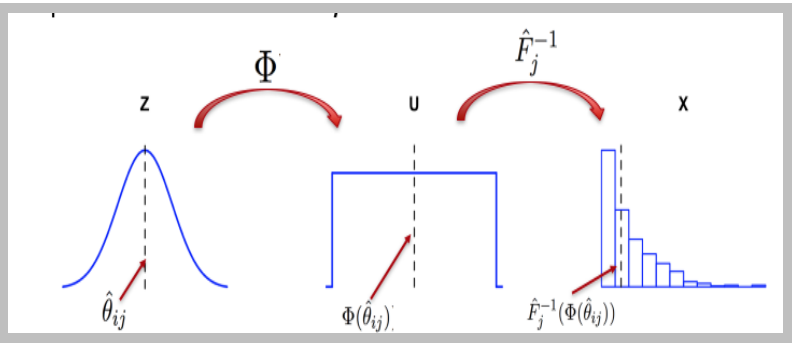
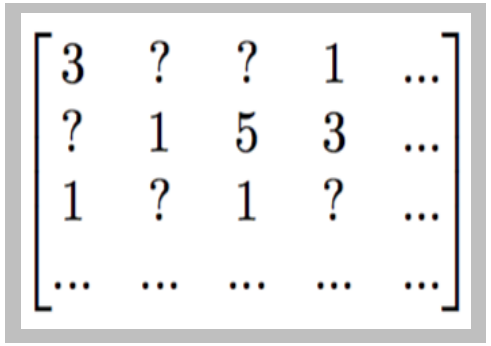
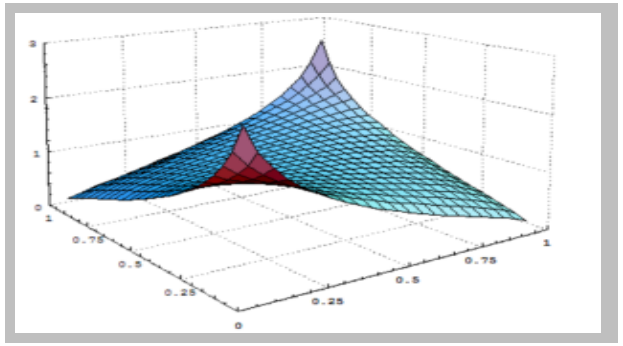
SAND2022-4859C

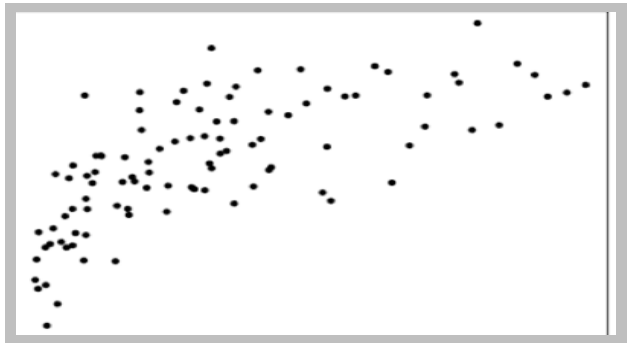
Exceptional service in the national interest

Sandia National Laboratories









XPCA: Copula-based Decompositions for Ordinal Data

Clifford Anderson-Bergman, Kina Kincher-Winoto and Tamara Kolda

Presented by: Clifford Anderson-Bergman

SDSS 2018

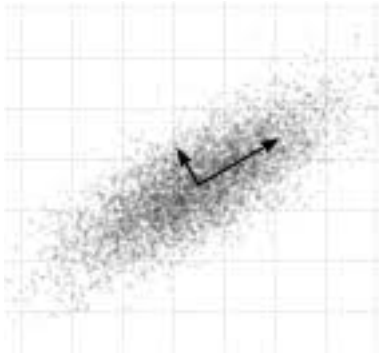
U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

