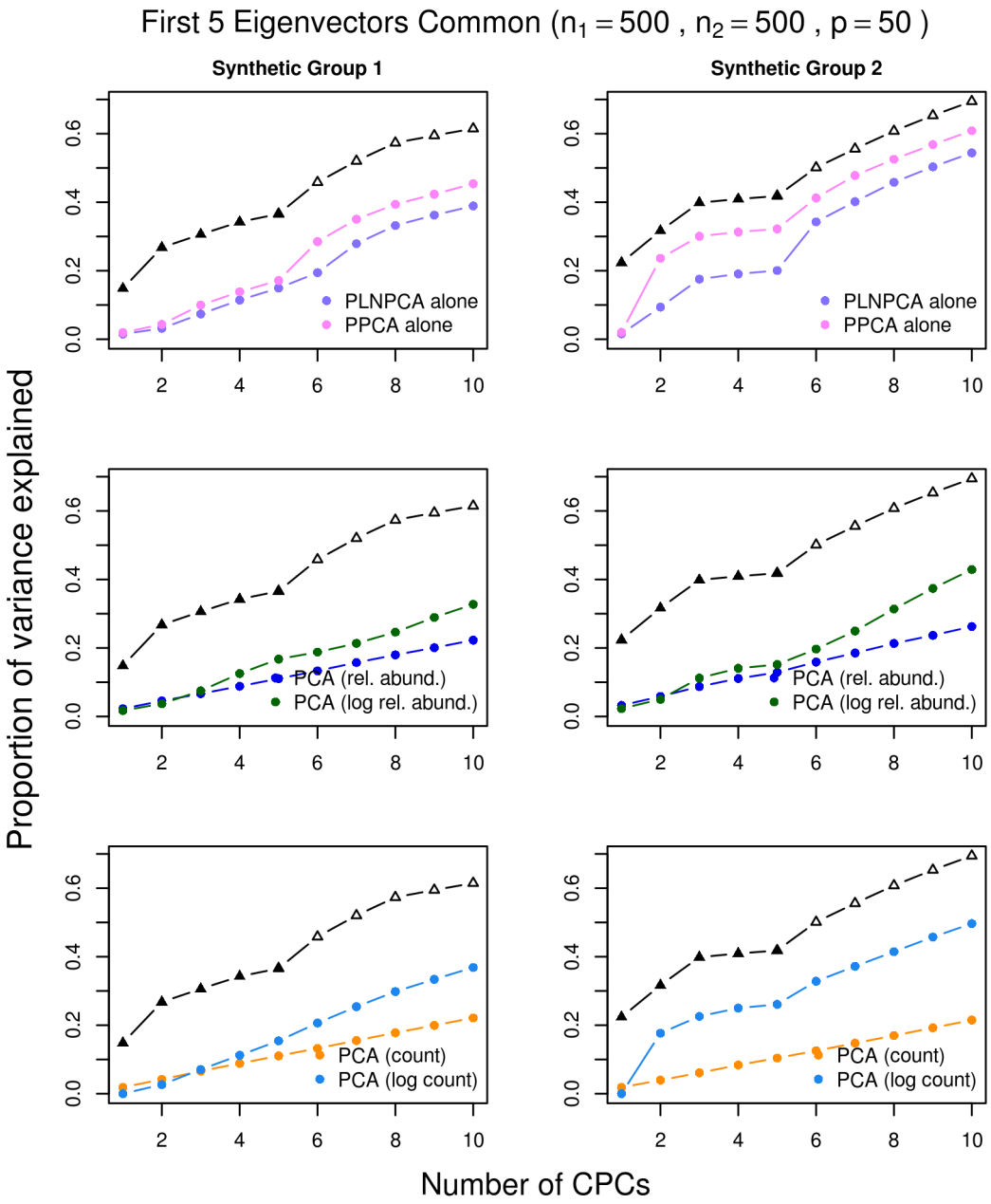

Figure A.24: Simulation results single-group methods for non-decreasing eigenvalues and five common eigenvectors, with SDC for PoissonPCA/PLNPCA; p=50, $n_1 = 500$, $n_2 = 500$.

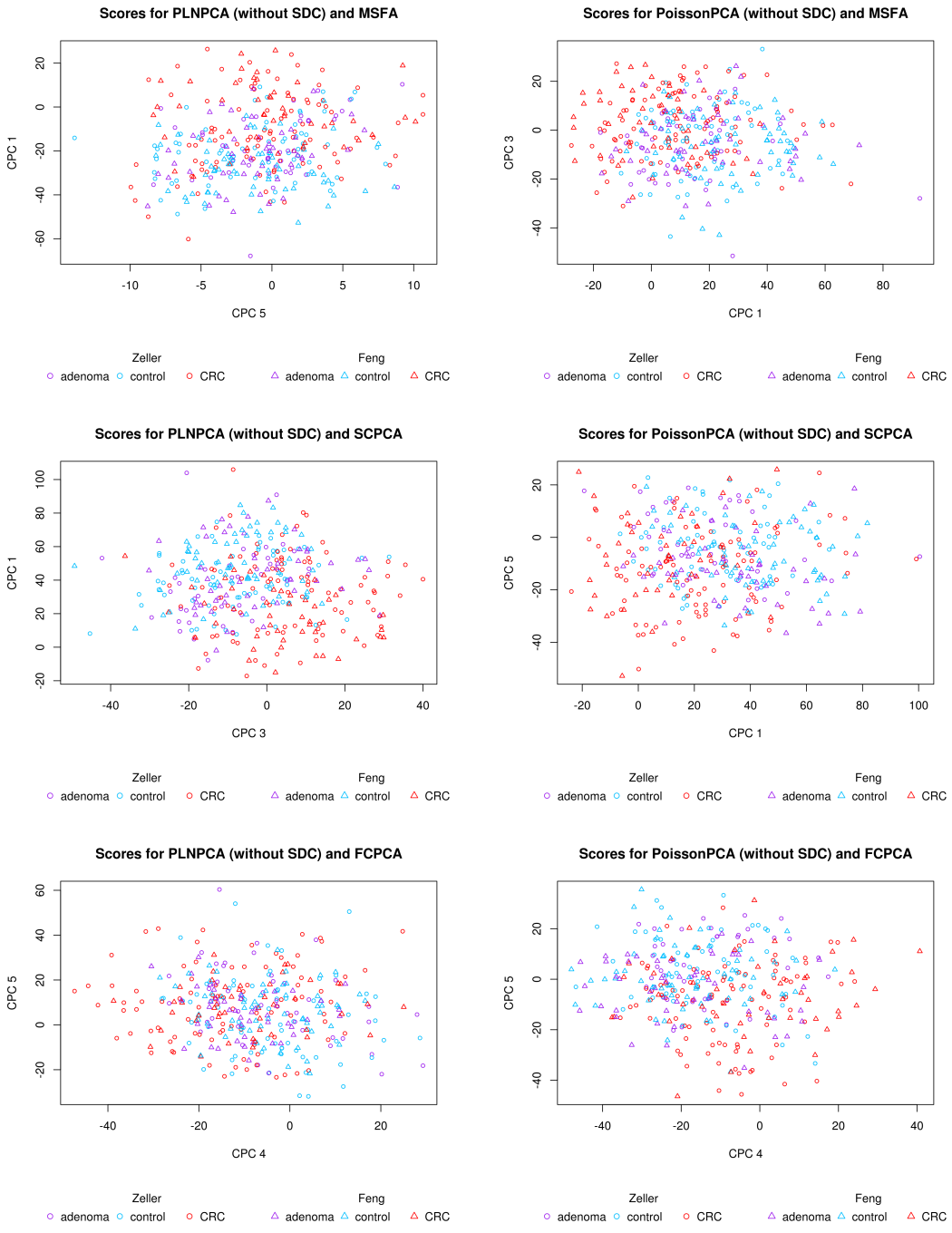

Figure A.25: Scores from ensemble methods without SDC by disease state.

