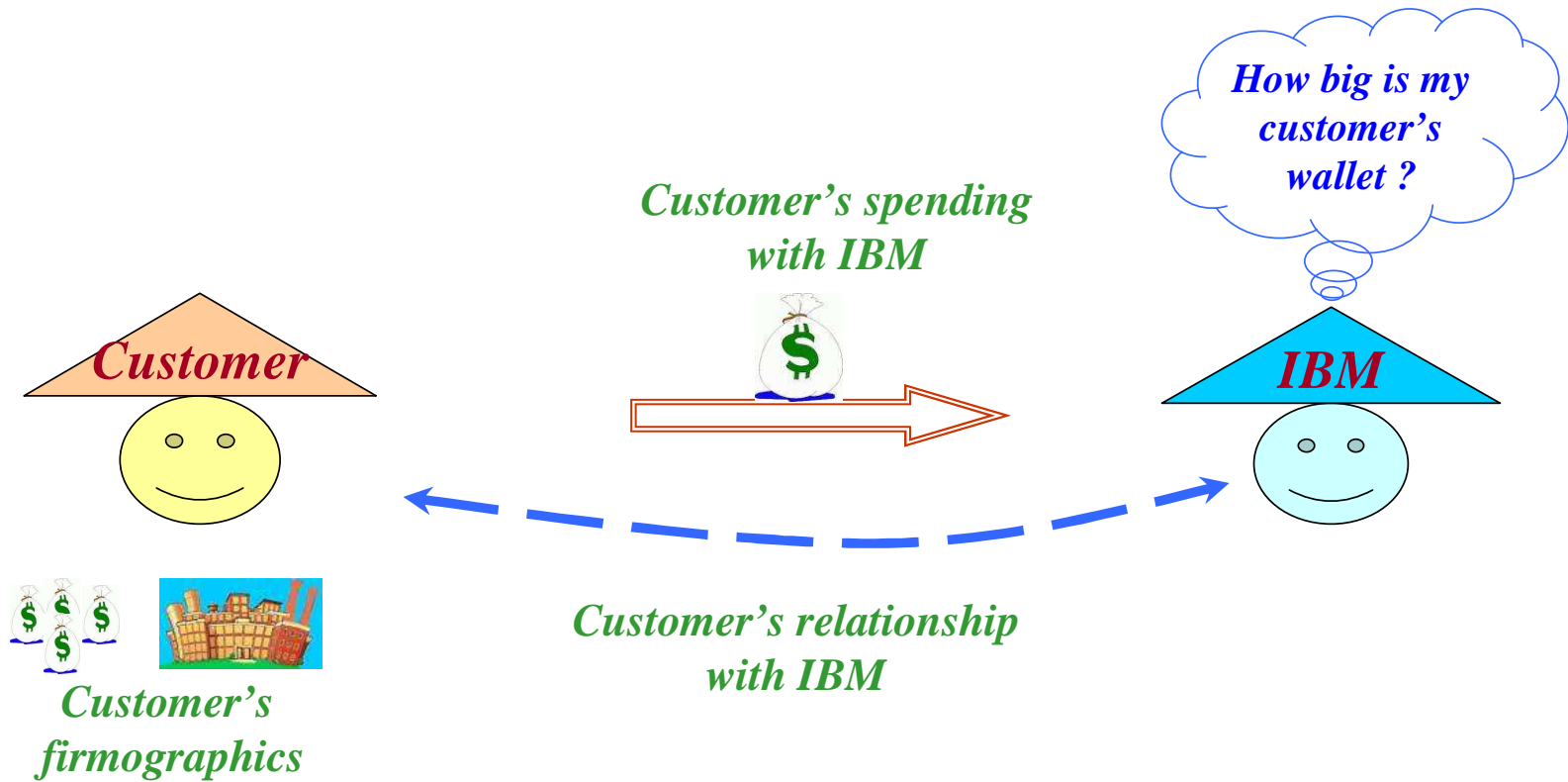

Multi-View Learning via Latent Variable Modeling with an Application to Customer Wallet Estimation

