Bias-variance analysis

Our study and others have shown that perturbation prediction evaluations can give different results depending on the choice of baseline and the choice of evaluation metric. In our view, many of these differences can be explained by conventional bias-variance tradeoffs. Below we provide bias-variance analyses under simple assumptions about perturbation effects. Our results show that:

- The full-data mean can yield lower MSE than the the controls-only mean even when estimating unperturbed expression.
- Models with no perturbation effect can yield lower predictive MSE than the true model even when predicting perturbation outcomes.
- MSE on a subset of highly variable genes, or on a subset of perturbation-responsive genes, or in a low-dimensional subspace, can select a different model than MSE computed across all genes.

Estimators for unperturbed gene expression

We first analyze estimators for unperturbed gene expression. A conventional estimator would be to include negative controls only. However, controls are often a small fraction of perturbation transcriptomics data, and perturbation effects are often weak or sparse, raising the possibility that even perturbed samples could help estimate control gene expression. Our study and others performing cross-validation on perturbation transcriptomics data have found especially strong predictive performance of the mean of the training data (including perturbed samples). Here we provide a brief bias-variance analysis. Due to a combination of weak effects, high noise, and low numbers of controls, the full-data mean can yield lower MSE than the the controls-only mean.

Assumptions

Consider a simple model of a single regulator that triggers a downstream transcriptional response. For gene j in sample i, denote the log-scale expression y_{ij}. Let x_{ij} indicate the activity level of a regulator of gene j in sample i. Suppose regulation is linear, so that $y_{ij} = \beta_j x_{ij} + \epsilon_{ij}$ where ϵ_{ij} is Gaussian with mean 0 and variance σ_j^2. Suppose that $x_{ij} = 0$ in control samples, $x_{ij} = 0$ in ineffective treatments (perturbations with no effect on gne j, and $x_{ij} = 1$ in effective treatments (perturbations with an effect on gne j). Suppose that the study includes n_c controls, n_{et} effective treatments, and n_{it} ineffective treatments. Controls will be listed first: if $i <= n_c$, the sample is a control sample. Let $n = n_c + n_{it} + n_{et}$.

Controls-only estimator

Suppose mean expression is estimated as $\sum_{i=1}^{n_c} y_{ij}/n_c$. The true mean is 0, so the error of the controls-only estimator is $\sum_{i=1}^{n_c} y_{ij}/n_c$, the bias of the controls-only estimator is 0, the variance of the

controls-only estimator is σ^2/n_c, and the MSE of the controls-only estimator is σ^2/n_c.

Full-data estimator

Suppose mean expression is estimated as $\sum_{i=1}^{n} y_{ij}/n$. The true mean is 0, so the error of the controls-only estimator is $\sum_{i=1}^{n} y_{ij}/n$, the bias of the full-data estimator is $n_{et}\beta_j/n$, the variance of the full-data estimator is σ^2/n, and the MSE of the full-data estimator is $\sigma^2/n + (n_{et}\beta_j/n)^2$.

Comparison

The full-data estimator has lower MSE when

$$\sigma^2/n + (n_{et}\beta_j/n)^2 < \sigma^2/n_c$$

. This simplifies to

$$(n_{et}\beta_j/n)^2 < \sigma^2(1/n_c - 1/n)$$

and then

$$n_{et}^2\beta_j^2/n^2 < \sigma^2(n - n_c)/nn_c$$

. For the datasets we study, $(n - n_c)/n \approx 1$, so the inequality is roughly

$$n_{et}^2\beta_j^2/n^2 < \sigma^2/n_c$$

. The final conclusion is that the full-data mean can yield lower MSE than the controls-only mean due to a combination of weak effects (low β_j), sparse effects (low n_{et}/n), high noise (σ^2), or low numbers of controls (n_c).