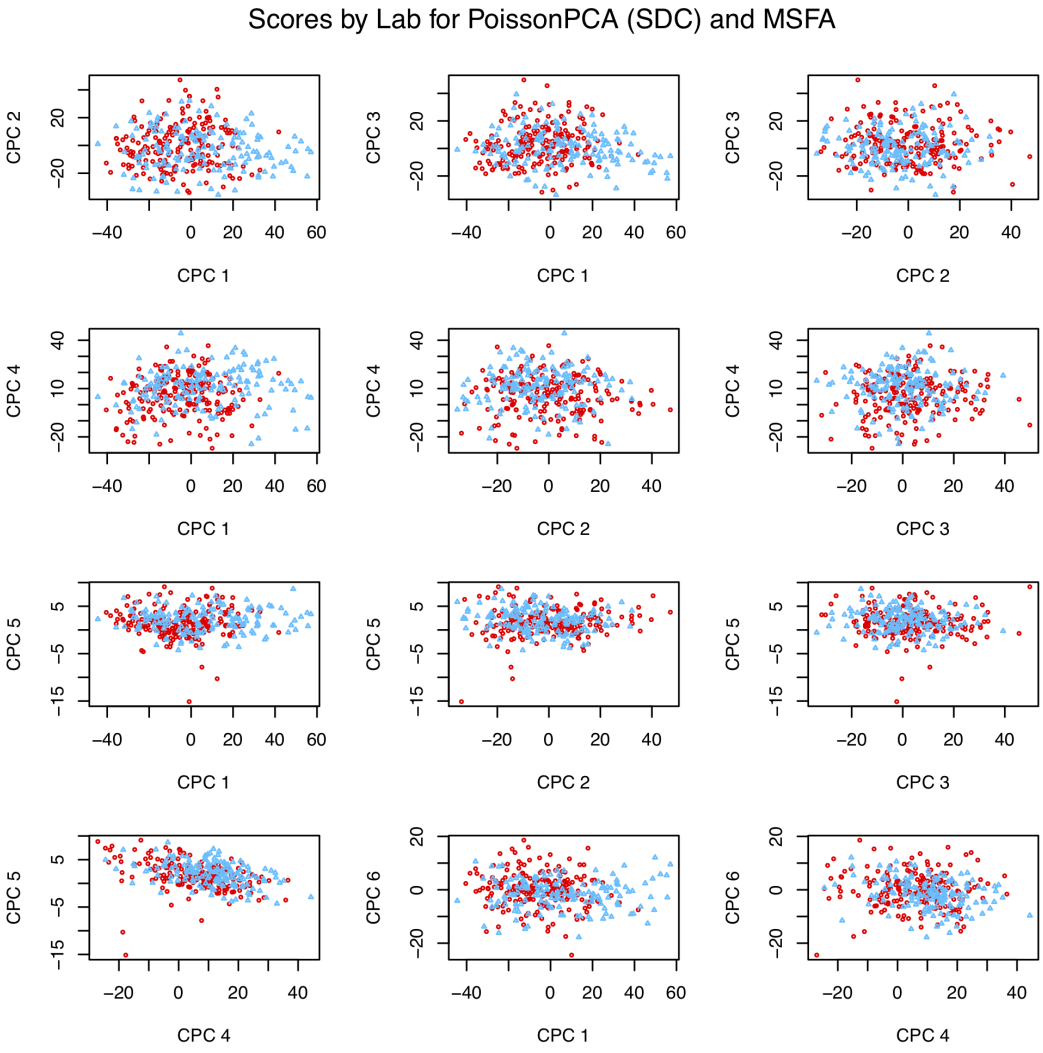

Figure A.26: Scores from PoissonPCA (SDC) and MSFA by study of origin.

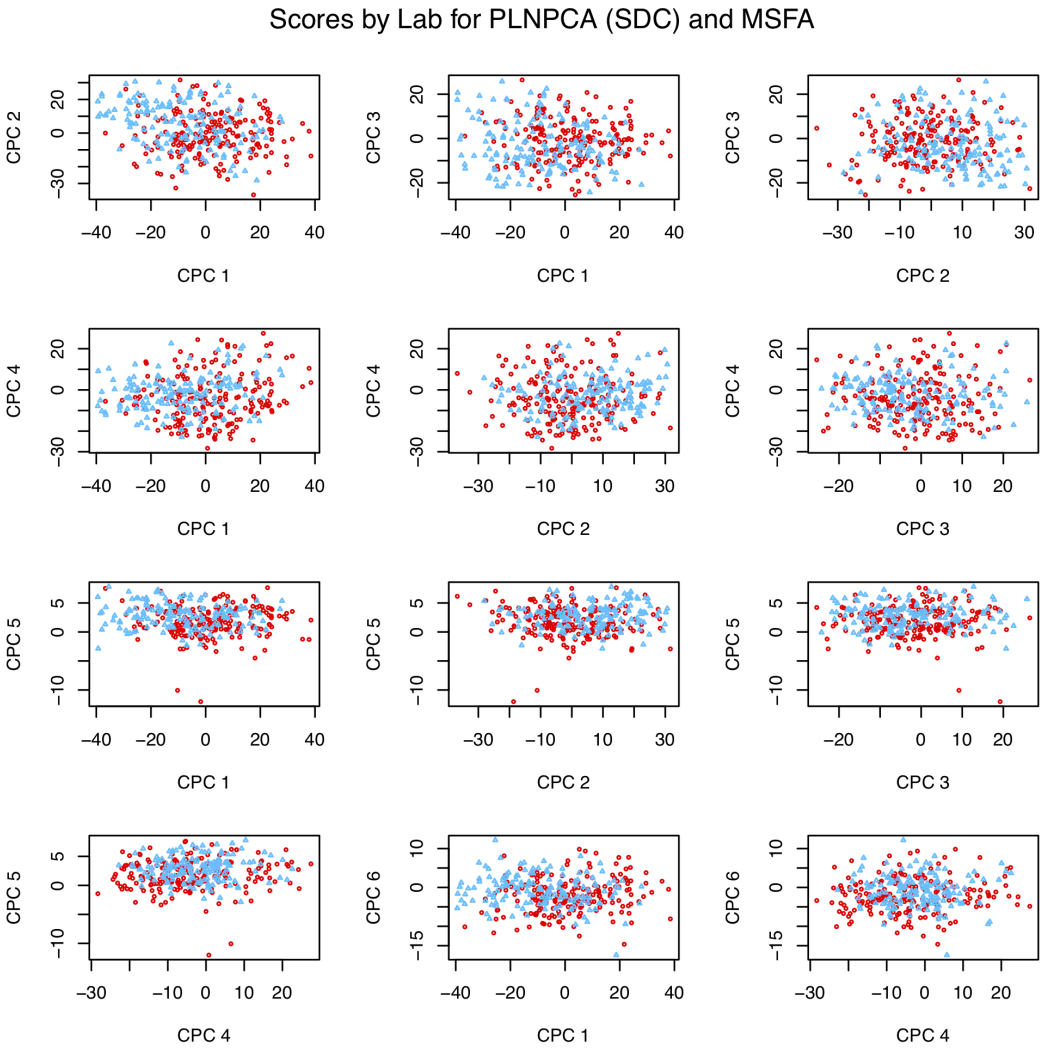

Figure A.27: Scores from PLNPCA (SDC) and MSFA by study of origin.

