Robust Probabilistic Multivariate Calibration Model

Yi Fang

Department of Computer Science
Purdue University
West Lafayette, IN 47907
(*fangy@cs.purdue.edu*)

Myong K. Jeong

Center for Operations Research &
Department of Industrial and Systems Engineering
Rutgers, The State University of New Jersey
Piscataway, NJ 08853
(*mkjeong@rutcor.rutgers.edu*)

In this article we propose a robust probabilistic multivariate calibration (RPMC) model in an attempt to identify linear relationships between two sets of observed variables contaminated with outliers. Instead of the Gaussian assumptions that predominate in classical statistical models, RPMC is closely related with the multivariate Student *t*-distribution over noises and latent variables. Thus RPMC diminishes the effect of outlying data points by regulating the thickness of the distribution tails. RPMC is essentially a robustified version of the supervised probabilistic principal component analysis (SPPCA) that has emerged recently. We show that RPMC encompasses probabilistic principal component analysis and SPPCA as limiting cases. We also derive an efficient EM algorithm for parameter estimation in RPMC. Based on a probabilistic description of latent variables, we present a procedure for the detection of outliers. The experimental results from both simulated examples and real life data sets demonstrate the effectiveness and robustness of our proposed approach.

KEY WORDS: Expectation-maximization algorithm; Latent variable model; Multivariate Student *t*-distribution; Outlier detection; Probabilistic principal component analysis; Supervised probabilistic principal component analysis.

1. INTRODUCTION

Outliers are observations that seem extreme or unusual with respect to the rest of the data and to previous knowledge about what values are plausible (Ghoshdastider and Schafer 2003). In recent years, the need to deal with huge amounts of data has become a common problem in research laboratories and in industrial operations across most research disciplines and their commercial applications. It is well known that all large data sets contain outliers (Wold, Berglund, and Kettaneh 2002), and that the manual evaluation of these outliers is extremely difficult. Most conventional multivariate calibration methods are sensitive to outliers because they are based on least squares or a similar loss function in which even a single outlier has a huge impact on the development of a model. Thus automatic outlier detection and robust methods are of paramount importance in a multivariate calibration model.

Some simple approaches to screening outliers exist, including trimming and winsorizing (Tukey 1962). These intuitive methods have a common major drawback in their discarding of data. These missing data in turn create a situation likely to lead to biased estimates. Many sophisticated robust approaches have been proposed as alternatives to these simple techniques. Among the alternatives, considerable attention has been given to replacing the nonrobust least squares estimate with a robust estimate. Notable examples of these robust estimates include M-estimates (Huber 1964), the Stahel–Donoho estimate (Stahel 1981; Donoho 1982), least median of squares (Rousseeuw 1984), and S-estimators (Davies 1987; Lopuhaa 1989); a recent overview of these robust estimates has been given by Moller, Frese, and Bro (2005). Most of these methods are attempts to improve the robustness of common multivariate regression techniques, such as principal component regression (PCR) (Chatterjee and Price 1977) and partial least squares (PLS) (Wold 1966). An alternative to giving such attention to data-analytic activities is to use the capability of an experimental design that explicitly includes consideration of instrumental and environmental factors (Thomas and Ge 2000). Although these approaches are sometimes effective, none defines a generative model or normalized probability for the data.

Another notorious problem with high-dimensional data is what Bellman (1961) termed the *curse of dimensionality*, in which both the number of computations required for a predictive model and the amount of data required for calibration grow exponentially with the increased dimensionality of the feature vectors. Although PCR and PLS do reduce data dimensionality, they do not define a proper probability model underlying the data generation; in other words, they cannot use the model to tell us how well new data fit.