Bias-variance analysis of simple perturbation prediction models

To explain when different evaluation criteria would favor simpler or more complex models, we conduct a conventional bias-variance tradeoff analysis. This analysis demonstrates that biased, overly simple models can have lower MSE than the true model due to a combination of weak effects, limited biological variation, or strong noise. It shows that selecting highly-variable genes or genes that respond to perturbation, then calculating MSE on only those genes, might favor the correct model even when calculating MSE on all genes does not. A follow-up analysis shows that MSE on all genes can favor different models than MSE computed in low-dimensional space, either because of how genes are weighted or because of how well the perturbation effects align with the low-dimensional space. Low-dimensional summaries will be more likely to favor the **unbiased** model when the low-dimensional space gives more weight to highly-variable genes or perturbation-responsive genes. Low-dimensional summaries will be more likely to favor the **unbiased** model when the low-dimensional space aligns with the perturbation effects. Low-dimensional summaries will be more likely to favor the

biased model when the low-dimensional space gives more weight to genes with small effects and high noise, or when it is orthogonal to the perturbation effects.

Assumptions

Consider a simple model of a single regulator that triggers a downstream transcriptional response. For gene j in sample i, denote the log-scale expression y_{ij}. Let x_{ij} indicate the activity level of a regulator of gene j in sample i. Suppose regulation is linear, so that $y_{ij} = \beta_j x_{ij} + \epsilon_{ij}$ where ϵ_{ij} is Gaussian with mean 0 and variance σ_j^2.

Unbiased model MSE

Given data $y_j = [y_{1j}, ..., y_{nj}] \in \mathbb{R}^n$, $x_j = [x_{1j}, ..., x_{nj}] \in \mathbb{R}^n$, suppose β is estimated by least squares as

$$(x_j^T x_j)^{-1} x_j^T y_j$$

. The error in estimating β_j is

$$\epsilon_{\beta_j} = \beta - (x_j^T x_j)^{-1} x_j^T y_j$$

and it is Gaussian with mean 0 and variance $\sigma_j^2 (x_j^T x_j)^{-1}$. Given a new observation $y_{n+1,j}, x_{n+1,j}$, the prediction is

$$\hat{y} = x_{n+1,j} (x_j^T x_j)^{-1} x_j^T y_j$$

. The residual for the unbiased model is

$$
\begin{aligned}
r_j &= y_{n+1} - \hat{y}_j \\
r_j &= y_{n+1} - x_{n+1,j} (x_j^T x_j)^{-1} x_j^T y_j \\
r_j &= x_{n+1,j} \beta_j + \epsilon_{n+1} - x_{n+1,j} (x_j^T x_j)^{-1} x_j^T y_j \\
r_j &= x_{n+1,j} (\beta_j - (x_j^T x_j)^{-1} x_j^T y_j) + \epsilon_{n+1} \\
r_j &= x_{n+1,j} (\epsilon_{\beta_j}) + \epsilon_{n+1}
\end{aligned}
$$

The residual for the unbiased model has mean $E[r_j] = 0$ and variance $\sigma_j^2 + x_{n+1,j}^2 \sigma_j^2 (x_j^T x_j)^{-1}$. The MSE for the unbiased model equals the variance.

Biased model MSE

Consider also a biased model where β_j is always fixed to 0. Given a new observation $y_{n+1,j}, x_{n+1,j}$, the prediction is 0. The residual for the biased model is y_{n+1}. The residual variance for the biased model is σ_j^2 and the bias is $-x_{n+1,j} \beta$. The MSE is $\sigma_j^2 + x_{n+1,j}^2 \beta^2$.

Comparison

For genes where

$$x_{n+1,j}^2 \beta_j^2 < x_{n+1,j}^2 \sigma_j^2 (x_j^T x_j)^{-1}$$

,

the biased model will have lower MSE. This occurs when the effect β is very weak, the noise σ_j^2 is very strong, or the training-data variance in regulator activity levels $x^T x$ is low. Selecting the top 20 or top 100 genes with strong responses to perturbation is selecting for large β. Selecting 1,000 genes with high dispersion is selecting for large $x_j^T x_j$ and small σ_j^2. Either of these changes could favor the unbiased model.