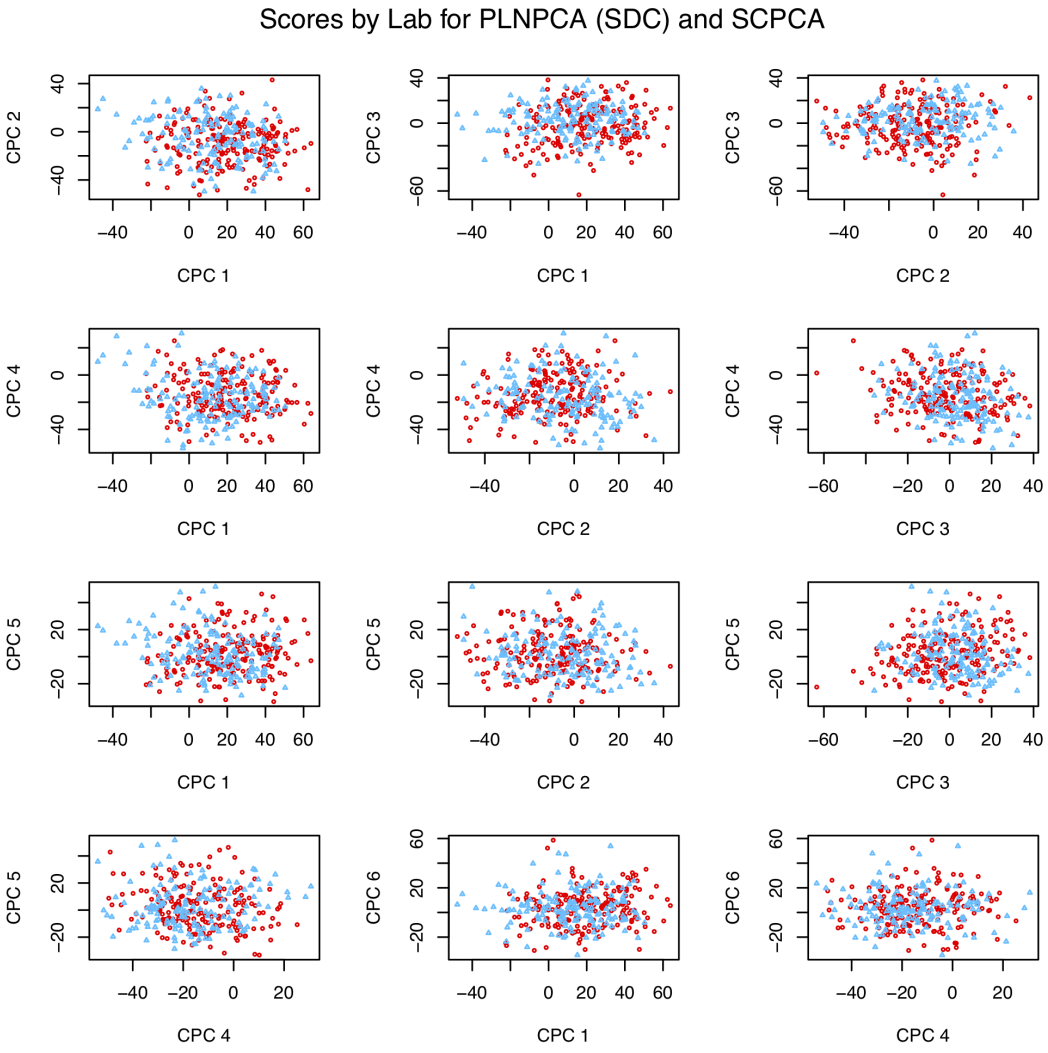

Figure A.28: Scores from PLNPCA (SDC) and SCPCA by study of origin.

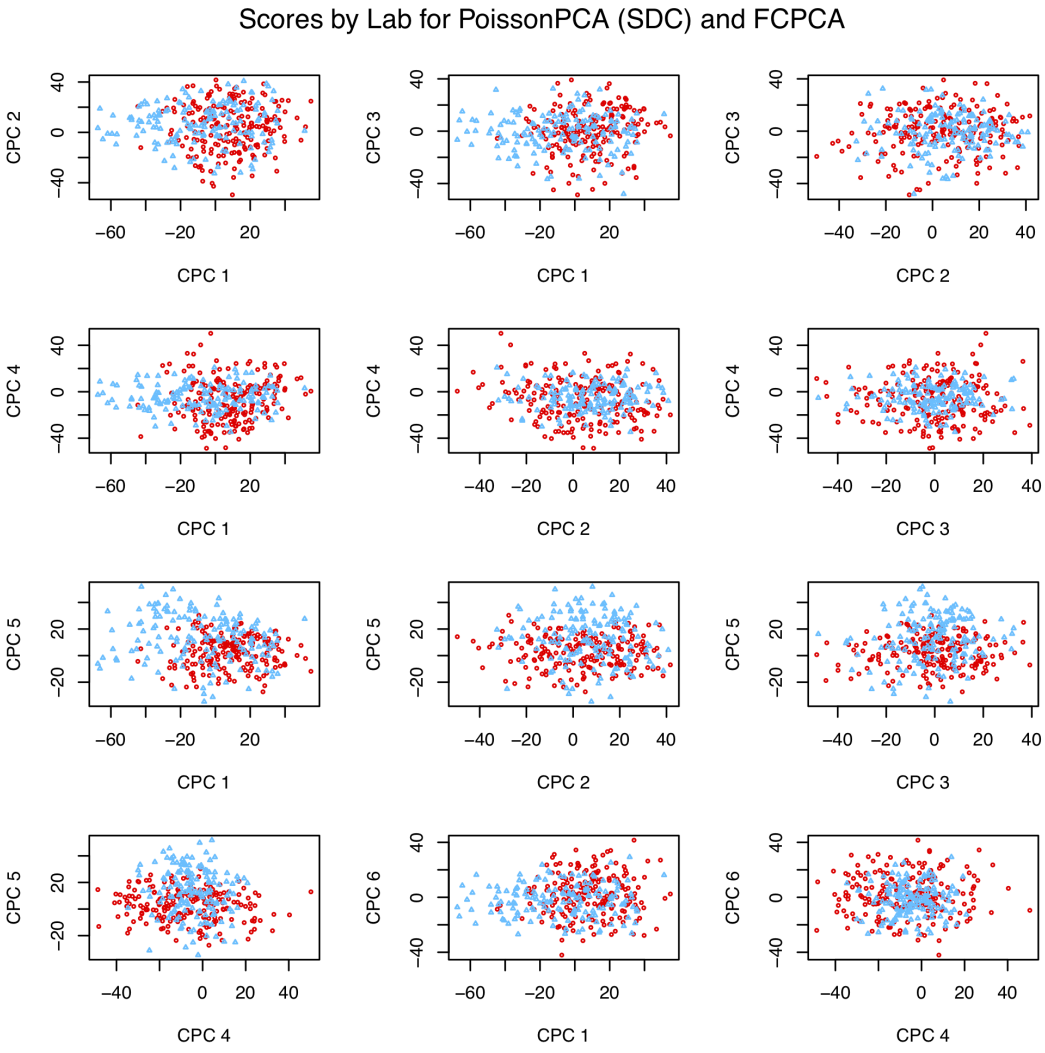

Figure A.29: Scores from PoissonPCA (SDC) & FCPCA by study of origin.

