Distillation of figure 2 so far:

Pt 1.

Cross-validation and SV component analysis (need better phrase for this) are more sensitive measures to detect number of dimensions/signal in the data. [not sure if this needs to be main text figure or if we can just say it and direct to the supplement]

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 magnitude of the signal component can change

- spread of mutants/conditions
- number of true dimensions
-

— ways in which magnitude of the noise component can change

- measurement noise level itself
- correlation/spread of the measurement noise

[somewhere, need to show how good predictions are]

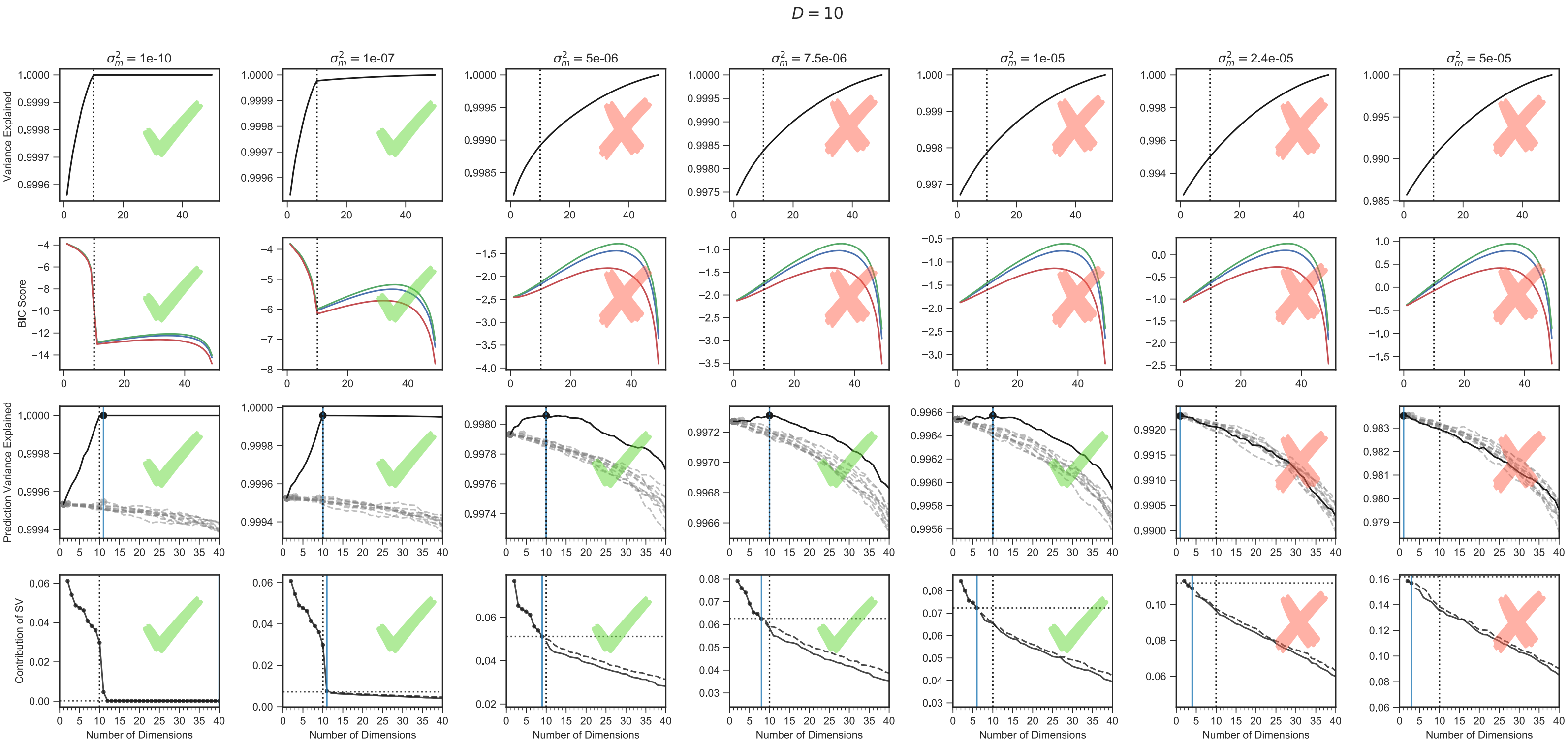
[Later figure will show that signal component also influenced by weird uncles + the like]