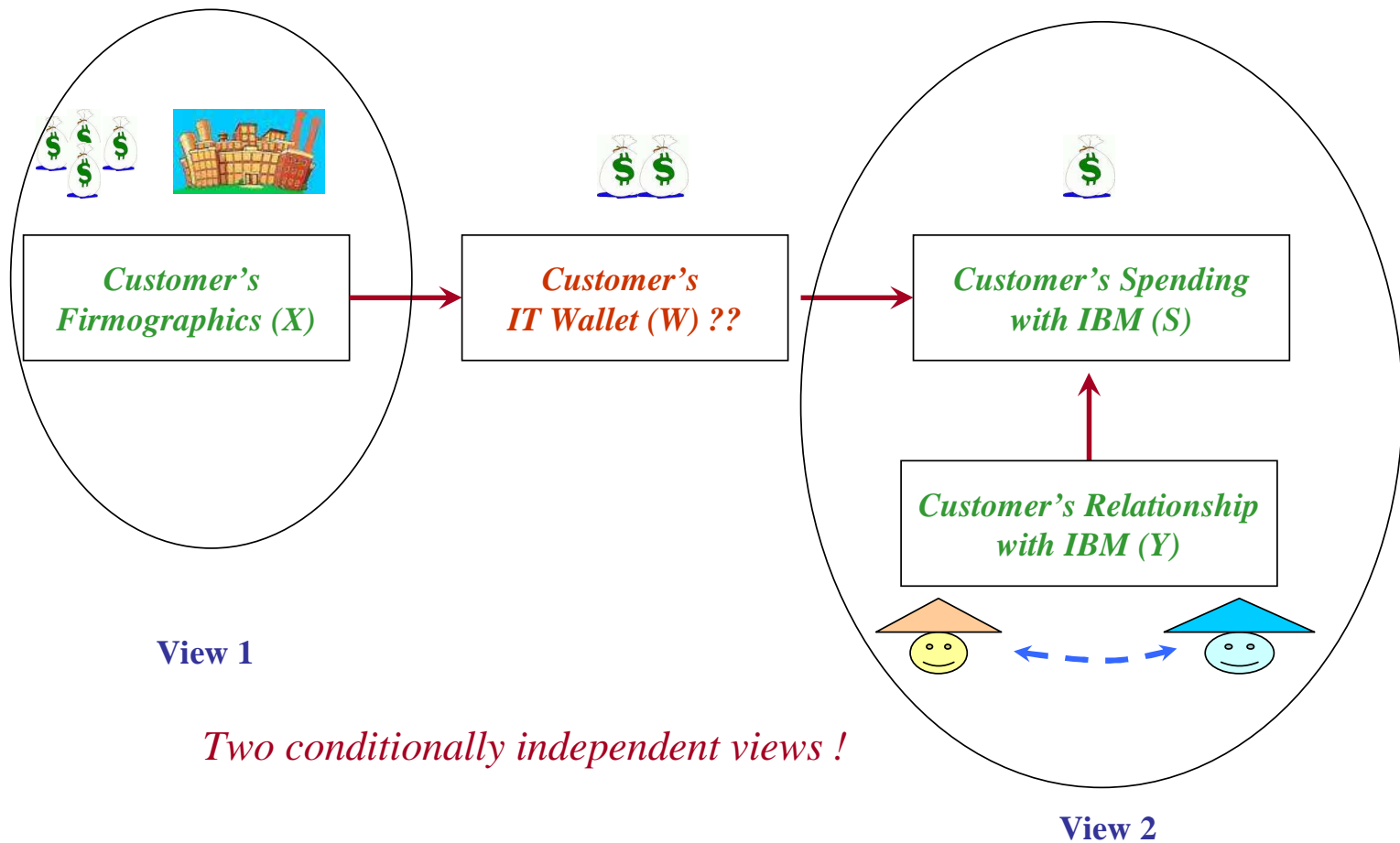
Srujana Merugu , Saharon Rosset, Claudia Perlich

srujana@gmail.com, (srosset,cperlich)@us.ibm.com

Wallet Estimation Problem



Reasonable Graphical Model for *SERVED* Wallet



Two conditionally independent views !

