Bayesian Visual Analytics: BaVA

FODAVA Technical Report

Leanna House, Scotland Leman, and Chao Han

Abstract

Large, high dimensional datasets generally contain information in small, concentrated regions of the data space. To extract this information, it is necessary to draw on several fields and use a variety of tools. We develop a new analytics framework that merges two areas of research, Bayesian Statistics and Visual Analytics. Mathematical and statistical disciplines rely on model based formulations which make use of structured parameterizations; whereas, visualizations of high dimensional data provide a means for non-quantitative experts to make sense of the data. However, coherent organization of data displays is often difficult. In the Bayesian Visual Analytics (BaVA) paradigm, we synthesize these tools to make cohesive visualizations that are adjustable. We consider display adjustments to be reliable feedback concerning the underlying analytical approach and necessary for "sense-making". This paper will focus on presenting the BaVA process, the formal descriptions of cognitive and parametric feedback, and some illuminating examples.

Keywords: Bayesian, Visual Analytics, Elicitation, Sequential Updating, Sense-making, Data Mining, High Dimensional Data, Statistical Visualization

Acknowledgements: This research was funded by the National Science Foundation, Computer and Communications Foundations; #0937071. The authors thank Chris North, Dipayan Maiti and Alex Endert for their insights and helpful suggestions.

1 Introduction

Visual Analytics (VA) is "the science of analytical reasoning facilitated by interactive visual interfaces" (Thomas and Cook, 2005) and combines research in data management and mining, statistics, information visualization, human cognition, and human-computer interaction. It emerged from the effort to assess massive amounts of data efficiently in order to identify interesting structures and uncover new information. Crucial to any data assessment is "sense-making" (Lederberg, 1989; Thomas and Cook, 2005); i.e., the combining of external information in data with expert judgement. The goal of this paper is to develop a framework for analyzing data that relies on both the knowledge/capabilities of experts and complex quantitative data assessments.

We develop a Bayesian Visual Analytics (BaVA) framework to create visualizations that enable experts to interact with data, test hypotheses, and formulate knowledge instantly. The inspiration for our work is that sense-making as defined by Lederberg (1989) and Thomas and Cook (2005) involves assimilating new information with old and parallels the formulation of a Bayesian model; a Bayesian model includes prior judgements, a model for external data, and an assessment of posterior distributions. Additionally, similar to sense-making, posterior distributions may be updated seamlessly when new or more information becomes available.

Mechanistically, BaVA entails characterizing complex datasets probabilistically and displaying relevant posterior results for experts to assess. If the results fail to reveal known or intuitive data structure, experts may adjust displays accordingly. The novelty is that we consider display adjustments to be reliable, expert feedback or additional data concerning the underlying Bayesian model. Thus, we quantify and express the feedback in probabilistic form so that we can use makes Bayesian sequential updating (Spiegelhalter and Lauritzen, 1990; West and Harrison, 1997) to adjust the underlying probabilistic model and, subsequently,

the display.

Critical to the BaVA paradigm is characterizing data using parametric, probabilistic models while reducing its dimension for visualisation. For example, factor analyses (Cattell, 1965; Press and Shigemasu, 1989; Lewin-Koh and Amemiya, 1998) characterize high dimensional data $\boldsymbol{d}$ as a function of latent reduced dimensional parameters $\boldsymbol{r}$. Taking a Bayesian approach to assess $\boldsymbol{d}$, we recognize that $\boldsymbol{r}$ is uncertain, assign an appropriate or noninformative (Yang and Berger, 1997) prior distribution, and display a highly probable *a posteriori* value for $\boldsymbol{r}$ in a two- or three- dimensional graph. The graphical axes designate a portion of the parameter space for $\boldsymbol{r}$, so any coordinate within the axes may represent a realistic or plausible value for $\boldsymbol{r}$. In turn, we have the capacity to interpret visual feedback (e.g., observation adjustments) parametrically based upon the underlying probabilistic model.

Our quantification of feedback is similar in spirit to the specification of subjective prior distributions (Buxton, 1978; Goldstein, 2006) in that we parametrize expert judgments (in our case, display adjustments). Although, unlike standard prior elicitation procedures, we need not have a facilitator (Garthwaite et al., 2005) and we avoid any ambiguity concerning the communication between statisticians and experts. Namely, psychological and knowledge-base barriers may prevent experts from understanding fundamental statistical concepts, including expectation and variance, which are arguably essential for specifying realistic prior distributions (Kadane and Wolfson, 1998; Daneshkhah, 2004). For BaVA, experts need not understand basic statistics to provide useful feedback. Provided intuitive data visualizations, they only need to understand their field and make judgements at the observation level; e.g., express their judgments concerning pairwise relationships between observations.

For example, a factor modeling approach known as Probabilistic Principal Component Analysis (PPCA) (Tipping and Bishop, 1999) projects high dimensional data in the direction with the largest variance to a lower dimensional latent space (similar to Principal Component Analysis). If a dataset contains clusters of observations, but the within cluster variance is

smaller than the between cluster variance, a visualization of the low dimensional latent data will not reveal the clusters. However, if experts believe that two observations belong to the same (or different) cluster(s) and proximity on the computer screen is a measure of similarity, experts have an opportunity to inject their judgements into the statistical analysis by dragging the observations together (or apart).

The outline for our explanation of BaVA is as follows. We start by reiterating the motivation for our work in Section 2 based on a simulated example. We then explain Bayesian fundamentals and establish notation in Section 3. In Section 4, we detail the BaVA process steps which result in malleable, adaptable displays and apply them to the simulated example in Section 5. In Section 6, we exemplify the benefits of BaVA for two real-world applications that are similar in nature to the simulated example and concern the cost of education and functional genomics. In this section we also propose two procedures that may assist experts to adjust displays wisely. Since BaVA is a framework and not an application-specific VA tool, we provide another use of BaVA within the context of Multi-Dimensional Scaling (MDS) (Torgerson, 1958; Kruskal and Wish, 1978) in Section 7. In this section, we develop an appropriate probabilistic model that relies on MDS machinery and the means to inject and parametrise feedback for real-world example. We conclude with a discussion of our work in Section 8.

2 Motivation for BaVA: Simulated Example

Current visualizations tend to display inflexible, deterministic transformations of data that inherently separate data visualization from the visual synthesis process. Analysts cannot manipulate displays to inject domain-specific knowledge into the image and assess the merger of their expert judgment with the data formally. This may inhibit sense-making when the data transformation masks known or intuitive data structure.

For example, we simulated a three dimensional dataset $\boldsymbol{d}$ where $\boldsymbol{d} = [d_1 ... d_n]$ and $d_i =$

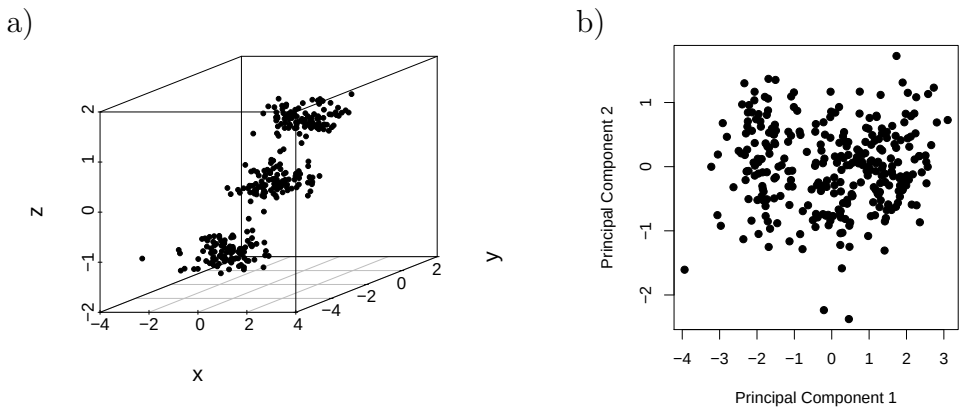

Figure 1: Figure a) is a three dimensional scatter plot of simulated data that contains three clusters. Plot b) plots the top two Principal components of the data displayed in plot a).

$[x_i, y_i, z_i]'$ ($\boldsymbol{d}$ is an $n \times 3$ data matrix and d_i is a 3×1 observation vector). These data contain three clusters which are easily seen in Figure 1a). Suppose that we want to discover the clustering structure visually using a display with fewer dimensions. Since a common data mining (DM) approach used to reduce data dimensions is Principal Components Analysis (PCA) (Pearson, 1901; Jolliffe, 2002; Torokhti and Friedland, 2009), we plot the top two principal components in Figure 1b). Alas, the clusters in the two-dimensional plot are indistinguishable.

Projection based methods are not formally clustering nor structure-discovering algorithms, yet they are still often used for high dimensional visualization. Traditional PCA projects data in the directions with the largest sample variability, and, for this example, is not appropriate for uncovering structure. The variance of the simulated data within clusters is larger than the variance between clusters so that the direction in which the data are projected results in occluded clusters. What can we do?

We propose to use the new BaVA framework for PCA so that the user may manipulate/interact with the data and, possibly, help PCA to reveal the hidden structure. Since BaVA relies fundamentally on Bayesian statistics, we provide a brief summary of Bayes while

establishing notation in the next section.

3 Formal Bayesian Analysis

Bayesian statistics is founded on rich, philosophical principles (Ramsey, 1926; Savage, 1954; Jeffreys, 1961; Good, 1983; Jaynes, 1983) that we do not discuss here. Rather, we focus on the fundamentals of Bayesian inference and how Bayesian statistical models may facilitate the sense-making process.

In Bayesian statistics, as in classical statistics, the first step is to specify a probability model or sampling distribution $\pi(\boldsymbol{d}|\boldsymbol{\theta})$ for data $\boldsymbol{d}$ that depends upon unknown parameter $\boldsymbol{\theta}$. One of many reasons for which the model is important is that it specifies important features of the data that are uncertain. In turn, experts may focus on these features and make assessments of the unknowns that may also be characterized by a prior probability model, $\pi(\boldsymbol{\theta})$. Given data $\boldsymbol{d}$, we apply Bayes' Rule and update the prior distribution $\pi(\boldsymbol{\theta})$ to derive the posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{d})$.

Inference about $\boldsymbol{\theta}$ follow from summaries and interpretations of $\pi(\boldsymbol{\theta}|\boldsymbol{d})$. One such summary is a comparison between the prior and posterior distributions. When they are similar, we could infer that the data support the current understanding of $\boldsymbol{\theta}$; and, when they differ, the data suggest a need to change the current understanding of $\boldsymbol{\theta}$. Thus, built into the Bayesian paradigm is a means to assess and update judgements of $\boldsymbol{\theta}$ which lends itself nicely to further updating when more information about $\boldsymbol{\theta}$ becomes available.

