The Best-Basis algorithm for linear regressions, a quantitative improvement over stepwise regression and its applications.

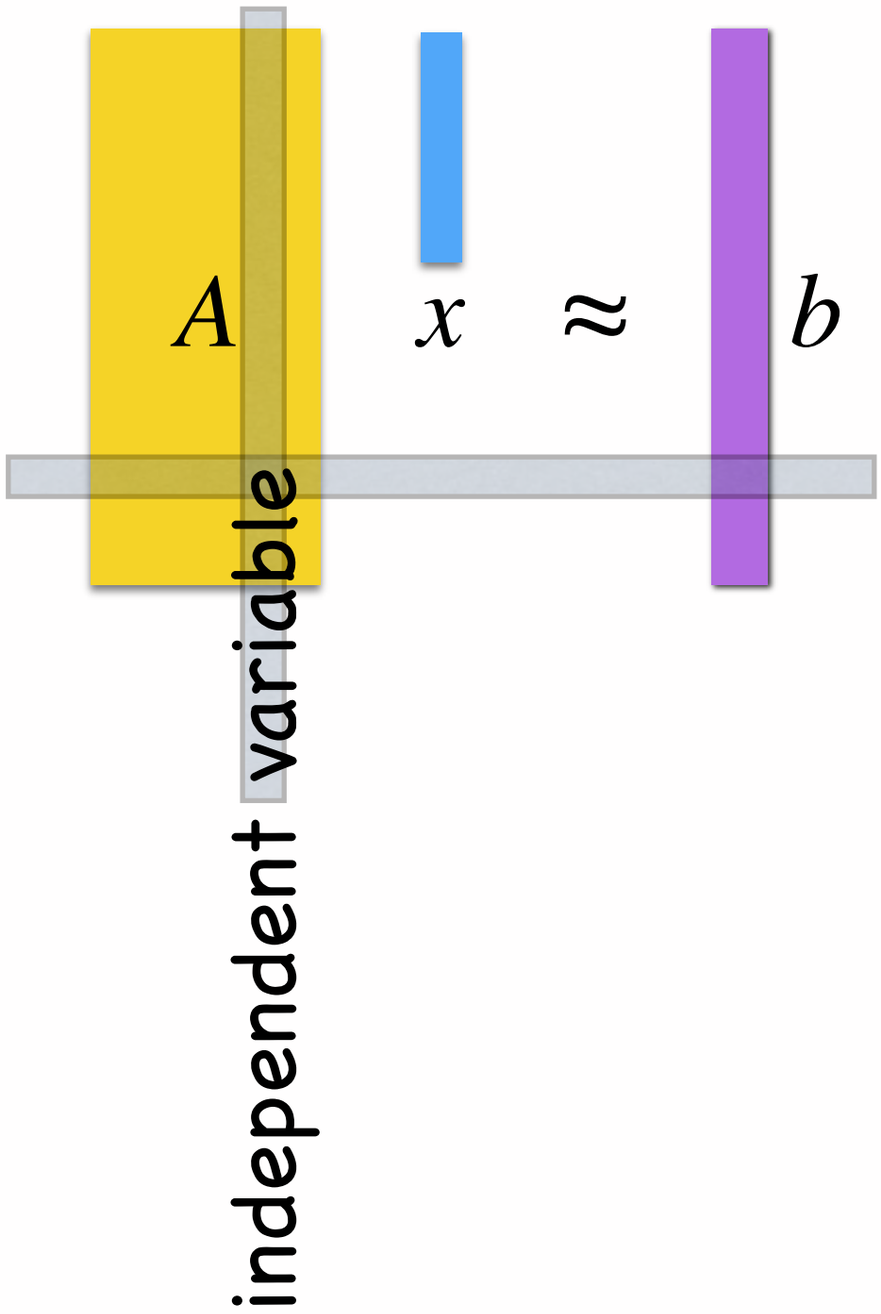
Gaston H. Gonnet
Informatik, ETH Zurich

Fusion 2014
Wednesday July 9th, 2014
Salamanca, Spain

Outline

- Comparison of Least Squares, SVD/PCA, Stepwise regression, Best Basis
- Approximate algorithms for Best Basis.
- Example I: reducing the number of independent variables
- Example II: Confidence analysis by adding random variables
- Final thoughts