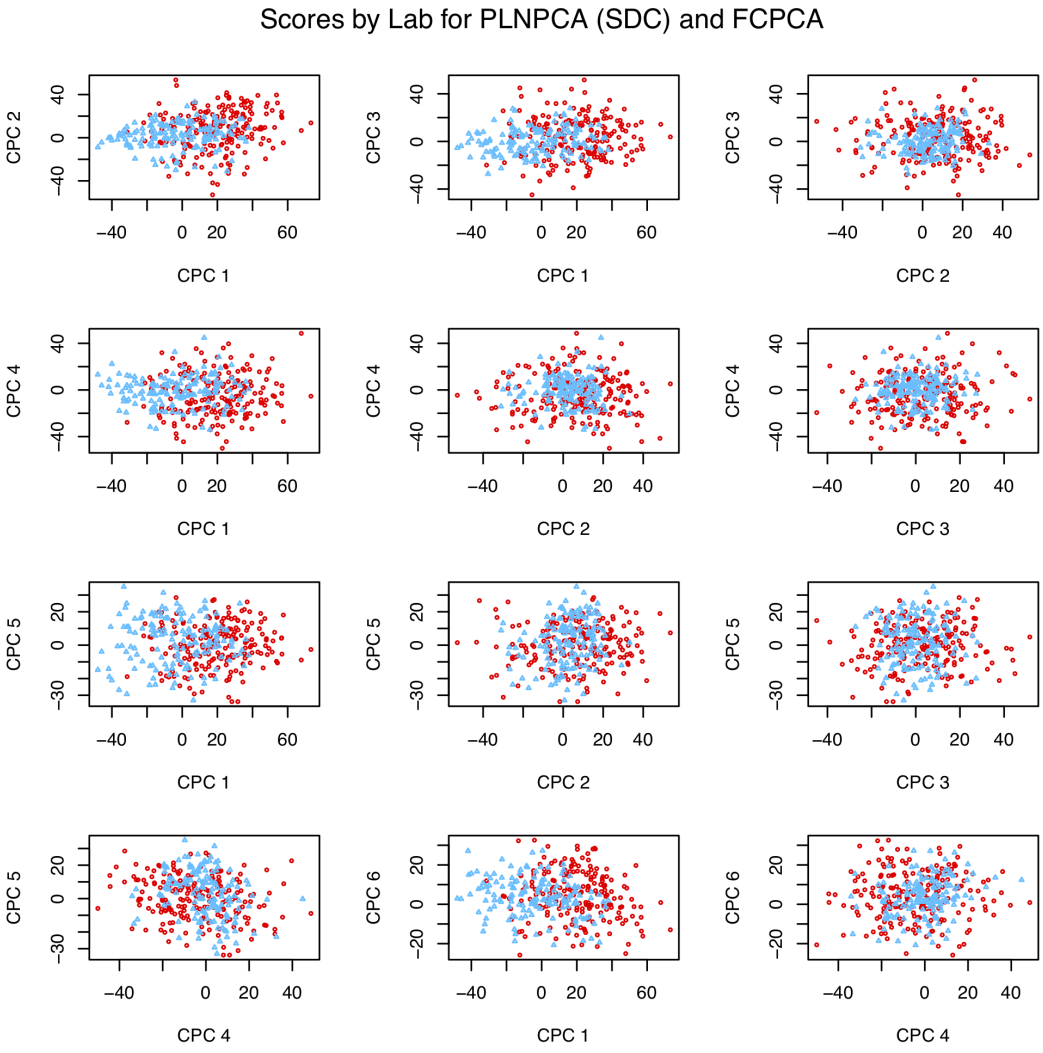

Figure A.30: Scores from PLNPCA (SDC) and FCPCA by study of origin.

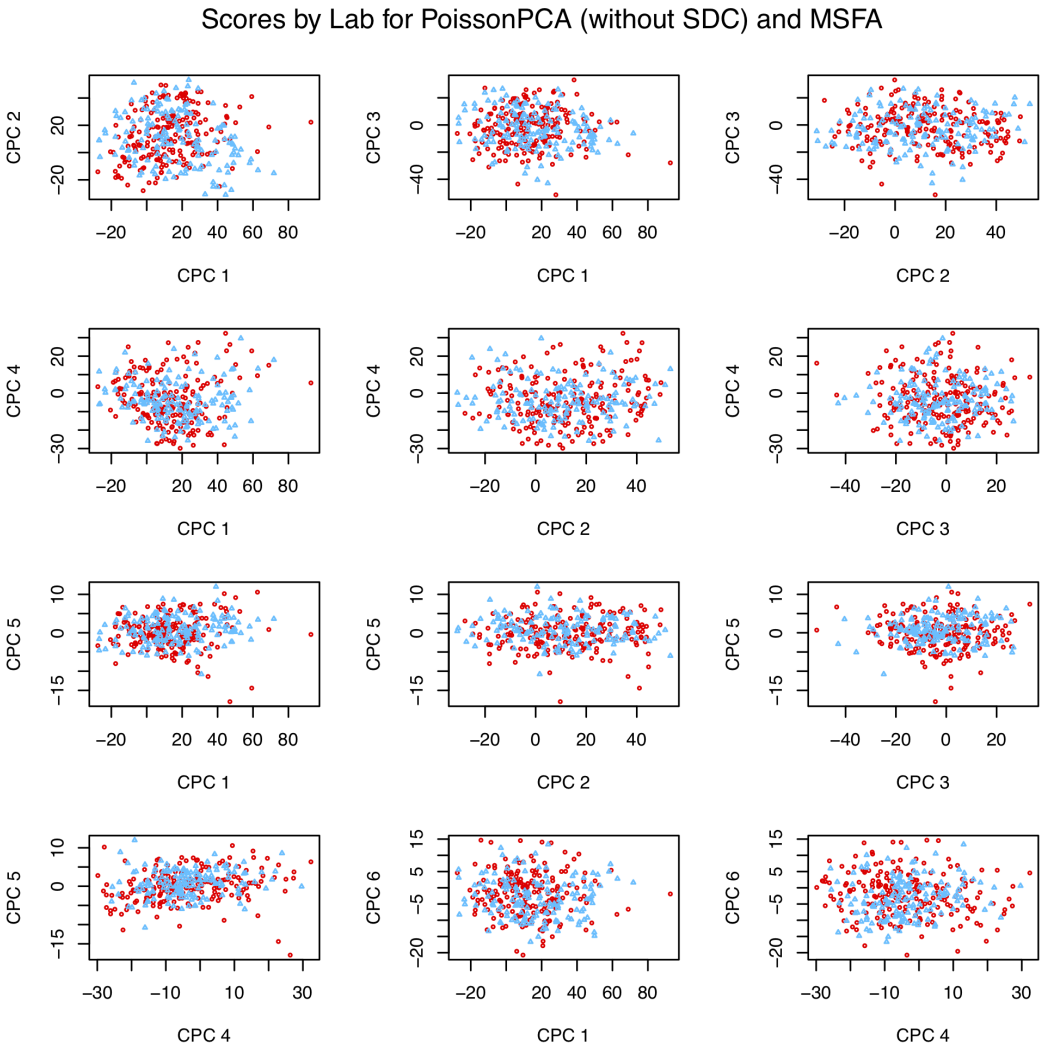

Figure A.31: Scores from PoissonPCA (no SDC) & MSFA by study of origin.

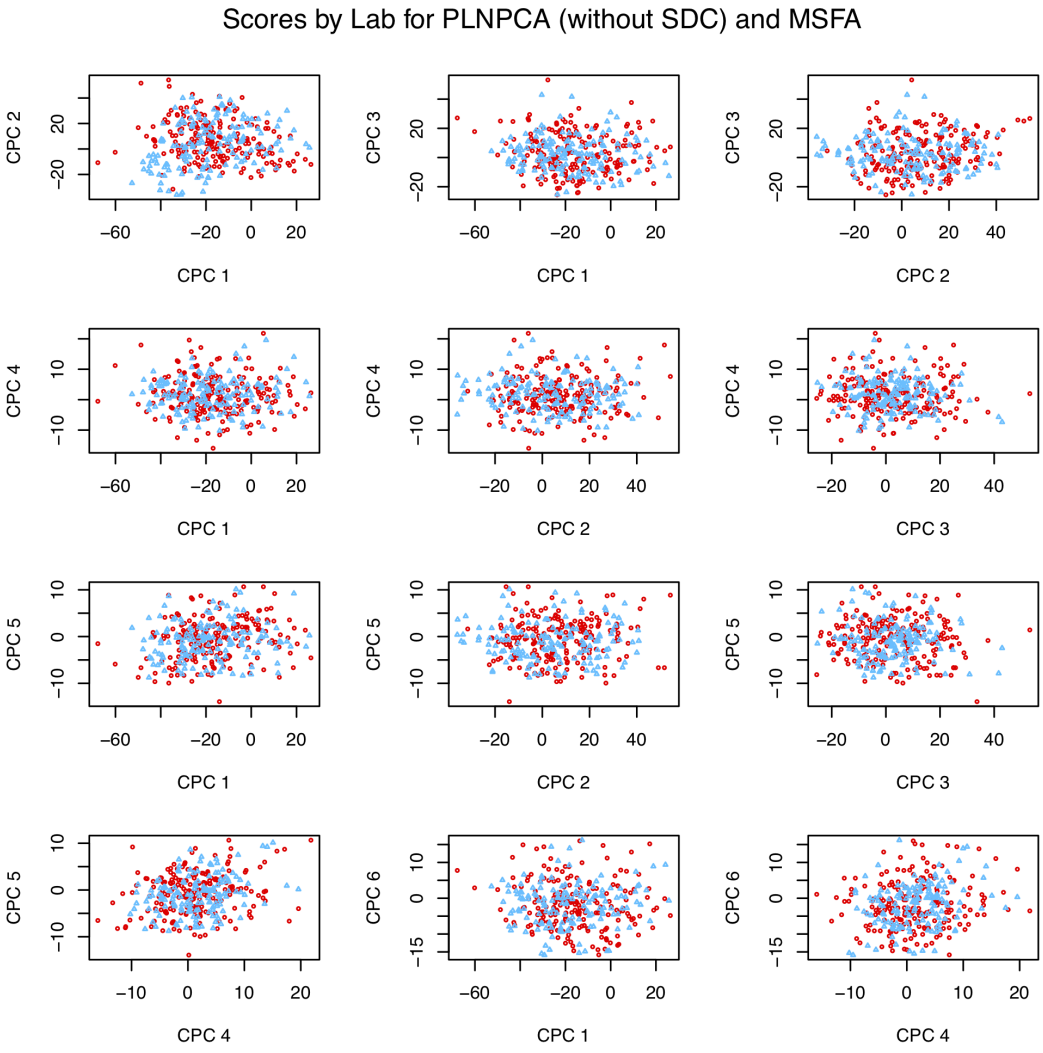

Figure A.32: Scores from PLNPCA (no SDC) and MSFA by study of origin.