Alternative approach (that we could do/may be more straightforward):

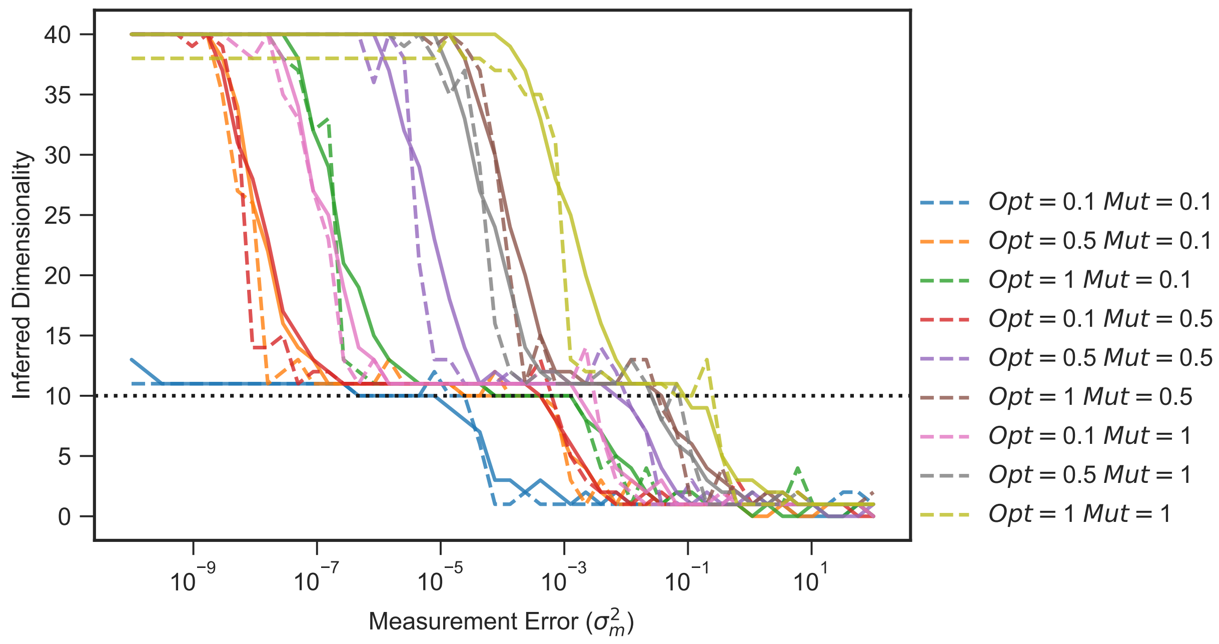
Have separate figures.

Fig 2. Fix signal - how can the noise change?

Fig 3. Fix noise - how can the signal change? [this would include everything here, as well as weird uncles, etc.]

