

A New Multi-View Regression Approach with an Application to Customer Wallet Estimation

Srujana Merugu
Dept. of Electrical and Computer Eng.
The University of Texas at Austin
Austin, TX 78712
srujana@gmail.com

Saharon Rosset, Claudia Perlich
IBM T.J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598
{srosset, perlich}@us.ibm.com

ABSTRACT

Motivated by the problem of customer wallet estimation, we propose a new setting for multi-view regression, where we learn a completely unobserved target (in our case, customer wallet) by modeling it as a “central link” in a directed graphical model, connecting multiple sets of observed variables. The resulting conditional independence allows us to reduce the discriminative maximum likelihood estimation problem to a convex optimization problem for parametric forms corresponding to exponential linear models. We show that under certain modeling assumptions, in particular, when we have two conditionally independent views and the noise is Gaussian, we can reduce this problem to a single least squares regression. Thus, for this specific, but widely applicable setting, the “unsupervised” multi-view problem can be solved via a simple supervised learning approach. This reduction also allows us to test the statistical independence assumptions underlying the graphical model and perform variable selection. We demonstrate our approach on our motivating problem of customer wallet estimation and on simulation data.

Categories and Subject Descriptors

H.4.8 [Database Management]: Database Applications — Data Mining; I.2.6 [Artificial Intelligence]: Machine Learning

General Terms

Algorithms

Keywords

Multi-view learning, Bayesian networks, Regression

1. INTRODUCTION

In standard predictive modeling methodology, an observed “target” variable of interest is modeled as a function of a collection of predictors. The ultimate goal is predicting the target variable in future cases when we only observe the

predictors. In this paper, we are interested in an “unsupervised” situation where a specific target variable exists, but is never observed, and we still want to build a prediction model for it. The only information available is in the form of domain knowledge that indicates the existence of multiple views, which provide “independent” information about the unobserved target. This domain knowledge allows us to obtain a directed graphical model by formalizing the independence relations and sets a framework for inference about the target.

One example of such an application is the problem of customer wallet estimation, which is of great practical interest to us at IBM. One definition of a customer’s wallet for a specific product category (for example, Information Technology (IT)) is the customer’s *total budget* for purchases in this product category across various vendors. As an IT vendor, IBM observes the amount its customers (which are almost invariably companies) spend with it, but does not typically have access to the customers’ budget allocation decisions, their spending with competitors, etc. Information about the customers’ wallet, as an indicator of their potential for growth, is considered extremely valuable for marketing, resource planning and other tasks. For a detailed survey of the motivation, problem definition, and some alternative solution approaches, see [15]. For our purpose, the important aspect of this problem is that the desired target, i.e., the customer wallet, is completely unobserved, but we have access to two sources of related information: IBM’s internal databases, which tell us about IBM’s relationship with the customer, including the current and past sales by product; and publicly available firmographics about the customer company, including its revenue, industry, location, etc.

Let us now take a closer look at the IT purchase process. One can reasonably argue that this involves two stages: the first where the customer company’s executives decide on the company’s IT wallet W based on the company’s situation and needs, which are captured by firmographics X , and the second, where the IT department decides on the portion of the wallet that is spent on IBM products S depending on their relationship with IBM captured by Y . The causal relations emerging from this purchase model can be readily represented in the form of a Bayesian network as shown in Figure 1 where X and (S, Y) are *conditionally independent* of each other given W . Additional domain knowledge can then be used to identify the appropriate parametric forms

for each of the causal relations in the Bayesian network. Given all of these, the unobserved wallet can be treated as missing data and estimated via a maximum likelihood approach, e.g., using EM.

A similar picture can be argued to apply for other business and scientific problems, e.g., estimating an online advertiser’s share of customers’ clicks, where the click behavior is unobserved, but some customer characteristics affecting it are known.

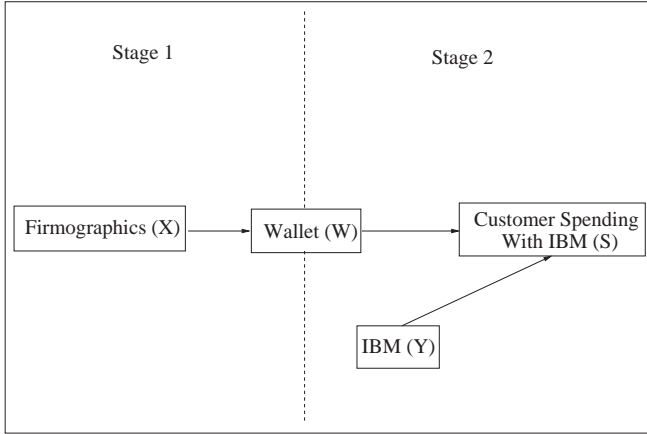


Figure 1: Causal relations between customer wallet and observed predictors

In this paper, we consider unsupervised learning problems that follow the special structure described above and various solution approaches for them. In Section 2, we develop a formal description of the “multi-view” problem with conditional independence; pose the discriminative learning problem that arises from it in terms of likelihood maximization; and show how it can be solved using the standard EM methodology [12]. Further, we show that when the conditional distributions in the graphical model follow parametric forms arising from exponential linear models, the likelihood maximization reduces to a convex optimization problem so that the EM algorithm converges to a global optimum.

In Section 3, we concentrate on the special case of two views and linear models with Gaussian noise. This paper’s main result is that in this case, the problem can be solved by reducing it to a *supervised* learning problem that involves fitting the surrogate response (corresponding to S in Figure 1) on the observed predictors. In addition to being computationally favorable, this also allows us to harness the inferential power of linear modeling, including variable selection and ANOVA-based hypothesis testing, which can be used to test the validity of our conditional independence assumptions.

In Section 4, we demonstrate the applicability and performance of our framework on simulation data. On our simulated examples, the predictive performance of the multi-view learning approach *with no labeled data* (i.e., target is unobserved) turns out to be comparable to that of the standard supervised learning approach (i.e., when we do observe the

target, but do not make use of conditional independence) to the same problem with significant amount of training data. In Section 5, we apply our learning framework to our motivating problem of wallet estimation. Although we have no direct wallet observations to validate our predictions, we present several pieces of indirect evidence on the success of our models.