Problem Setting

Unsupervised learning scenario

- Unobserved target variable
- Observations on multiple predictor variables
- Domain knowledge suggesting that the predictors form multiple conditionally independent views

Goal: To predict the target variable

Main Contributions

- Analysis of a relevant class of latent variable models
 - Markov blanket can be split into conditionally independent views
 - For exponential linear models, the maximum likelihood estimation reduces to convex optimization problem
- Solution approaches for Gaussian likelihoods
 - Reduction to single linear least squares regression
 - ANOVA for testing conditional independence assumptions
- Empirical evaluation
 - Comparable to supervised learning with significant amount of training data
 - Case study on wallet estimation

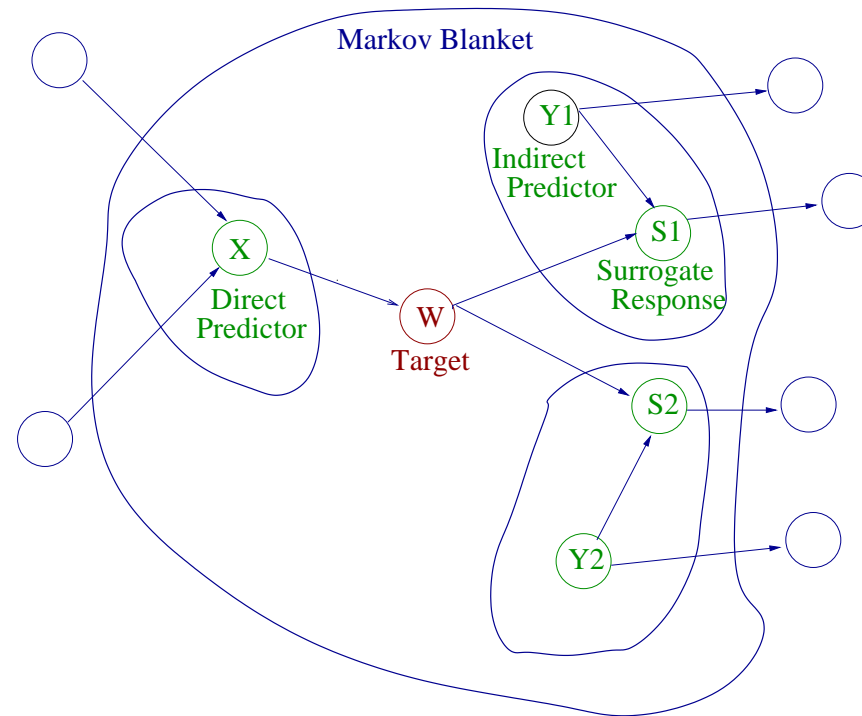
Variable grouping

First step is grouping the observed variables into three classes:

- **Direct predictors**, which directly influence the target, i.e., antecedents of a causal relation with the target, *e.g., firmographics in wallet estimation*
- **Surrogate responses**, which are directly affected by the target, i.e., the consequents of a causal relation with the target, *e.g., customer's actual spending with IBM*
- **Indirect predictors**, which influence a certain surrogate response without directly affecting the target, i.e., antecedents of the surrogate response variables, *e.g., IBM's relationship with the customer*

Special Class of Latent Variable Models

- Graphical models where the Markov blanket of target can be partitioned into the above three classes with no extra dependencies



Direct predictors \rightarrow Bayesian parents, Surrogate responses \rightarrow Bayesian children,
Indirect predictors \rightarrow Parents of Bayesian children

- Multiple conditionally independent views with the target as the central link

Maximum Likelihood Formulation

- **Given:** Directed graphical model and parametric form of the conditional distributions of nodes given their parents
- **Goal:** Predict the target W using the parameter estimates Θ^* that are **most likely** given the observed data and the graphical model

$$\Theta^* = \max_{\Theta} \log p_{D, \Theta}(S_1, \dots, S_{N_c} | X, Y_1, \dots, Y_{N_c}),$$