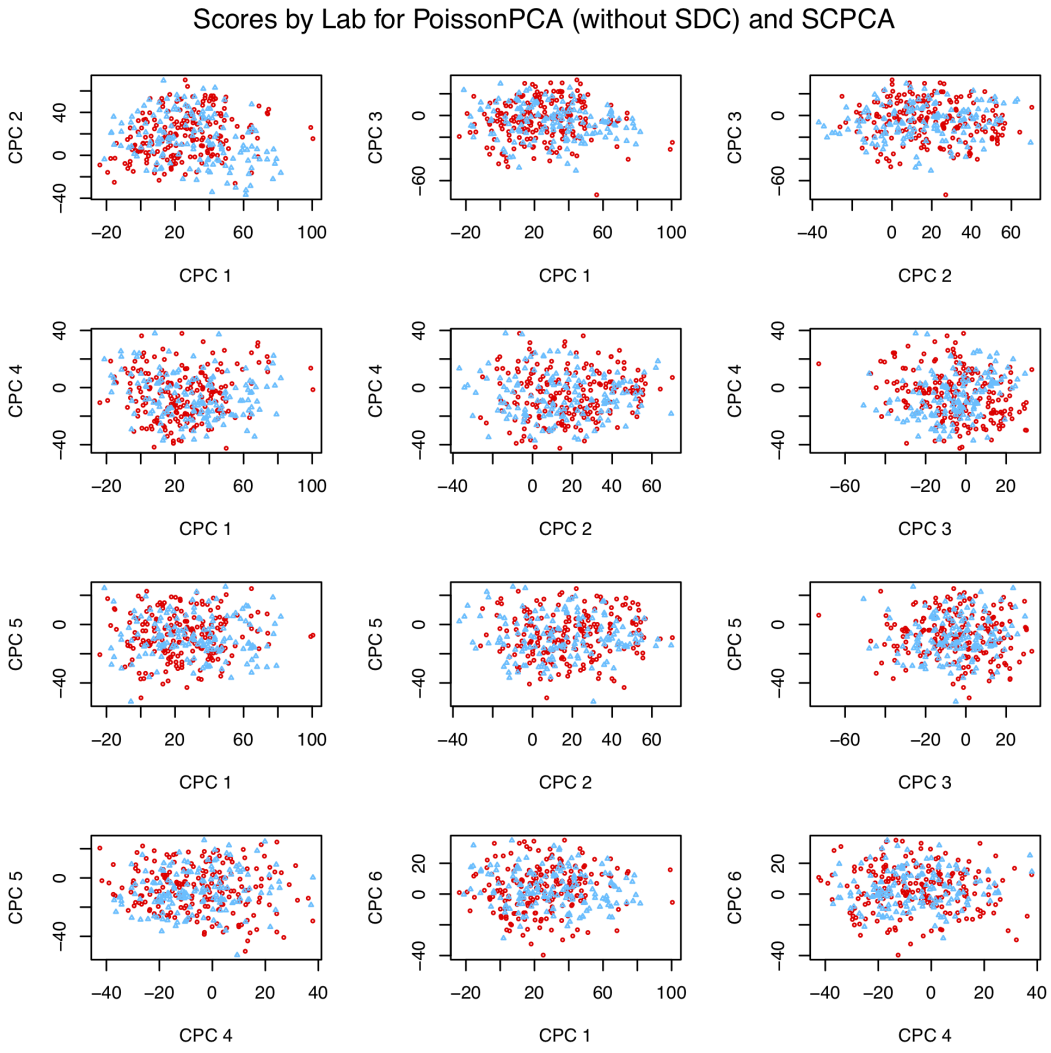

Figure A.33: Scores from PoissonPCA (no SDC) & SCPCA by study of origin.

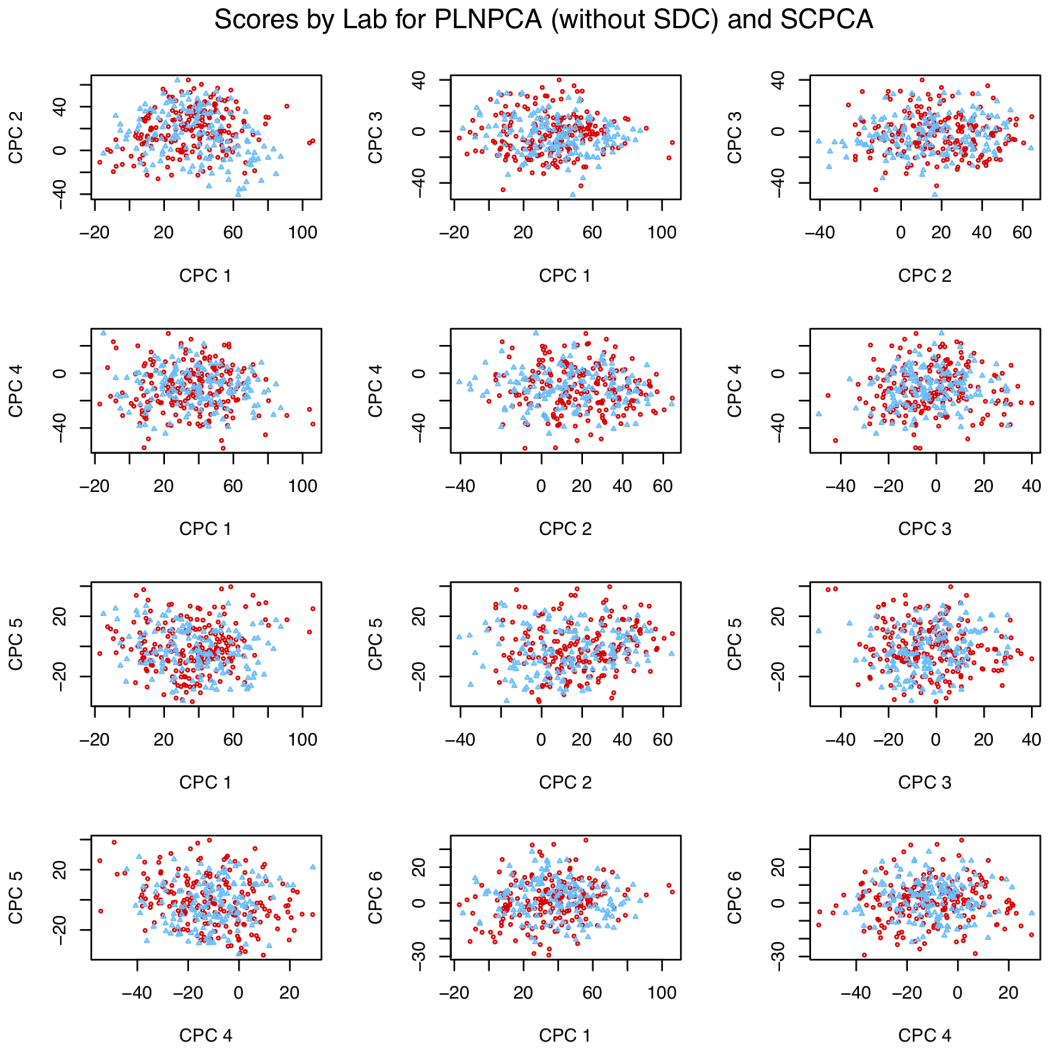

Figure A.34: Scores from PLNPCA (no SDC) & SCPCA by study of origin.