Thus, a conventional bias-variance tradeoff explains why using MSE as an error metric could favor simple models even when they are biologically wrong. It also explains how certain types of gene selection could yield different conclusions about which model performs best.

Summaries of transcriptome state

Some methods, notably CellOracle and GeneFormer, have been demonstrated primarily for predictions of overall cell state or cell type rather than gene-by-gene fold change. Consider a generative model with $G = 1,000$ genes, and suppose they can be modeled independently. Suppose these genes are mapped into a one-dimensional summary, as in a simple pseudotime analysis. Suppose the mapping $U \in \mathbb{R}^{G,1}$ is fixed, known, and linear. Suppose the MSE is calculated on Uy instead of y. Then the MSE is defined as

$$E[(Uy - U\hat{y})^2]$$

.

Unbiased model MSE

The residual for the unbiased model is $\sum_{j=1}^{1,000} U_j r_j$. The bias for the unbiased model remains 0 (by linearity). The predictive variance for the unbiased model is

$$\sum_j U_j^2 (\sigma_j^2 + x_{n+1,j}^2 \sigma_j^2 (x_j^T x_j)^{-1})$$

and the MSE for the unbiased model equals the predictive variance.

Biased model MSE

The residual for the biased model is $\sum_{j=1}^{1,000} U_j y_j$. The bias for the biased model is $-\sum_j U_j x_{n+1,j} \beta_j$. The residual variance for the biased model is $\sum_j U_j^2 \sigma_j^2$. The MSE is $(\sum_j U_j x_{n+1,j} \beta_j)^2 + \sum_j U_j^2 \sigma_j^2$.

Comparison

The biased model will have lower MSE when

$$(\sum_j U_j x_{n+1,j} \beta_j)^2 < \sum_j U_j^2 x_{n+1,j}^2 \sigma_j^2 (x_j^T x_j)^{-1}$$

. This displays many of the same effects discussed above: small β_j's, small $x_j^T x_j$, and large σ_j will favor the biased model. However, MSE after projection by U is not equivalent to MSE for two reasons. First, the values of U could increase or decrease the weight assigned to specific genes. Second, comparing to the inequality above, which was

$$x_{n+1,j} \beta_j^2 < x_{n+1,j}^2 \sigma_j^2 (x_j^T x_j)^{-1}$$

,

, the MSE after projection by U is also not equivalent to a linear combination of both sides with coefficients U_j^2. A simple linear combination would yield

$$\sum_j U_j^2 x_{n+1,j}^2 \beta_j^2 < \sum_j x_{n+1,j}^2 \sigma_j^2 x_j^T x_j^{-1}$$

whereas the actual calculation yields

$$(\sum_j U_j x_{n+1,j} \beta_j)^2 < \sum_j x_{n+1,j}^2 \sigma_j^2 x_j^T x_j^{-1}$$

. The right-hand side (unbiased model) is equal, but the left-hand side (biased model) differs. In some cases, the actual calculation yields a higher left-hand side than the reweighting of per-gene MSE by U_j^2. For example, this happens when $U_j x_{n+1,j} \beta_j$ is positive for all j or negative for all j. In other cases, the the actual calculation yields a lower left-hand side than the reweighting of per-gene MSE by U_j^2. For example, this happens when $\sum_j U_j x_{n+1,j} \beta_j = 0$ and any individual terms of the sum are not zero: in other words, when U is orthogonal to the perturbation effects.

To summarize, MSE can favor different models than MSE computed in low-dimensional space because of how genes are weighted. There are also scenarios where even weighted MSE designed to mimic MSE computed in low-dimensional space can favor different models than MSE computed in low-dimensional space. Low-dimensional summaries will be more likely to favor the **unbiased** model

when the low-dimensional space aligns with the perturbation effects. Low-dimensional summaries will be more likely to favor the **biased** model when the low-dimensional space is orthogonal to the perturbation effects.