$P_{D, \Theta}(\cdot)$: data likelihood, N_c : number of children

- **Solution:** Expectation-Maximization (EM) algorithm
 - Converges to a local optimum in general

Formulation: More Details

- Partial discriminative likelihood to be maximized:

$$\begin{aligned} L_D(\Theta) &= \log p_{D,\Theta}(S_1, \dots, S_{N_c} | X, Y_1, \dots, Y_{N_c}) \\ &= \log \left(\int_W p_{D,\theta_0}(W|X) \prod_{k=1}^{N_c} p_{D,\theta_k}(S_k|W, Y_k) \right) \end{aligned}$$

where $\Theta = (\theta_0, \dots, \theta_k)$: parameters for conditional likelihoods in graphical model

- “Posterior” of W given ML parameter estimates Θ^* :

$$\begin{aligned} p_{\Theta^*}(W|M) &= p_{\Theta^*}(W|X, S_1, \dots, S_{N_c}, Y_1, \dots, Y_{N_c}) \\ &= c_{\Theta^*} p_{\theta_0^*}(W|X) \prod_{k=1}^{N_c} p_{\theta_k^*}(S_k|W, Y_k), \end{aligned}$$

W can now be estimated as posterior **mode** or **mean**

Resulting EM Algorithm

Input: Dataset D consisting of predictors $(X, S_1, \dots, S_{N_c}, Y_1, \dots, Y_{N_c})$,
parametric forms $p_{\theta_0}(W|X)$ and $p_{\theta_k}(S_k|W, Y_k)$, $[k]_1^{N_c}$

Output: Target distribution $\tilde{p}(W)$, (“posterior” for W)

Method:

Initialize Θ at random

repeat

Expectation Step

$$\tilde{p}(W) = p_{\Theta}(W|M) = c_{\Theta} p_{\theta_0}(W|X) \prod_{k=1}^{N_c} p_{\theta_k}(S_k|W, Y_k)$$

where c_{Θ} is a normalizing factor.

Maximization Step

$$\theta_0 \leftarrow \operatorname{argmax}_{\theta_0} E_{\tilde{p}}[\log p_{D, \theta_0}(W|X)]$$

$$\theta_k \leftarrow \operatorname{argmax}_{\theta_k} E_{\tilde{p}}[\log p_{D, \theta_k}(S_k|W, Y_k)], [k]_1^{N_c}$$

until convergence

return \tilde{p}

Theoretical Results: Exponential Linear Models

Theorem: When the conditional distributions $p(W|X)$ and $p(S_k|W, Y_k)$, $[k]_1^{N_c}$ correspond to exponential linear models with matching link functions, the incomplete discriminative log-likelihood

$$L_D(\Theta) \equiv \log p_{D,\Theta}(S_1, \dots, S_{N_c} | X, Y_1, \dots, Y_{N_c})$$

is a concave function of the parameters Θ

- Maximum likelihood estimation reduces to a convex optimization problem
- EM algorithm converges to the globally optimal solution

Modeling with Gaussian Distributions (Two Views)

● **Dataset:** $D = \{(w_i, \mathbf{x}_i, \mathbf{y}_i, s_i)\}_{i=1}^n$, $w_i, s_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^{m_1}$, $\mathbf{y}_i \in \mathbb{R}^{m_2}$

● **Coupled Gaussian linear models**

● Conditional distribution of target W given direct predictors X

$$w_i - \boldsymbol{\alpha}^t \mathbf{x}_i = \epsilon_w, \epsilon_w \sim \mathcal{N}(0, \sigma_w^2), [i]_1^n, \boldsymbol{\alpha}, \sigma_w : \text{parameters}$$

● Conditional distribution of surrogate response S given W and indirect predictors Y

$$s_i - w_i - \boldsymbol{\beta}^t \mathbf{y}_i = \epsilon_s, \epsilon_s \sim \mathcal{N}(0, \sigma_s^2), [i]_1^n, \boldsymbol{\beta}, \sigma_s : \text{parameters}$$