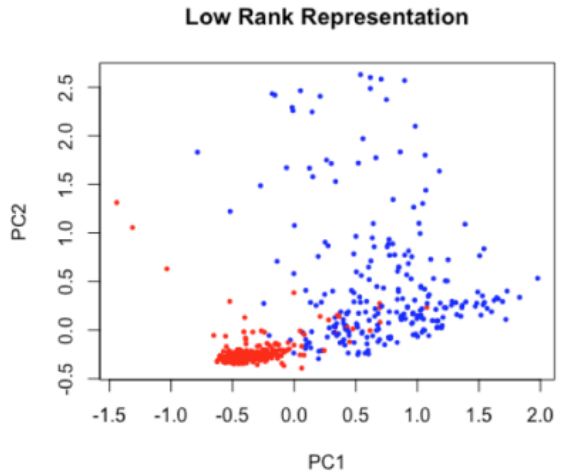
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND NO. 2017-XXXXP

Introduction: PCA

- *Principal Components Analysis* fundamental tool for data analysis
  - *Exploratory data analysis*
  - *Dimension reduction*
  - *Data compression*
  - *Missing data imputation*



$$\begin{bmatrix} 3 & ? & ? & 1 & \dots \\ ? & 1 & 5 & 3 & \dots \\ 1 & ? & 1 & ? & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$



Introduction: Low-rank PCA

- PCA procedure can described as

$$\arg\min_{U,V} ||X - UV^T||_2$$

Low-rank representation of relation between columns: *p x k*

Original Data: *n x p*

Low-rank representation of rows: *n x k*

- Can be shown to be maximum likelihood estimate of model

$$X \sim \text{MVN}(UV^T, \sigma^2)$$

(MVN = multivariate normal)

- We will work this interpretation of PCA, but there are others

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis.

Introduction: Copulas

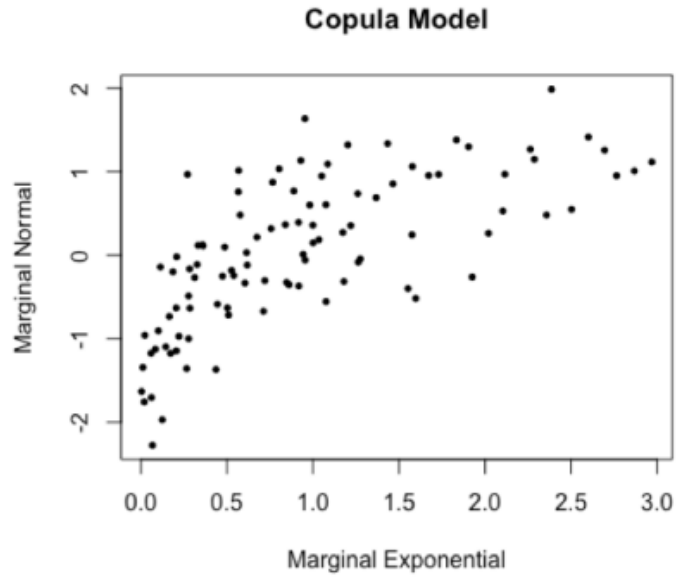
- *Copulas* are methods for generically modeling joint distribution between differing variable types
- Copulas create a function that joins the CDF for different variables

$$P(X_1 > x_1, ..., X_p > x_p) = C(F_{X_1}(x_1), ..., F_{X_p}(x_p))$$

Marginally uniform(0,1)

Copula function that defines relation between marginal uniform(0,1) variables

- Separates modeling into two steps:
  - Modeling marginal distributions
  - Modeling relation between marginal uniform RV's



Introduction: Gaussian Copula

- Gaussian copula popular model

$$C(u_1, ..., u_p|R) = \Phi_R(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_p))$$

CDF of multivariate Normal with correlation matrix *R*

Inverse CDF of standard Normal

- Relation between *ordered values* of variables from each column same as *ordered values* of multivariate normal
- Relatively simple to work with
  - Push original values through marginal CDF + inverse CDF of standard normal
  - Treat as multi-variate normal

XPCA: PCA for non-normal ordinal data

- Recall that PCA maximum likelihood estimate *under assumption of multivariate normal data*
- We want to relax assumption of multivariate normality by using Copula models to model joint distribution across columns
- Copula Component Analysis (COCA) took similar approach, but made implicit assumption of all variables being continuous
- We extend results to discrete ordinal variables

Ma, J., & Sun, Z. (2007). Copula component analysis.