Least squares regression



$$Ax \approx b$$

$$\|Ax - b\| \quad \text{is min}$$

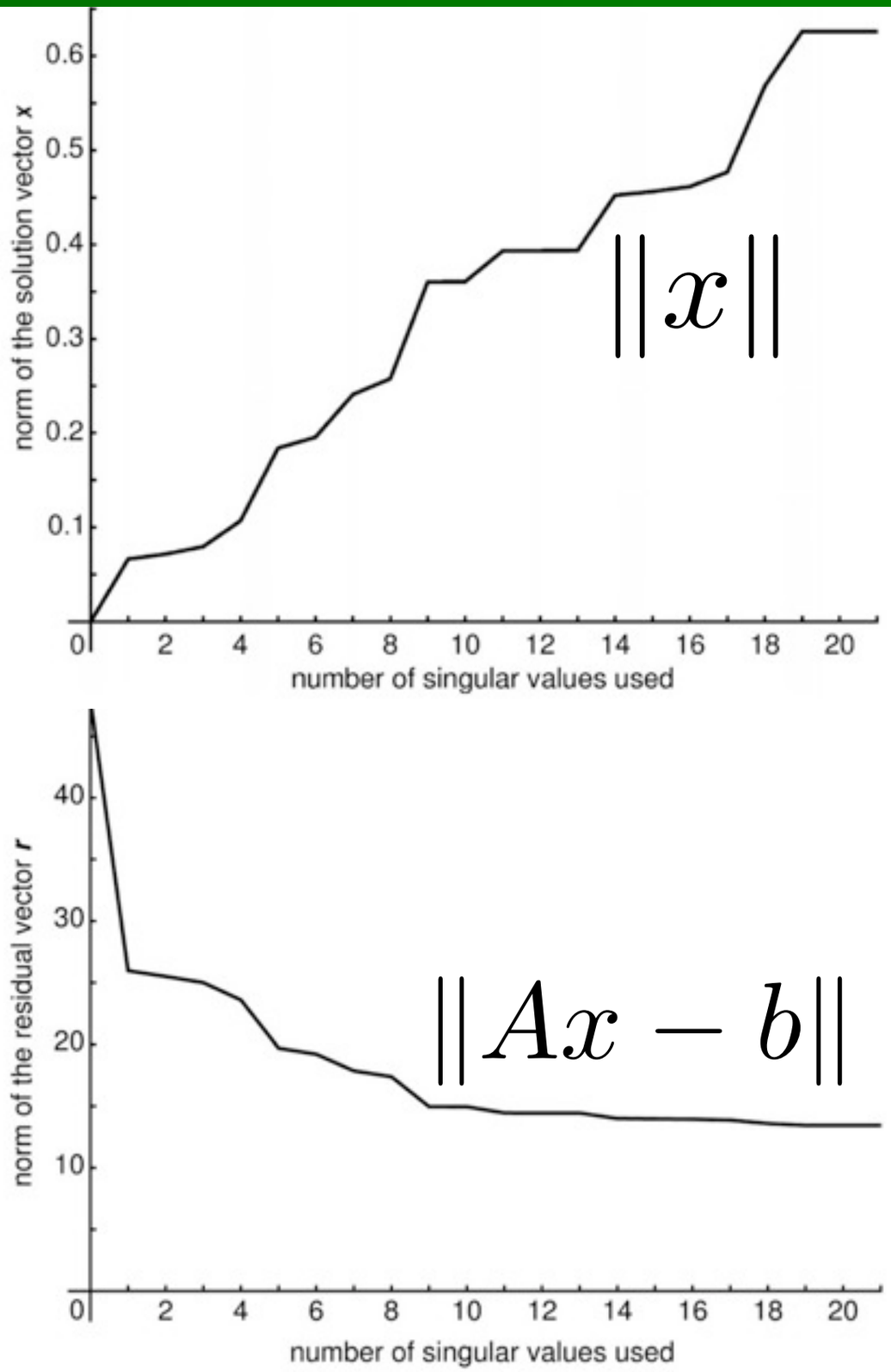
SVD vs PCA vs Best Basis

SVD (Singular Value Decomposition) allows us to solve the Least Squares regression at various tradeoff points.