The procedure of Bayesian sequential updating is straightforward and allows experts to incorporate new information into a current analysis. For example, let $\boldsymbol{d^{(1)}}$ and $\boldsymbol{d^{(2)}}$ represent datasets that contain information about $\boldsymbol{\theta}$ and were collected at, say, time 1 and 2 respectively. After time 1, a posterior distribution for $\boldsymbol{\theta}$ is derived, $\pi(\boldsymbol{\theta}|\boldsymbol{d^{(1)}})$. After time

2, we may assess $\boldsymbol{\theta}$ as follows:

$$\pi(\boldsymbol{\theta}|\boldsymbol{d^{(1)}}, \boldsymbol{d^{(2)}}) = \pi(\boldsymbol{d^{(2)}}|\boldsymbol{\theta}, \boldsymbol{d^{(1)}})\pi(\boldsymbol{\theta}|\boldsymbol{d^{(1)}})/\pi(\boldsymbol{d^{(2)}}|\boldsymbol{d^{(1)}}), \tag{1}$$

When $\boldsymbol{d^{(1)}}$ and $\boldsymbol{d^{(2)}}$ are independent conditional on $\boldsymbol{\theta}$, $\pi(\boldsymbol{d^{(2)}}|\boldsymbol{\theta}, \boldsymbol{d^{(1)}}) = \pi(\boldsymbol{d^{(2)}}|\boldsymbol{\theta})$. Bayesian sequential updating is often used for streaming data or in the event that an experiment is repeated. However, an original dataset, $\boldsymbol{d^{(1)}}$, need not be of the same form as the subsequent data $\boldsymbol{d^{(2)}}$; e.g., the two datasets may differ in type or units. Within the context of sense-making, the additional source of information $\boldsymbol{d^{(2)}}$ could consist of additional data collected later in time or, simply, additional expert judgement.

When experts have the opportunity to synthesize their judgements with data *a posteriori*, they may wish to inject additional information into the statistical analysis. Using Bayesian sequential updating machinery, we develop BaVA so that experts can assess aspects of the posterior distribution before and after feedback to test new hypotheses and ultimately form inferences that "make sense."

4 The BaVA Process

To create and interpret malleable visualizations, we propose a five step procedure that is displayed in Figure 2. Provided data $\boldsymbol{d}$, the first step is to characterize it with a probability model that depends upon $\boldsymbol{\theta}$ and derive the posterior distribution of $\boldsymbol{\theta}$. The second step is to display a posterior estimate(s) of $\boldsymbol{\theta}$ in a meaningful, adjustable visualization which we denote as v. The third step prompts experts to inject feedback by adjusting the display, if desired. We refer to this manual or visual feedback as *cognitive feedback* and denote it as $f^{(c)}$. Since the display is based on the probability models, we can interpret $f^{(c)}$ quantitatively and parameterize it so that its distribution is a function of $\boldsymbol{\theta}$. We denote the parameterized feedback as $f^{(p)}$ and consider its specification to be step 4 in Figure 2. The final step is

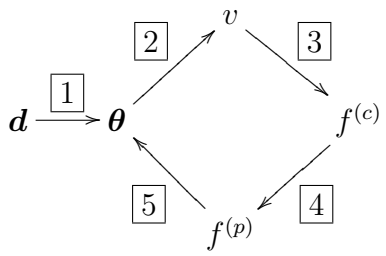

Figure 2: Schematic illustration of the BaVA process.

to update the current probability model using Bayesian sequential updating which, in turn, allows the procedure to repeat. For the remainder of this section, we describe each step in detail.

4.1 Step 1: Form Bayesian Inferences

The Bayesian visualization process begins like any typical analysis that is described in Section 3; given a sampling distribution $\pi(\boldsymbol{d}|\boldsymbol{\theta})$ and a set of prior beliefs $\pi(\boldsymbol{\theta})$, the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{d})$ is formed through Bayes' rule. Although, unlike typical analyses, we must consider the next step (visualization) in the BaVA process when choosing a reasonable model for the data. That is, in order for BaVA to be a success, we must select a probability model that both characterizes the data well and relies on unknowns $\boldsymbol{\theta}$ (or a function of the unknowns) that are able to be graphed in three or fewer dimensions. Often, the dimension of $\boldsymbol{\theta}$ is far less than the dimension of the raw data $\boldsymbol{d}$, thus sensible, dimension reduction models are important to the BaVA process.

4.2 Step 2: Construct a Visual Representation

A BaVA display v must satisfy two fundamental criteria. First, the visualization must be easy to understand and adjust in ways that are intuitive to the user (Icke and Sklar, 2009). Second, any adjustment to the visualization must relate directly to both the data and inferences formed/tested about $\boldsymbol{\theta}$. For example, we advocate selecting a highly probable

value (e.g., the posterior mean or maximum *a posteriori* (MAP) estimate of $\boldsymbol{\theta}$) which we denote as $\hat{\boldsymbol{\theta}}$ to display. For this paper, we define the visualization v to be a deterministic transformation of $\hat{\boldsymbol{\theta}}$, $v = g(\hat{\boldsymbol{\theta}})$. Thus, v conditional on $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ is known with probability one. As we discuss in Section 8, future work will explore including the uncertainty in v in the analysis of $\boldsymbol{\theta}$.

4.3 Steps 3 and 4: Enable User Feedback

Using a malleable visualization, an expert may wish to inject additional knowledge into the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{d})$ because 1) the probabilistic model is inadequate for constructing reasonable inferences; 2) the transformation $g()$ guiding the visualization masks important information contained in the data, and/or 3) the user wants to explore alternative visualizations. We define the information that an expert may wish to inject as feedback f which we decompose into cognitive $f^{(c)}$ and parametric $f^{(p)}$ feedback; $f = \{f^{(c)}, f^{(p)}\}$. We consider f to be an random variable with a distribution that is equal to the joint distribution, $\pi(f^{(c)}, f^{(p)}|v, \boldsymbol{\theta})$.

4.3.1 Cognitive Feedback

How an expert interacts with a visualization is inherently random. For example, an expert might choose to adjust a display while considering a set of comparable movements. Thus, when experts alter a display, they must provide a measure of certainty or a weight to which we should consider their feedback in the analysis of $\boldsymbol{\theta}$ with respect to the current assessment. Let κ represent the expert specified weight, where $\kappa \in [0, 1]$. We consider the display adjustment and κ as partial prior specifications (Goldstein and Woof, 2007) for the distribution $\pi(f^{(c)}|v)$. It is important to note that we do not need to know the mathematical form of this distribution.

4.3.2 Parametric Feedback

We link $f^{(c)}$ to the parameter set $\boldsymbol{\theta}$ through a transformation $h()$ and define parametric feedback as $f^{(p)} = h(f^{(c)})$, where $f^{(p)}$ has distribution $\pi(f^{(p)}|f^{(c)}, \boldsymbol{\theta})$. We cannot state $h()$ explicitly because it is application specific, but the specification of $h()$ stems from the following rational. Information is lost when low dimensional representations, i.e., displays of $\boldsymbol{\theta}$, are used to portray high dimensional datasets, and the loss is greater for some data dimensions than others. Adjustments to low dimensional displays suggest a need to re-weight the data dimensions for the analysis of $\boldsymbol{\theta}$. Thus, crudely, the task of specifying $h()$ entails identifying the degree to which dimensions in the high-dimensional dataset are represented in displays and up- or down- weighting the dimensions according to the cognitive feedback.

When selecting a distribution or model for $f^{(p)}$, we consider two issues. First, we must provide a reasonable model for $f^{(p)}$ that complies with the nature of the expert-provided information; e.g., the model has an appropriate domain, expectation, and variance for the data collected from the expert feedback. Second, we must select a model for $f^{(p)}$ that eases subsequent computation in Section 4.4 so that adjustments to the visualization are instantaneous. Time consuming Markov Chain Monte Carlo (MCMC) methods are not ideal.

4.4 Step 5: Update the Model Based on Feedback

At the conclusion of step 4, we have expert feedback f that we would like to include in our posterior analysis of $\boldsymbol{\theta}$. Thus, we use Bayesian sequential updating as described in Section 3 to assess $\pi(\boldsymbol{\theta}|f, v, \boldsymbol{d})$,

$$\begin{aligned}
\pi(\boldsymbol{\theta}|f, v, \boldsymbol{d}) &= \frac{\pi(f, v|\boldsymbol{\theta}, \boldsymbol{d})\pi(\boldsymbol{\theta}|\boldsymbol{d})}{\int \pi(f, v|\boldsymbol{\theta}, \boldsymbol{d})\pi(\boldsymbol{\theta}|\boldsymbol{d})d\boldsymbol{\theta}} \\
&\propto \pi(f|v, \boldsymbol{\theta})\pi(v|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{d}) \\
&= \pi(f^{(p)}|f^{(c)}, \boldsymbol{\theta})\pi(f^{(c)}|v)\pi(v|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{d})
\end{aligned}$$

Since we stated earlier that, for this paper, we have a deterministic method to create the visualization v ($\pi(v|\boldsymbol{\theta}) = 1$) and $f^{(c)}|v$ is independent of θ, $\pi(\boldsymbol{\theta}|f, v, \boldsymbol{d})$ is

$$\pi(\boldsymbol{\theta}|f, v, \boldsymbol{d}) \quad \propto \quad \pi(f^{(p)}|f^{(c)}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{d}).$$

If $\pi(f^{(p)}|f^{(c)}, \boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{d})$ are conjugate, MCMC is obviated and $\pi(\boldsymbol{\theta}|f, v, \boldsymbol{d})$ is straightforward to derive.

The BaVA process is iterative and may continue until the experts are satisfied with their exploration of the data. To denote the iterations, we may add super scripts to the feedback and visualizations so that f and v may take values

$$\{v^{(1)}, v^{(2)}, \dots\}, \qquad \{f^{(1)}, f^{(2)}, \dots, \},$$

and, at iteration i, we have the posterior $\pi(\boldsymbol{\theta}|f^{(1)}, \dots, f^{(i)}, v^{(1)}, \dots, v^{(i)}, \boldsymbol{d})$. Convergence of the BaVA process is application and expert specific and marks, possibly, the point of cognition or the acquisition of knowledge based on the complete merger of judgement and information in the data. In Sections 5 - 7, we demonstrate how to create a malleable visualization using the BaVA process in three applications.

5 Simulated Example Continued

In our simulated example, experts need only find the direction in which to rotate the data and project so that the cluster structure is visible in two dimensions. Since the simulated data span only $p = 3$ dimensions, we could use any available software (e.g., `ggobi` as described in Swayne et al. (2003); Cook and Swayne (2007)) to view and rotate the data until a useful projection is found. For high dimensional data ($p \geq 4$) however, viewing and rotating the data is not possible.

For $p \geq 4$, we could employ one of several projections methods available including Grand

Tour (Asimov, 1985), Projection Pursuit (Friedman and Tukey, 1974), and VizRank (Leban et al., 2006). All three methods have the potential to reveal the best direction in which to rotate the simulated data, but due to high computational demands, they may not scale well to assess high-dimensional datasets. For example, Grand Tour provides a sequence of two-dimensional projections that circumvent the entire data space so that an expert may explore the data visually from multiple directions. Projection Pursuit and VizRank return only one projection, but consider every possible data projection first. Specifically, they associate a measure of "interesting-ness" to *every* two-dimensional projection on the data dimension axes and select the projection with the best measure. The benefit of BaVA is that the choice of projection to display is guided by experts and computational power is minimized in comparison to standard projection methods.

In our BaVAtized version of Figure 1, we allow the expert to drag data points together or apart using `ggobi` and we develop a method that transforms the adjustments into information regarding the direction and magnitude of a data rotation. In doing so, we re-weight the raw marginal data variances using Bayesian sequential updating and select the appropriate direction to project using standard PPCA machinery. The precise BaVA procedure is stated below.