Section 6 is devoted to a survey of related work in several different areas. It is worth noting here briefly the close relationship and interesting differences between our work and two of these areas. First, the area of co-training and its variants [17, 3, 13] deal with a similar problem of modeling a target that is rarely observed in the presence of conditionally independent views. However, these works make a compatibility (or learnability) assumption that while powerful, is also very limiting. Our approach makes no such assumptions. Second, the area of latent variable modeling [2] considers multiple views with conditional independence much in the same spirit as this work, although typically with a different goal in mind, of modeling the observed data better. The main difference between the latent variable graphical models and our approach are that they do not have observed variables causally affecting the unobserved ones (like our $X \rightarrow W$ connection in Figure 1). This detail turns out to have a major effect on the resulting algorithms. In particular, in the case of linear models with Gaussian noise, latent variable modeling also reduces to a simple problem, like our result in Section 3. However, that problem turns out to be Principal Components Analysis (i.e., a maximum eigenvalue problem) as compared to the reduction to a least squares that we obtain in Section 3.

Notation. Sets such as $\{x_1, \dots, x_n\}$ are enumerated as $\{x_i\}_{i=1}^n$ and an index i running over the set $\{1, \dots, n\}$ is denoted by $[i]_1^n$. Vectors are denoted using bold lower case letters, e.g., \mathbf{x} with the corresponding sub-scripted plain letters, e.g. x_i denoting the components. Matrices are denoted using bold upper case letters, e.g., \mathbf{X} with the corresponding lower case bold letters, e.g., \mathbf{x}_i denoting the column vectors. Transpose of a matrix \mathbf{X} is denoted by \mathbf{X}^t . Random variables are denoted by plain upper case letters X and the corresponding distributions are denoted by $p(X)$ using subscripts to resolve any ambiguity.

2. UNSUPERVISED LEARNING VIA BAYESIAN MODELING

We first describe our multi-view learning setting and the associated directed graphical model. Then, we provide a formal definition of the unsupervised learning problem in terms of maximizing the observed discriminative likelihood.

Our modeling approach is based on grouping the predictor variables in the learning problem into three classes:

- (a) *Direct predictors*, which directly influence the target, or in other words, the antecedents of a causal relation with the target, e.g., firmographics (X) in the wallet estimation problem
- (b) *Surrogate responses*, which include the variables directly affected by the target, or in other words, the consequents of a causal relation with the target, e.g., customer’s actual spending with IBM (S) in the wallet estimation problem
- (c) *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 (Y) in the wallet estimation problem

In Bayesian network terms, these three groups correspond to the parents of the target, children of the target and other parents of the children of the target, respectively.

2.1 Directed Graphical Model

We are interested in the case that all relevant variables in our graphical model, i.e., predictors in the Markov blanket [8] of the target can be disjointly partitioned into these three categories with no dependencies other than the ones specified above. The reason we focus on this class of configurations is because they result in multiple views that are conditionally independent of each other given the target, which in turn enables us to obtain a more “learnable” parametric form for the joint distribution, i.e., one with fewer degrees of freedom.

Using the same notation as in our wallet example, let W denote the (unobserved) target and X denote all the direct predictors or the Bayesian parents of W . Further, let $\{S_k\}_{k=1}^{N_c}$ be the surrogate responses or the children of W and $\{Y_k\}_{k=1}^{N_c}$ their corresponding indirect predictors or Bayesian parents where N_c is the number of children of W .

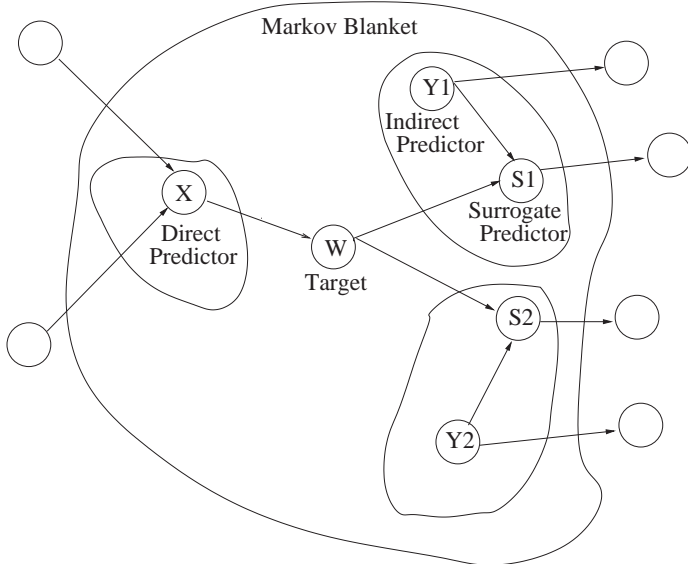


Figure 2: Bayesian network corresponding to our multi-view (here, three-view) learning setting

Figure 2 shows the corresponding Bayesian network. The Markov blanket of the target W is given by the set $M = \{X\} \cup \{S_k\}_{k=1}^{N_c} \cup \{Y_k\}_{k=1}^{N_c}$. Since the target variable is independent of all other factors given the predictors in M , we focus only on this set. First, we observe that since the three sets of predictor (direct, surrogate and indirect) are disjoint, the views corresponding to the sets $\{X\}$, $\{S_1, Y_1\}$, \dots , $\{S_{N_c}, Y_{N_c}\}$ are all conditionally independent of each other given W . Hence, W forms a central link connecting all these multiple views as in Figure 2 and has to be reconstructed so as to be consistent with all the $N_c + 1$ views.

In addition to the conditional independence encoded by the Bayesian network, we also need to specify the parametric forms of conditional distributions of each node given its antecedents or Bayesian parents, which in turn determines the parametric form of the joint distribution of all variables.

2.2 Maximum Likelihood Formulation

We now consider the problem of predicting the unobserved target W given the predictors. When W is observed, i.e., when we have access to training data on W , we can use domain knowledge to specify a parametric form for the conditional distribution of W given all the predictors and estimate the parameters that maximize the discriminative likelihood $p(W|M)$ where M is the Markov blanket of W . In the absence of training data on W , we can still specify the parametric forms for the various conditional distributions using the causality information, but we cannot compute the discriminative likelihood $p(W|M)$. The best one can do is to predict the target using the parameter estimates that are most “consistent” with the observed data as well as the Bayesian network assumptions. A natural way to quantify this consistency is in terms of the incomplete data likelihood, i.e., likelihood of the observed predictors. Since the main objective is to estimate only the unobserved target, one needs to only consider the incomplete discriminative likelihood corresponding to the surrogate responses, i.e., those that are influenced by the target. The learning approach, therefore, consists of two steps:

- (i) Estimate the parameters that correspond to the maximum incomplete discriminative likelihood.
- (ii) Obtain the target using the parametric form of the conditional distribution $p(W|M)$ and the maximum likelihood estimates.