In this article we propose an approach that we term *robust probabilistic multivariate calibration* (RPMC) in the framework of latent variable models. RPMC models outliers by probabilistic description instead of by simply removing them. RPMC also yields more robust estimates at contaminated data. Using these accurate estimates, we provide a statistically principled way to identify outliers. Moreover, we analyze the relationships between RPMC and other recently proposed latent variable models. Finally, we derive an efficient expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) for parameter estimation in RPMC. Section 2 summarizes previous work with latent variable models. Section 3 presents the RPMC model and the EM algorithm for learning parameters, and Section 4 gives experimental results obtained

TECHNOMETRICS, AUGUST 2008, VOL. 50, NO. 3
DOI 10.1198/004017008000000073

from various data sets. Section 5 contains some concluding remarks and maps out some directions for future studies.

2. LATENT VARIABLE MODELS

Latent variables, also called hidden variables, are variables that are not observed directly but instead are inferred from other variables that are measured directly. A latent variable model is a statistical model used to investigate the dependence of a set of manifest (observed) variables on a set of latent variables (Everitt 1984). A general latent variable model has the form

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{t})h(\mathbf{t})\,d\mathbf{t},$$

where $\mathbf{x} = [x_1, \dots, x_M]^T$ represents the observable variables and $\mathbf{t} = [t_1, \dots, t_P]^T$ represents the latent variables. The number of latent variables, P, is usually far fewer than the number of observable variables, M. In essence, all latent variable models assume that $\mathbf{x}$ has a joint probability distribution conditional on $\mathbf{t}$, denoted by $p(\mathbf{x}|\mathbf{t})$. Based on some assumptions, we can infer the density functions, p and h, from the known or assumed density of $\mathbf{x}$ to explore how the manifest variables depend on the latent variables. Latent variable models rest on a key assumption of conditional independence. This assumption holds that observable variables are independent of one another, given the values of latent variables. In other words, the observed interdependence among the observable variables results wholly from their common dependence on the latent variables. Once the latent variables are fixed in place, the behavior of the observable variables is essentially random. Mathematically, this can be expressed as

$$p(\mathbf{x}) = \int h(\mathbf{t}) \prod_{i=1}^{M} p(x_i|\mathbf{t})\,d\mathbf{t}.$$

Depending on the different assumptions made about latent variables, different classes of latent variable models can be constructed. The best known of these is factor analysis, which was initially developed by psychologists (Spearman 1904). Recent research has found that many popular multivariate statistical techniques are closely related to latent variable models; notable examples include vector quantization, independent component analysis (ICA) models, Kalman filter models, and hidden Markov models (HMMs) (Roweis and Ghahramani 1999).

2.1 Probabilistic Principal Component Analysis

Principal component analysis (PCA) (Pearson 1901; Hotelling 1936) is a widely used statistical tool for reducing the dimensionality of high-dimensional data sets for ease of analysis. Although PCA originates from the analysis of data variances, recently it has been connected to the maximum likelihood solution for a generative latent variable model, which is called probabilistic PCA (PPCA) (Tipping and Bishop 1999b) and is defined as

$$\mathbf{x} = \mathbf{W}_\mathbf{x}\mathbf{t} + \boldsymbol{\mu}_\mathbf{x} + \boldsymbol{\varepsilon}_\mathbf{x}, \tag{1}$$

where $\mathbf{x} \in \Re^M$ is the observed variable, $\mathbf{t} \in \Re^P$ is the latent variable, $\mathbf{W}_\mathbf{x}$ is a $M \times P$ matrix called factor loading, and $\boldsymbol{\varepsilon}_\mathbf{x}$ defines a noise process. In addition, we have the parameter $\boldsymbol{\mu}_\mathbf{x}$ which allows nonzero means for the data. In this probabilistic model, the latent variable $\mathbf{t}$ is conventionally assumed to satisfy a standard multivariate Gaussian distribution [i.e., $\mathbf{t} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$], and $\boldsymbol{\varepsilon}_\mathbf{x}$ takes an isotropic Gaussian form as $\boldsymbol{\varepsilon}_\mathbf{x} \sim \mathrm{N}(\mathbf{0}, \sigma_\mathbf{x}^2\mathbf{I})$. Tipping and Bishop (1999b) showed that the maximum likelihood estimator of $\mathbf{W}_\mathbf{x}$ is given as

