Notes from meeting with Elizabeth feb 20:

Reorganize like

I Intro

II case study

A methods

B results

III Simulated data

A methods

B results

Intro:

1. Istead of increasingly multivariate data, increasing interest in infering processes from patterns. Want to test hypotheses where there are too many variables to do manipulative experiments.
2. Curse of dimensionality, so one traditional solution is to reduce dimensionality with something like PCA and then use PCs as predictors in modeling. But this is problematic because PCA is an unsupervised analysis and agnostic to your response variable.
3. Unsupervised vs. supervised. Give heuristic example like plant communities along elevational gradient. Doing PCA regression is asking “What plant species contribute most to the variation among sites, and do those species vary along elevation” Supervised analysis would answer the question “What species respond most strongly to elevation”.

Send monthly climate data to Elizabeth tonight/tomorrow

Send simulated data methods and results by sunday

1. Introduction
   1. There is an increasing need for multivariate statistics in ecology that can deal with the sorts of data we collect
   2. Supervised vs. unsupervised analyses answer different questions
   3. Supervised multivariate analyses used to be limited due to computing power
      1. “curse of dimensionality”, missing values, etc.
      2. Ecologists did a lot of PCA-regression type things
   4. But, PLS-R is built to handle multivariate ecological data and should be used over PCA regression if your question is appropriate
   5. We will demonstrate that PLS and PCA-R answer subtly, but importantly different questions giving different answers with the same data.
2. Methods
   1. A case study (mourning cloaks?)
   2. Simulated data scenarios
      1. Control—variables that contribute to variation also discriminate between groups
      2. Red Herring—variables that co-vary strongly discriminate weakly and those that co-vary weakly discriminate strongly
      3. Needle in a haystack—Few variables that discriminate between groups but do not co-vary at all within groups. Many variables co-vary but do not discriminate between groups.
      4. Null—some co-variation but not correlated to group membership.
   3. Software
      1. Wrote a package to simulate multivariate datasets
      2. All code and data available in supplementary
3. Results and Discussion
   1. Case Study
      1. Describe PCA and PLS
         1. %variation explained
         2. R2Y for PLS
         3. appearance of score plots
      2. Loadings
         1. Loading plots or table of loadings
         2. Describe differences
      3. Both find a pattern but answer slightly different questions and provide slightly different answers in terms of loadings.
   2. Simulated data
      1. PLS-DA and PCA regression give similar results under “control” scenario.
         1. Still important to note they are answering different questions. We constructed the data so the main axis of variation also explains differences between groups
         2. Interpreting PLS results
            1. R2Y
            2. Q2 or other cross-validation
            3. permutation testing
            4. VIP scores
      2. PLS-DA picks up on variables most important to discrimination
         1. Red herring scenario is most similar to case study above.
         2. PLS finds discriminating variables while PCA finds co-variation
      3. PLS-DA finds significant separation when PCA sometimes cannot
         1. Needle in a haystack PCA finds no separation between groups in first 2 PCs
         2. But PLS model is highly significant and shows strong separation
         3. Worley et al say to use PCA as validation for PLS models, but not a good idea always
      4. Best practices for PLS reporting
         1. Watch out! Even in the “null” scenario, PLS score plots can look promising.
         2. Always do cross-validation
         3. don’t plot non-significant models
         4. Don’t show score plots without some other reason besides “hey, look, the circles don’t overlap!). E.g. loading plot, more than two groups, etc.
4. Conclusion
   1. PCA regression type analyses may be appropriate in some cases, but not for “direct hypothesis testing”
   2. PLS is just one of several supervised multivariate techniques with a lot of advantages for typical ecological data
   3. But also check out variations on PLS and other things like RDA and PERMANOVA