● Maximizing the incomplete discriminative likelihood $L_D(\Theta)$ using EM yields the optimal parameter estimates $\hat{\Theta}_{MLE} = (\hat{\boldsymbol{\alpha}}_{MLE}, \hat{\boldsymbol{\beta}}_{MLE}, \hat{\sigma}_{w_{MLE}}, \hat{\sigma}_{s_{MLE}})$

Reduction to Simple Linear Regression

- Reduced linear regression model: $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ and $\boldsymbol{\gamma}^t = [\boldsymbol{\alpha}^t, \boldsymbol{\beta}^t]$

$$s_i - \boldsymbol{\gamma}^t \mathbf{z}_i = \epsilon_{ws}, \quad \epsilon_{ws} \sim \mathcal{N}(0, \sigma_{ws}^2), \quad [i]_1^n, \quad \sigma_{ws}^2 = \sigma_w^2 + \sigma_s^2$$

- Optimal parameters are the least squares estimates $(\hat{\boldsymbol{\alpha}}_{LS}, \hat{\boldsymbol{\beta}}_{LS})$
- **Equivalence Theorem:** When \mathbf{Z} is a full column rank matrix, the maximum likelihood estimates of the coupled Gaussian linear models are unique and identical to the least squares estimates of the reduced linear model, i.e.,

$$(\hat{\boldsymbol{\alpha}}_{MLE}, \hat{\boldsymbol{\beta}}_{MLE}) = (\hat{\boldsymbol{\alpha}}_{LS}, \hat{\boldsymbol{\beta}}_{LS}).$$

Otherwise, the set of optimizers of the two problems are identical.

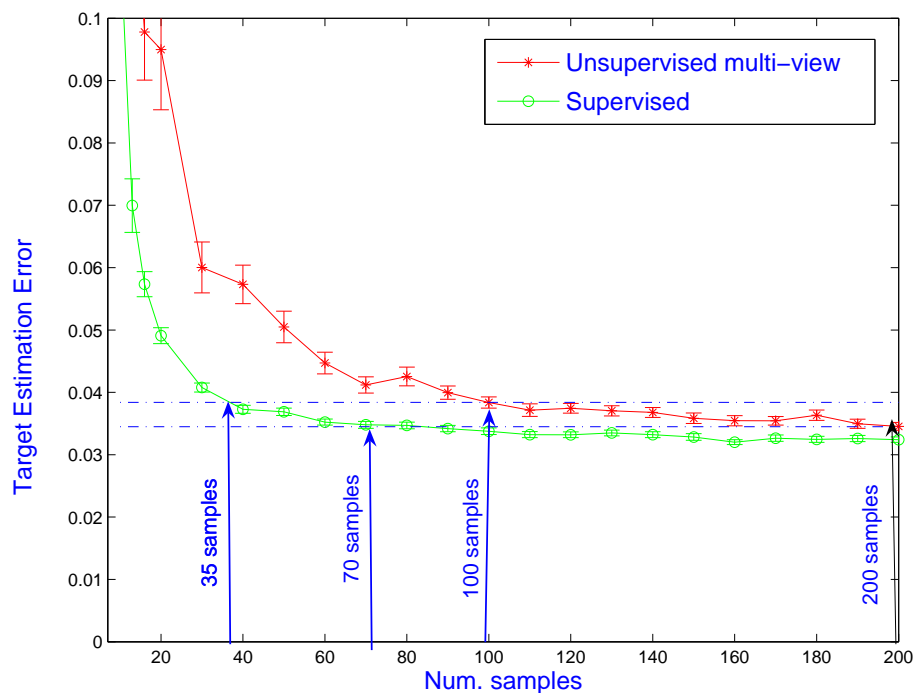
More on Linear Regression Reduction

- Computationally and conceptually simple
- Parameter estimates $\hat{\alpha}$, $\hat{\beta}$ are **consistent** estimators (from ML result)
- Linear regression implies **unbiased** estimates of α , β , hence of W
- Can use linear regression **inference** machinery:
 - Variable selection
 - ANOVA to test independence and parametric assumptions (to some extent)
- Some issues glossed over:
 - Full rank requirement of \mathbf{Z} means remaining, unestimable degree of freedom (if we want intercept in both formulae)
 - As turns out, cannot estimate σ_w , σ_s separately