$$\tilde{\mathbf{W}}_\mathbf{x} = \mathbf{U}_P(\mathbf{E}_P - \sigma_\mathbf{x}^2\mathbf{I}_P)^{1/2}\mathbf{R}, \tag{2}$$

where $\mathbf{U}_P$ is the matrix of the P principal eigenvectors (corresponding to the P largest eigenvalues) of the sample covariance matrix $S_\mathbf{x} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \tilde{\mathbf{x}})(\mathbf{x}_i - \tilde{\mathbf{x}})^T$, $\mathbf{E}_P \in \mathbb{R}^{P\times P}$ is the diagonal matrix of the corresponding eigenvalues, $\mathbf{I}_P \in \mathbb{R}^{P\times P}$ is the P-dimensional identity matrix, and $\mathbf{R}$ is an arbitrary $P \times P$ orthogonal matrix. The expected projection $\tilde{\mathbf{t}}$ for new observation $\mathbf{x}^*$ is given as

$$\tilde{\mathbf{t}} = \mathbf{R}^T(\mathbf{E}_P - \sigma_\mathbf{x}^2\mathbf{I}_P)^{1/2}\mathbf{E}_P^{-1}\mathbf{U}_P^T(\mathbf{x}^* - \boldsymbol{\mu}_\mathbf{x}).$$

PCA is recovered when the covariance of the noise becomes infinitesimally small. This probabilistic formulation provides additional advantages over conventional PCA, including a principled way of handling missing values, a fast EM learning procedure, and the availability of a Bayesian treatment. In addition, PPCA has strong connections to factor analysis (Tipping and Bishop 1999b).

2.2 Supervised Probabilistic Principal Components Analysis

In many applications, each data observation is associated not only with input $\mathbf{x}$, but also with output $\mathbf{y} = [y_1, \dots, y_K]^T \in \mathbb{R}^K$. Therefore, an unsupervised learning method, such as PCA or PPCA, may not be able to project the data into useful subspaces. Many supervised learning methods have been proposed to make use of output information, including PCR, PLS, linear discriminant analysis (LDA), and supervised principal component methods (Bair, Hastie, Paul, and Tibshirani 2006). Based on latent variable models, supervised probabilistic PCA (SPPCA) was recently introduced (Yu, Yu, Tresp, Kriegel, and Wu 2006). Like PPCA, SPPCA uses the key assumption of latent variable models that all of the observations are conditionally independent, given the latent variables. In SPPCA, the observed data $(\mathbf{x}, \mathbf{y})$ are generated from a latent variable model as

$$\begin{aligned} \mathbf{x} &= \mathbf{W}_\mathbf{x}\mathbf{t} + \boldsymbol{\mu}_\mathbf{x} + \boldsymbol{\varepsilon}_\mathbf{x}, \\ \mathbf{y} &= \mathbf{W}_\mathbf{y}\mathbf{t} + \boldsymbol{\mu}_\mathbf{y} + \boldsymbol{\varepsilon}_\mathbf{y}. \end{aligned} \tag{3}$$

A unit-isotropic Gaussian distribution is assumed for the P-dimensional latent variable $\mathbf{t}$ and independent for the error terms $\boldsymbol{\varepsilon}_\mathbf{x}$ and $\boldsymbol{\varepsilon}_\mathbf{y}$, that is, $\mathbf{t} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\varepsilon}_\mathbf{x} \sim \mathrm{N}(\mathbf{0}, \sigma_\mathbf{x}^2\mathbf{I})$, and $\boldsymbol{\varepsilon}_\mathbf{y} \sim \mathrm{N}(\mathbf{0}, \sigma_\mathbf{y}^2\mathbf{I})$. It can be shown that the maximum likelihood solutions of $\mathbf{W}_\mathbf{x}$ and $\mathbf{W}_\mathbf{y}$ are given by Yu et al. (2006) as

