**Distillation of figure 2 so far:**

Pt 1.

Cross-validation and SV component analysis (need better phrase) are more sensitive measures to detect number of dimensions/signal in the data.

Pt 2.

Detectability is influenced by the relative difference between (1) the contribution of *smallest* component of “signal” and (2) the contribution of the *largest* component of the “noise”.

— ways in which the signal can be larger/smaller

- spread of mutants/conditions

- number of true dimensions

-

— ways in which the noise can be larger/smaller

- measurement noise level itself

- correlation/spread of the measurement noise

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Some snippets of writing (in rough shape but potentially usable)**

To infer the number of fitness-relevant phenotypes affected by a collection of mutants, we use two related inference methods. The first is an explicit model of phenotypic evolution grounded in Fisher’s geometric model. [explain model]

The second inference method uses Singular Value Decomposition to reduce the fitness data observed into low-rank representations of that data. [explain more].

To assess the ability of these inference techniques to adequately infer fitness-relevant phenotypes, we simulate data according to a phenotype model and check how the inference matches the true underlying simulated data. In a set of simple simulations, where the optimum representing each condition is pulled from a uniform ball around the origin and the mutants are pulled from a uniform ball around the ancestor, we can test the ability to accurately predict dimensionality and fitness data, as a function of the number of phenotypes contributing to fitness for the collection of mutants and measurement error.

One challenge commonly faced for these types of inference techniques is knowing when to set the cutoff for the number of dimensions. Taking how well the model captures the data is prone to overfitting, since measurement error will confound the real signal. One commonly used metric to set a cutoff for Singular Value Decomposition is the “elbow” method. This involves plotting the predictability (R^2 here) as a function of the dimensionality. Once a clear break occurs (R1C1), you use this as the inferred number of dimensions. However, note that as measurement error increases, the elbow method becomes unable to capture the true dimensionality.

Another common method is the use of information criteria that explicitly impose a penalty for additional parameters, only selecting a more complex model if its explanatory power is significantly better than the penalty imposed by the additional parameters. Multiple information criteria have been proposed [cite the paper] in the context of Singular Value Decomposition. All of these criteria behave similarly to the “elbow” method - accurately inferring dimensionality when error is very low, but ultimately unable to for intermediate levels of error.

A third method of selecting the model cutoff is cross-validation. In this method, a portion of data is first used to “train” the model, and the remaining data is used as an independent test set to evaluate the fit of the model. If measurement error is truly random, once the model begins to fit measurement error instead of the true underlying signal signal, then prediction ability begins to decline. Thus the maximum value of prediction in the test set should identify the model that captures the most signal without overfitting to measurement error. This method has more power to detect the true dimensionality than the “elbow” method or using information criteria.

A fourth method infers dimensionality by comparing the magnitude of the singular values to the largest singular value generated by measurement error alone. For a given observed set of data, each data point is a combination of true signal and measurement error. Matrices that consist only of measurement error can have structure, and thus singular value decomposition can identify the structure in these random matrices [cite Sengupta and Mitra]. For the simple case where there is an identical level of measurement error for every datapoint, theory predicts the magnitude of the largest singular value, or contribution of the largest “dimension” that is only measurement error. This sets the limit of detectability for inference using singular value decomposition.

Our ability to detect the dimensionality of the data depends on the *contribution* of each dimension to the data that we observe. For a given spread of data (distances between mutants and conditions) and set level of measurement error, increasing the dimensionality of the decreases our ability to detect all of the dimensions, leading these methods to infer fewer dimensions.

[non-independent error? -> non-uniform error works (just average error is good enough)]

The ability to detect the correct number of dimensions depends the relative contribution of the last dimension to the fitness data and the contribution of the largest dimension from

Detection depends on: (1) contribution of each dimension, which depends on spread of mutants, conditions, # dimensions, (# mutants, # conditions).