The tradeoff is between the norm of x and the norm of Ax-b

$$\|x\| \;\; vs. \;\; \|Ax - b\|$$

SVD tradeoff points



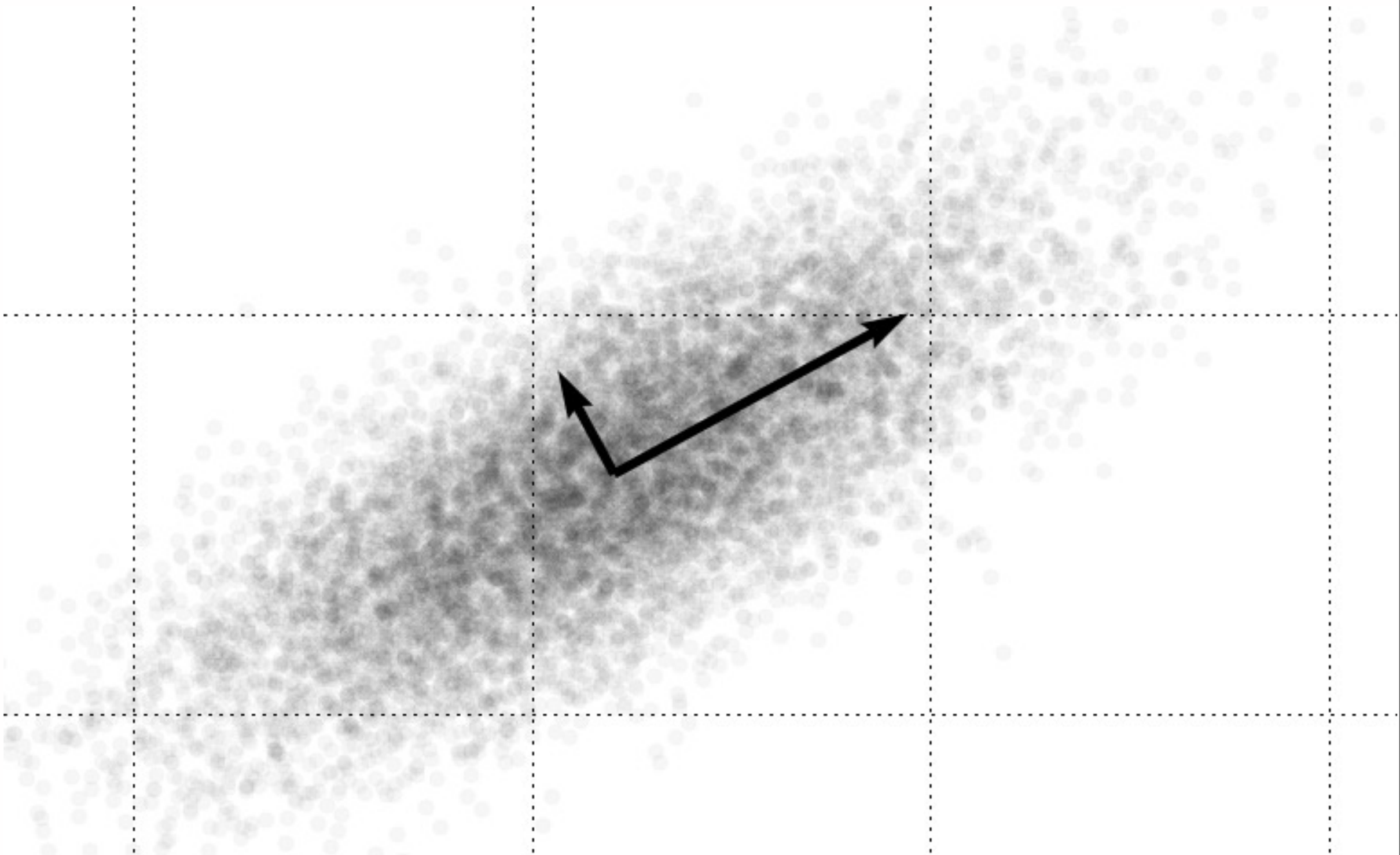
SVD tradeoff points

The tradeoff points are determined by the eigenvalue/eigenvector decomposition of:

$$A^T A = W \Lambda W^{-1}$$

and hence are independent of b

SVD example, n=2



SVD and PCA

PCA is identical to SVD over a matrix A with columns normalized to have average 0.

Record number of re-discoveries/versions? 15: principal component analysis, discrete Karhunen–Loève transform, Hotelling transform, proper orthogonal decomposition, singular value decomposition, eigenvalue decomposition, factor analysis, Eckart–Young theorem, Schmidt–Mirsky theorem, empirical orthogonal functions, empirical eigenfunction decomposition, empirical component analysis, quasiharmonic modes, spectral decomposition and empirical modal analysis.

Problems with SVD/PCA

The tradeoff points are independent of b

The results are dependent on scaling of the columns of A

Solutions x_i are seldom 0 or very close to 0 even when the column is irrelevant.

Problems with SVD/PCA (II)

The solution is often non-intuitive, e.g.

$$x = (1.2, -0.7, -0.09, 0.13, ...)$$

If new points are to be estimated, and the independent columns are the results of experiments, we really want to minimize the number of independent columns needed.

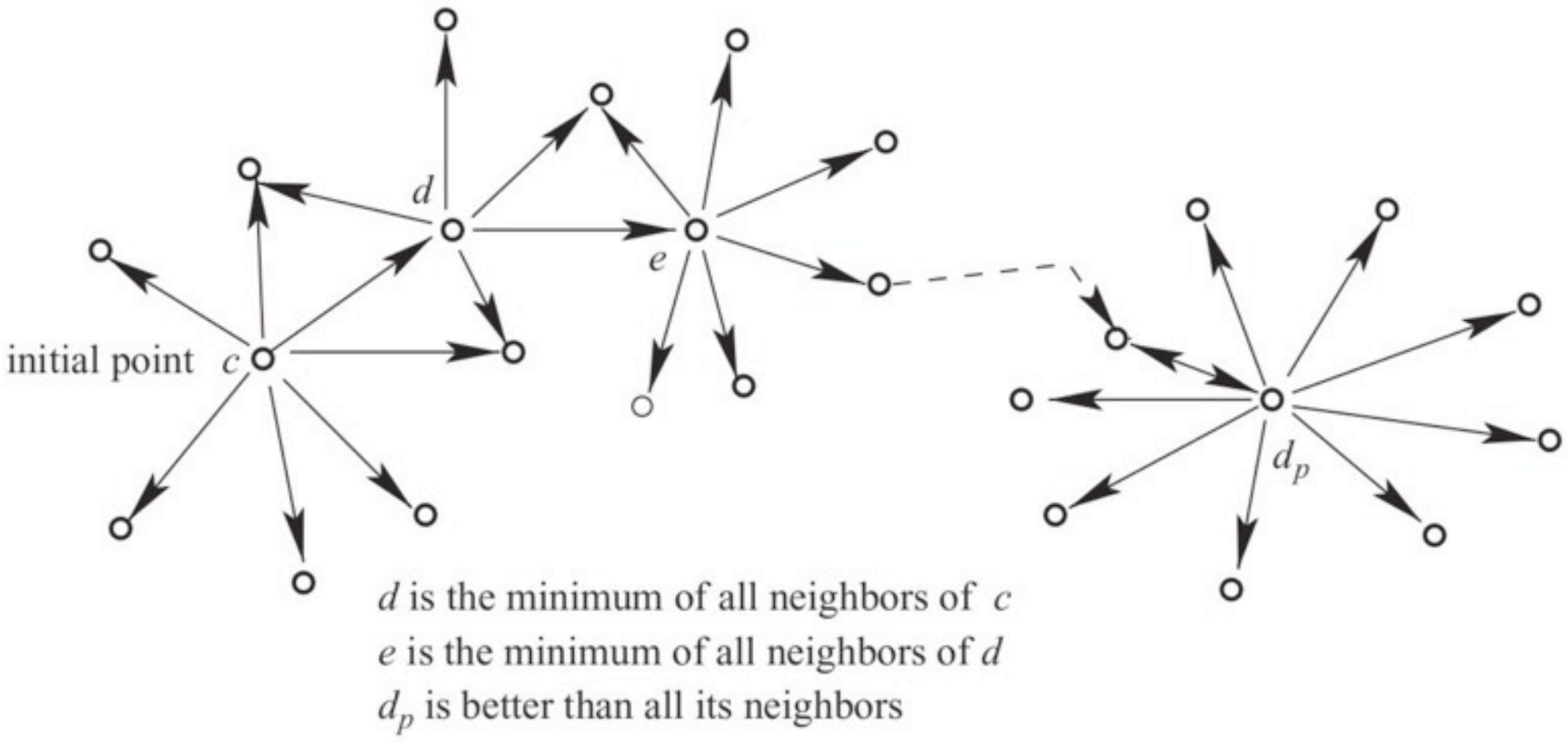
Best Basis and Stepwise regression

Stepwise regression (SR) and Best Basis (BB) compute a vector x where only k entries are non-zero which minimizes ||Ax-b||

SR uses a greedy algorithm starting from an empty solution (forward) or from the full solution (backward)

Best Basis algorithm

BB approximates the optimal solution using CHE (Constructor-Heuristic-Evaluator) methods. It normally finds solutions much better than SR.