$$\tilde{\mathbf{W}}_\mathbf{x} = \sigma_\mathbf{x}\mathbf{U}_M(\mathbf{E}_P - \mathbf{I}_P)^{1/2}\mathbf{R}$$

and

$$\tilde{\mathbf{W}}_\mathbf{y} = \sigma_\mathbf{y}\mathbf{U}_K(\mathbf{E}_P - \mathbf{I}_P)^{1/2}\mathbf{R},$$

where $\mathbf{U}_M$ ($\mathbf{U}_K$) contains the first M (or last K) rows of eigenvectors of the normalized sample covariance matrix $\mathbf{S}$ for centered observations $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, $\mathbf{E}_P \in \mathbb{R}^{P\times P}$ is the diagonal matrix of the corresponding eigenvalues, $\mathbf{I}_P \in \mathbb{R}^{P\times P}$ is the P-dimensional identity matrix, and $\mathbf{R}$ is an arbitrary $P \times P$ orthogonal matrix. The projected latent variable $\mathbf{t}^*$ for centered new input $\mathbf{x}^*$ is given by

$$\mathbf{t}^* = \frac{1}{\sigma_{\mathbf{x}}}\mathbf{R}^T(\mathbf{E}_P - \mathbf{I}_P)^{1/2}[\mathbf{U}_M^T\mathbf{U}_M + (\mathbf{E}_P - \mathbf{I}_P)^{-1}]^{-1}\mathbf{U}_M^T\mathbf{x}^*. \tag{4}$$

When $K > 0$, SPPCA explains not only the intracovariances of inputs $\mathbf{S}_{\mathbf{x}}$ and output $\mathbf{S}_{\mathbf{y}}$, but also the intercovariance between input and output, $\mathbf{S}_{\mathbf{xy}}$ and $\mathbf{S}_{\mathbf{yx}}$. In contrast to SPPCA, PCA only explains the covariance of inputs; as for PLS, it finds the maximal covariance between inputs and outputs but ignores the intracovariance of both inputs and outputs.

3. ROBUST PROBABILISTIC MULTIVARIATE CALIBRATION MODEL

3.1 Multivariate Student's *t*-Distribution

Both PPCA and SPPCA take advantage of the Gaussian assumption about noise, based on the fact that the convolution of two independent Gaussian-distributed quantities also is Gaussian-distributed. This nice analytical property of Gaussian distributions often yields tractable algorithms for linear Gaussian models. One major limitation of such Gaussian models, however, is their sensitivity to outliers. This can be readily understood by recalling the linear Gaussian regression models in which the maximization of likelihood function is equivalent to finding the least squares solution, which is well known for its lack of robustness (Svensen and Bishop 2004).

The Student *t*-distributions have heavier tails compared with a Gaussian distribution (Fig. 1). By assigning higher probability densities to outliers, the effect of outliers on model development is diminished. A *t*-distribution is commonly used in robust regression (Lange, Little, and Taylor 1989). Previous work has replaced Gaussian distributions with *t*-distributions (Archambeau, Delannay, and Verleysen 2006) as a way to increase the robustness of PPCA and probabilistic canonical correlation analysis (Bach and Jordan 2005). The *t*-distributions also have proven effective in computer vision and mixture modeling (Peel and McLachlan 2000; Torre and Black 2001; Archambeau 2005).