We now proceed to obtain the incomplete discriminative likelihood. Let D be a dataset consisting of n i.i.d. tuples of the observed predictors $(X, S_1, \dots, S_{N_c}, Y_1, \dots, Y_{N_c})$ with W being unobserved. The joint likelihood of the relevant part of the Bayesian network can be readily obtained as follows:

$$P(W|M) = p_D(X)p_D(W|X) \prod_{k=1}^{N_c} p_D(Y_k) \prod_{k=1}^{N_c} p_D(S_k|W, Y_k). \quad (2.1)$$

Since S_1, \dots, S_{N_c} are surrogate responses, the incomplete discriminative likelihood corresponds to conditional distribution $p(S_1, \dots, S_{N_c}|X, Y_1, \dots, Y_{N_c})$. Therefore, assuming that $p(W|X)$ follow the parametric form $p_{\theta_0}(W|X)$ and let $p(S_k|W, Y_k)$ follow the parametric form $p_{\theta_k}(S_k|W, Y_k)$ for all $[k]_1^{N_c}$, the incomplete discriminative log-likelihood becomes:

$$\begin{aligned} L_D(\Theta) &= \log(p_{D,\Theta}(S_1, \dots, S_{N_c}|X, Y_1, \dots, Y_{N_c})) \\ &= \log \left(\int_W p_{D,\theta_0}(W|X) \prod_{k=1}^{N_c} p_{D,\theta_k}(S_k|W, Y_k) \right) \end{aligned} \quad (2.2)$$

where $\Theta = (\theta_0, \theta_1, \dots, \theta_{N_c})$ and D in the sub-script denotes that the likelihood is evaluated on the dataset D .

Our unsupervised learning problem, therefore, reduces to the optimization problem:

$$\max_{\Theta} L_D(\Theta). \quad (2.3)$$

The resulting maximum likelihood estimates Θ^* can now be plugged into the conditional distribution of the target given all the predictors to obtain

$$\begin{aligned} p_{\Theta^*}(W|M) &= p_{\Theta^*}(W|X, S_1, \dots, S_{N_c}, Y_1, \dots, Y_{N_c}) \\ &= c_{\Theta^*} p_{\theta_0^*}(W|X) \prod_{k=1}^{N_c} p_{\theta_k^*}(S_k|W, Y_k), \end{aligned} \quad (2.4)$$

where c_{Θ^*} is a normalizing factor that ensures that the probability mass under the conditional distribution sums up to 1. The target W can then be estimated either as the most likely value or the expected value of this distribution.

For the special case where the conditional distributions $p(W|X)$ and $p(S_k|W, Y_k)$ correspond to generalized linear models with matching link functions, the incomplete discriminative log-likelihood $L_D(\Theta)$ turns out to be a concave function of the parameters Θ taking values on a convex domain. As a result, the likelihood maximization problem (2.3) reduces to a (not always strict) convex optimization problem with a unique global optimum. Theorem 1 states this result more formally.

Theorem 1 *For the Bayesian network described in Sec 2.1, let W and S_k , $[k]_1^{N_c}$ be real-valued and let the conditional distributions $p(W|X)$ and $p(S_k|W, Y_k)$, $[k]_1^{N_c}$ correspond to exponential linear models, i.e., satisfy the following parametric forms:*

$$\begin{aligned} p(W|X) &= \exp(W\theta_0^t X - \psi(\theta_0^t X)) \\ p(S_k|W, Y_k) &= \exp(S_k\theta_k^t W + S_k\theta_k^t Y_k - \psi(W + \theta_k^t Y_k)) \end{aligned}$$

where ψ is the log-partition function. Then, the incomplete discriminative log-likelihood $L_D(\Theta)$ is a concave function of $\Theta = (\theta_0, \dots, \theta_{N_c})$.

*Proof Sketch:*¹ The proof makes use of the fact that each parameter θ_k , $[k]_1^{N_c}$ occurs in a single conditional distribution contributing to $L_D(\Theta)$ in (2.2), each of which corresponds to an exponential distribution known to be log-concave [1] in the natural parameters, in this case $\theta_0^t X$ and $(W + \theta_k^t Y_k)$, $[k]_1^{N_c}$. \square

2.3 Expectation-Maximization based Solution

Given the specific form of the likelihood maximization problem (2.3), one could use a suitable optimization technique for solving it. For instance, the special case considered in Theorem 1 can be addressed using any convex optimization method. For the general case, we now outline an expectation-maximization based solution which makes use of the fact that the objective function in (2.3) corresponds to an incomplete log-likelihood.

Following [12], we consider the negative free energy function $F_D(\tilde{p}, \Theta)$ corresponding to the likelihood maximization problem defined as:

$$F_D(\tilde{p}, \Theta) = E_{\tilde{p}}[\log p_{D, \Theta}(W, S_1, \dots, S_{N_c})] + H(\tilde{p}), \quad (2.5)$$

¹Detailed proofs have been omitted for brevity. Please see [11] for details.

where \tilde{p} is the posterior distribution of the hidden variable W given the observed ones, $H(\cdot)$ is Shannon's entropy and the first term is the expected complete likelihood of W and the surrogate predictors. Using the property that for every local maximizer Θ^* of $L_D(\cdot)$, the pair $(\Theta^*, p_{\Theta^*}(W|M))$ is a local maximizer of $F_D(\tilde{p}, \Theta)$ where $p_{\Theta^*}(W|M)$ is determined by (2.4), we can now restate learning problem in terms of maximizing the negative free energy function:

$$(\tilde{p}^*, \Theta^*) = \underset{(\tilde{p}, \Theta)}{\operatorname{argmax}} F_D(\tilde{p}, \Theta). \quad (2.6)$$

The above problem can be readily addressed using the EM approach where the alternate E and M steps involve optimizing $F_D(\tilde{p}, \Theta)$ with respect to \tilde{p} and Θ respectively keeping the other argument fixed. Algorithm 1 shows the various steps in the EM-based approach, which is guaranteed to converge to a locally optimal solution. For the special case considered in Theorem 1, it converges to a global optimum. The maximizing posterior distribution \tilde{p}^* resulting from the algorithm can be directly used to estimate the target W .