5.1 Form Bayesian Inferences and Construct Visualization

We start by modeling the data using PPCA (Tipping and Bishop, 1999). PPCA is similar to PCA in that the both approaches estimate low dimensional projections of high dimensional datasets, but PPCA relies on probability models rather than deterministic data transformations. Consider the following probability model for data d_i conditional on a reduced or q dimensional vector r_i,

$$d_i = \boldsymbol{W} r_i + \mu + \epsilon_i, \qquad \epsilon_i \sim \text{No}(\mathbf{0}, \mathbf{I}_p \sigma^2) \tag{2}$$

where μ represents a p-vector and the mean of $\boldsymbol{d}$; r_i is a q-vector; $\boldsymbol{W}$ is a $p \times q$ transformation matrix; $\mathbf{I}_p$ is a $p \times p$ identity matrix; and ϵ_i represents an error term that has a Multivariate Normal distribution with mean $\mathbf{0}$ and isotropic variance $\mathbf{I}_p\sigma^2$. Parameters r_i and $\boldsymbol{W}$ represent the *latent factors* of d_i and *factor loadings* of $\boldsymbol{d}$ respectively. In Appendix A, we show that the factors $\boldsymbol{r} = \{r_1, ..., r_n\}$ are scaled, noise-free versions of the top q (for the simulated example, $q = 2$) principal components (Tipping and Bishop, 1999).

If we assign a Normal prior distribution with mean zero and variance $\mathbf{I}_q$ for r_i ($\pi(r_i) = \text{No}(\mathbf{0}, \mathbf{I}_q)$ for all i), the conditional posterior distribution is

$$\pi(r_i|\boldsymbol{d}, \boldsymbol{W}, \sigma^2) = \text{No}(\eta, \boldsymbol{\Sigma_r}), \tag{3}$$

where $\eta = (\boldsymbol{W}'\boldsymbol{W} + \sigma^2\mathbf{I}_q)^{-1}\boldsymbol{W}'(d_i - \mu)$ and $\boldsymbol{\Sigma_r} = (\sigma^{-2}\boldsymbol{W}'\boldsymbol{W} + \sigma^2\mathbf{I}_q)^{-1}$. Based on model (3), $\boldsymbol{\eta} = \{\eta_1, ..., \eta_n\}$ is a natural, low dimensional parameter to plot for visualization v because each η_i equals a centered version of data d_i that is projected to a q dimensional space by $P_{\boldsymbol{r}} = (\boldsymbol{W}'\boldsymbol{W} + \sigma^2\mathbf{I}_q)^{-1}\boldsymbol{W}'$,

$$\eta_i = P_{\boldsymbol{r}}(d_i - \mu),$$

conditional on $\boldsymbol{W}$, σ^2, and μ. To select initial values for these parameters, we specify reference priors for each and set $\boldsymbol{W}$, σ^2, and μ to their respective MAP estimators, $\hat{\boldsymbol{W}}$, $\hat{\sigma^2}$, and $\hat{\mu}$ (which incidentally are equivalent to maximum likelihood estimators (Tipping and Bishop, 1999)). We describe the full hierarchical model in Appendix A.

Comparable to PCA, the projection $P_{\boldsymbol{r}}$ depends heavily on the dimensions in $\boldsymbol{d}$ with the highest variance. To see this, consider $\mu = 0$ (without loss of generality) and the marginal distribution of $\boldsymbol{d}$,

$$\pi(d_i|\boldsymbol{W}, \sigma^2, \mu) \;\; = \;\; \int \text{No}(d_i|\boldsymbol{W}r_i + \mu, \mathbf{I}_p\sigma^2)\text{No}(r_i|\mathbf{0}, \mathbf{I}_q)d\boldsymbol{r} = \text{No}(\mu, \boldsymbol{\Sigma_d}),$$

where, $\boldsymbol{\Sigma_d} = \boldsymbol{W}\boldsymbol{W}^T + \mathbf{I}_p\sigma^2$. Notice that both $\boldsymbol{\Sigma_d}$ and $P_{\boldsymbol{r}}$ rely on parameters $\boldsymbol{W}$ and σ^2, and $\boldsymbol{W}$ contains the top eigenvectors of $\boldsymbol{\Sigma_d} - \mathbf{I}_p\sigma^2$. The correspondence between $\boldsymbol{\Sigma_d}$ and $P_{\boldsymbol{r}}$ suggests that the primary data unknown, or characteristic of $\boldsymbol{d}$ that influences the structure we see in v, is the variance. Changes in $\boldsymbol{\Sigma_d}$ will impact both $P_{\boldsymbol{r}}$ and v and visa versa. Thus, $v = g(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}})$, where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$ represents an estimate for $\boldsymbol{\Sigma_d}$ and $g(\cdot)$ represents the means by which we use $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$ to solve for η or another highly probable value for $\boldsymbol{r}$ that we can visualize. Let manipulations to v reflect feedback concerning $\boldsymbol{\Sigma_d}$.

In the next section, we describe how experts inject feedback and the method we use to parameterize the feedback, but we must first assess $\boldsymbol{\Sigma_d}$ probabilistically. If we specify $\pi(\boldsymbol{\Sigma_d}) \propto 1$ *a priori*, $\pi(\boldsymbol{\Sigma_d}|\boldsymbol{d})$ is an Inverse Wishart (IW) distribution,

$$\pi(\boldsymbol{\Sigma_d}|\boldsymbol{d}) \;\; = \;\; \text{IW}(n\boldsymbol{S_d}, p, n-p-1) \tag{4}$$

where $\text{IW}(a,b,c) = |a|^{-c/2}|\boldsymbol{\Sigma_d}|^{-(c+b+1)/2}\exp\{-\text{tr}(a\boldsymbol{\Sigma_d}^{-1})/2\}/2^{cb/2}\Gamma_b(c/2)$ ($\Gamma_b(\cdot)$ is a multivariate gamma function), $\boldsymbol{S_d} = \frac{1}{n}\sum_{i=1}^{n}(d_i - \mu)(d_i - \mu)^T$, and the MAP of $\boldsymbol{\Sigma_d}$ is $\boldsymbol{S_d}$. For the visualization, we set $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \boldsymbol{S_d}$. Figure 3a) provides an example.

5.2 Cognitive Feedback

Experts may choose to manipulate Figure 3a) if observations they expected to be different are close in proximity or observations they expected to by similar are separate. Thus, for this BaVA example, we allow experts to select two observations and either drag them apart or together as shown in Figure 3b). For example, suppose experts selected points j and k and moved them apart as depicted in Figure 3c). Let $\tilde{r_j}$ and $\tilde{r_k}$ represent the new, low dimensional locations of points j and k. The new locations suggest that, despite the current display, the experts are informed well enough to make the judgement that these points belong in separate clusters. If the experts believe strongly in the separation, they should specify κ close to one; close to 0.5 if they are apprehensive; and close to zero otherwise. Note that

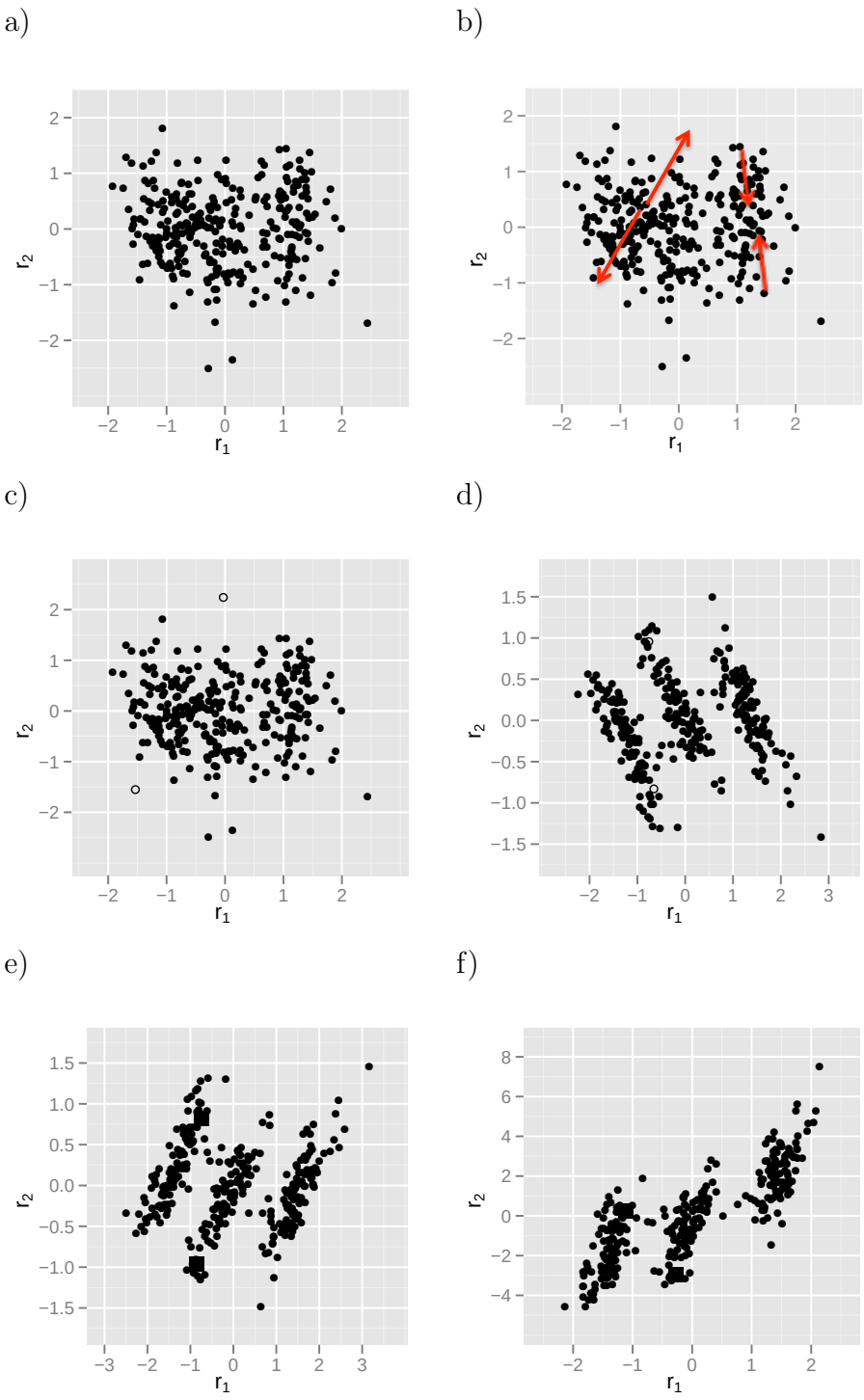

Figure 3: Figure a) plots the posterior mean of $\boldsymbol{r}$Figure b) shows different adjustments an expert can make to the visualization including moving points apart or together. Figure c) displays the adjustment we made; we moved two points apart and mark them by $\circ$ figure to exemplify BaVA and Figures d)-f) present updated versions of the display for κ equal to 0.5, 0.7, and 0.9 respectively.

specifying $\kappa = 0$ is comparable to opting not to manipulate the display.

In this paper, we restrict the experts to moving only two points. Future BaVA research will include more elaborate schemes for injecting cognitive feedback.

5.3 Parametric Feedback

We learned in Section 5.1 that the model parameter driving the structure (or lack there of) we see in visualization v is the variance matrix $\mathbf{\Sigma_d}$ and that Figure 3a) was created conditional on $\mathbf{\Sigma_d} = S_{\boldsymbol{d}}$. Thus, we must interpret adjustments to v as proposed estimates for the variance of $\boldsymbol{d}$ where $f^{(p)} = h(f^{(c)})$ is a semi-definite, $p \times p$ matrix.