Why use XPCA?

- Decomposition is more robust to outliers in individual variables than PCA
  - COCA has same feature
- If all data is continuous, resulting factors will be nearly identical between COCA and XPCA
- XPCA makes better use of discrete variables, i.e., binary being an extreme case
  - COCA does not and has shown no improvement over PCA with discrete variables
- We also present method for deriving full conditional distribution of missing data

XPCA: Basic Model

- Begin with (unobserved) low-rank PCA model in *copula space*
- Latent variables pushed through standard normal CDF so they are marginally uniform(0,1)
- Values are then pushed through different inverse CDF functions for each variable observed

Latent PCA variable

$$Z \sim \mathrm{MVN}(UV^T, \sigma^2)$$

Transforming latent value to observed value

$$U_{ij} = \Phi(Z_{ij})$$

$$X_{ij} = F_j^{-1}(U_{ij})$$

Final observed value

Inverse marginal CDF for column *j*

Likelihood Function: Continuous Variables

- *If* all variables are *continuous*, likelihood can be written as

$$\ell(\Theta = UV^T\sigma, F_j|X) = \sum_{(i,j)\in\Omega} \log\left(\phi(\Phi^{-1}(F_j(X_{i,j})), \mu = \theta_{ij}, \sigma = \sigma)\right)$$

Likelihood function used by COCA

- Problem: if column *j* is discrete, then a *range* of values of *U* lead to the same value of *X*
- Hoff (2007) showed ignoring this can lead to heavy bias in estimated correlations

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation.

Likelihood Function: Discrete variables

- Define:

$$l_{ij} = \min\{u_{ij} : F^{-1}(u_{ij}) = x_{ij}\}$$

$$r_{ij} = \max\{u_{ij} : F^{-1}(u_{ij}) = x_{ij}\}$$

- Likelihood function is then:

Integrating over range of latent variables that leads to observed outcome

$$\ell(\Theta = UV^T, \sigma, F_j | X) = \sum_{(i,j)\in\Omega} \int_{l_{ij}}^{r_{ij}} \log\left(\phi(\Phi^{-1}(t), \mu = \theta_{ij}, \sigma = \sigma)\right) dt$$

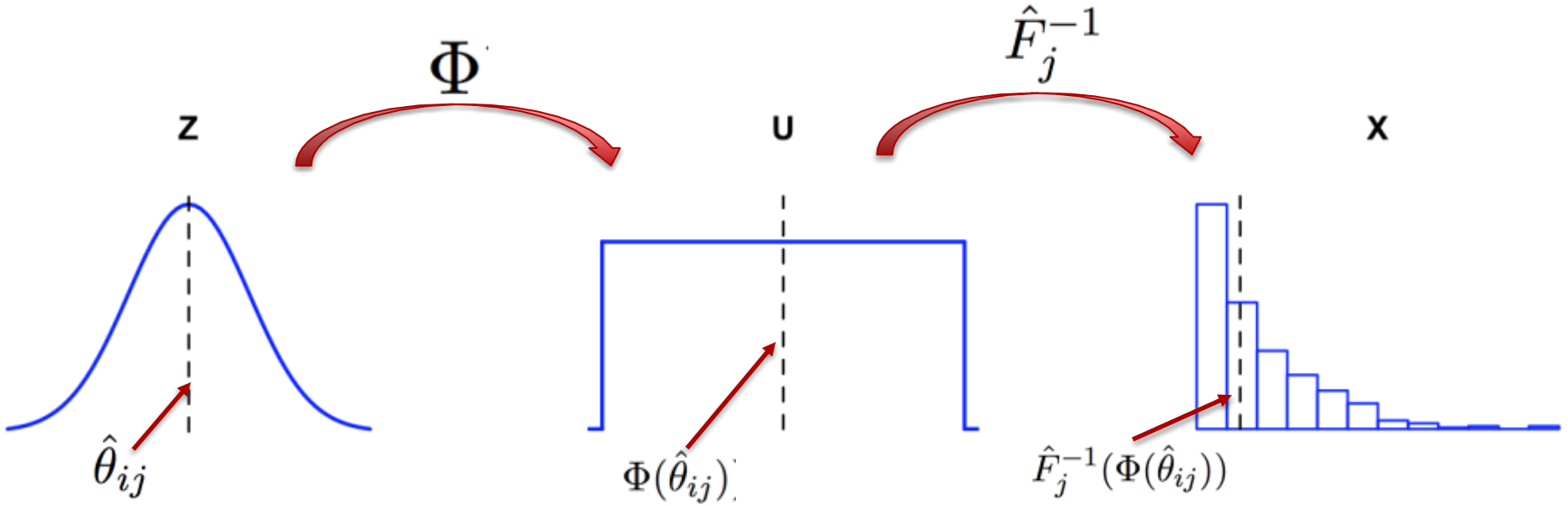
$$= \sum_{(i,j)\in\Omega} \log\left[\Phi\left(\frac{r_{ij} - \theta_{ij}}{\sigma}\right) - \Phi\left(\frac{l_{ij} - \theta_{ij}}{\sigma}\right)\right]$$