Algorithm 1 EM algorithm for multi-view learning

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 $\tilde{p}(W)$, (local optimizer of (2.6))
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}
 $\theta_0 \leftarrow \underset{\theta_0}{\operatorname{argmax}} E_{\tilde{p}}[\log p_{D, \theta_0}(W|X)]$
 $\theta_k \leftarrow \underset{\theta_k}{\operatorname{argmax}} E_{\tilde{p}}[\log p_{D, \theta_k}(S_k|W, Y_k)]$, $[k]_1^{N_c}$
until convergence
return \tilde{p}

3. GAUSSIAN LINEAR MODELS AND THE REDUCTION TO LINEAR REGRESSION

Gaussian linear models are one of the most widely used parametric models. In this section, we present a detailed analysis of the case where the conditional distributions in (2.1) correspond to Gaussian linear models. For simplicity, we restrict our analysis to the case where there is a single surrogate response and demonstrate that the unsupervised prediction problem (2.3) for this case can be reduced to a single linear least squares regression.

Let the dataset $D = (X, Y, W, S)$ consist of n tuples of the form $(\mathbf{x}_i, \mathbf{y}_i, w_i, s_i)$ where the unobserved target w_i and surrogate s_i are real-valued while the direct predictor $\mathbf{x}_i \in \mathbb{R}^{m_1}$ and indirect predictor $\mathbf{y}_i \in \mathbb{R}^{m_2}$. Further, let the target W be distributed according to a Gaussian linear model based on X , i.e.,

$$w_i - \boldsymbol{\alpha}^t \mathbf{x}_i = \epsilon_w, \quad \epsilon_w \sim \mathcal{N}(0, \sigma_w^2), \quad [i]_1^n \quad (3.7)$$

where $\boldsymbol{\alpha}$ and σ_w are the model parameters that need to be estimated. Similarly, let the surrogate response S be distributed according to a Gaussian linear model based on the target W and the indirect predictor Y , such that the

coefficient of W equals 1, i.e.,

$$s_i - w_i - \beta^t \mathbf{y}_i = \epsilon_s, \quad \epsilon_s \sim \mathcal{N}(0, \sigma_s^2), \quad [i]_1^n, \quad (3.8)$$

Putting together (3.7) and (3.8), we can now compute the incomplete likelihood $L_D(\Theta)$ and the negative free energy function $F_D(\tilde{p}, \Theta)$ where $\Theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_w, \sigma_s)$ corresponds to the model parameters and $\tilde{p} = \{\tilde{p}_i(w_i)\}_{i=1}^n$ consists of the posterior distributions of the unobserved target W . As mentioned earlier, the Expectation-Maximization algorithm (Algorithm 1) progressively maximizes the likelihood to converge to an optimizer $(\hat{\Theta}, \hat{p})$, which can be characterized by the following result.

Theorem 2 *Let \mathbf{w} and \mathbf{s} denote the vectors $[w_1, \dots, w_n]^t$ and $[s_1, \dots, s_n]^t$ respectively and let \mathbf{X} and \mathbf{Y} denote the matrices $[\mathbf{x}_1, \dots, \mathbf{x}_n]^t$ and $[\mathbf{y}_1, \dots, \mathbf{y}_n]^t$. Then, Algorithm 1 converges to an optimizer $(\hat{\Theta}, \hat{p})$ that satisfies the following conditions:*

- (a) *The posterior distribution \hat{p}_i is a Gaussian distribution with mean $\hat{w}_i \equiv E_{\hat{p}_i}[w_i]$ and variance $\hat{\sigma}_i^2 \equiv E_{\hat{p}_i}[(w_i - \hat{w}_i)^2]$ given by*

$$\begin{aligned} \hat{w}_i &= \hat{\eta}(\hat{\boldsymbol{\alpha}}^t \mathbf{x}_i) + (1 - \hat{\eta})(s_i - \hat{\boldsymbol{\beta}}^t \mathbf{y}_i), \quad [i]_1^n \\ \hat{\sigma}_i^2 &= \hat{\eta}(1 - \hat{\eta})(\hat{\sigma}_s^2 + \hat{\sigma}_w^2), \quad [i]_1^n. \end{aligned}$$

$$\text{where } \hat{\eta} = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_w^2 + \hat{\sigma}_s^2}.$$

- (b) *When $[\mathbf{X}, \mathbf{Y}]$ is full column rank matrix, $\hat{\eta} = \frac{1}{2}$ and the parameter estimates are uniquely determined by the following system of equations:*

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= [\mathbf{X}^t \mathbf{H}_Y \mathbf{X}]^{(-1)} \mathbf{X}^t \mathbf{H}_Y \mathbf{s} \\ \hat{\boldsymbol{\beta}} &= [\mathbf{Y}^t \mathbf{H}_X \mathbf{Y}]^{(-1)} \mathbf{Y}^t \mathbf{H}_X \mathbf{s} \\ \hat{\sigma}_s^2 &= \hat{\sigma}_w^2 = \frac{1}{4n} \|\mathbf{s} - \mathbf{X} \hat{\boldsymbol{\alpha}} - \mathbf{Y} \hat{\boldsymbol{\beta}}\|^2 \end{aligned}$$

$$\text{where } \mathbf{H}_Y = (\mathbf{I} - \mathbf{Y}(\mathbf{Y}^t \mathbf{Y})^{(-1)} \mathbf{Y}^t) \text{ and } \mathbf{H}_X = (\mathbf{I} - \mathbf{X}(\mathbf{X}^t \mathbf{X})^{(-1)} \mathbf{X}^t).$$

Proof Sketch: Part (a) follows from the E-step of Algorithm 1 and the observation that product of two Gaussian probability density functions results in Gaussian distribution centred at a mean weighted by the variances. Part (b) follows from the first order necessary conditions for optimizing the likelihood function in the M-step of Algorithm 1. \square

3.1 Reduction to Linear Least Squares Regression

We now consider the linear least squares regression problem obtained by eliminating the unobserved response W from (3.7) and (3.8). Let $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ and $\boldsymbol{\gamma}^t = [\boldsymbol{\alpha}^t, \boldsymbol{\beta}^t]$. Then, we have

$$s_i - \boldsymbol{\gamma}^t \mathbf{z}_i = \epsilon_{ws}, \quad \epsilon_{ws} \sim \mathcal{N}(0, \sigma_{ws}^2), \quad [i]_1^n, \quad (3.9)$$

where the error ϵ_{ws} is the sum of the two independent errors ϵ_w and ϵ_s so that $\sigma_{ws}^2 = \sigma_w^2 + \sigma_s^2$.