Our interpretation of $f^{(c)}$ depends upon the type of manipulation chosen by the expert. For example, if experts move points j and k apart, we should conclude that the variance in the dimensions that are least represented in the display should be larger than what is estimated by $S_{\boldsymbol{d}}$. Similarly, if experts move two points together, the variance in the dimensions represented largely in v is over-stated. Let S_a and S_t represent respectively, the interpreted variance matrices for the apart and together move-types. We assign $f^{(p)}$ to be a weighted average S_a and S_t

$$f^{(p)} = \omega S_a + (1 - \omega) S_t,$$

where $\omega \in [0, 1]$. For the remainder of this section, we explain how to specify ω, S_a, S_t and the distribution for $f^{(p)} | f^{(c)}, \mathbf{\Sigma_d}$.

5.3.1 Specifying ω

The weight ω is determined directly from the ratio of the inter-point distances before and after the display adjustment $\tilde{f}$,

$$\tilde{f} = \frac{||\tilde{r_j} - \tilde{r_k}||_2}{||r_j - r_k||_2}.$$

When points are dragged apart, $\tilde{f} \geq 1$; and, when points are pushed together, $\tilde{f} \leq 1$. Since $\tilde{f} \in R$, we map $\tilde{f}$ to $[0, 1]$ by $\omega = 2\pi^{-1} \arctan(\tilde{f})$.

5.3.2 Specifying S_a

We start by learning which dimensions are least explained in v by comparing the raw and projected marginal discrepancies between points j and k. Let Δ_l, $\Delta_l^{(0)}$, and $\Delta_l^{(P)}$ represent the raw, marginalized (in vector form), and projected discrepancies in dimension l ($l \in [x, y, z]$ for this example) between points j and k,

$$\begin{aligned} \Delta_l &= d_{j,l} - d_{k,l} \\ \Delta_l^{(0)} &= (0, \ldots, 0, \overbrace{\Delta_l}^{l^{th}\text{position}}, 0 \ldots, 0)^T \\ \Delta_l^{(P)} &= P_{\boldsymbol{r}} \Delta_l^{(0)} \end{aligned}$$

where $\Delta_l^{(0)}$ is the result of multiplying the scalar Δ_l by the l^{th} unit vector. The ratio of the vector lengths of $\Delta_l^{(P)}$ and $\Delta_l^{(0)}$ measures the percent of the raw, high dimensional discrepancy for which the visualization accounts. In turn, $\Delta_l^{(u)}$ where

$$\Delta_l^{(u)} = \Delta_l (1 - \frac{||\Delta_l^{(P)}||_2}{||\Delta_l^{(0)}||_2})$$

represents the amount of the original discrepancy in dimension l that remains unexplained by the visualization. For example, if $\Delta_l^{(u)}$ is close to zero, the visualization captures the interpoint discrepancy in dimension l; and, if $\Delta_l^{(u)}$ is comparable to Δ_l the visualization fails to display the interpoint discrepancy in dimension l.

Given $\Delta = \{\Delta_1, ..., \Delta_p\}$ and $\Delta^{(u)} = \{\Delta_1^{(u)}, ..., \Delta_p^{(u)}\}$, we select one of the directions needed to project data $\boldsymbol{d}$ into q dimensions. We denote this direction by $v^{(u)}$ and define it as

$$v^{(u)} = \frac{\Delta + \Delta^{(u)}}{||\Delta + \Delta^{(u)}||_2}$$

This definition adds weight (as much as two times) to the dimensions that are least explained

in the visualization.

To select the remaining $q-1$ directions in which to project data $\boldsymbol{d}$, we proceed in the spirit of PCA. Note for $q=2$, we need only resolve one direction. We suggest calculating the $q-1$ directions that are both orthogonal to $v^{(u)}$ and account for the maximum variation in $\boldsymbol{d}$. Within the context of the simulated example, we must find a new orthogonal direction $v^{(o)}$ that satisfies

$$v^{(o)} = \underset{v^{(o)}}{\operatorname{argmax}}\{\operatorname{Var}[v^{(o)\prime}\boldsymbol{d}]\} \quad \text{s.t. } v^{(o)\prime}v^{(o)}=1 \text{ and } v^{(o)\prime}v^{(u)}=0. \tag{5}$$

The solution for $v^{(o)}$ is the largest principal direction of $\boldsymbol{\Sigma_d}$ that is orthogonal to $v^{(u)}$. The proof is found in Appendix B. Based on $v^{(u)}$ and $v^{(o)}$, we define S_a as

$$S_a = [v^{(o)}, v^{(u)}][v^{(o)}, v^{(u)}]'.$$

5.3.3 Specifying S_t

When experts move points j and k together, the implication is that they have similarities in the high dimensional space that are being lost in the projected space. To guarantee an ideal projection, we start again with the raw, high dimensional discrepancies Δ. Since the vector Δ runs directly through the points d_j and d_k, we can map d_j and d_k to the same set of coordinates in a lower dimensional space if we project the data in the direction of Δ. Thus, Δ is an orthogonal vector to the ideal projection plane that is embedded in the high dimensional space. To identify the projection plane, we solve for two vectors that are both orthogonal to one another and to Δ. The solution for $\{v^{(1)}, v^{(2)}\}$ in

$$0 = \Delta' v^{(1)} = \Delta' v^{(2)} = v^{(1)\prime} v^{(2)}$$

defines an orthogonal basis for the projection plane (proof in Appendix C). Given $\{v^{(1)}, v^{(2)}\}$, we set S_t to

$$S_t = [v^{(1)}, v^{(2)}][v^{(1)}, v^{(2)}]',$$

Note that both S_t and S_a have, by definition, only two eigenvectors with non zero eigenvalues, and these eigenvectors are the defined basis sets.

5.3.4 Specify $\pi(f^{(p)}|f^{(c)}, \mathbf{\Sigma_d})$

Since Wishart (Wi) distributions are defined over the space of semi-definite matrices, we model $f^{(p)}$ by

$$f^{(p)}|f^{(c)}, \mathbf{\Sigma_d} \sim \text{Wi}(\frac{\mathbf{\Sigma_d}}{\nu}, p, \nu), \tag{6}$$

where, $\text{Wi}(a,b,c)=|f^{(p)}|^{(c-b-1)/2} \exp\{-\text{tr}(f^{(p)}a^{-1})/2\}/(2^{bc/2}\Gamma_b(c/2))$, the conditional expectation of $f^{(p)}$ is $\mathbf{\Sigma_d}$, and $\nu = \kappa n/(1-\kappa)$. This model choice is both practical for this application and, as explained in the following section, computationally convenient.

5.4 Update the Model Based on Feedback

Since we chose to parametrize the feedback using a Wishart Distribution, the sequential updating step for the distribution of $\mathbf{\Sigma_d}$ is straightforward,

$$\pi(\mathbf{\Sigma_d}|\boldsymbol{d}, f) \quad = \quad \text{IW}(nS_{\boldsymbol{d}} + \nu f^{(p)}, p, n + \nu - p - 1).$$

The updated MAP estimator for $\mathbf{\Sigma_d}$ is a weighted average of the parametric feedback $f^{(p)}$ and the original MAP estimator $S_{\boldsymbol{d}}$,

$$\text{E}[\mathbf{\Sigma_d}|\boldsymbol{d}, f] = \frac{\nu}{\nu+n} f^{(p)} + \frac{n}{\nu+n} S_{\boldsymbol{d}}.$$

Notice that $\kappa = \frac{\nu}{\nu+n}$ because of our definition for ν in Equation (6).

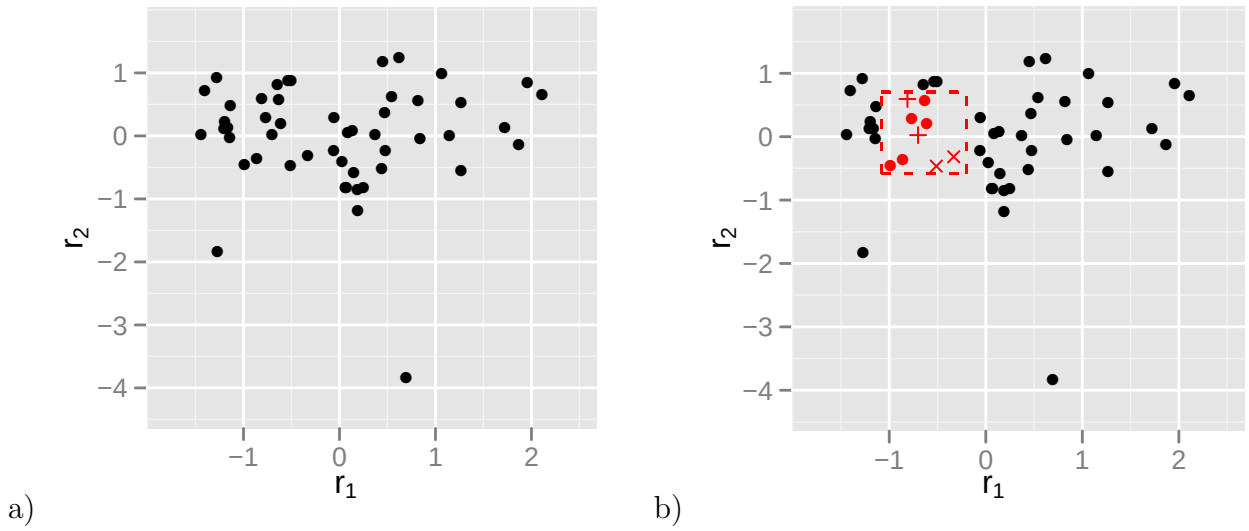

Figure 4: In Figure a) the first two latent dimensions from PPCA are plotted for the education dataset with seven variables: SAT, EXP, FAC, SAL, PER, HSG, and INC. Figure b) is identical to a) except points of interest have been selected by the square *brush*. The two points that appear to be the most similar high dimensionally in the brush are marked by '+'; and the two points that appear to be the most different are marked by '×'. The measures of similarity and difference are stated in Section 6.1.

6 Two Real-World Examples

In this section, we apply the BaVA methods described in Section 5 to assess two real world datasets. A vital point in Section 5 is that the success of BaVA relies critically on sound expert judgement concerning the relationship between two observations. We are confident that for some applications experts will have immediate, reliable judgments, but there will be applications when experts need assistance. For the latter cases, we present two approaches to assess the relationship between two or more points. We explain each approach within the context of two proceeding examples.

6.1 Cost and Quality of Education

A sensitive issue for tax payers, parents, children, educators, and policy makers is whether an increase in money devoted to education will increase education quality. Money provides

a means to buy modern textbooks, employ experienced teachers, and provide a variety of classes and/or extra curricular activities. Although, do the students who benefit from these high-priced resources actually improve academically?

In 1999, Dr. Deborah Guber compiled a dataset for pedagogical purposes that addresses this question (Guber, 1999). Based on the following variables, the dataset summarizes the academic success, the educational expenses, and other related variables in 1997 for each U.S. state: the average exam score on the Standard Aptitude Test (SAT); the average expenditure per pupil (EXP); the average number of faculty per pupil (FAC); the average salary for teachers (SAL); and the percentage of students taking the SAT (PER). To increase the complexity of the dataset slightly, we added two variables from the National Center for Education Statistics (`http:nces.ed.gov`): the number of high school graduates (HSG) and the average household income (INC). To assess these data, we investigate the possibility of observation clusters which might be explained by variables in the dataset. We start by visualizing the data in two dimensions using PPCA. The initial PPCA projection in Figure 4a) however, does not reveal the presence of any data structure. Thus, we apply PPCA using the BaVA framework to navigate the dataset.