Schematic diagram of the steepest descent algorithm.

Neighbouring solutions

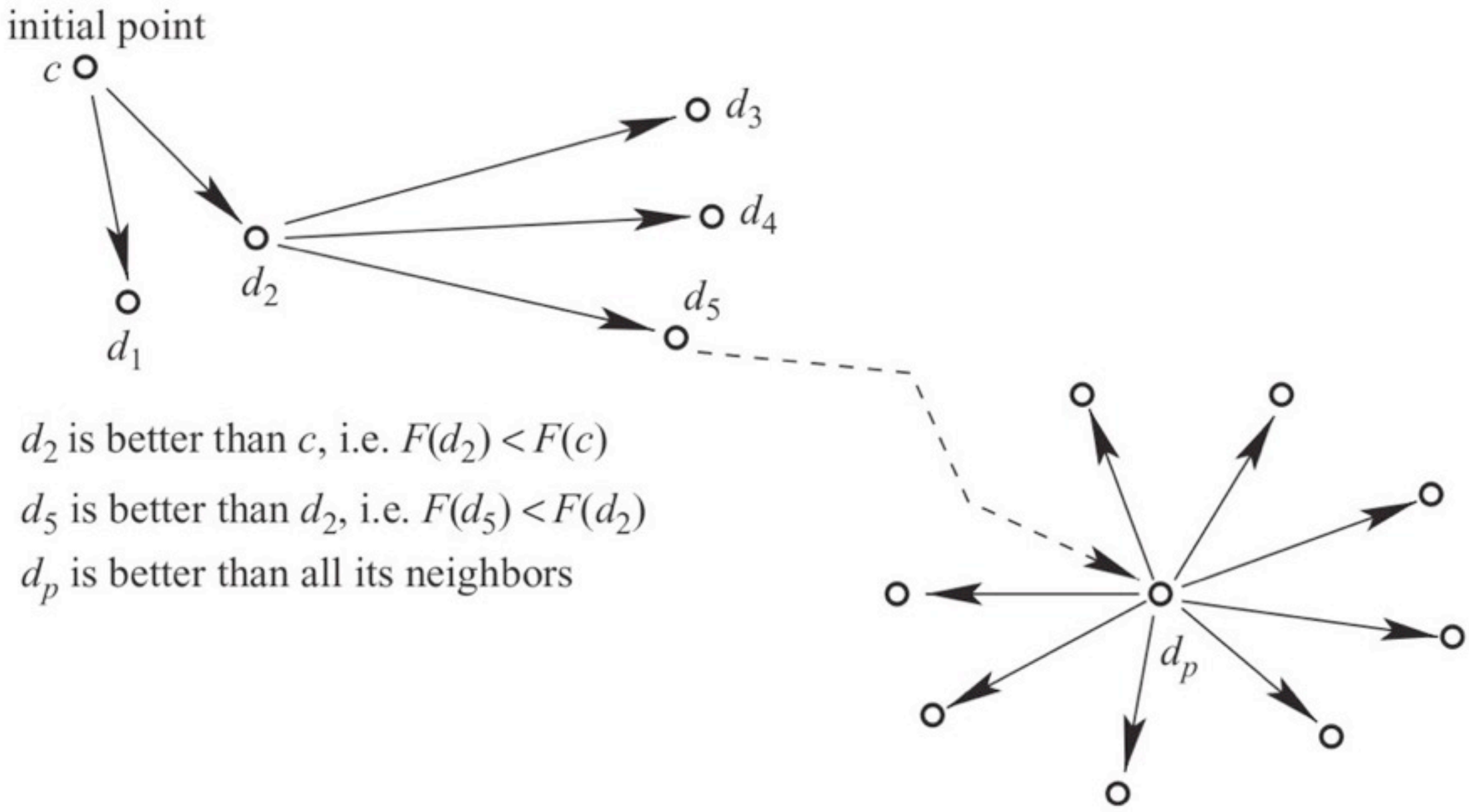
For the BB algorithm, two candidate solutions are neighbours if they differ in only one independent variable.

$$k = 3, \;\; n = 10$$

$$(x_2, x_5, x_9) \;\text{ and }\; (x_1, x_2, x_9)$$

are neighbours. Each solution has k(n-k) neighbours.

Best Basis, EA algorithm



d_2 is better than c, i.e. $F(d_2) < F(c)$

d_5 is better than d_2, i.e. $F(d_5) < F(d_2)$

d_p is better than all its neighbors

Schematic diagram of the early abort algorithm.

Best Basis, key ingredients

- Discrete Steepest Descent / Early abort
- Computation of norm of residuals of neighbour solutions in $O(k^2)$ time
- Remembering previous optimization points
- Parallel execution

Best Basis, efficiency

Consequently we can routinely solve problems with hundreds of columns, tens of selected columns, and unlimited number of rows (the number of rows does not affect the BB computation)

Best Basis, efficiency

BB uses:

$$A^T A$$

$$A^T b$$

$$b^T b$$

Best Basis, purpose

What is the use of this?

BB discovers which are the k most significant/ relevant independent variables for a particular problem.

Best Basis, general dependencies

Dependency does not mean correlation. True.

But most non-linear dependencies will still give significant regression values.