ANOVA for Testing Independence Assumptions

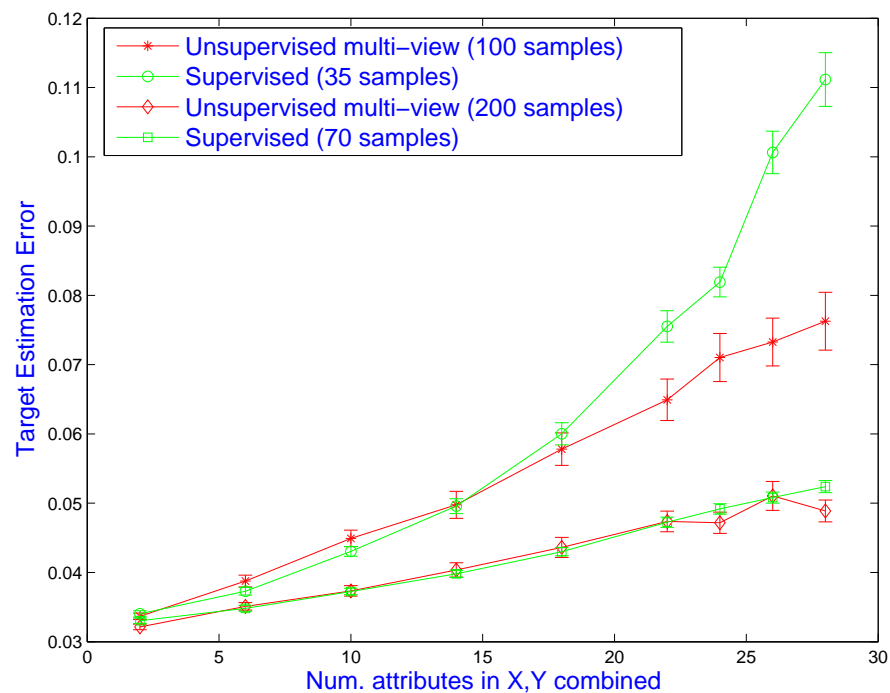
- **ANOVA:** Variance-based analysis for determining the goodness of fit for nested linear models
- **Nested models**
 - **Model A:** Linear model with only variables in X, Y and no interactions
 - **Model B:** Linear model with interactions only within X and Y
 - **Model C:** Linear model with interactions between variables in X and Y
- **Key Idea:** Model C is superior to model B \Rightarrow conditional independence and parametric assumptions are not valid