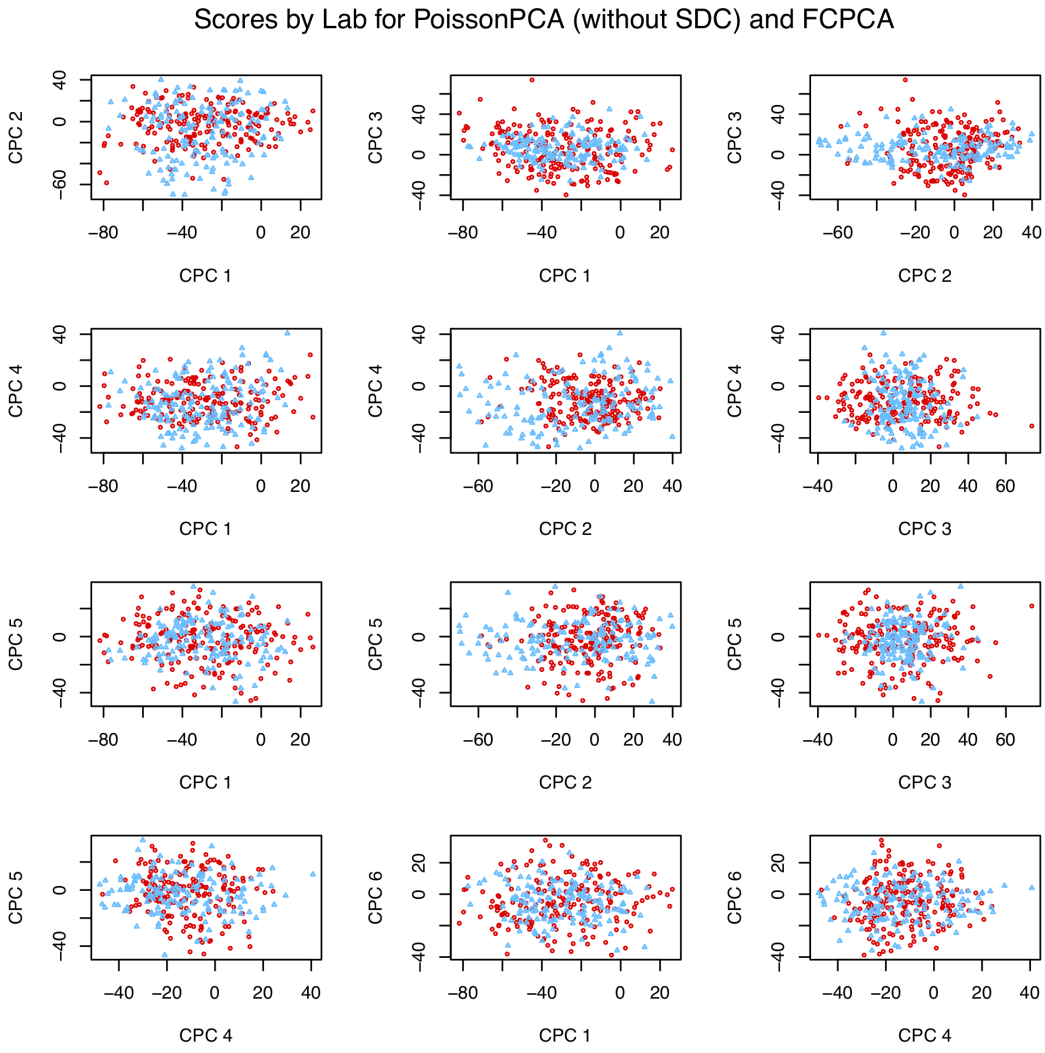

Figure A.35: Scores from PoissonPCA (no SDC) & FCPCA by study of origin.

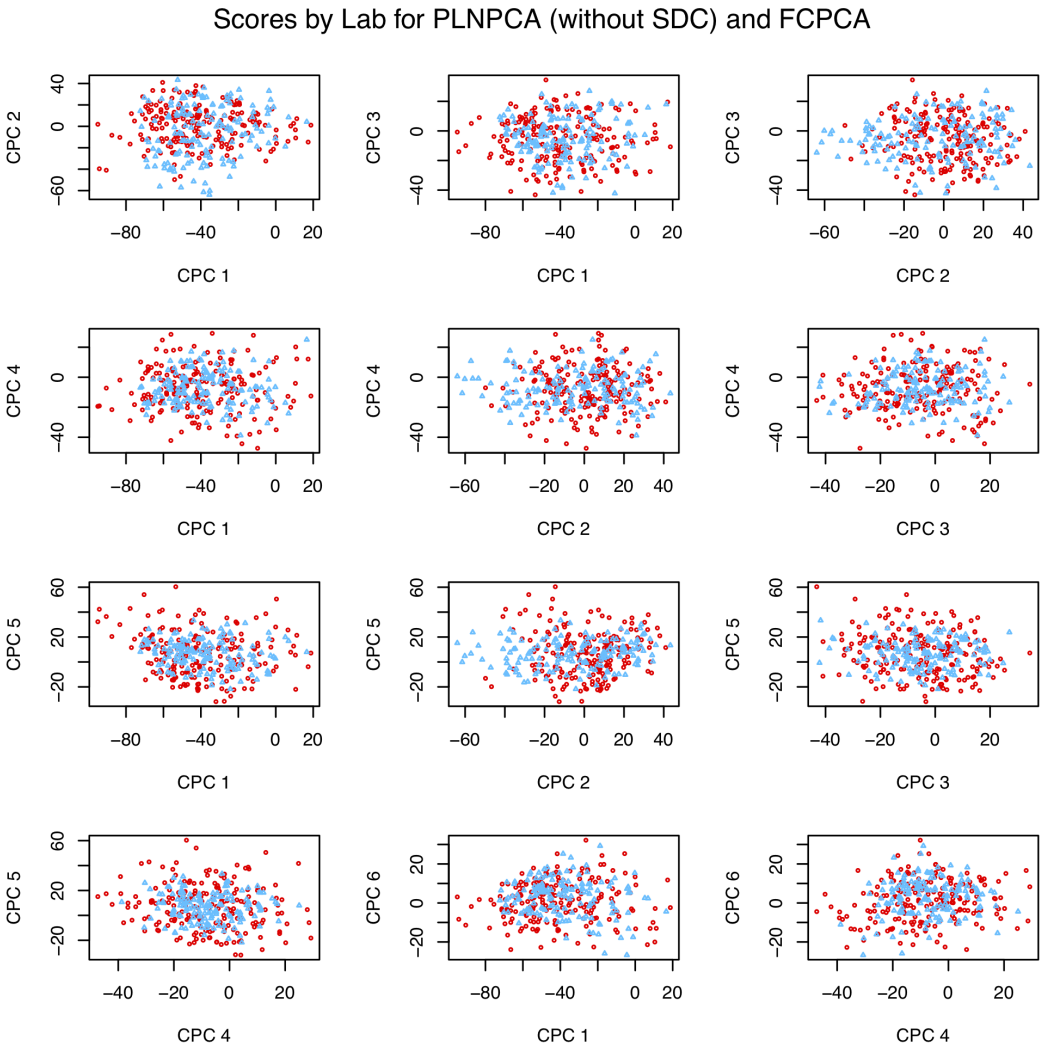

Figure A.36: Scores from PLNPCA (no SDC) & FCPCA by study of origin.