Specifically, the *t*-distribution for the M-dimensional variable $\mathbf{x}$ is defined as (Svensen and Bishop 2004)

$$S(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\nu/2 + M/2)|\boldsymbol{\Sigma}|^{-1/2}}{\Gamma(\nu/2)(\nu\pi)^{M/2}}\left(1 + \frac{\Delta^2}{\nu}\right)^{-(\nu+M)/2},$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance matrix of $\mathbf{x}$ $\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$, and $\Gamma(\cdot)$ denotes the gamma function, that is, $\Gamma(z) = \int_0^\infty y^{z-1}e^{-y}\,dy$. The parameter $\nu > 0$ is the degree of freedom, which controls the thickness of the distribution tails and thus regulates the degree of robustness to outliers. Maximum likelihood estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is robust in the sense that outliers with large squared Mahalanobis distance are downweighted (Lange et al. 1989). When parameter ν goes to infinity, *t*-distributions approach Gaussian distributions.

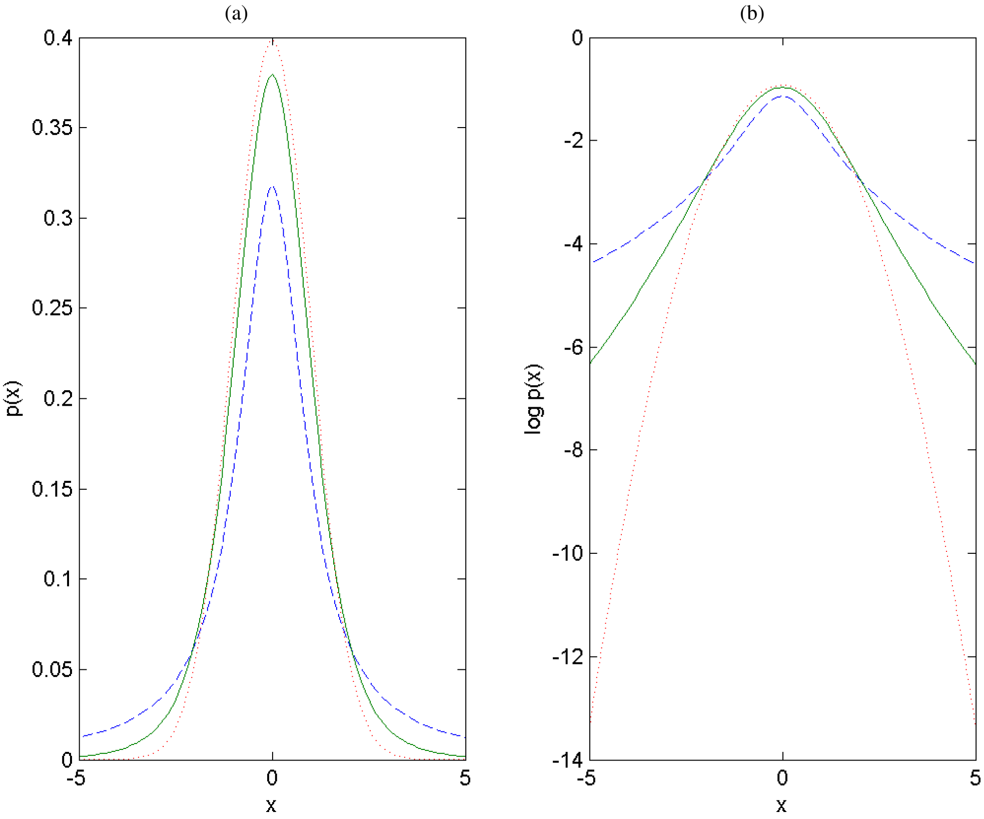

Figure 1. Heavy tails of *t*-distributions $S(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with fixed $\boldsymbol{\mu} = \mathbf{0}$ and various ν on (a) a normal scale and (b) a log scale (--- $\nu = 1$; —— $\nu = 5$; ······ $\nu \to \infty$).