Some examples with a random A matrix and a vector b which is non-linearly computed from two columns of A. (n=20, m=200)

Best Basis, examples general dependencies

arctan $b_i = \tan^{-1} A_{i,3}/A_{i,13}$

```
Results of Best Basis
2000 data points, 20 independent variables, 2 used here
Norms of:     raw data    residuals
                   908       618.65
2 singular values used:  1901  1992
        variable       coeff/stdev    norm decrease
               3   0.3869 +- 0.0127     287.3
              13  -0.0397 +- 0.0125     3.116
```

Shifted product $b_i = A_{i,3}A_{i,13} + 0.1$

```
Results of Best Basis
2000 data points, 20 independent variables, 2 used here
Norms of:     raw data    residuals
                 353.65       21.98
2 singular values used:  166.1  1179
        variable       coeff/stdev    norm decrease
               3   0.3765 +- 0.0061     41.37
              13   0.3735 +- 0.0062     40.54
```

Two important applications:

Two important applications:

Methods which suffer from additional dimensions

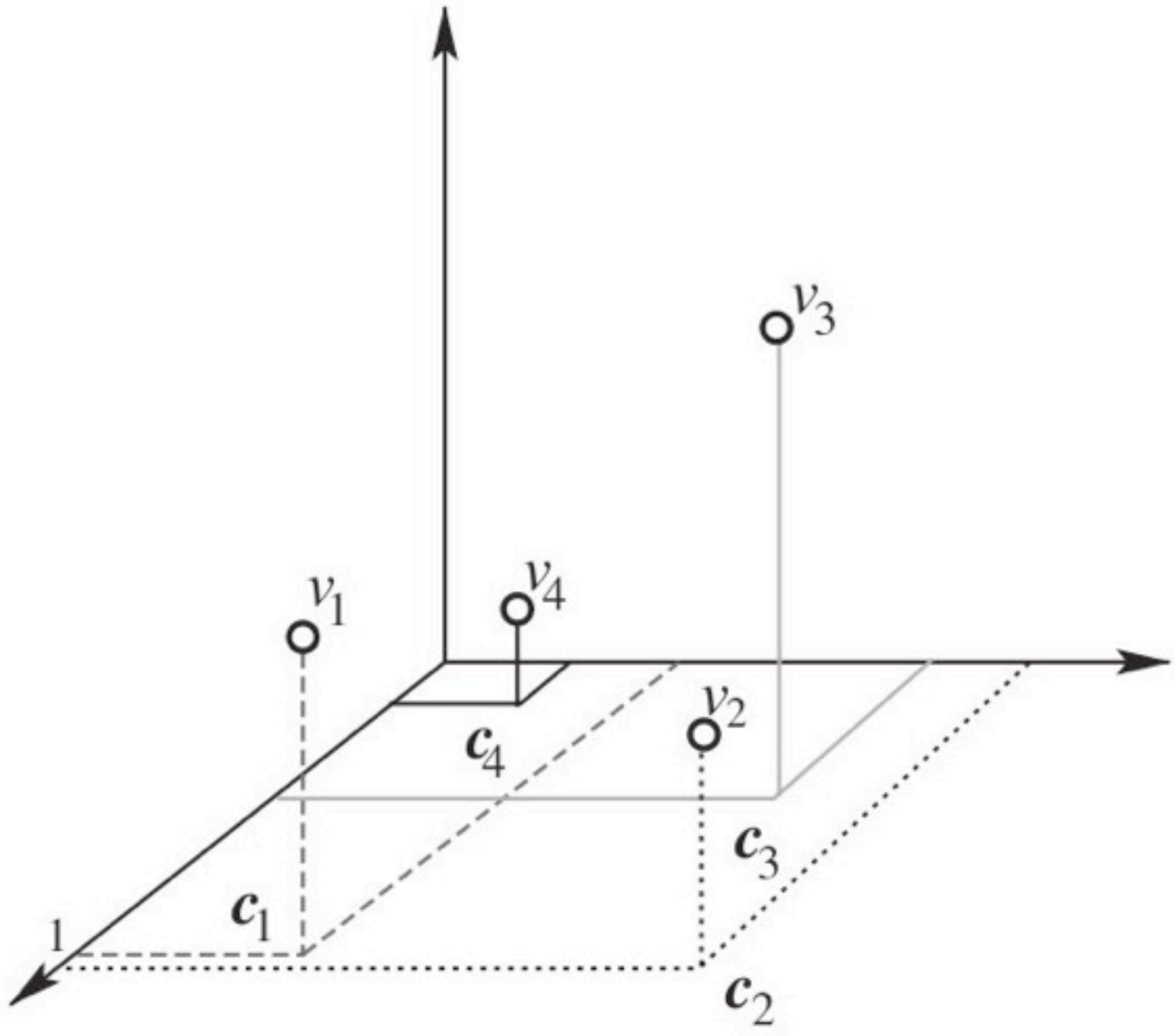
Two important applications:

Methods which suffer from additional dimensions

Confidence analysis by random columns

Nearest Neighbour

Nearest neighbour methods, in any of their variants, suffer from unrelated independent variables



Filtering independent variables

Filtering independent variables

Best Basis can be used to select the k most promising variables for more sophisticated methods (e.g. nearest neighbours, random trees)

Filtering independent variables

Best Basis can be used to select the k most promising variables for more sophisticated methods (e.g. nearest neighbours, random trees)

We can include every imaginable independent variable and let BB narrow them down.