To navigate the dataset, we need to make judgements concerning the relationship between two or more observations, yet we are not education experts. Thus, we apply a tool to which we refer as the "brush" that identifies observations within a small region which way be good candidates to adjust. Specifically, the brush selects two pairs of observations based on a measure m that seem to be the most similar and different in the high dimensional dataset relative to the low dimensional display. Let vectors $\boldsymbol{\delta} = \{\delta_{ij}\}_{i<j<b}$ and $\boldsymbol{\gamma} = \{\gamma_{ij}\}_{i<j<b}$ represent respectively the distance between every pair of observations in the high and low dimensional spaces, where b represents the number of brushed observations. To compare the vectors, we divide each vector by its maximum so that every element in $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ is between zero and one. We then calculate $m_{ij} = \delta_{ij}/\sqrt{\gamma_{ij}}$ for all (i, j) pairs. If the measure m_{ij} is

small, the low dimensional display exaggerates the true distance between observations i and j in the high dimensional space. Similarly, if the measure m_{ij} is large, the distance between points i and j in the high dimensional space is greater than what it seems in the display. The pairs of observations with the minimum and maximum measure m are identified by the brush. Note, we explored defining m_{ij} as either $m_{ij} = d_{ij}/r_{ij}$ or $m_{ij} = d_{ij} - r_{ij}$, but such measures were impractical because they were either too sensitive or not sensitive enough to extreme low dimensional distances (e.g., $r_{ij} \approx 0$).

For this application, we placed the brush in an area that seemed to be between two possible clusters in Figure 4b). The observations with minimum m are marked by '+' and observations with maximum m are marked by '$\times$'. We now have two options: 1) we could drag the observations with low m closer or 2) we could spread the points with large m apart. We opted to do the latter and obtained a BaVA-updated view that is displayed in Figure 5a). There are two clusters in Figure 5a). These clusters correspond perfectly with SAT scores above and below the SAT median.

Those that advocate increasing education budgets might suspect that the clustering structure in SAT relates to EXP. However, when we re-plot Figure 5a) and label the upper and lower EXP 50% quantiles in Figure 5b), EXP does not explain the clusters. We repeated this re-labeling exercise for every variable in the dataset. When we mark the observations above and below the empirical PER median in Figure 5c), we see that PER and SAT clearly relate to the formation of clusters in the dataset. Thus, further analyses of SAT and EXP, must control for PER.

6.2 Functional Genomics

For this section, we consider a microarray dataset (Eisen et al., 1998; Brown et al., 2000; Leban et al., 2006) that was collected to assess the function of 186 yeast genes in the *Saccharomyces cerevisiae* genome based on comparisons in expression for 79 hybridization experiments. Eisen et al. (1998) suggested that genes with similar expression profiles under varying

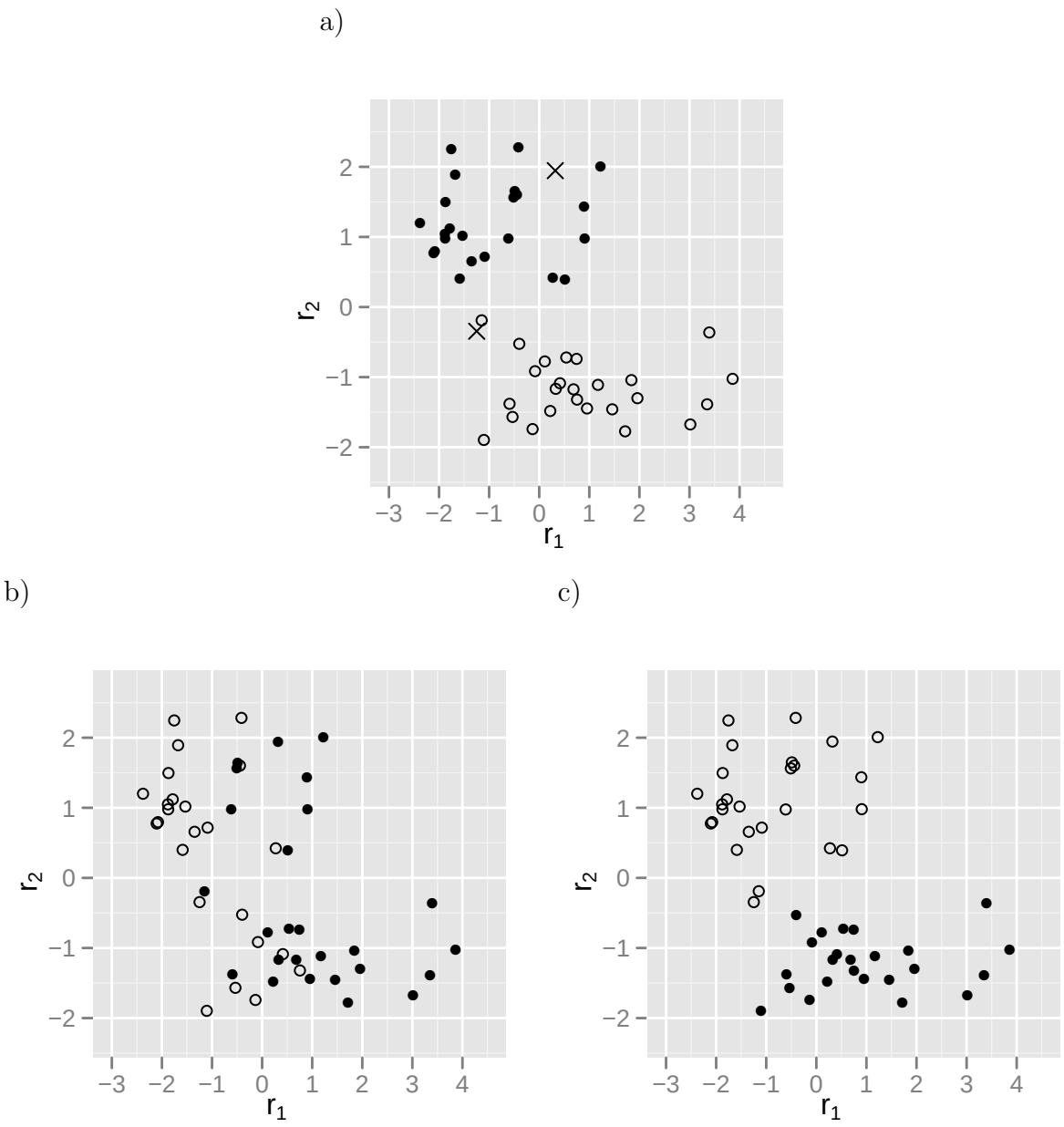

Figure 5: After injecting expert feedback into Figure 4, we obtain Figures a)-c). For frame of reference, we marked the two points moved to inject feedback by '×' in Figure a). The configuration of points in each graph are identical, but the observations are labeled differently. In Figure a), symbols '•' and '○' mark the upper and lower 50% quantiles for SAT respectively; in Figure b), symbols '•' and '○' mark the upper and lower 50% quantiles for EXP respectively; and in Figure c), symbols '•' and '○' mark the upper and lower 50% quantiles for PER respectively. Notice the clusters in each graph correspond with SAT and PER, but not EXP.

conditions have the potential to serve the same function. Thus, we may make reasonable semi-supervised (Chapelle et al., 2006; MacInnes et al., 2010) predictions for the functions of every gene in the experiment based on a) the known functions for a subset of genes in the experiment and b) the identification of gene clusters based on similar experimental, expression profiles.

Semi-supervised DM approaches aim to develop classification rules based on the *ground-truth* for a subset of the data. Using the ground-truth for only a data subset is clearly less work than either relying on the ground-truth for all observations or taking a fully unsupervised data-mining approach (MacInnes et al., 2010). Additionally, knowing the ground-truth for a subset of the data may assist experts to adjust BaVA visualizations appropriately. Within the context of genomics, semi-supervised learning methods are often reasonable because public databases are readily available that store detailed information concerning known genes. For this example, the Munich Information Center of Protein Sequences Yeast Genome Database (MYGD) lists, among other characteristics, the functions for every gene in the *Saccharomyces cerevisiae* genome.

According to the MYGD, the 186 genes of interest belong to one of three functional classes to which are referred as cytoplasmic respiration (f1), ribosomes (f2), and proteasome (f3). Yet, when we project the data using PPCA in Figure 6a), only two possible clusters appear. To adjust the display, we take a semi-supervised approach and use the MYGD to look-up the functions for 25 genes selected at random from the dataset. The ground-truth for these genes are shown in Figure 6b) where '•' denotes f1, '+' denotes f2, and '■' denotes f3. The genes with function f2 cluster nicely, but the genes with functions f1 and f3 do not separate well in the visualization. Thus, we select and separate two genes with functions f1 and f3 that are highlighted in red in Figure 6b). Given the separation, we update the visualization using BaVA machinery which we display in Figure 7.

In Figure 7a) we see that with the exception of two observations (one of which was pre-

selected and we know its function), three clusters separate clearly. We predict the functions for each cluster using the ground-truths of the 25 preselected genes and label them in Figure 7a). Since the MYGD includes the functions for all 186 genes, we can compare our predictions to the true genetic functions depicted in Figure 7b). Excluding the point which we do not label, only one gene is predicted erroneously; our true prediction rate is 0.994.

7 Multidimensional Scaling Example

Thus far, BaVA has been exemplified using projection based methods to reduce data dimensionality. In this section, we present the BaVA process for another dimension reduction technique known as multidimensional scaling (MDS) (Torgerson, 1958; Kruskal and Wish, 1978). The purpose of this section is two fold: 1) to present a different use of BaVA and 2) to exemplify that BaVA is a framework and not a method- or application- specific analytical approach. With thought, we believe that BaVA can be used for a variety of data structure-seeking techniques in the future.

This section has three subsections. The first subsection will explain MDS and an extension of MDS known as Weighted MDS (WMDS) (Carroll and Chang, 1970; Schiffman et al., 1981) briefly. The second subsection will develop the BaVA steps in Section 4 within the context of MDS. The third subsection will exemplify the MDS version of BaVA to assess a real-world dataset.