Fitting the model

- Fit marginal CDF with Empirical Distribution function
  - Results in *all* variables being treated as discrete, *even if originally continuous*
  - Consistent estimator of any CDF (including continuous variables)
  - If all columns are continuous, XPCA solution approaches COCA solution
- Fit PCA parameters using maximum likelihood estimation
- Two algorithms implemented
  - Generic L-BFGS implementation
  - Custom Alternating Newton's method
    - On average, L-BFGS *slightly* faster (about 2x)
    - Alternating Newton's more robust

Imputing Data: Median Estimate

- According to XPCA model, estimated median of Z_{ij} is $\hat{\theta}_{ij}$
- We want estimate of X_{ij}
- Because Φ and $\hat{F}_j^{-1}$ are monotonic functions, estimated median of X_{ij} is $\hat{X}_{ij} = \hat{F}_j^{-1}(\Phi(\hat{\theta}_{ij}))$
- Imputation method used by COCA



Imputing Data: Mean Estimates

- Once model is fit, can compute

$$P(X_{ij} = x|\hat{U}, \hat{V}, \hat{\sigma}, \hat{F}_j)$$

Characterizes full conditional distribution!

- Estimated expected value is then

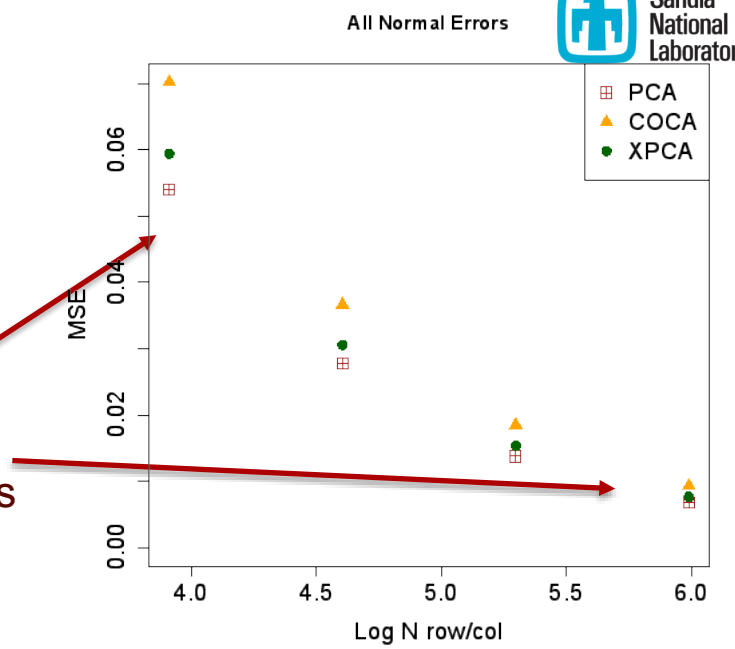
$$\sum xP(X_{ij} = x|\hat{U}, \hat{V}, \hat{\sigma}, \hat{F}_j)$$