Two important applications:

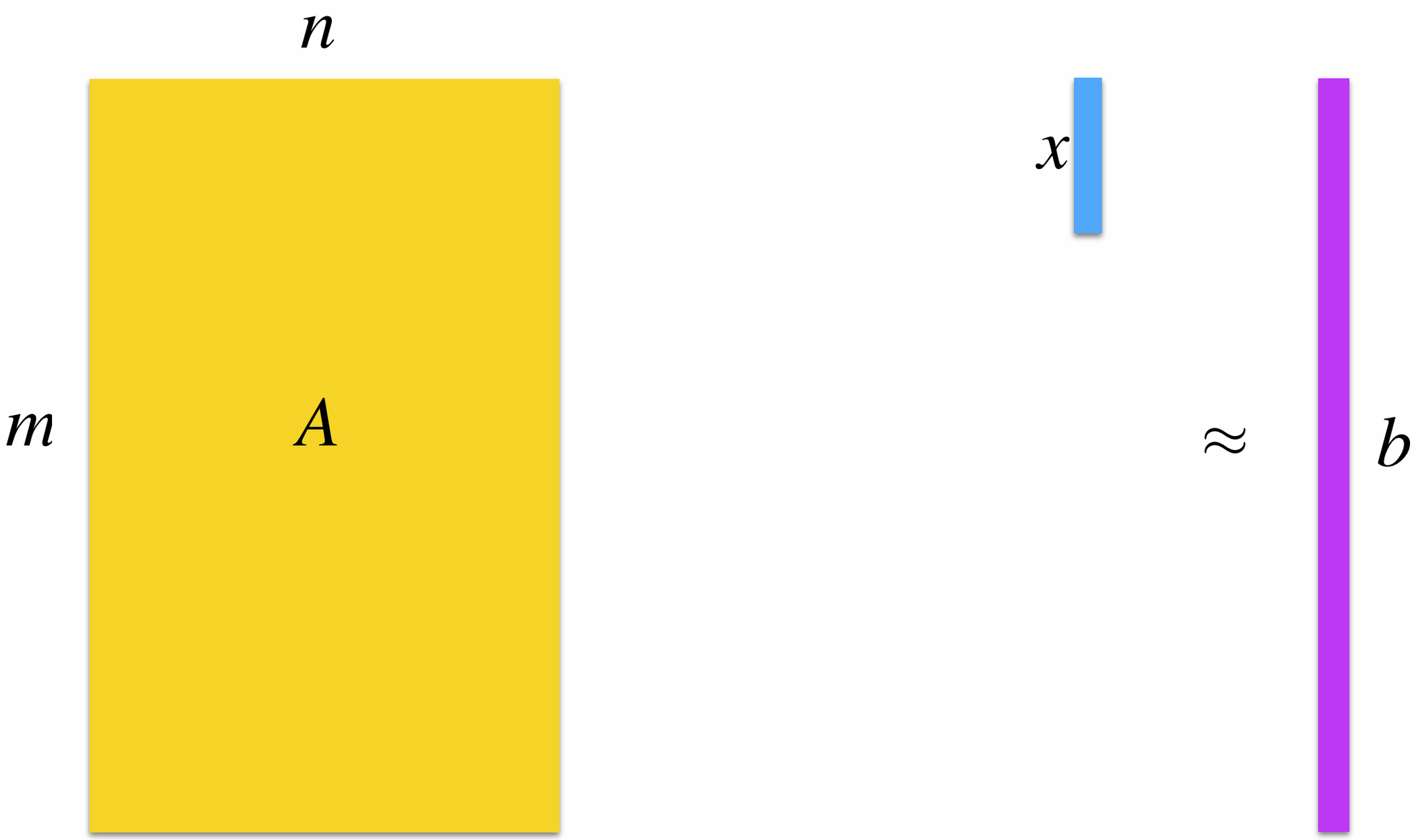
Confidence analysis by random columns

Validation by random columns

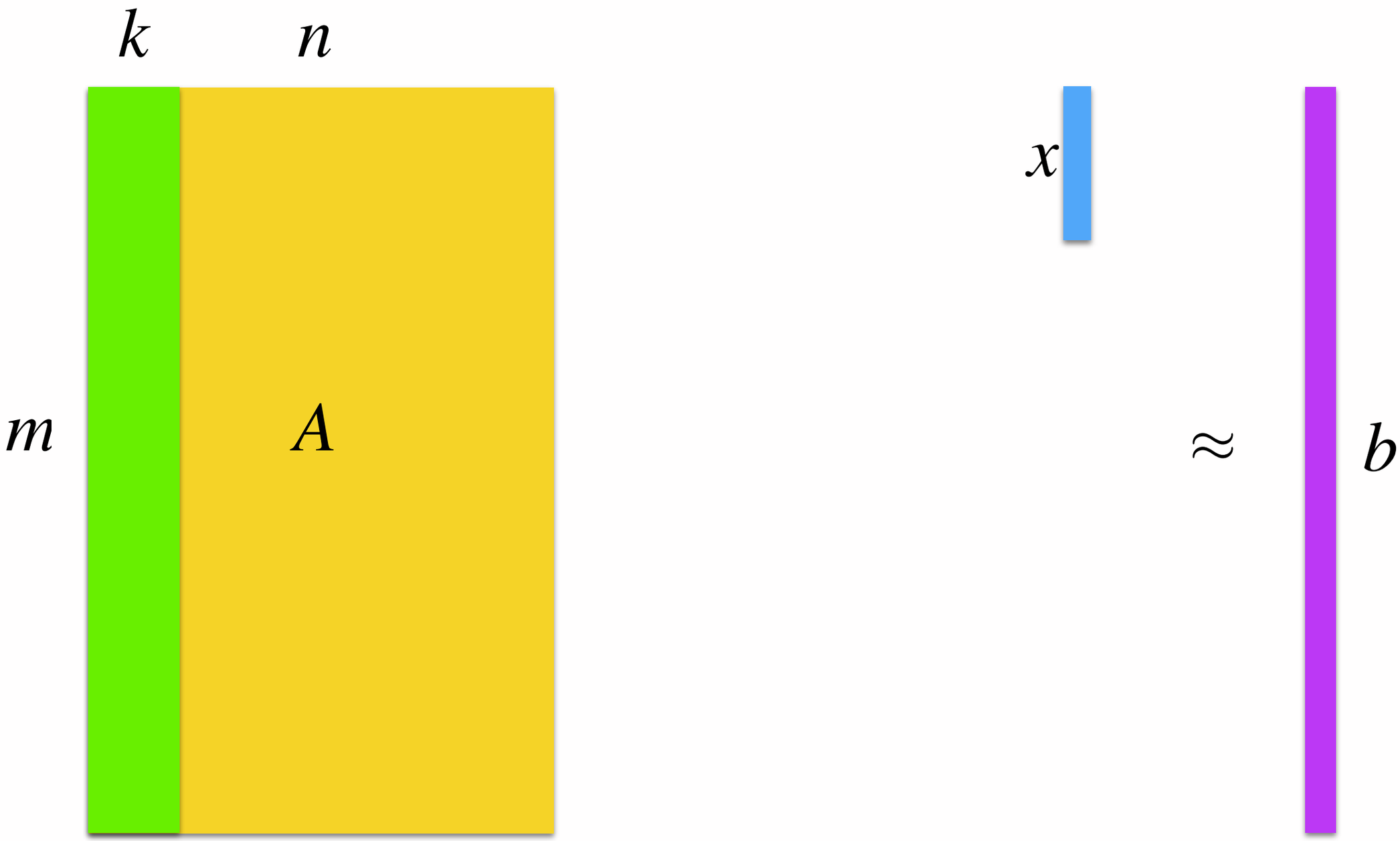
For regressions or problems where the regression shows a strong signal, the BB algorithm allows an elegant validation technique