Simulation Experiments: Gaussian Linear Models



Prediction error with varying number of samples, 6 attributes and variances

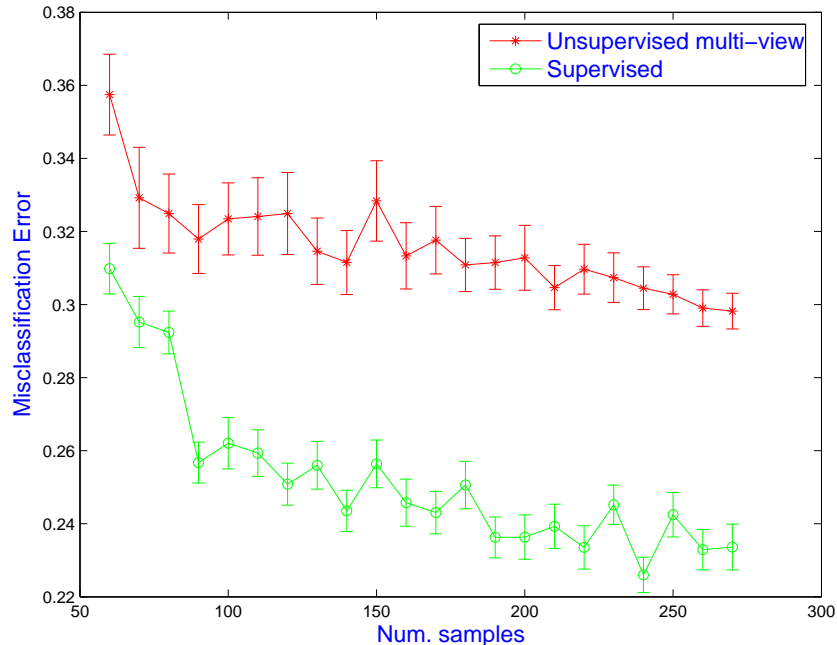
$$\sigma_s = \sigma_w = 0.5$$



Prediction error with varying number of attributes in X and Y and variances

$$\sigma_s = \sigma_w = 0.5$$

Simulation Experiments: Logistic Regression



Misclassification error with varying number of samples and 20 attributes



Generative Model



W, S are binary valued and
 $X \in \mathbb{R}^{m_1}, Y \in \mathbb{R}^{m_2}$



$\text{logit}(p(W = 1|X)) = X\alpha$



$\text{logit}(p(S = 1|W, Y)) = W + Y\beta$



Parameter estimation using EM

Case Study: Wallet Estimation

- W : unobserved wallet, S : customers' spending with IBM, X : customer firmographics, Y : IBM relationship variables
- **Modeling equations:** (monetary values \rightarrow log scale)

$$\log(w_i) = f_\alpha(\mathbf{x}_i) + c_w + \epsilon_w, \quad \epsilon_w \sim \mathcal{N}(0, \sigma_w^2), \quad [i]_1^n$$

$$\log(s_i) - \log(w_i) = g_\beta(\mathbf{y}_i) + c_{sw} + \epsilon_s, \quad \epsilon_s \sim \mathcal{N}(0, \sigma_s^2), \quad [i]_1^n$$

c_w, c_{sw} : constants, f_α, g_β : parametric functions

Wallet Estimation: ANOVA Results

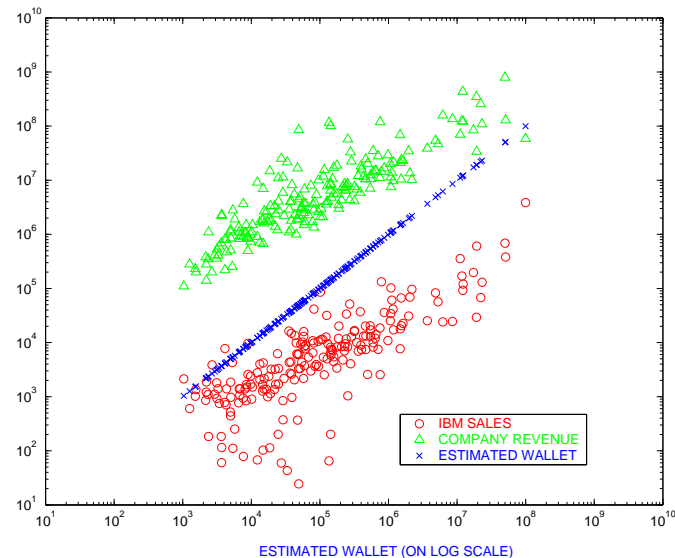
Model (f_α, g_β)	RSS	Deg. Freedom	F statistic	P value
No interaction (A)	10736.7	21		
Within-group (B)	10033.5	75	1.74	0.0001
All interaction (C)	9382.5	100	1.21	0.081

Model C is not superior to model B indicating that the conditional independence hypothesis is reasonable

Wallet Estimation: Choosing Intercept

- Condition that predictor matrix be full column rank \Rightarrow only sum of intercepts $(c_w + c_{sw})$ can be estimated
- **Estimating \hat{c}_w :** Use additional information, e.g., the ordering between customer's spending with IBM (S), IT wallet (W) and revenue(R)

$$\hat{c}_w = \underset{c}{\operatorname{argmax}} \left| \{ [i]_1^n : r_i \geq \hat{w}_i \exp(c) \geq s_i \} \right|$$



Wallet predictions on test data preserve the desired ordering $R \geq W \geq S$