Theorem 3 *Let $(\hat{\boldsymbol{\alpha}}_{LS}, \hat{\boldsymbol{\beta}}_{LS})$ be the least squares estimators for the linear regression model in (3.9) and let $(\hat{\boldsymbol{\alpha}}_{MLE}, \hat{\boldsymbol{\beta}}_{MLE})$ be the maximum likelihood estimators in Theorem 2. Then, the estimators $(\hat{\boldsymbol{\alpha}}_{LS}, \hat{\boldsymbol{\beta}}_{LS})$ are identical to $(\hat{\boldsymbol{\alpha}}_{MLE}, \hat{\boldsymbol{\beta}}_{MLE})$ when $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ is a full column rank matrix. $[\mathbf{X}, \mathbf{Y}]$ is not a full column rank matrix, the optimal parameter estimates for the linear regression model in (3.9) are not unique, but they are still identical to the optimal estimates of the maximum likelihood problem in Theorem 2.*

Proof Sketch: The above equivalence follows from the fact that the maximum likelihood estimates $\hat{\sigma}_w^2 = \hat{\sigma}_s^2$ as in Theorem 2, which ensures that incomplete log-likelihood $L_D(\Theta)$ in (2.3) is linearly related to the least squares error of the linear model in (3.9). \square

The above equivalence also results in certain nice properties for the maximum likelihood estimators as the following Corollary shows.

Corollary 1 *The maximum likelihood estimators $\hat{\boldsymbol{\alpha}}_{MLE}$ and $\hat{\boldsymbol{\beta}}_{MLE}$ are unbiased as well as consistent estimators of the true parameters.*

Proof Sketch: The result follows directly from the observation that the true parameters in the joint parametric model (2.1) identical to that of the linear least squares model (3.9) and that the least squares regression estimators are unbiased as well as consistent estimators of the true parameters [10]. \square

Since the posterior distribution $\tilde{p}_i(w_i) [i]_1^n$ is Gaussian, the ML estimate for the target is just the expected value of the distribution, i.e., \hat{w}_i . Using Corollary 1, we can now prove the unbiasedness of this ML estimator as well.

Corollary 2 *The maximum likelihood estimator for the unobserved response \hat{w} in Theorem 2 is unbiased with respect to the true values.*

Proof Sketch: The above result follows from the relations $E[\mathbf{w}] = \mathbf{X} \boldsymbol{\alpha}^0$ and $E[\mathbf{s}] = E[\mathbf{w}] - \mathbf{Y} \boldsymbol{\beta}^0$, where $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$ are true parameters of (3.7) and (3.8). \square

Theorem 3 shows that we can solve the problem of estimating the unobserved target via a supervised learning approach on the surrogate target. This is of course beneficial from a computational perspective, as it allows us to harness the full power of linear regression methodology [16]. Among the things we can now do are variable selection methodologies, such as forward and backward selection, and analysis of variance (ANOVA) for testing goodness of fit for nested models.

The use of ANOVA is particularly interesting, since it allows us, to some extent, to test the conditional independence implied by our graphical model. Equation (3.9) defines the predictor matrix \mathbf{Z} as a concatenation of the columns of \mathbf{X}

and \mathbf{Y} . What if we wanted to extend the predictor matrix as $\tilde{\mathbf{Z}} = [\mathbf{X}^2, \mathbf{Y}^2]$, where we use \mathbf{X}^2 to denote a matrix of size $n \times m_1^2$ containing of all interactions between variables in \mathbf{X} , and similarly for \mathbf{Y}^2 ? Such a model would be completely consistent with both our linear model assumption and the graphical model in Figure 2, it would just be a more elaborate model, and an ANOVA can determine whether it is supported by the data.

But what if we also wanted to add interactions between variables in \mathbf{X} and variables in \mathbf{Y} ? That would be a violation of the conditional independence assumption inherent in Figure 2, since it defies the additive representation in (3.7, 3.8). Thus, if an ANOVA would tell us that a model with interactions between variable in \mathbf{X} and \mathbf{Y} is superior, that would cast a severe doubt on our independence assumptions and/or our parametric assumptions. In Section 5, we show an example of such an ANOVA on our customer wallet prediction problem, and demonstrate that the additivity hypothesis — and hence, our conditional independence and parametric assumptions — cannot be rejected.

4. SIMULATION EXPERIMENTS

We now present results on simulation data to demonstrate the effectiveness of our unsupervised learning methodology. We consider two learning tasks: one involving linear least squares regression suitable for a continuous real valued target, and a second one involving logistic regression tailored for a binary valued target. In both cases, we show that even without any training data on the target, one can obtain good prediction accuracy by exploiting the conditional independence relations between the various predictors. Synthetic data was used for both sets of experiments in order to ensure that the underlying generative models satisfy the desired conditional independence requirements.

4.1 Gaussian Linear Models

The first task involves predicting a continuous real valued target W using predictors S , X and Y of similar type. We assume a generative model identical to the one described in Section 3 and evaluate the performance of our unsupervised multi-view approach, which in this case was shown to be equivalent to a single least squares regression involving S , X and Y .

For each run of our experiments, we generated data using the coupled linear models in (3.7) and (3.8) after randomly generating the attribute sets X , Y and various model parameters $\alpha, \beta, \sigma_w, \sigma_s$. Table 1 shows the details of the dataset generation mechanisms and experimental setups. Using this data, we compared the performance of our “unsupervised” multi-view approach with unobserved target with standard least squares regression that requires training data on W , and directly builds a linear model on all the predictors. In both the cases, we assumed Gaussian linear models so as to match the original generative model. The quality of prediction in each case was measured in terms of mean squared error of the target on a hold out data not used for training.

Multi-view approach vs. supervised regression

Figure 3 shows the prediction accuracy of the unsupervised multi-view approach and that of regular least squares regression approach with varying number of samples, in the

Gaussian1 setting of Table 1. For instance, in the figure, we can notice that the multi-view method with about 200 samples provides an accuracy similar to the supervised approach with 70 labeled samples. This ratio of about 3 times more data needed to achieve comparable performance in the unsupervised vs. supervised setting is roughly maintained throughout the range of sample sizes. Thus, the multi-view unsupervised approach behaves much in the same way as a supervised learning approach would, and an increase in the number of samples reduces the variance in the parameter estimates and results in a better prediction model.