- *Estimated means are a kernel-smoothed weighted average of all observed values*
  - Automatically range respecting
  - For properly coded binary variables, mean estimate is probability = 1

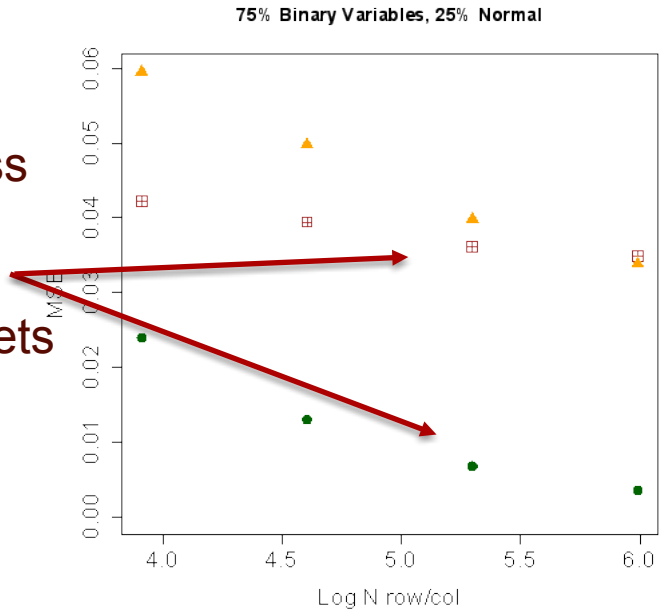
Simulated Comparisons

- Simulated low-rank data and examined MSE in recovering mean structure
  - N rows/cols = 50, 100, 200, 400
  - Rank = 5
  - Error terms
    - Scenario 1: all Normal
    - Scenario 2: 75% Probit model (i.e. binary), 25% Normal
- Note that MSE error metric favors PCA (directly attempts to minimize)

With all Gaussian data, PCA does best, but advantage decreases with data size



XPCA has much less error when binary variables included, even as data size gets larger



Basketball Data

- Season Statistics for 2015-2016 NBA basketball
- 476 players
- 39 variables
  - Season summary statistics
    - Games played, # wins for player's team, etc.
    - Shots made, assists, rebounds, etc.
  - Draft Round/Number
    - Draft Numbers 1-30 -> Round 1, 31-60 -> Round 2 + undrafted
    - Binary variable: was 1st round pick?
  - Others mix of count and continuous data

Data compiled by Justin Jacobs

Imputed data: Point estimates

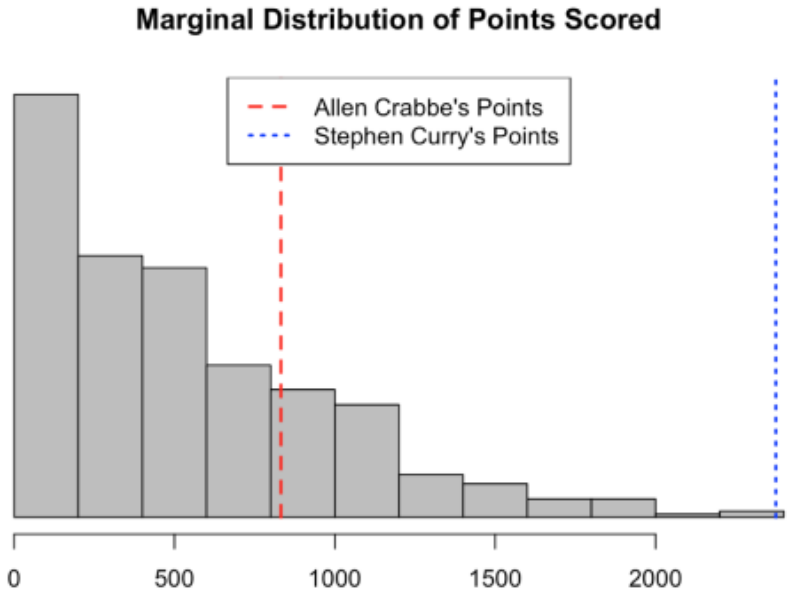
- Masked all draft information for 2 players
  - Stephen Curry: 1st pick of 1st round
  - Allan Crabbe: 1st pick of 2nd round
- Computed rank 3 decomposition and extracted imputed "Is first round pick?"
- PCA provides linear estimates of imputed data
  - Not constrained to be realistic values
- COCA provides median estimate
  - Realistic value but not very informative for binary data!
- XPCA can extract full distribution
  - Easily summarized by median/mean/whatever statistic desired