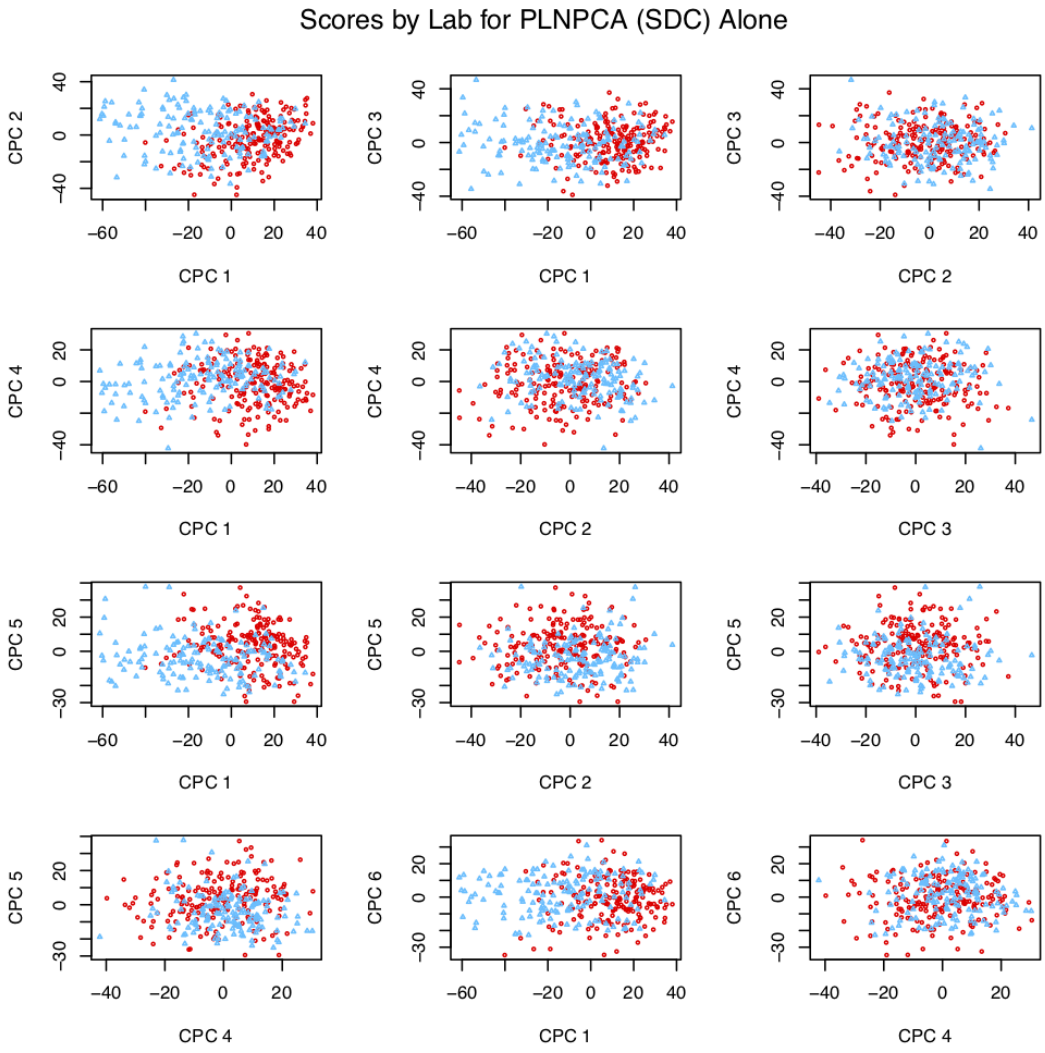

Figure A.37: Scores from PLNPCA alone with SDC by study of origin.

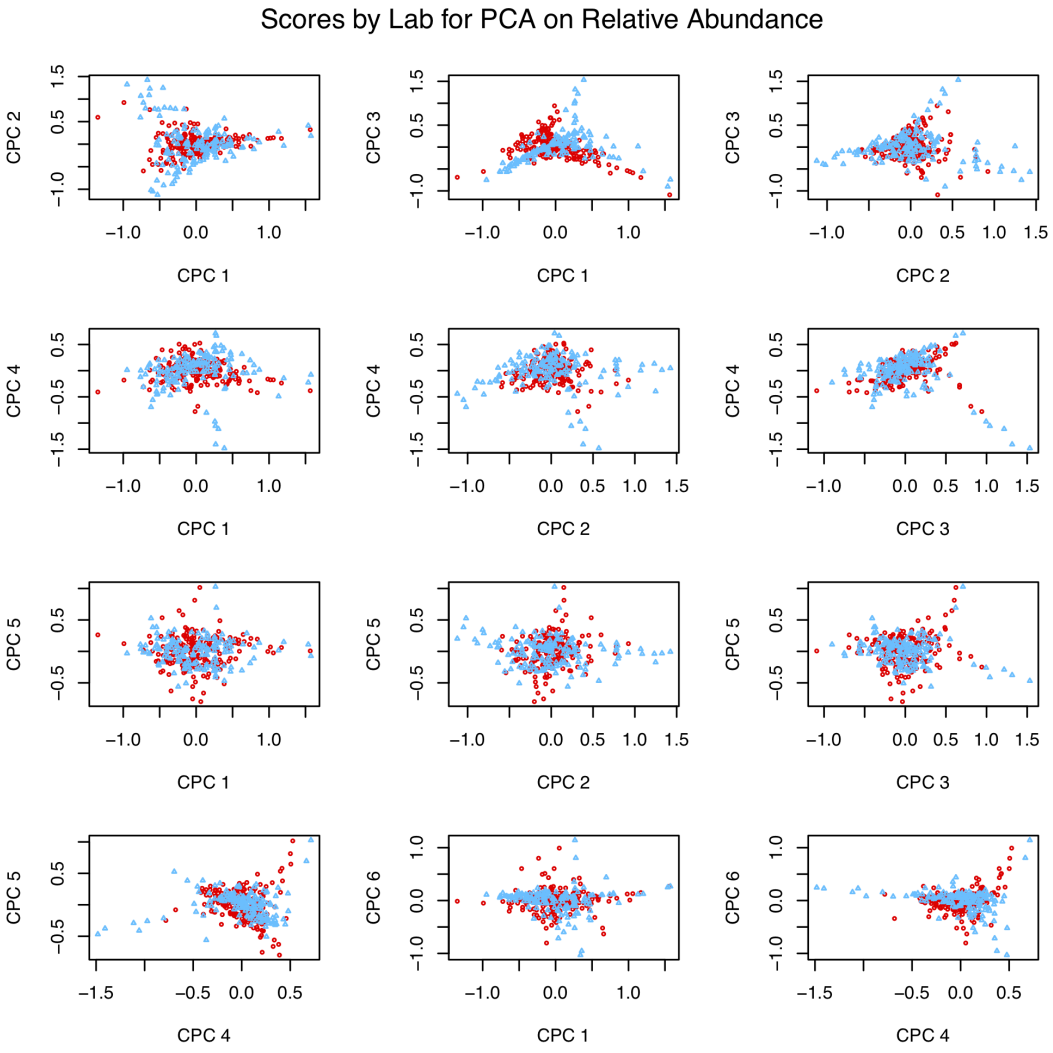

Figure A.38: Scores from PCA of relative abundance by study of origin.

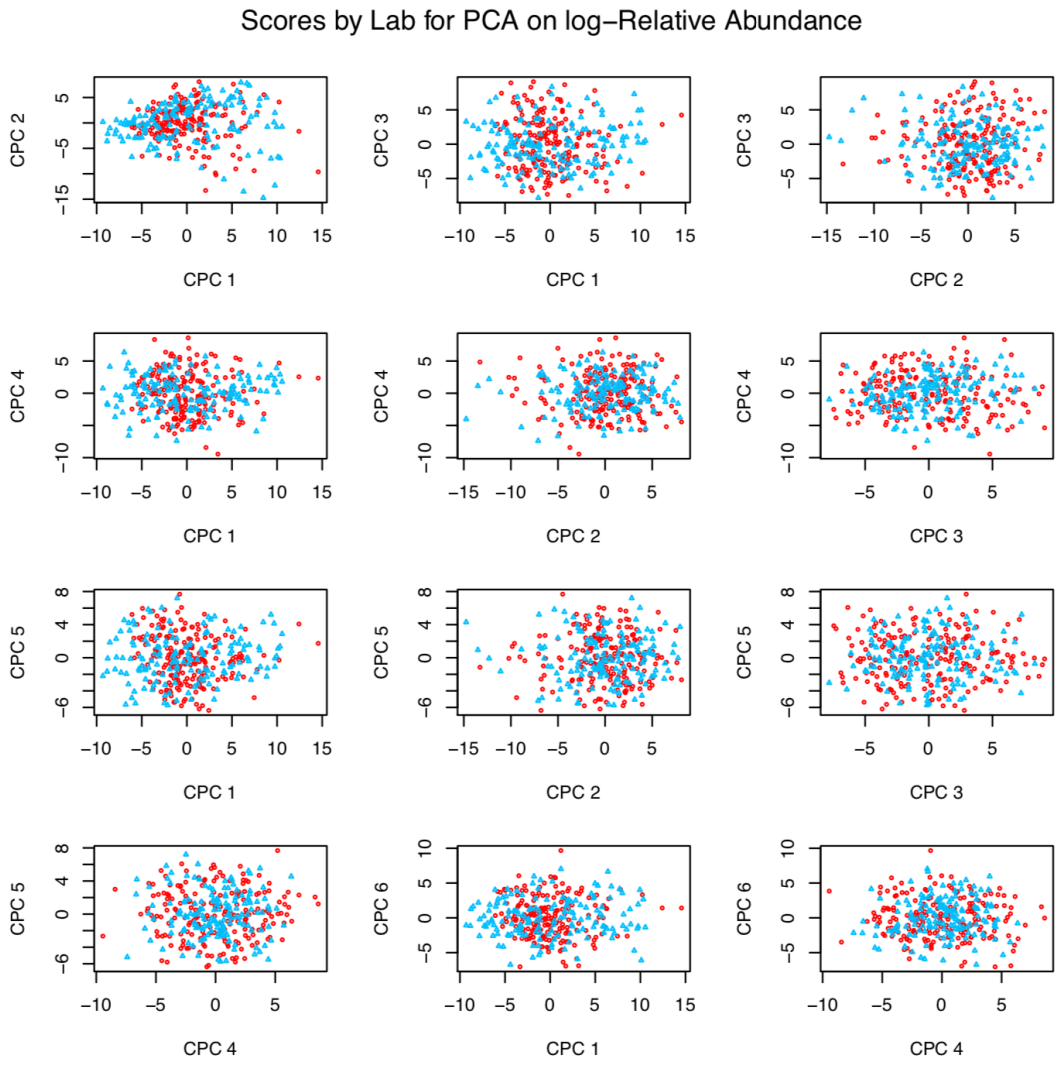

Figure A.39: Scores from PCA of log relative abundance by study of origin.