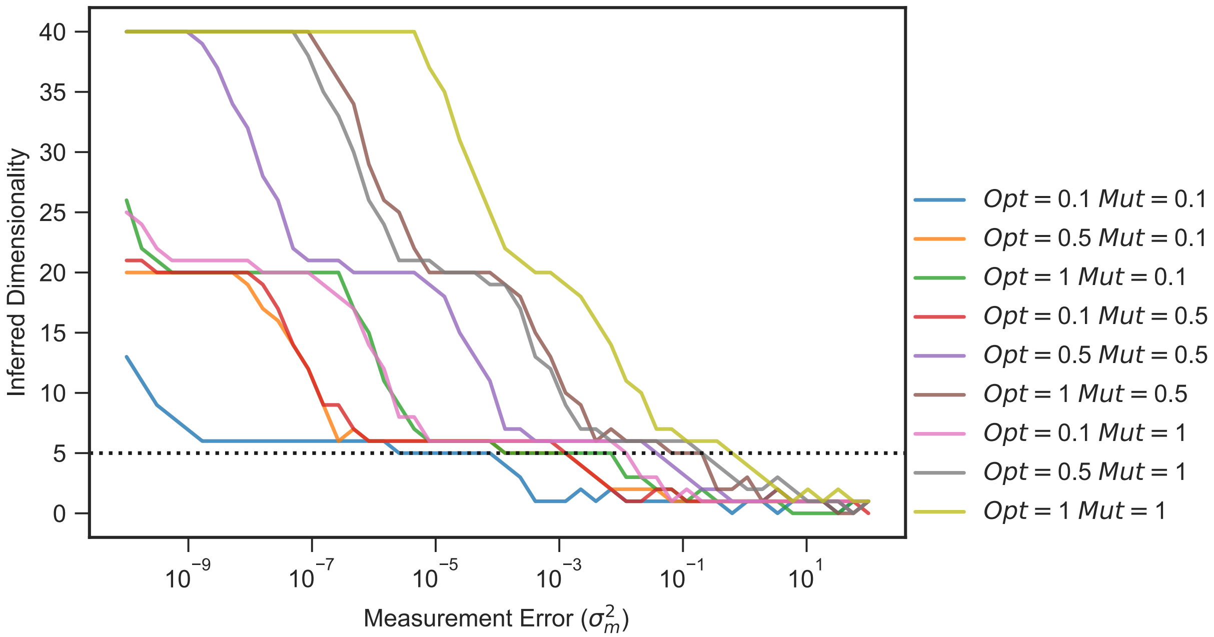


2a. Cross-validation and maximum noise singular value accurately predict at intermediate error levels. The ability to detect dimensions for 4 methods of inference: "elbow" method (row 1), information criteria (row 2), cross-validation (row 3), and maximum singular value with simulated noise (row 4) as a function of measurement error (columns). For extremely low measurement error, all four methods can detect the simulated number of dimensions. As error increases, "elbow" method and information criteria are unable to accurately detect, whereas cross-validation and SV method are fairly accurate (column 3). For high measurement error, all methods infer ~1 dimension. At intermediate levels of error, cross-validation and SV method predict an intermediate number of dimensions.