7.1 MDS and WMDS

Using the same notation from previous sections, let $\boldsymbol{d} = \{d_1, ...d_n\}$ where $\boldsymbol{d} \in R^p$ and $\boldsymbol{r}$ represent a low dimensional analog of $\boldsymbol{d}$ where $\boldsymbol{r} = \{r_1, ...r_n\}$, $\boldsymbol{r} \in R^q$, and $q \leq 3 < p$. In a typical MDS scheme, one seeks to select points $\boldsymbol{r}$ with pairwise distances that approximate the same pairwise distances of $\boldsymbol{d}$. Explicitly, $\boldsymbol{r}$ is the solution to the following optimization

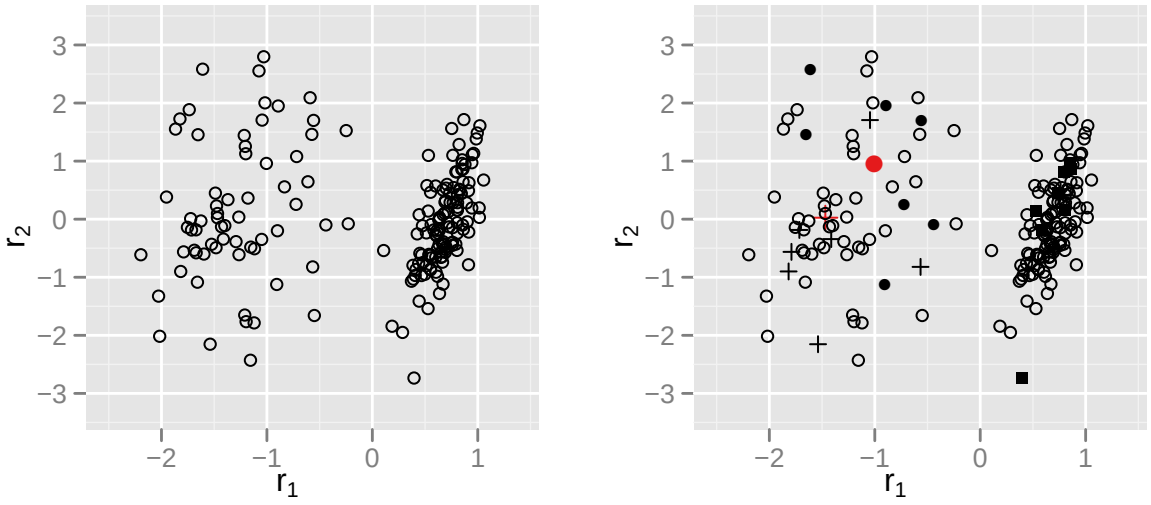

Figure 6: Both figures display the top two principal components from the *Saccharomyces cerevisiae* dataset which contains the expression profiles for 186 genes across 79 hybridization experiments. In Figure b) we include the true function for 25 genes: '•' denotes f1, '+' denotes f2, and '■' denotes f3. Based on these truths, we opt to adjust two observations highlighted in red.

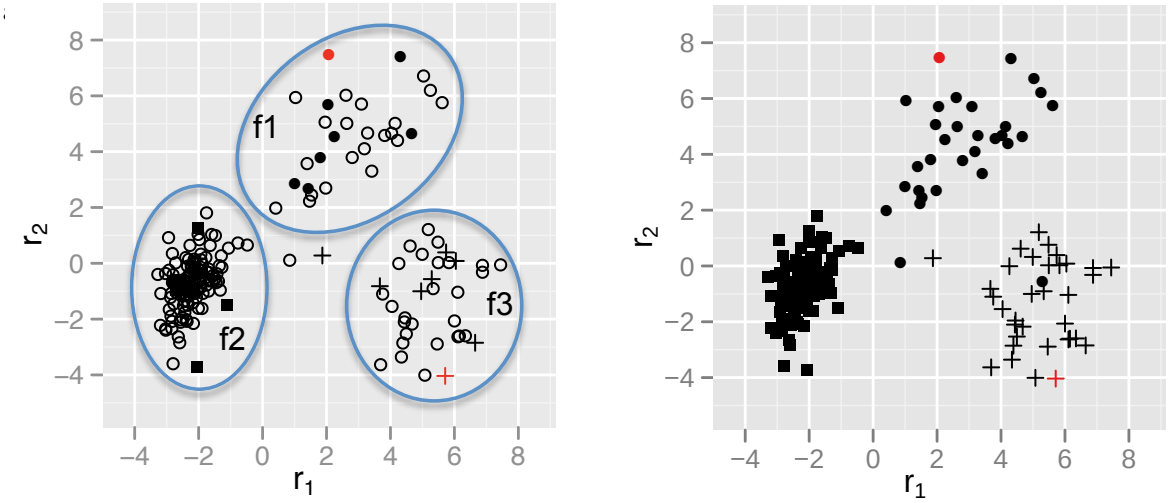

Figure 7: After injecting expert feedback into Figure 6b), we received the above displays. Figure a) includes the ground-truths for the 25 pre-selected genes: f1='•', f2='+', and f3='■'. Figure a) also highlights in red the observations adjusted to inject feedback. The large ovals in Figure a) represent the natural clusters that appear. We labeled each oval based on the semi-supervised prediction procedure. Figure b) labels each observation by their functions according to the MYGD. Notice that only one observation is predicted incorrectly in Figure a).

problem:

$$\boldsymbol{r} = \underset{r_1,\ldots,r_n}{\operatorname{argmin}} \sum_{i<j} |\|\boldsymbol{r}_i - \boldsymbol{r}_j\| - \delta_{ij}|,$$

where δ_{ij} represents $\|\boldsymbol{d}_i - \boldsymbol{d}_j\|$ and $\|\cdot\|$ represents a vector norm. The solution is invariant to rotations and reflections and the scale of $\boldsymbol{r}$ is mostly arbitrary. One advantage for MDS is that the pairwise distances in $\boldsymbol{r}$ help us to understand the pairwise relationships between observations in the original high dimensional vector space. A pair of observations that are distant are less related to one another than a pair of observations that are close to one another.

The distance measure or vector norm $\|\cdot\|$ used in MDS may influence the solution for $\boldsymbol{r}$ if the original distances δ_{ij} are sensitive to the vector norm selection as well. For our purposes, we use the L_2 norm where

$$\delta_{ij} = \sqrt{\sum_{k=1}^{p} (d_{i,k} - d_{j,k})^2}. \tag{7}$$

This choice is arbitrary and can be exchanged easily for other norms. The pairwise distances in $\boldsymbol{r}$ are computed similarly to Equation (7). Let γ_{ij} represent the distance between observations i and j in the low dimensional space.

WMDS (Carroll and Chang, 1970; Schiffman et al., 1981) is similar to MDS, but the dimensions of the data are weighted in order to express the impact of each dimension on δ. For example, if dimension 1 has a large impact on δ for all i and j, dimension 1 will have a large weight or impact on the solution for $\boldsymbol{r}$. Since we use the L_2 norm, the weighted distances are computed as follows:

$$\delta_{ij}^{(\omega)} = \sqrt{\sum_{k=1}^{p} \omega_k (d_{i,k} - d_{j,k})^2}, \tag{8}$$

where $\delta_{ij}^{(\omega)}$ represents the weighted distance between points d_i and d_j; ω_k represents the weight for dimension k; and the weights are constrained by $\sum_k^p \omega_k = 1$. The weighted solution for $\boldsymbol{r}$ results from

$$\boldsymbol{r} = \underset{r_1,...,r_n}{\operatorname{argmin}} \sum_{i<j} |\|r_i - r_j\| - \delta_{ij}^{(\omega)}|, \tag{9}$$

When $\omega_i = 1/p$ for $i \in [1, ..., p]$ the solution for $\boldsymbol{r}$ is identical to that of classical MDS.

When the dimensions that are relevant for resolving important structure in the data $\boldsymbol{d}$ are unclear, including $\omega = \{\omega_1, ...\omega_p\}$ is a powerful adaptation of MDS. The weight vector w also provides an excellent avenue to inject feedback when considering BaVA. Experts may provide information regarding ω directly or use a BaVA version of WMDS and manipulate observations.

7.2 The BaVA Process Steps for WMDS

The BaVA process includes five steps. We implement the five steps for WMDS.

7.2.1 Steps 1: Form Bayesian Inference

Recall that we use the vector ω to weight important dimensions appropriately when solving for γ. However, the degree to which dimensions are declared important is confounded by dimension variance. Thus, we standardize each dimension of data $\boldsymbol{d}$ before we calculate all pairwise distances $\delta_{ij}^{(\omega)}$, where $\omega_i = 1/q$ $(i \in [1, ..., p])$,

We start with a well defined probability model for each high dimensional distance $\delta_{ij}^{(\omega)}$ given γ_{ij}. While a reasonable probability model already exists (Oh and Raftery, 2001), we develop our own to ease the visual updating procedure. Our model simply adds truncated Gaussian noise with variance σ^2 to the discrepancy of the high and low dimensional distances,

$$\pi(\delta_{ij}^{(\omega)}|\gamma_{ij}, \sigma^2) = \mathbb{1}_{[\delta_{ij}^{(\omega)}>0]}\mathrm{No}(\gamma_{ij}, \sigma^2), \tag{10}$$

where $\mathbb{1}_{[\cdot]}$ has the same meaning as the Dirac Delta function. For this effort, we consider $\pi(\sigma^2) \propto 1/\sigma^2$ and plug-in the MAP estimator $\widehat{\sigma^2}$. The joint probability model across all pairwise distances is

$$\pi(\boldsymbol{\delta}^{(\omega)}|\boldsymbol{\gamma}, \widehat{\sigma^2}) \propto \mathbb{1}_{[\boldsymbol{\delta}^{(\omega)} > \mathbf{0}]} \exp\{-\frac{\widehat{\sigma^2}}{2} \sum_{1 \leq i < j \leq n} (\delta_{ij}^{(\omega)} - \gamma_{ij})^2\}, \tag{11}$$

where $\boldsymbol{\delta}^{(\omega)} = \{\delta_{ij}^{(\omega)}\}_{1 \leq i < j \leq n}$ and $\boldsymbol{\gamma} = \{\gamma_{ij}\}_{1 \leq i < j \leq n}$.

Given the prior $\pi(\gamma_{ij}) \propto \mathbb{1}_{[\gamma_{ij} > 0]}$, the posterior distribution for γ_{ij} is

$$\pi(\gamma_{ij}|\delta_{ij}^{(\omega)}, \widehat{\sigma^2}) = \mathbb{1}_{[\gamma_{ij} > 0]} \mathrm{No}(\delta_{ij}^{(\omega)}, \widehat{\sigma^2}),$$

so that the joint distribution across all pairs has the same form as Equation (11) and the MAP equals $\boldsymbol{\delta}^{(\omega)}$.

7.2.2 Steps 2: Construct Visualization

We do not model $\boldsymbol{r}$ explicitly for visualization. Rather, we assess a posterior estimate of the visual distances, $\widehat{\boldsymbol{\gamma}}$ and solve for $\boldsymbol{r}$ using an expression similar to Equation (9),

$$\boldsymbol{r} = \underset{r_1, \ldots, r_n}{\operatorname{argmin}} \sum_{i<j} |\|r_i - r_j\| - \widehat{\gamma_{ij}}|. \tag{12}$$

A natural choice for $\widehat{\boldsymbol{\gamma}}$ is the map of $\pi(\boldsymbol{\gamma}|\boldsymbol{\delta}^{(\omega)}, \widehat{\sigma^2})$. In reference to Section 4.2, the visualization v displays $\boldsymbol{r}$ which is a function of $\widehat{\boldsymbol{\gamma}}$; $v = g(\widehat{\boldsymbol{\gamma}})$.

7.2.3 Steps 3: Cognitive Feedback

Observations that are far apart in the high dimensional space should appear far in the low dimensional display v, and similarly, observations that are close high dimensionally should appear close in v. If the display contradicts the judgements of experts for at least $l = 3$ observations, experts may adjust the l observations accordingly and provide a measure κ

($\kappa \in [0,1]$) as defined in Section 4.3.1.

Unlike the projection based BaVA in Section 5, experts may manipulate more than two points. In fact, $l \geq 3$ is needed to make noticeable changes in the updated display. Let $\tilde{\boldsymbol{r}}$ and $\tilde{\boldsymbol{\gamma}}$ define $f^{(c)}$ and represent the set of l manipulated observations and the $\binom{l}{2}$ pairwise distances respectively.