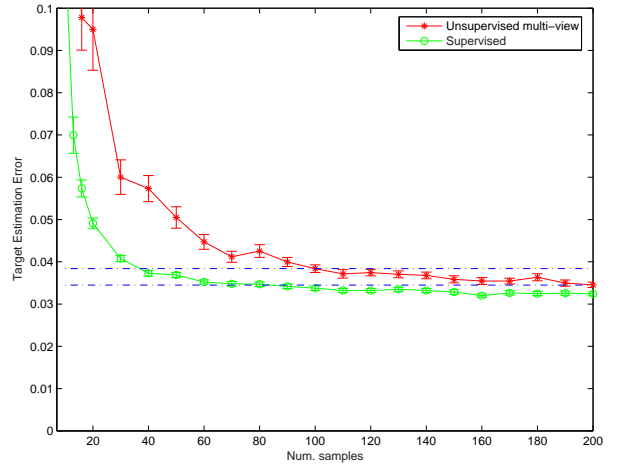


Figure 3: Target prediction error using multi-view approach and regular least squares regression with varying number of samples. Num. Attributes =3 and variances $\sigma_s = \sigma_w = 0.5$.

Variation with number attributes

Figure 4 shows the performance of both approaches in the settings *Gaussian2*, *Gaussian3*: we fix the number of samples as shown in Table 1 and vary the number of attributes in X and Y . In this case, the average prediction error as well as the variability in the errors goes up as the number of attributes (and hence parameters estimated) increase. The curves for both methods generally track each other closely, indicating that this ratio of approximately 3 times as much data for comparable performance is not strongly affected by the number of parameters. This relation breaks down when the number of parameters gets close to 35, which is the number of data points in the smaller supervised sample, and the corresponding supervised least squares problem approaches singularity.

4.2 Logistic Regression Model

To illustrate the generality of our framework, we consider a second task that corresponds to a classification scenario where the goal is to predict a binary-valued target W using predictors S , X and Y where S is binary valued while X and Y are set of continuous real valued attributes. For this case, we assume the following logistic generative model:

$$\begin{aligned} \text{logit}(p(W = 1|X)) &= X\alpha \\ \text{logit}(p(S = 1|W, Y)) &= W + Y\beta \end{aligned}$$

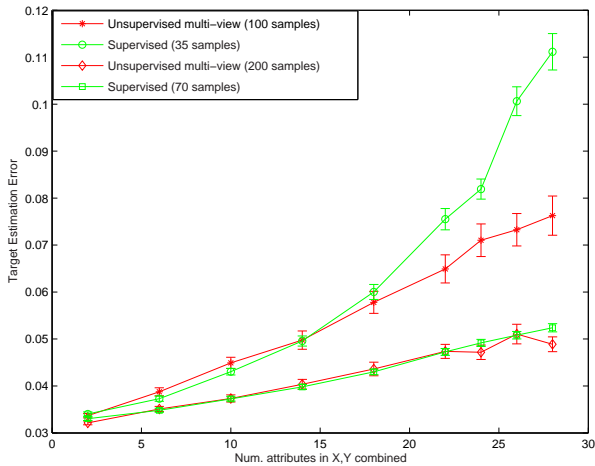


Figure 4: Target prediction error using multi-view approach with varying number of attributes in X and Y . Variances $\sigma_s = \sigma_w = 0.5$.

Since the parameter estimation problems corresponding to the above coupled models cannot be reduced to a simpler form as in case of Gaussian linear models, we follow the EM-based algorithm outlined in Section 2 to estimate the unobserved target W in terms of the other variables. As in the previous case, the data was generated using the above logistic models after randomly choosing the attribute sets X , Y and various model parameters α, β , and the performance of our “unsupervised” multi-view approach was compared with that of a regular logistic regression model based on all predictors, i.e., S, X and Y . Since this is a classification task, the quality of prediction was measured in terms of the misclassification error on a hold out set.

Figure 5 shows the misclassification error for the unsupervised multi-view approach and that of regular logistic regression approach with varying number of samples. As in the case of the least squares regression task, we find that the unsupervised multi-view approach can provide good accuracy without using the target information or class labels.

5. CASE STUDY: CUSTOMER WALLET ESTIMATION

In this section, we provide a more detailed description of our motivating customer wallet estimation problem, which is the main focus of our earlier work [15]. We discussed the business motivation and the general problem setting in great detail in this earlier paper, and also reviews solution approaches based on quantile regression, which model a somewhat different definition for wallet than we are attempting to model here (*REALISTIC* versus *SERVED*). We refer the reader to [15] for more details, and concentrate here on the current formulation. In this section, we apply the linear regression methodology of Section 3 to a dataset of IBM customers, use an ANOVA test to validate the conditional independence assumption under these modeling assumptions, and discuss some of the practical issues involved in obtaining useful predictions.

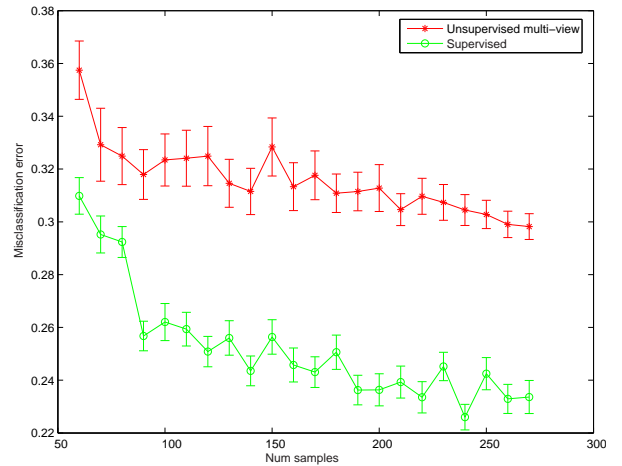


Figure 5: Misclassification error using multi-view approach and regular logistic regression with varying number of samples. Num. Attributes = 10.

We have described our generic view of IT purchase process previously in Sections 1. Recall that we have a set of firmographic variables X , an unobserved wallet W , and customer’s actual spending with IBM S and a set of IBM relationship variables Y . In this case study, the variables in X are publicly available data from Dun & Bradstreet², containing information about a company’s industry, size, financials, etc. The customer’s actual spending with IBM S and the IBM relationship variables Y comes from IBM’s proprietary data warehouse, containing detailed information on IBM’s sales by product and customer and other relationship indicators. For brevity and confidentiality reasons, we omit the detailed description of these variables.

We first assume the causal and conditional independence relationships described in Figure 1. As in the setup of Section 3, we additionally assume that the discriminative models for the target $p(W|X)$ and the surrogate response $p(S|W, Y)$ are linear in an appropriate representation of the variables and have Gaussian noise.