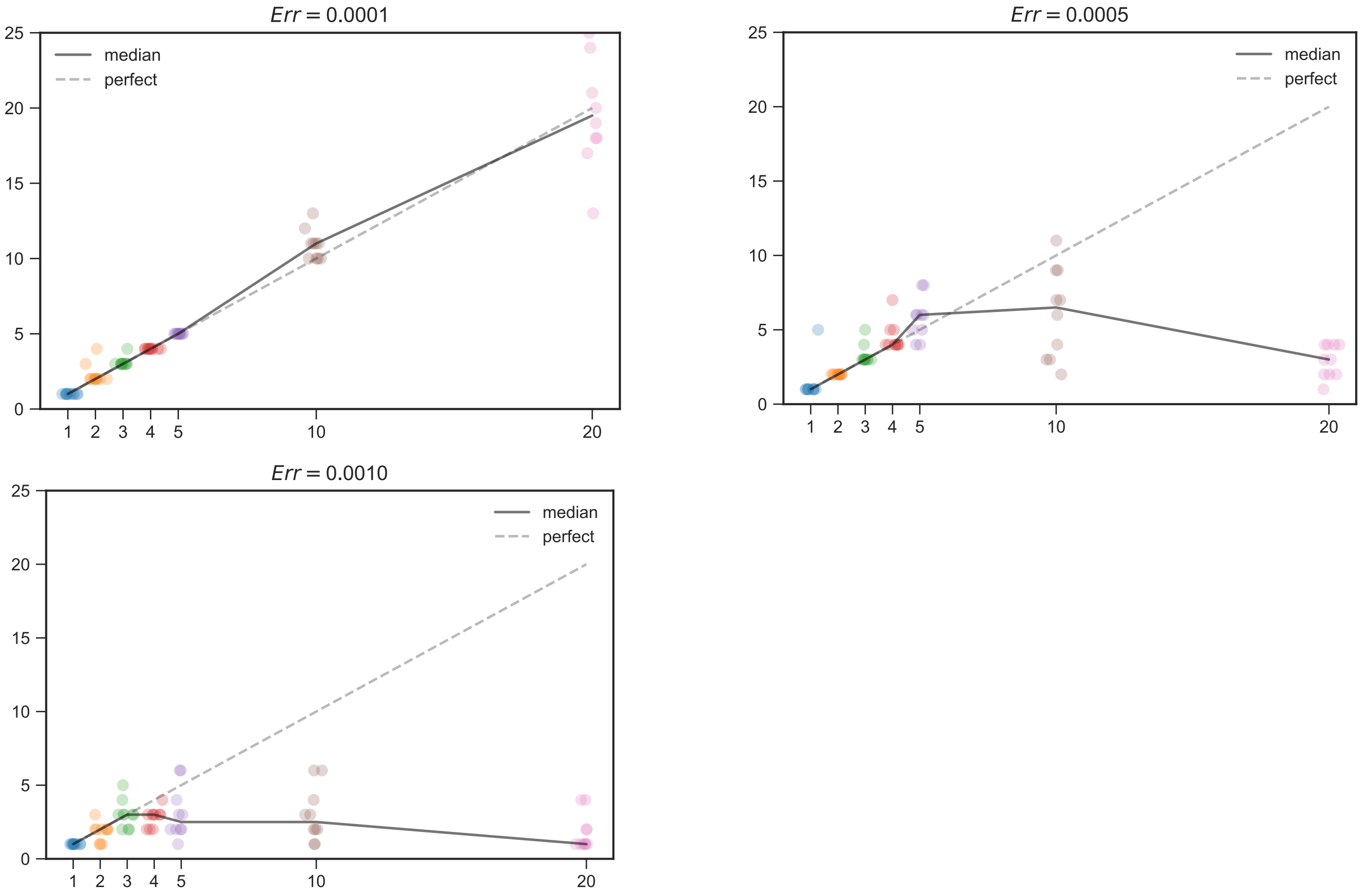


2b. Cross-validation and “SV” give similar levels of prediction AND Smaller “spreads” of conditions and mutants need more precise measurements. Dashed lines represent cross-validation inferred dimensionality, solid lines represent SV. When mutants and optima are more different from each other, the methods accurately predict the correct dimensionality with higher measurement error. 5 error regimes are shown: (1) “maxed out”, where the inferred dimensionality is the most possible given the number of conditions (and folds), (2) transition between “maxed out” and (3) accurate prediction, (4) “intermediate” between accurate prediction and (5) no information. [Not sure why (1) and (2) are there… my intuition is that only (3), (4), and (5) should exist - may be something to do with non-linear model and/or machine precision issues - though technically this is “inferred dimensions”, in all practical purposes it looks something like column 1 in previous figure.] In this particular model, spread of mutants and conditions are inter-changeable and affect critical measurement error values similarly.