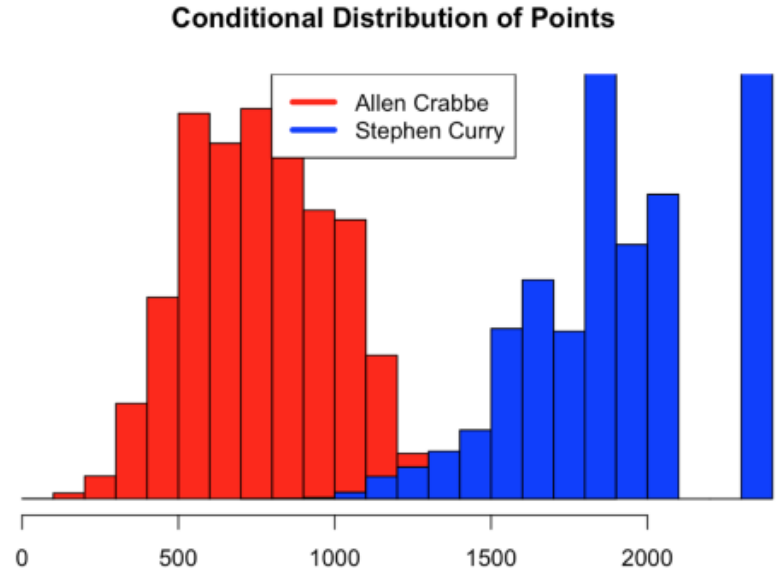
Method	Stephen Curry	Allan Crabbe
PCA	36.9	-0.38
COCA	1.0	0
XPCA (mean method)	1.00*	0.12
True Value	1.0	0.0

*Numerically equivalent to 1

Imputed data: Conditional Distribution

- Can use XPCA to extract estimated full conditional distribution of missing values





Basketball Data: Cross Validated Rescaled MSE

- Evaluating methods by 20x cross-validated Rescale Mean Squared Error
  - MSE of each column rescaled by column variance

Rank	XPCA	PCA	COCA	Mean
1	**0.424**	0.463	0.482	1.004
2	**0.301**	0.353	0.346	1.004
3	**0.249**	0.307	0.275	1.004
4	**0.239**	0.283	0.266	1.004
5	**0.238**	0.345	0.259	1.004
6	**0.247**	1.018	0.287	1.004
7	**0.256**	0.765	0.285	1.004
8	**0.247**	1.699	0.283	1.004
9	**0.293**	4.234	0.440	1.004
10	**0.275**	4.682	0.442	1.004

Lowest overall Rescaled MSE for XPCA

PCA error explodes with overfitting

XPCA + COCA also prone to overfitting, but because range restricted, damage to MSE more limited

Future Directions

- Accelerating algorithm
  - Randomized methods?
- Adding penalization
  - Improving out of sample error
  - Can be simpler to pick penalty than pick rank of decomposition
- Allowing for non-constant σ^2
  - Setting σ^2 as fixed across columns (a la PCA model) assumes each column is equally predictable
  - May not be a reasonable assumption for heterogeneous data; instead allow σ^2 to change by column
- Adding standard errors