Throughout our analysis, we transform all monetary variables — in particular, $S = \text{IBM SALES}$, used as response — to the log scale, due to the fact that these numbers have very long tailed distributions. It has often been observed that monetary variables have exponential-type distributions and behave much better when log-transformed (cf. the oft-cited “pareto rule”).

Thus, our two fundamental modeling equations are:

$$\log(w_i) = f_\alpha(\mathbf{x}_i) + \epsilon_i, [i]_1^n \quad (5.10)$$

$$\log(s_i) - \log(w_i) = g_\beta(\mathbf{y}_i) + c_0 + \delta_i, [i]_1^n \quad (5.11)$$

where c_0 is a constant and the second equation is conditional on W , and $\epsilon_1, \dots, \epsilon_n$ and $\delta_1, \dots, \delta_n$ are i.i.d Gaussian random variables denoting the noise (as we showed in Section 3, our setup inevitably leads to assuming equal noise variance). When equations (5.10) and (5.11) are added together, we

²<http://www.dnb.com>

Datasets	#Attributes in X & Y	#Train Samples	#Test Samples	Other Params	#Runs
<i>Gaussian1</i>	6	from 8 to 200	100	$\sigma_s = \sigma_w = 0.5$	100
<i>Gaussian2</i>	between 1 and 14	200 (unlabeled) 70 (labeled)	100	$\sigma_s = \sigma_w = 0.5$	100
<i>Gaussian3</i>	between 1 and 14	100 (unlabeled) 35 (labeled)	100	$\sigma_s = \sigma_w = 0.5$	100
<i>Logistic</i>	5	between 60 and 270	100		100

Table 1: Details of datasets and experiments.

obtain the equivalent linear regression, as discussed in Section 3.1:

$$\log(s_i) = f_\alpha(\mathbf{x}_i) + g_\beta(\mathbf{y}_i) + (\epsilon_i + \delta_i), [i]_1^n \quad (5.12)$$

From Theorem 3, we observe that a maximum likelihood solution $\hat{\alpha}, \hat{\beta}$ of (5.12), obtained through linear least squares regression, is also a maximum likelihood solution for (5.10, 5.11).

We consider two forms for f_α and g_β :

- (A) Simple linear model with no interactions, i.e., $f_\alpha(\mathbf{x}_i) = \alpha^t \mathbf{x}_i$ and similarly for g_β .
- (B) Linear model that includes interactions only within the variables in X and Y , but not between variables in both groups. Thus, we assume $f_\alpha(\mathbf{x})$ only consists of factors $= x_k \cdot x_l$, and similarly for g_β .

Finally, we also consider a linear model (C) with all possible interactions between X and Y . This model is not consistent with the representation in 5.12, and we use this model to validate the conditional independence assumption via an analysis of variance (ANOVA). In practice, we also apply variable selection, leading to use of other models. However for our illustration purposes we limit the discussion here to these three models.

Table 2 shows an ANOVA table resulting from fitting these three nested models to our data. As we can see, the within-view interaction model (model B) is clearly a better fit than the no interaction model(model A) (F-statistic p value of about 10^{-4}), i.e., the data supports the usefulness of within- X and within- Y interactions compared to the no-interaction model. On the other hand, the all-interaction model (model C) does not significantly improve our fit over model B (p value of 0.08)³. If the improvement were significant, it could be taken as evidence against the conditional independence we assume (and the validity of the graphical model), since one possible reason for the model C being better would be if the errors in (5.10) were not independent of the variables in Y or the errors in (5.11) were not independent of the variables in X . As it is, the results can be taken as validation (although clearly not as proof) of both our conditional independence assumptions, and our parametric model assumptions.

³While this p value is quite low, consider we are using a large number of observations (2000), while the most complex model has less than 200 DF. Thus, we can assume that this test is quite powerful against many reasonable alternatives, and that non-rejection of the sufficiency of model B is by no means a trivial outcome.

Model	RSS	DF	F statistic	P value
No interaction (A)	10736.7	21		
Within-group interaction (B)	10033.5	75	1.7448	0.00011
All interaction (C)	9382.5	100	1.2114	0.081

Table 2: ANOVA table for linear models.

We next want to recover wallet predictions from our model, and investigate their quality, which we have limited ability to do since the actual wallet is never observed. First, we have to consider an issue, which we have glossed over in our model description so far, relating to the existence of intercept (often referred to as “bias” in machine learning) in our basic models (5.10, 5.11). Recall that our main result in Theorem 3 requires that the matrix $[f_\alpha(\mathbf{X})g_\beta(\mathbf{Y})]$ be of full column rank. Consequently, we cannot estimate separate intercepts for (5.10, 5.11) through the linear regression in (5.12), but rather can only estimate the sum of the two intercepts by adding an intercept to (5.12)⁴. Assuming we want to allow an intercept in both of (5.10, 5.11) (which we typically want to do), we need to use additional, external information to estimate \hat{c}_w , the intercept in (5.10), leading us to predict $\log(\hat{w}_i^{new}) = \log(\hat{w}_i) + \hat{c}_w$ where \hat{w}_i is the old estimate. In the case of wallet estimation, the additional piece of information we typically use is based on our expectation that every customer company’s IT wallet should always be smaller than that customer’s revenue (which is known since it is included in X — let’s denote it by R) and larger than the customer’s IT spending with IBM S . Thus, if we denote every customer’s revenue by r_i , and given estimates $\hat{f}_\alpha, \hat{g}_\beta, \hat{c}_0$ from (5.12), we can estimate \hat{c}_w as the value which minimizes the number of such order violations:

$$\hat{c}_w = \arg \max_c \{ |[i]_1^n : r_i \geq \hat{w}_i \exp(c) \geq s_i | \}$$

In Figure 6, we show the results of applying this full estimation methodology to our wallet data. For this purpose, we re-fit model B and estimated \hat{c}_w using 1500 observations only. The plot shows wallet estimates *on the 500 hold out data samples not used for fitting* compared to the observed customer revenue R and IBM sales S . The sanity check of preserving the order $R \geq \hat{W} \geq S$ holds very well, as only a handful of predictions give $S > \hat{W}$ and only one gives $\hat{W} > R$.