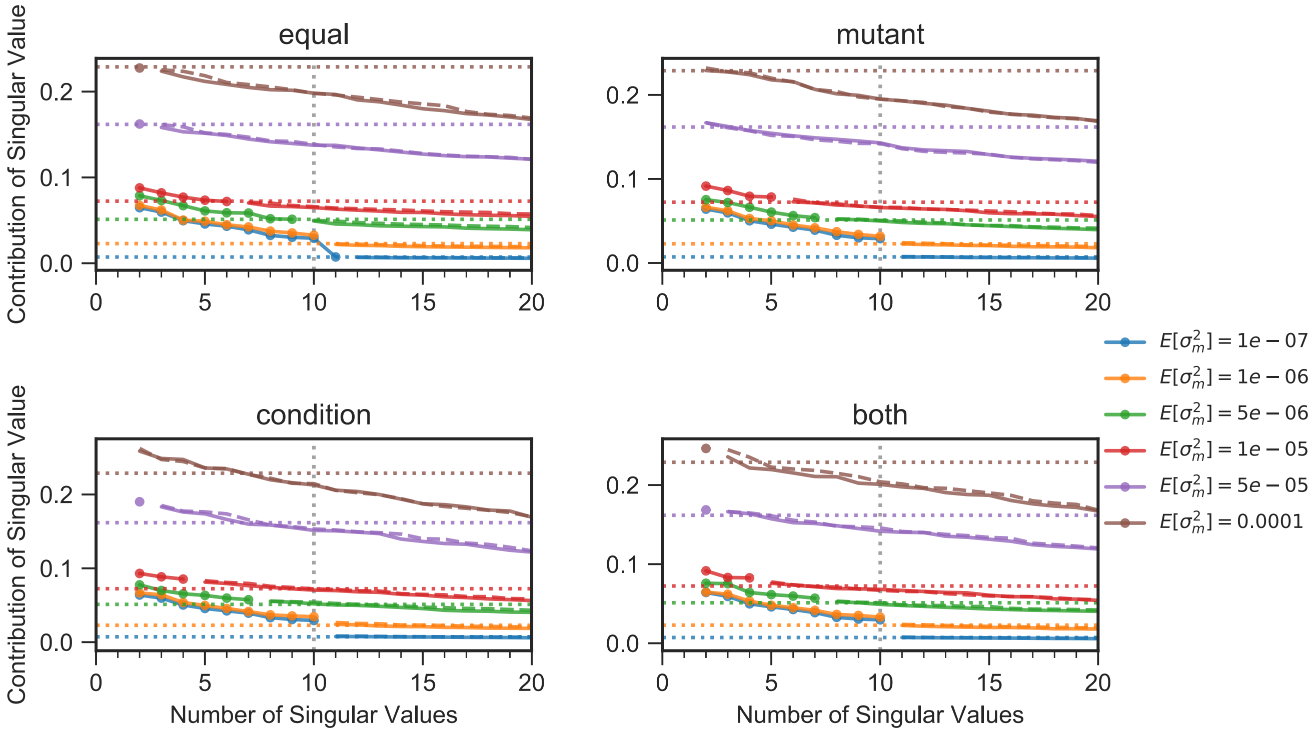
SEE FIGURE in “Supplemental Information” for explicit/detailed version of the spreads.

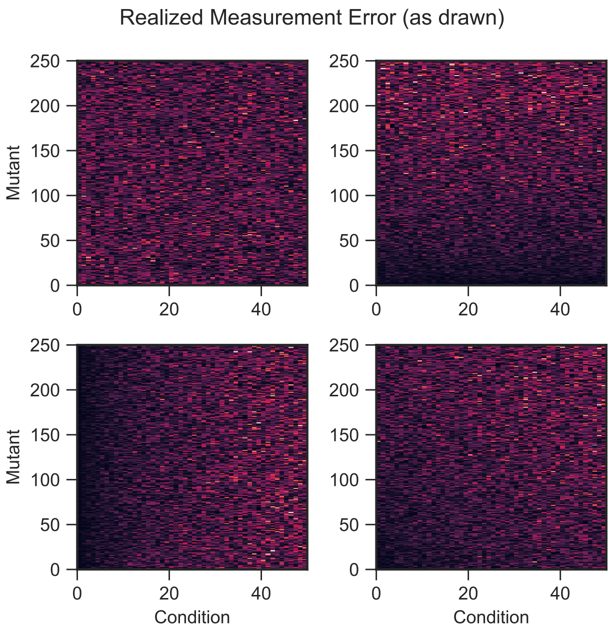


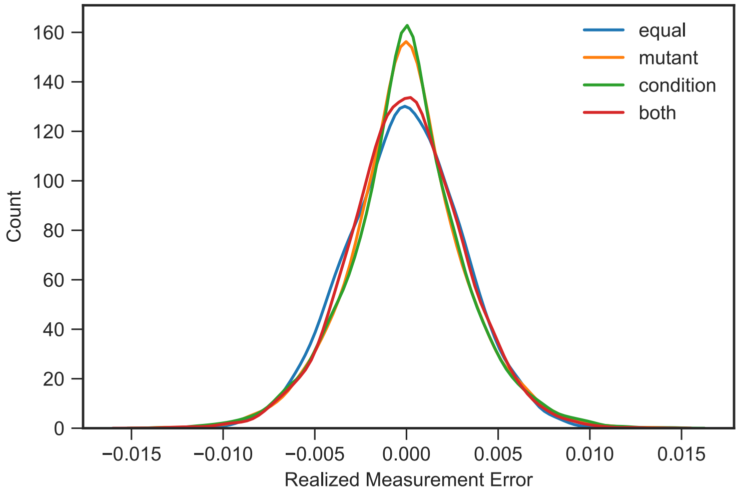
2c. For low dimensionality, SV shows weird behavior. Showing only SV here. Not sure exactly why, but when truly 5 dimensions, there's another intermediate error regime that predicts 20 dimensions. Maybe some non-linearities that are actually picked up? (One consideration with all of this is that I'm simulating with Gaussian fitness function which is non-linear and not "truly" 5 dimensional - whatever that means).