Suppose we have found a subset of k ($x_1..x_k$) independent variables out of n, which appear to be significant in explaining b.

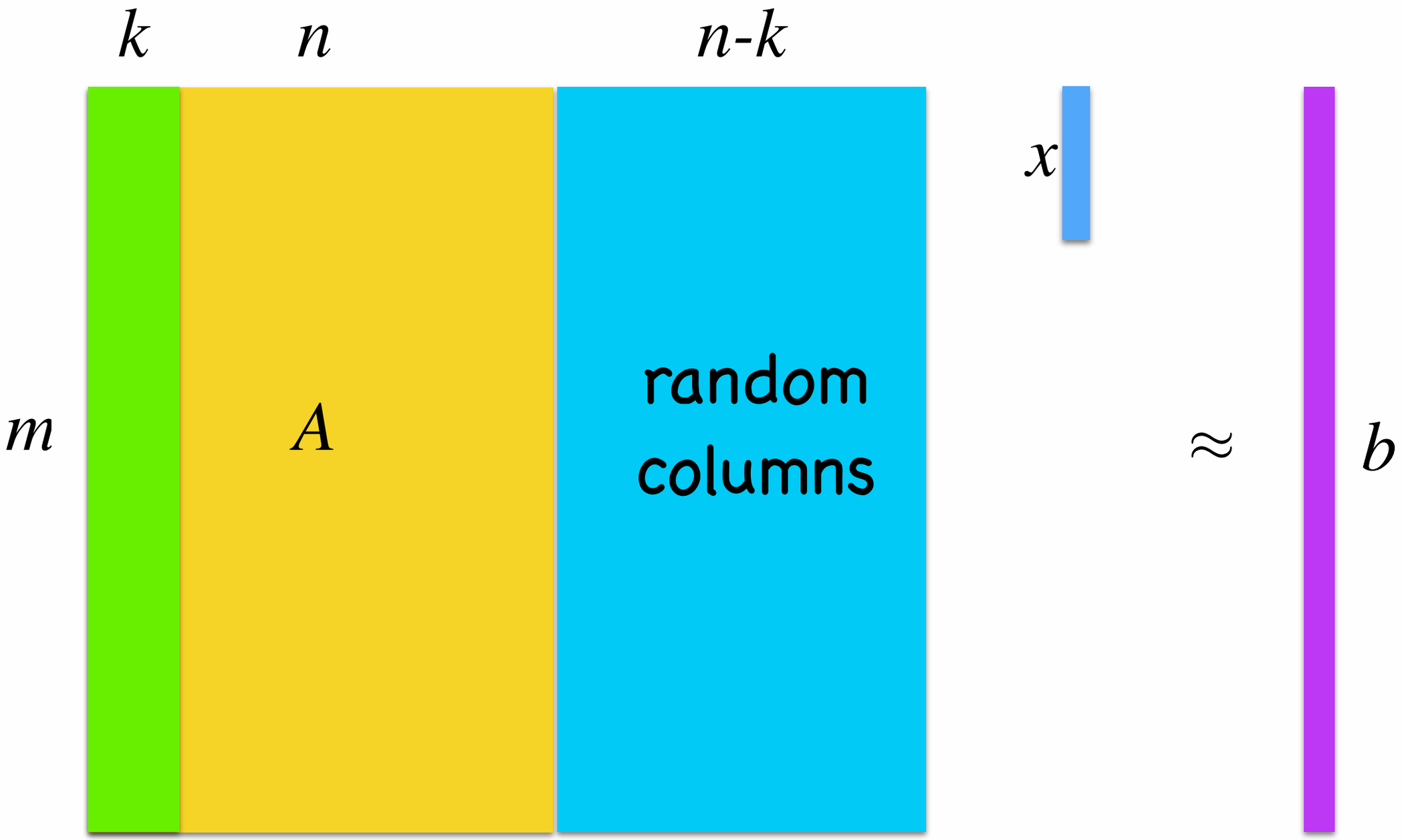
Validation by random columns (II)



Validation by random columns (II)



Validation by random columns (II)



Validation by random columns (III)

The random columns can be purely random, if a suitable distribution is available, or random permutations of columns of A.