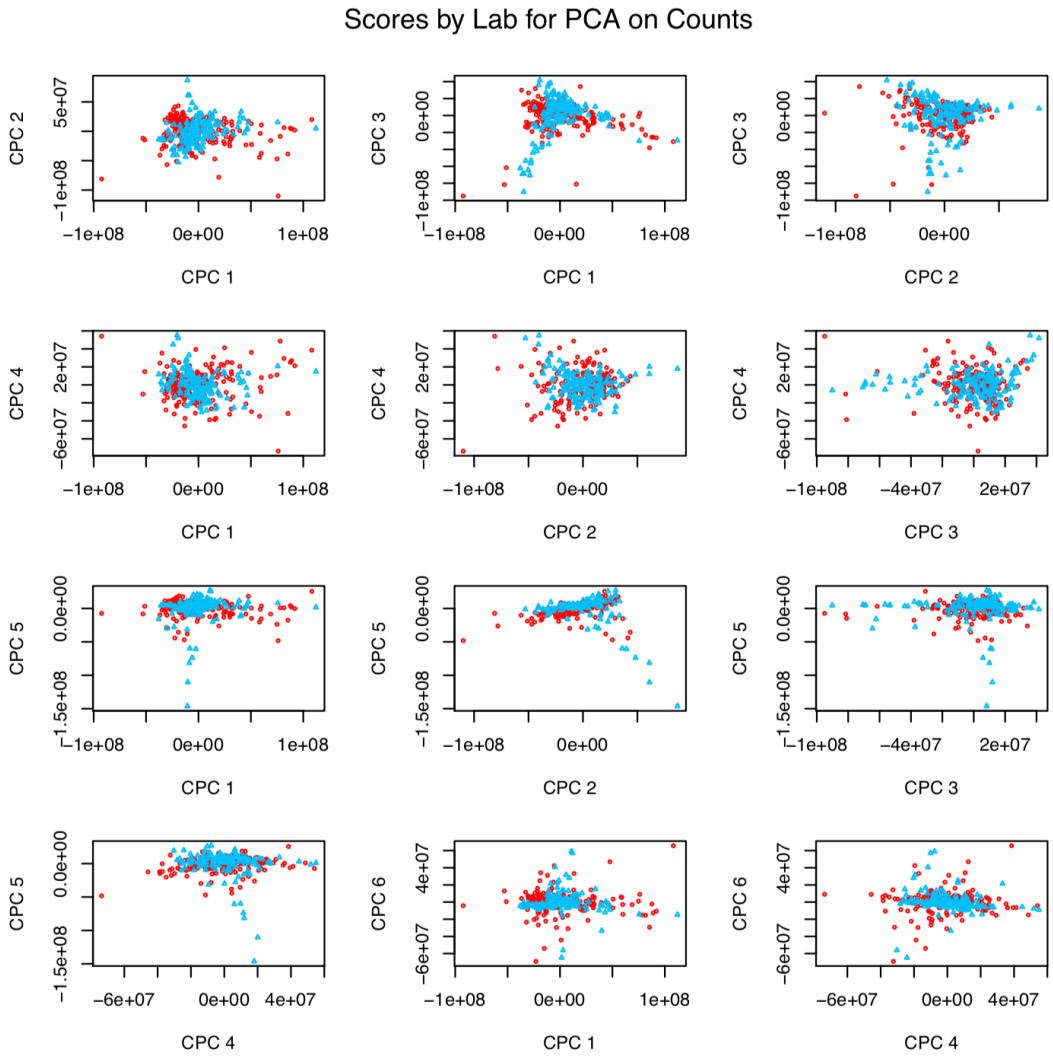

Figure A.40: Scores from PCA of counts by study of origin.

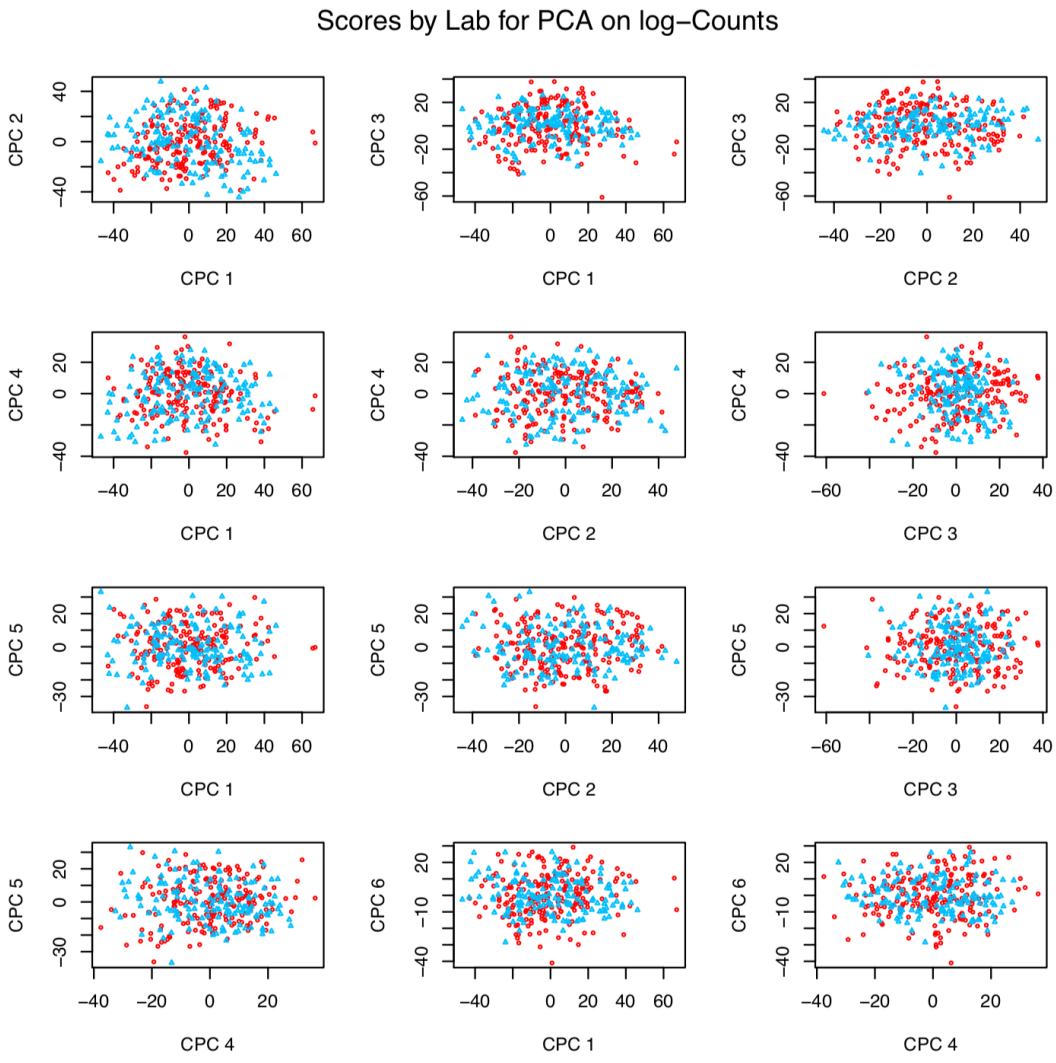

Figure A.41: Scores from PCA of log counts by study of origin.