2d. High-dimensional data is more difficult to accurately infer. For a fixed spread of mutants and conditions, higher dimensions are more difficult to detect, and thus need more precise measurement (lower measurement error). This is because the spread of those mutants and conditions is spread equally (in these simulations) across all the dimensions, meaning the contribution of each dimension is much smaller and less likely to exceed largest contribution of noise.

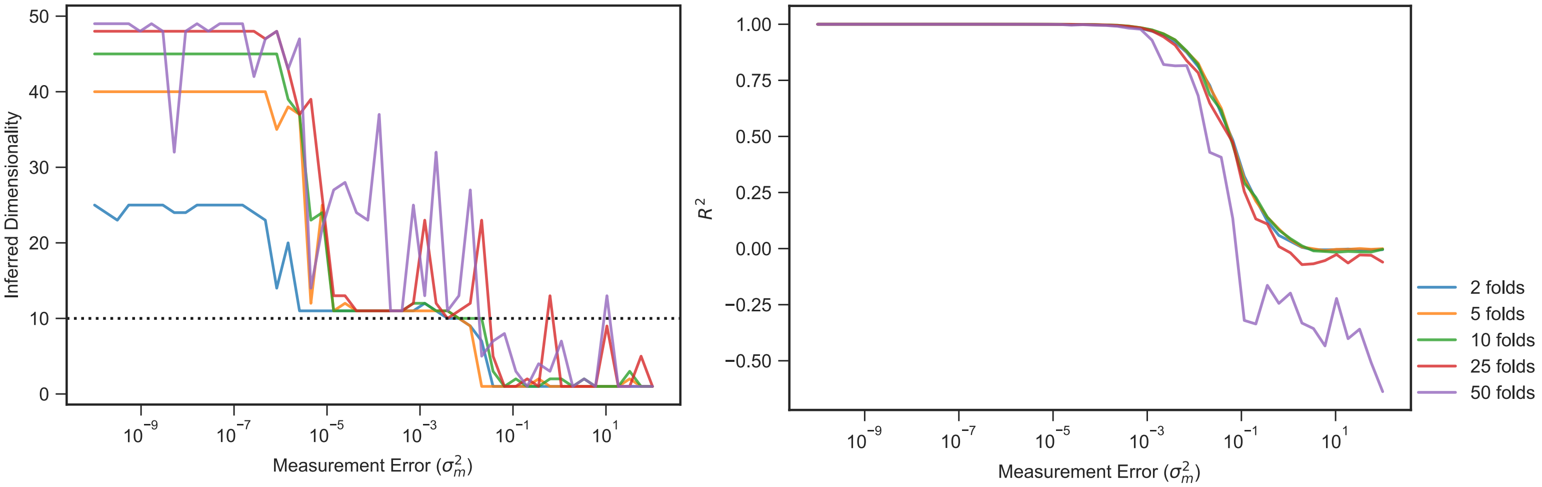




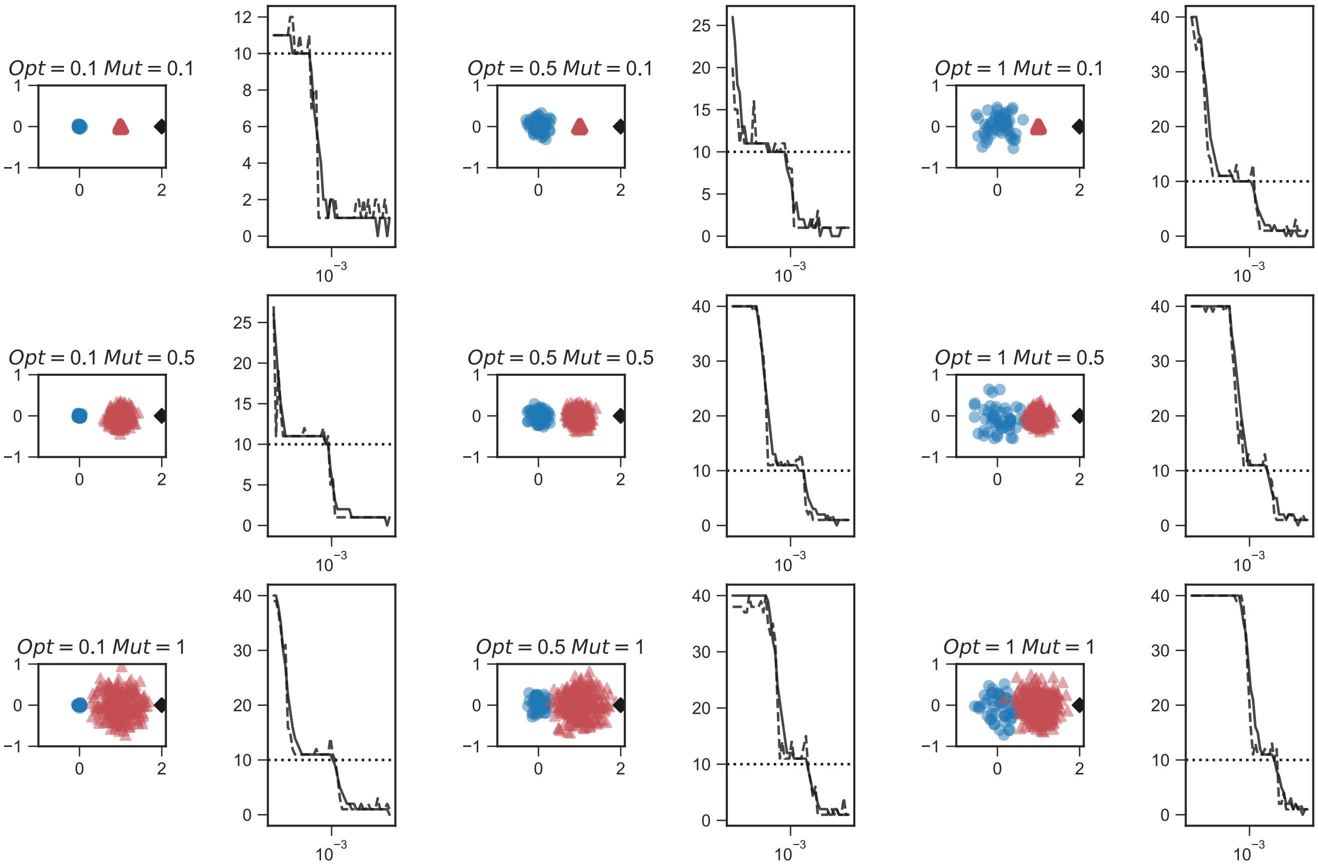


2e. Uneven measurement error makes detection more difficult. If there is uneven amount of measurement error around the data, then the largest component represented by measurement noise (figures B,C,D of top panel) is comparatively larger than the corresponding noise component with uniform amounts of error (figure A of top panel). However, the empirical amount of measurement noise (dashed line) still works well [this is kind of cheating, since I'm using the *exact* error here - maybe simulate many separate draws with "known" error values and find expected largest value?], even when theory doesn't apply to these cases. [need a way to express uncertainty in these types of figures - it will/does depend on the particulars of the actual simulated uncertainty] Second panel shows absolute value of realized measurement errors (actual drawn for each of the 4 scenarios. Third panel shows kernel density estimates for data in panel 2. [do we want to do this in a different way so that these kdes all look identical (same data), but we just change how they are distributed in the matrix to make a more fair comparison?]

Supplementary Information



SI 2a/b. Number of folds used in cross-validation does not matter. (A) Ability to detect number of dimensions and (B) prediction accuracy of those dimensions does not change substantially based on number of folds used. In general, fewer folds give more smooth inference and better predictions, though this is with simulated data that is roughly homogenous. For standard D=10 simulation. [why not just use 2 folds always? (I've been using 5 as default)] < (and maybe we should use just 2 folds)



SI 2c. Spread and predictability. Explicitly showing the spreads/predictability summarized in figure 2b.