6. RELATED WORK

⁴The remark after Theorem 3 clarifies that this is a fundamental non-estimability property of the probabilistic setup, not a problem in the reduction to linear regression

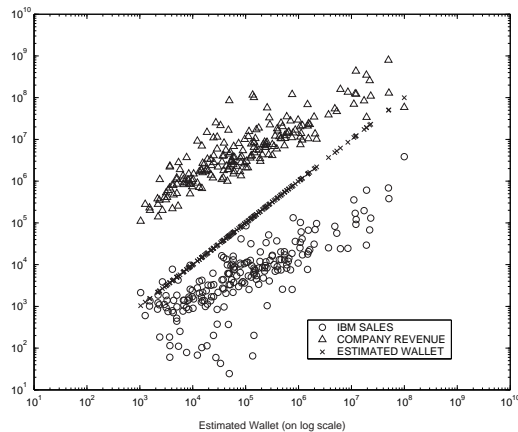


Figure 6: Evaluating wallet predictions on hold out data

Our current work is primarily related to four main areas — (i) statistical market analysis, (iii) Bayesian modeling and maximum likelihood estimation, (ii) multi-view learning, and (iv) least squares regression analysis. In particular, our work is motivated by the customer wallet estimation problem, which is an important market analysis problem while the methodology we adopt and theoretical results we demonstrate are closely related to Bayesian network learning and least squares regression.

Statistical Market Analysis. Most of the classical market analysis approaches such as life time value modeling [14] focus on sales history. Recent work [7, 9] shows that the share-of-wallet is a better indicator of the customer growth potential. However, there has been relatively little work on designing principled statistical methods for estimating the share-of-wallet or equivalently, the wallet itself. Most of the existing wallet modeling work [6, 5] involves building predictive models using self-reported wallets of the customers, which are often unreliable. A recent work [15] presents novel predictive techniques for estimating the “realistic” wallet, i.e., defined as a certain high percentile of the spending distribution using quantile regression and k -nearest neighbor approaches. The current work is also an effort in the same direction, but it differs from [15] in the definition of the wallet and the modeling assumptions.

Multi-view Learning. Recently, there has been much interest in the area of multi-view learning, which deals with learning from multiple sets of features that provide “independent” information about the desired target. Most of the existing work in this area such as Yarowsky’s algorithm[17], co-training [3], co-EM [13] focuses on classification, usually in semi-supervised setting. Our learning methodology is similar to the co-training and its variants in the sense that we assume the same conditional independence relations among multiple views. However, unlike co-training, we do not make any additional compatibility or learnability assumptions. Instead, we quantify the consistency of the views in terms of the observed discriminative likelihood of a Bayesian network that conforms to the conditional independence relations among those views and learn the desired

target by optimizing this quantity.

Bayesian Network Learning Our work is intimately connected with Bayesian network learning [8] as the core idea in our “unsupervised” regression methodology is to exploit the causal relations between the target and predictors to obtain a Bayesian network that satisfies a special property, i.e., the Markov blanket of the target can be partitioned into direct, surrogate and indirect predictors. This transformation allows us to employ the standard Bayesian network learning methods for estimating the unobserved or missing target values using a maximum likelihood formulation and EM-based algorithm [4]. Further, due to the special structure of the Bayesian networks we consider, the resulting maximum likelihood estimation problem turns out to be convex for a large class of parametric models so that the EM converges to a global optimum unlike in the general Bayesian network learning case where we can only obtain local optimum. Though our approach is similar in spirit to other maximum likelihood estimation based unsupervised learning techniques focused on classification, for example, learning mixtures of models [2], there is an important difference since we do not assume knowledge of the parametric form of the conditional distribution of the predictors given the target unlike in a classification scenario where it is relatively straightforward to model the class conditional distributions.

Least Squares Regression. Least squares regression has been known [16] to be equivalent to performing maximum likelihood estimation using a Gaussian linear model. In our current work, we demonstrate that this equivalence extends to maximum likelihood estimation over coupled Gaussian linear models. This reduction enables us to directly apply the extensive model selection and variable selection methods developed for least squares regression to our “unsupervised” regression task.

7. CONCLUSION

We propose a fairly general multi-view learning methodology for unsupervised settings where the target is not observed. Our approach exploits the causal relations between an unobserved target and different subsets of the available predictors to obtain a Bayesian network with a special structure. Using this Bayesian network and domain-dependent distribution assumptions, we transform the regression problem into a standard Bayesian network learning problem, which can be solved using EM. We show that it converges to the global optimum for a large class of parametric distributions corresponding to exponential linear models.

We then present a detailed analysis of the specific, but widely applicable case involving two views and Gaussian linear models and show that it can be reduced to a single least squares regression problem. This reduction is practically significant as it allows us to perform variable selection and test the independence assumptions underlying the Bayesian network.

Experimental evaluation of our methodology on our motivating customer wallet estimation problem as well as on simulation data indicates the effectiveness and flexibility of our approach.

Although we have focused in this paper only on an “unsuper-

vised” setting, our methodology permits a natural extension to a semi-supervised setting where the target is observed for a subset of data samples.

We believe the proposed methodology and the reduction to least squares regression are likely to be useful for other real-life business and scientific applications, beyond our wallet estimation problem, which have the conditional independence and causality structure which our formulation supports.

8. REFERENCES

- [1] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.
- [2] C. M. Bishop. Latent variable models. In *Learning in Graphical Models*, pages 371–403. MIT Press, 1998.
- [3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory*, 1998.
- [4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
- [5] R. Du and W. Kamakura. Imputing customers share of category requirements., 2005.
- [6] epsilon.
- [7] R. Garland. Share of wallet’s role in customer profitability. *Journal of Financial Services Marketing*, 8(8):259–268, 2004.
- [8] D. Heckerman. A tutorial on learning with bayesian networks. Technical report, Microsoft Research, 1995, MSR-TR-95-06, 1995.
- [9] T. Keiningham, T. Perkins-Munn, and H. Evans. The impact of customer satisfaction on share-of-wallet in a business-to-business environment. *Journal of Service Research*, 6(1):37–50, 2003.
- [10] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2003.
- [11] S. Merugu, S. Rosset, and C. Perlich. A new multi-view regression method with an application to customer wallet estimation. Technical report, IBM T.J. Watson Research, 2006, 2006.
- [12] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. MIT Press, 1998.
- [13] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *CIKM*, pages 86–93, 2000.
- [14] S. Rosset, E. Neumann, U. Eick, N. Vatnik, and Y. Idan. Customer lifetime value modeling and its use for customer retention planning. In *KDD*, pages 332–340, 2002.
- [15] S. Rosset, C. Perlich, B. Zadrozny, S. Merugu, S. Weiss, and R. Lawrence. Wallet estimation models. In *Proceedings of the International Workshop on Customer Relationship Management: Data Mining Meets Marketing*, 2005.
- [16] S. Weisberg. *Applied Linear Regression*. Wiley, 1985.
- [17] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, 1995.