7.2.4 Step 4: Parametric Feedback

As mentioned previously, display manipulations reflect the distribution of weights across the dimensions. Thus, we use $f^{(c)}$ to solve for new weights $\tilde{\omega}$ and the parametric feedback, $f^{(p)}$. We use a constrained, gradient search (Mordecai, 1976) to solve for $\tilde{\omega}$ in

$$\tilde{\omega} = \underset{\tilde{\omega}}{\operatorname{argmin}} \sum_{i<j\leq l}^{p} \left| \tilde{\gamma_{ij}} - \sqrt{\sum_{k} \tilde{\omega_k}(d_{i,k} - d_{j,d})^2} \right|,$$

such that $\sum_k \tilde{\omega_k} = 1$. Given $\tilde{\omega}$, we re-calculate all high dimensional, weighted pairwise distances to derive the parametric feedback; i.e., $f^{(p)} = \boldsymbol{\delta}^{(\tilde{\omega})}$. By Equation (10) $f^{(p)}$ has the following distribution

$$\pi(f^{(p)}|\boldsymbol{\gamma}, \alpha) = \mathbb{1}_{[f^{(p)}>0]}\mathrm{No}(\boldsymbol{\gamma}, \alpha),$$

where $\alpha = \widehat{\sigma^2}(1-\kappa)/\kappa$. We will justify the expression for α in the next section.

7.3 Step 5: Update the Model Based on Feedback

Recall from Section 7.2.2, we visualize $\boldsymbol{r}$ which we derive from the MAP of $\pi(\boldsymbol{\gamma}|\boldsymbol{\delta}^{(\omega)}, \widehat{\sigma^2})$. We apply Bayesian sequential updating to include $f^{(p)}$ in the posterior analysis of $\boldsymbol{\gamma}$,

$$\pi(\boldsymbol{\gamma}|\boldsymbol{\delta}^{(\omega)}, f^{(p)}, \widehat{\sigma^2}) \propto \pi(f^{(p)}|\boldsymbol{\gamma}, \alpha)\pi(\boldsymbol{\gamma}|\boldsymbol{\delta}^{(\omega)}) \propto \mathbb{1}_{[\boldsymbol{\gamma}>\mathbf{0}]}\mathrm{No}(\mu, \Sigma),$$

where

$$\mu = \frac{\widehat{\sigma^2} f^{(p)}}{\widehat{\sigma^2} + \alpha} + \frac{\alpha \boldsymbol{\delta}^{(\omega)}}{\widehat{\sigma^2} + \alpha}, \qquad \Sigma = \mathbf{I}_m \frac{\widehat{\sigma^2} + \alpha}{\widehat{\sigma^2} \alpha}, \tag{13}$$

$m = \binom{n}{2}$, and $\mathbf{I}_m$ represents an identify matrix of dimension m. Notice that the MAP estimator μ is a weighted average of $f^{(p)}$ and the data $\boldsymbol{\delta}^{(\omega)}$. In Section 7.2.4, we solved for α such that $\kappa = \widehat{\sigma^2}(\widehat{\sigma^2} + \alpha)^{-1}$.

Based on Steps 1-5 in this section, we have the foundation to apply BaVA MDS. Thus, we provide an application in the next section.

7.4 MDS Application

To exemplify the benefit of MDS in the BaVA framework, we consider the well known "Iris Data" (Fisher, 1936). This dataset includes three sets of 50 observations for the following iris species: *Iris setosa*, *Iris virginica*, and *Iris versicolor*. Four measurements were taken for each observation including sepal length, sepal width, petal length, and petal width. Since the species are known, this dataset is often used in both the statistical and machine learning literature as a benchmark dataset for supervised learning methods, including discriminant analyses and classification algorithms. Our goal is to take a supervised learning approach and discover visually (if possible) variables which separate the iris species well.

To meet our goal we apply the steps from Section 7.2. First, we standardize the data and calculate all pairwise distances $\boldsymbol{\delta}^{(\omega)}$ based on $\omega = \{0.25, 0.25, 0.25, 0.25\}$. Second, we solve for $\boldsymbol{r}$ which we display in Figure 8a) and calculate all low dimensional, pairwise distances $\boldsymbol{\gamma}$. Despite the presence of three iris species, Figure 8a) contains only two clear clusters. We cannot separate *Iris virginica* from *Iris versicolor* when we consider each dimension to be equally predictive of species. Third, we select six or seven observations (20 observations in total) at random from each species as displayed in Figure 8a) and inject feedback. Specifically, we cluster the selected observations by species in separate areas of the latent space

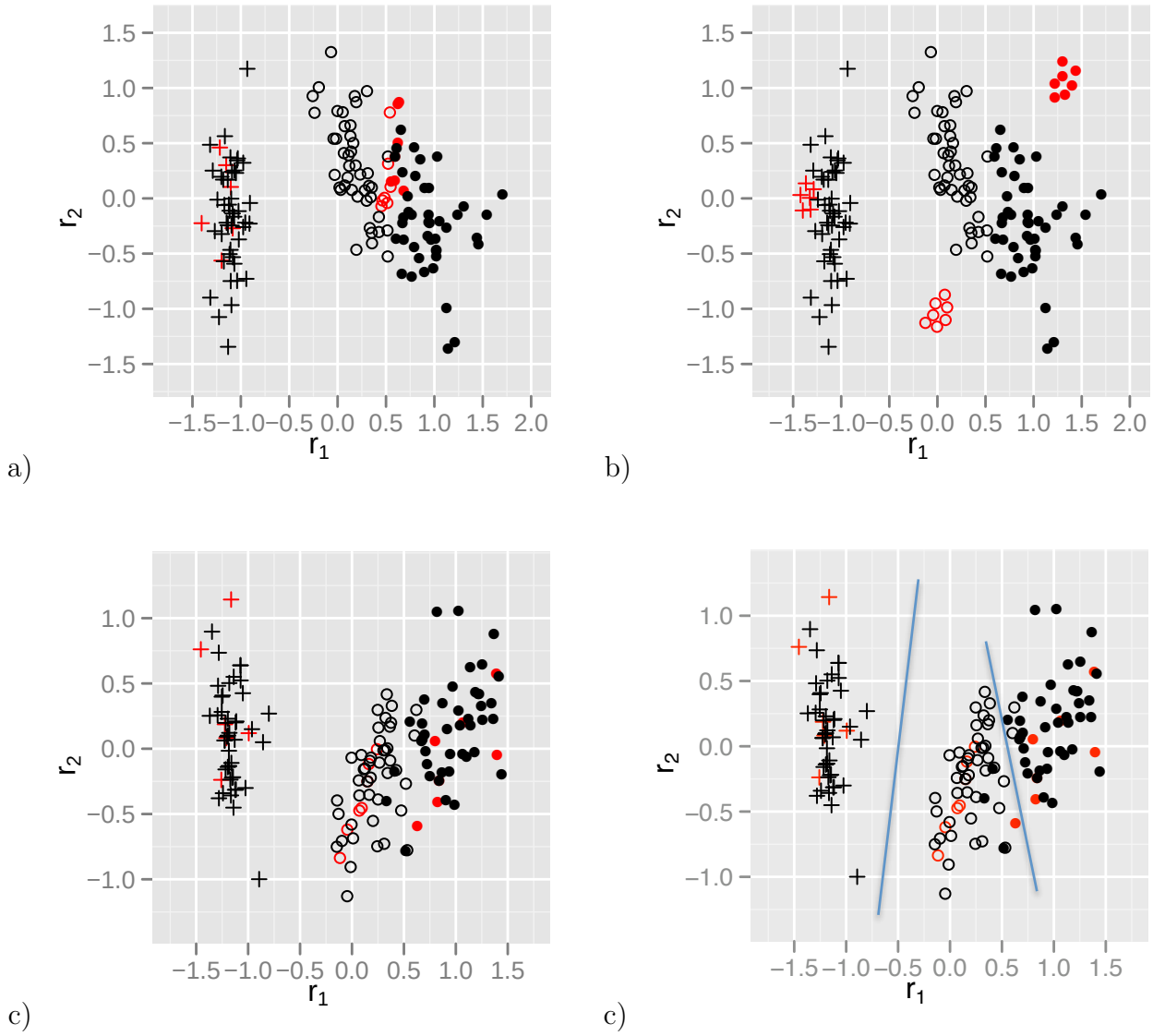

Figure 8: Figure a) displays the initial WMDS projection of the iris dataset where $\omega = \{0.25, 0.25, 0.25, 0.25\}$. The observations for *Iris setosa*, *Iris virginica*, and *Iris versicolor* and denoted by $+$, $\bullet$ and $\circ$ respectively. Figures a) also highlights in red the observations we decided to move to inject feedback. Figure b) displays where we opted to place the observations. Figure c) presents the updated iris display after injecting feedback. Notice three visible clusters in Figure c). We add lines to separate the clusters in Figure d).

(Figure 8b) and set $\kappa = 1$. Since this is a supervised learning strategy, we are very confident in our feedback. Fourth, we solve for $\tilde{\omega}$ and α to specify $\pi(f^{(p)}|\boldsymbol{\gamma}, \alpha)$. Fifth, we update $\pi(\boldsymbol{\gamma}|\boldsymbol{\delta})$ and display the MAP of $\pi(\boldsymbol{\gamma}|f^{(p)}, \boldsymbol{\delta})$ in Figure 8c).

We see in Figures 8c) and 8d) that, with the exception of five observations, *Iris virginica* and*Iris versicolor* separate nicely. This separation is due to the updated weights which solved previously to be $\tilde{\omega} = \{0.194, 0.0643, 0.742, 0.000\}$. According to $\tilde{\omega}$, the differences between the species is explained mainly by petal length (0.742) and sepal length (0.194), and independent of petal width (0.000).

8 Discussion

Johnson (2004) listed the current top ten scientific visualization research problems, two of which included the need for integrated problem solving and the development of tools to enhance human-computer interaction. BaVA is a novel solution for both problems. Unlike typical data displays that simply communicate analytical results, BaVA visualizations serve as a means for experts to synthesize information in the data, interact with the data if desired, and guide automated, analytical procedures.

In effect, experts and machines share the responsibility of knowledge discovery when using BaVA machinery. Since BaVA relies fundamentally on quantitative characterizations of the data (e.g., Bayesian statistical models), it has the potential to reveal both unexpected and expected data structure in visualizations. Experts may learn new information from unexpected data structures (Zhao et al., 2005; MacInnes et al., 2010) and validate the analytical procedure informally based on the identification of expected structure. The presence of expected structure gives experts confidence in the analytical approach. In the event that a visualization is missing expected structure, experts may include feedback in the analytical approach via display adjustments. These adjustments include intuitive manipulations of data points and are not limited to standard interactive procedures such as filtering, zoom-

ing, distorting, and linking/brushing observations as detailed in Keim (2002). Furthermore, experts need not understand the statistical underpinnings of BaVA to make adjustments.

With careful thought, the BaVA framework applies to a variety of statistical models, dimension reducing methods, and data mining techniques. In this paper, we provided three examples for two dimension reduction methods: PPCA and WMDS. To do so, we used probabilistic forms of the reduction methods and developed prototypes mainly in `ggobi`. These prototypes allow experts to adjust two points in the BaVA application of PPCA and three or more points in the BaVA application of WMDS. Other types of cognitive arrangements are possible and constitute future research.