We now run BB on the extended A matrix selecting for the best k variables. We run BB several times, for different random columns

Validation by random columns (IV)

Every time a different set of columns appears in the answer (it must contain a random column), it is a clear sign that the original k variables are not reliable.

A very reasonable (and safe) p-value can be obtained this way.

Complexity estimation

In this application we want to estimate the time complexity of various MSA (multiple sequence alignment) programs. The programs have to be treated like black boxes, it is VERY difficult to derive their complexity by inspection.

The approach we take is to run them for many different problems, and record the time and the complexity input parameters (n: number of sequences, l: total length)

Complexity estimation (II)

The independent variables are any reasonable mathematical combination of the input complexity measures, e.g.

$$\ln n,\ \sqrt{n},\ n,\ nl,\ l^2,\ l^3/n, \ldots$$

Complexity estimation (III)

The output often reveals properties that the authors themselves may not know

```
BioNJ -> +1.27032 + 6.9627e-11*n^2*l
Circular -> +2.8702e-05*n^1.5*l + 2.59653e-06*l^2/n^0.5
ClustalW -> +6.1305e-08*l^2 - 3.49493e-11*n*l^2
LST -> +0.00293021*n^2
Mafft -> +6.85704e-08*n*l^1.5
. . . .
```

Large number of independent variables

Final thoughts

Large number of independent variables

While we are able to use/filter large number of independent variables, is this good?

Proc Natl Acad Sci U S A. 2013 Nov 26;110(48):19313-7. 2013 Nov 11.
Revised standards for statistical evidence.
Johnson VE.

Makes the point that p-values, at large, are not correct, and suggests lower thresholds.

Large number of independent variables

The author is right, but the cause lies elsewhere. In my opinion, researchers try too many models and as soon as they find one that passes with a good p-value they think they are done.

Or in other words, what is the value of 1 (out of 100) models passing with a p-value of 1/100? None!

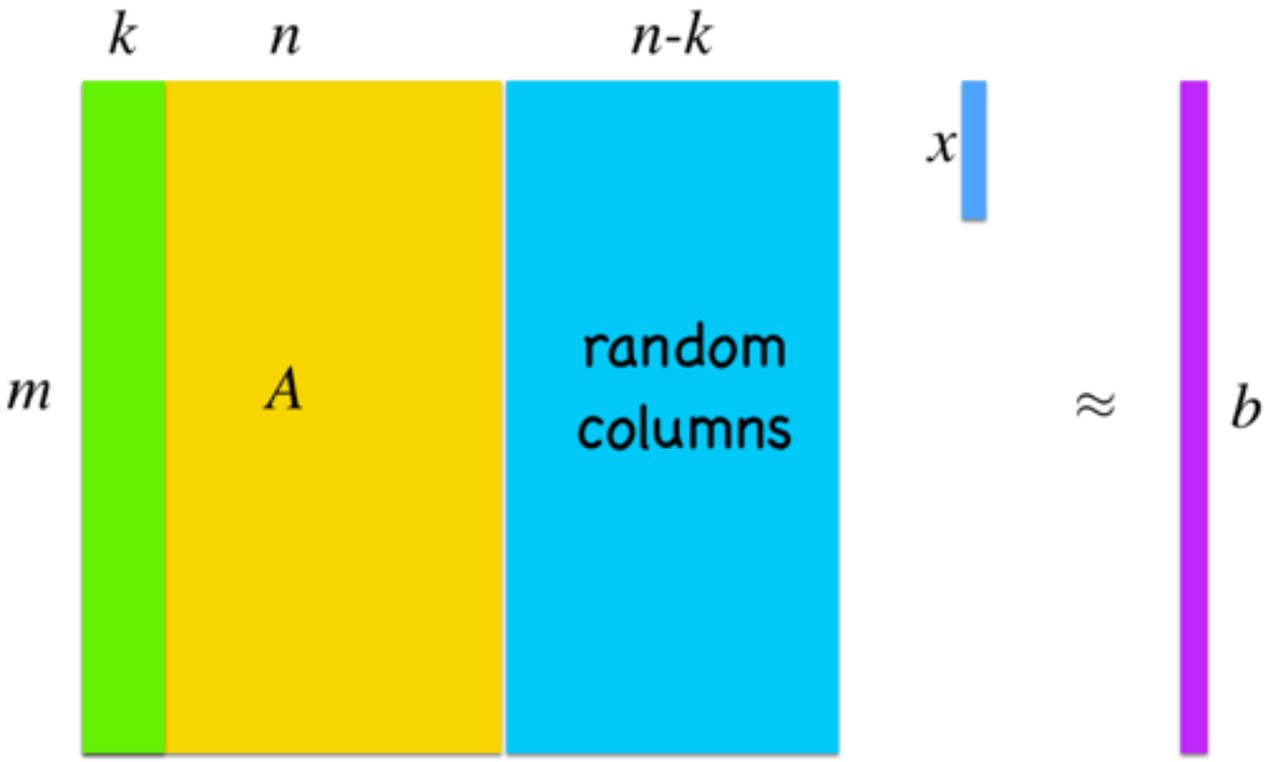
Large number of independent variables

Bonferroni-style corrections?

Model selection theory?

Large number: food for thought

A true weak dependence, may show a weak p-value, e.g. 1/100. By including 100 other independent variables or by trying 100 different models, we may blur the strength of the signal - and lose it.



Big thank you, organizers and audience

```
Results of Best Basis
2000 data points, 20 independent variables, 2 used here
Norms of:     raw data    residuals
                   908       618.65
2 singular values used:  1901  1992
        variable       coeff/stdev    norm decrease
               3   0.3869 +- 0.0127     287.3
              13  -0.0397 +- 0.0125     3.116
```