A continuing challenge is the development of BaVA data displays that incorporate visualization uncertainty. We mentioned in Section 4.2 that we condition on a chosen method to visualize posterior results, yet the visualization itself may contain uncertainty. The visualization v may have distributions $\pi(v|\boldsymbol{\theta})$ and $\pi(v|\boldsymbol{d})$ where

$$\pi(v|\boldsymbol{d}) = \int \pi(v|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{d})d\boldsymbol{\theta}.$$

Including visualization uncertainty in displays may influence experts to make adjustments that they would not have otherwise chosen.

A PPCA

Provided $\pi(\boldsymbol{W}) \propto 1$, $\pi(\mu) \propto 1$, and $\pi(\sigma^2) \propto \sigma^{-2}$, the posterior distribution of $\boldsymbol{r}$ is

$$\pi(\boldsymbol{r}|\boldsymbol{d}) = \mathrm{No}\Big((\hat{\boldsymbol{W}}'\hat{\boldsymbol{W}})^{-1}\hat{\boldsymbol{W}}'(\boldsymbol{r} - \hat{\mu}), \hat{\sigma}^2(\hat{\boldsymbol{W}}'\hat{\boldsymbol{W}})^{-1}\Big), \tag{14}$$

where $\hat{\boldsymbol{W}}$, $\hat{\mu}$, and $\hat{\sigma}^2$ represent the maximum likelihood estimate (MLE) for Equation (2). Under the isotropic model constraint ($\mathrm{Var}[\epsilon_i] = \mathbf{I}_p\sigma^2$ for $i \in [1, ..., n]$), we obtain the MLE

for $\boldsymbol{W}$ as

$$\hat{\boldsymbol{W}} = \boldsymbol{V}(\boldsymbol{\Lambda} - \sigma \mathbf{I}_q)^{1/2}\boldsymbol{R},$$

where $\boldsymbol{V}$ are the q largest eigenvectors in $\boldsymbol{S_d}$, $\boldsymbol{\Lambda}$ is a diagonal matrix of the associated eigenvalues ($\lambda_i,\ i \in [1, ..., q]$), and $\boldsymbol{R}$ represents an arbitrary rotation matrix, which for simplicity can equal the identity matrix. This shows a direct relationship between PCA and isotropic factor models, since there is a mapping through the orthogonal principal space that relates the $\boldsymbol{d}$ to the lower dimensional space which contains $\boldsymbol{r}$.

B Solving for $v^{(o)}$

Finding a solution for Equation 5 is equivalent to solving the Lagrange multiplier problem:

$$\{v^{(o)}, \lambda_1, \lambda_2\} = \underset{v^{(o)},\lambda_1,\lambda_2}{\operatorname{argmax}}\ f(v^{(o)}, \lambda_1, \lambda_2),$$

where

$$\begin{aligned} f(v^{(o)}, \lambda_1, \lambda_2) &= \operatorname{Var}[v^{(o)\prime}\boldsymbol{d}] - \lambda_1(v^{(o)\prime}v^{(o)} - 1) - \lambda_2 v^{(o)\prime}v^{(u)} \\ &= v^{(o)\prime}\Sigma v^{(o)} - \lambda_1(v^{(o)\prime}v^{(o)} - 1) - \lambda_2 v^{(o)\prime}v^{(u)} \\ \frac{\partial f}{\partial v^{(o)}} &= \Sigma v^{(o)} - \lambda_1 v^{(o)} - \lambda_2 v^{(u)} = 0. \end{aligned} \tag{15}$$

When we multiply both the center and right sides of Equation (15) by $v^{(u)\prime}$, we obtain $\lambda_2 v^{(u)\prime}v^{(u)} = 0$ (because $v^{(u)\prime}v^{(o)} = 0$ by definition) and deduce that $\lambda_2 = 0$. Substituting $\lambda_2 = 0$ in Equation (15) yields $\Sigma v^{(o)} = \lambda_1 v^{(o)}$. Hence, $v^{(o)}$ is an eigenvector of Σ with the eigenvalue λ_1. To determine the exact eigenvector, we recall that we are maximizing the quantity $v^{(o)\prime}\Sigma v^{(o)} = v^{(o)\prime}v^{(o)}\lambda_1 = \lambda_1$. Thus, $v^{(o)}$ corresponds to the eigenvector of Σ with the largest eigenvalue.

C Solve for Basis $(v^{(1)}, v^{(2)})$

Find a basis $(v^{(1)}, v^{(2)})$ that satisfies $\Delta' v^{(1)} = \Delta' v^{(2)} = v^{(1)'} v^{(2)} = 0$. Let $a = -(\sum_{i=1}^{p-1} \Delta_i)\Delta_p^{-1}$ and define $v^{(1)}$ as $[\mathbf{1}_{p-1}, a]'$ where $\mathbf{1}_{p-1}$ represents a $p-1$ vector of ones. With this definition, $\Delta' v^{(1)} = 0$. Similar to $v^{(1)}$, define $v^{(2)}$ as $[\mathbf{1}_{p-2}, b, c]'$, where the solution for b and c is determined by solving the following system:

$$\begin{bmatrix} \Delta_{p-1} & \Delta_p \\ 1 & a \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix} = - \begin{bmatrix} \sum_{i=1}^{p-2} \Delta_i \\ p-2 \end{bmatrix}.$$

References

Asimov, D. (1985), "The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM Journal on Scientific and Statistical Computing*, 6, 128–143.

Brown, M. P. S., Grundy, W. N., Lin, D., Sugnet, C. W., Furey, T. S., Ares, Manuel, J., and Haussler, D. (2000), "Knowledge-Bases Analysis of Microarray Gene Expression Data by using Support Vector Machines," *Proceedings of the National Academy of Sciences (PNAS)*, 97, 262–267.

Buxton, R. (1978), "The Interpretation and Justification of the Subjective Bayesian Approach to Statistical Inference (MR V57 14200)," *The British Journal for the Philosophy of Science*, 29, 25–38.

Carroll, J. D. and Chang, J. J. (1970), "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition," *Psychometrika*, 35, 238–319.

Cattell, R. B. (1965), "Factor Analysis: An Introduction to Essentials. I: The Purpose and Underlying Models," *Biometrics*, 21, 190–215.

Chapelle, O., Schölkopf, B., and Zien, A. (2006), *Semi-Supervised Learning*, The MIT Press, Cambridge, Massachusetts.

Cook, D. and Swayne, D. F. (2007), "Interactive and Dynamic Graphics for Data Analysis with R and GGobi," *Amstat News*, 364, 26–26.

Daneshkhah, A. (2004), "Psychological Aspects Influencing Elicitation of Subjective Probability," Tech. rep., University of Sheffield UK; BEEPs report.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences (PNAS)*, 95, 14863–14868.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

Friedman, J. H. and Tukey, J. W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers*, 23, 881–890.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005), "Statistical Methods for Eliciting Probability Distributions," *Journal of the American Statistical Association*, 100, 680–701.

Goldstein, M. (2006), "Subjective Bayesian Analysis: Princiles and Practice (Pkg: P403-472)," *Bayesian Analysis*, 1, 403–420.

Goldstein, M. and Woof, D. (2007), *Bayes Linear Statistics*, West Sussex: Jon Wiley and Sons Ltd.

Good, I. J. (1983), *Good Thinking: the Foundations of Probability and Its Applications*, University of Minnesota Press.

Guber, D. (1999)), "Getting What You Pay For: The Debate Over Equity in Public School Expenditures," *Journal of Statistics Education*, 7.

Icke, I. and Sklar, E. (2009), "Visual Analytics: A Multifaceted Overview," Tech. rep., City University of New York.

Jaynes, E. (1983), *Papers on Probability, Statistics, and Statistical Physics (ed. Rosenkrantz, R.D.)*, D. Reidel publishing Co., Dordrecth, Holland.

Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press, 3 edn.

Johnson, C. (2004), "Top Scientific Visualization Research Problems," *IEEE Computer Graphics and Applications*, 24, 13–17.

Jolliffe, I. (2002), *Principal Component Analysis*, John Wiley and Sons, Ltd, 2nd edn.

Kadane, J. B. and Wolfson, L. J. (1998), "Experiences in Elicitation," *The Statistician*, 47, 3–19.

Keim, D. A. (2002), "Information Visualization and Visual Data Mining," *IEEE Transactions On Visulations and Computer Graphics*, 7, 100–107.

Kruskal, J. B. and Wish, M. (1978), "Multidimensional Scaling," *Sage University Paper series on Quantitative Application in the Social Sciences*, 48, 07–011.

Leban, G., Zupan, B., Vidmar, G., and Bratko, I. (2006), "VizRank: Data Visualization Guided by Machine Learning," *Data Mining and Knowledge Discovery*, 13, 119–136.

Lederberg, J. (1989), *Excitement and Fascination of Science*, chap. Preface: Twelve-Step Process for Scientific Experiments: Epicycles of Scientific Discovery., Annual Reviews, Inc., Palo Alto, California.

Lewin-Koh, S.-C. and Amemiya, Y. (1998), "Latent Variable Modeling with Error Variances Depending on Latent Variables," in *ASA Proceedings of the Statistical Computing Section*, pp. 113–118, American Statistical Association.

MacInnes, J., Santosa, S., and Wright, W. (2010), "Visual Classification: Expert Knowledge Guides Machine Learning," *Computer Graphics and Applications, IEEE*, 30, 8 – 14.

Mordecai, A. (1976), *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, New Jersey.

Oh, M.-S. and Raftery, A. E. (2001), "Bayesian Multidimensional Scaling and Choice of Dimension," *Journal of the American Statistical Association*, 96, 1031–1044.

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, 6, 559–572.

Press, S. J. and Shigemasu, K. (1989), "Bayesian Inference in Factor Analysis," in *ASA Proceedings of the Social Statistics Section*, pp. 292–294, American Statistical Association.

Ramsey, F. (1926), "Truth and Probability," *Foundations: Essays in Philosophy, Logic, Mathematics, and Economics*.

Savage, L. (1954), *Foundation of Statistical Inference*, Wiley, New York.

Schiffman, S. S., Reynolds, M. L., and Young, F. W. (1981), *Introduction to Multidimensional Scaling: Theory, Methods, and Applications*, Academic Press, New York.

Spiegelhalter, D. J. and Lauritzen, S. L. (1990), "Sequential Updating of Conditional Probabilities on Directed Graphical Structures," *Networks*, 20, 579–605.

Swayne, D. F., Lang, D. T., Buja, A., and Cook, D. (2003), "GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization," *Computational Statistics & Data Analysis*, 43, 423–444.

Thomas, J. and Cook, K. (eds.) (2005), *Illuminiating the Path*, National Visualizations and Analytics Center.

Tipping, M. E. and Bishop, C. M. (1999), "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61, 611–622.

Torgerson, W. S. (1958), *Theory and Methods of Scaling*, John Wiley, New York.

Torokhti, A. and Friedland, S. (2009), "Towards theory of generic Principal Component Analysis," *Journal of multivariate analysis*, 100, 661–669.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer-Verlag Inc.

Yang, R. and Berger, J. O. (1997), "Catalog of Nonimformative Priors," Tech. Rep. 97-42, Department of Statistical Science (formally known as the Institute of Statistics and Decision Sciences), Duke University.

Zhao, K., Jiu, B., Tirpak, T. M., and Xiao, W. (2005), "A Visual Data Mining Framework for Convenient Identification of Useful Knowledge," in *Proceedings of the Fifth IEEE International Conference on Data Mining*, pp. 530–537, IEEE Computer Society Washington, DC, USA.