3.2 The Robust Probabilistic Multivariate Calibration Model

We now formally describe our proposed model, which we cal the *robust probabilistic multivariate calibration* (RPMC) approach. RPMC is a latent variable model whose components have a t-distribution instead of a Gaussian distribution. In RPMC, the observed input and output $(\mathbf{x}, \mathbf{y})$ are generated from a latent variable model defined as

$$\begin{aligned} p(\theta) &= G\left(\theta \middle| \frac{\nu}{2}, \frac{\nu}{2}\right), \\ p(\mathbf{t}|\theta) &= \mathrm{N}(\mathbf{t}|\mathbf{0}, \theta^{-1}\mathbf{I}_P), \\ p(\mathbf{x}|\mathbf{t}, \theta) &= \mathrm{N}(\mathbf{x}|\mathbf{W}_\mathbf{x}\mathbf{t} + \boldsymbol{\mu}_\mathbf{x}, \theta^{-1}\sigma_\mathbf{x}^2\mathbf{I}_M), \\ p(\mathbf{y}|\mathbf{t}, \theta) &= \mathrm{N}(\mathbf{y}|\mathbf{W}_\mathbf{y}\mathbf{t} + \boldsymbol{\mu}_\mathbf{y}, \theta^{-1}\sigma_\mathbf{y}^2\mathbf{I}_K), \end{aligned} \tag{5}$$

where $G(\cdot)$ represents a gamma distribution, that is, $G(\theta|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$.

Similar to in SPPCA, here $\mathbf{t}$ is the latent variable shared by observed variables $\mathbf{x}$ and $\mathbf{y}$; the difference is that RPMC defines an extra latent variable θ behind the observed variable $(\mathbf{x}, \mathbf{y})$ and the latent variable $\mathbf{t}$. Writing $\boldsymbol{\mu} = \binom{\boldsymbol{\mu}_\mathbf{x}}{\boldsymbol{\mu}_\mathbf{y}}$, $\boldsymbol{\Phi} = \left(\begin{smallmatrix} \sigma_\mathbf{x}^2\mathbf{I} & 0 \\ 0 & \sigma_\mathbf{y}^2\mathbf{I} \end{smallmatrix}\right)$, and $\mathbf{W} = \binom{\mathbf{W}_\mathbf{x}}{\mathbf{W}_\mathbf{y}}$, our goal is to infer the latent variables $H = \{\theta, \mathbf{t}\}$ and the parameters $\Omega = \{\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\mu}, \nu\}$ from the observed data $(\mathbf{x}, \mathbf{y})$. Graphical representations of PPCA, SPPCA, and RPMC are shown in Figure 2.

Next, we link RPMC with a t-distribution by Theorem 1 and show its close relationships with SPPCA and PPCA by Propositions 1 and 2 (see App. A for proofs).

Theorem 1. For the RPMC model, the marginal distribution of $\mathbf{t}$ and of $\mathbf{x}$ and $\mathbf{y}$ conditional on $\mathbf{t}$ is given by

$$\begin{aligned} p(\mathbf{t}) &= S(\mathbf{t}|\mathbf{0}, \mathbf{I}_P, \nu), \\ p(\mathbf{x}|\mathbf{t}) &= S(\mathbf{x}|\mathbf{W}_\mathbf{x}\mathbf{t} + \boldsymbol{\mu}_\mathbf{x}, \sigma_\mathbf{x}^2\mathbf{I}_M, \nu), \\ p(\mathbf{y}|\mathbf{t}) &= S(\mathbf{y}|\mathbf{W}_\mathbf{y}\mathbf{t} + \boldsymbol{\mu}_\mathbf{y}, \sigma_\mathbf{y}^2\mathbf{I}_K, \nu). \end{aligned} \tag{6}$$

Theorem 1 actually demonstrates an intimate relationship between RPMC and model (6) defined by the Student t-distribution, which assumes that the noise is drawn from a t-distribution, as is the latent variable $\mathbf{t}$. This model can be considered a robustified version of SPPCA by changing the Gaussian distribution on noise and latent variables to t-distributions. But the model (6) lacks a closed-form solution or even a tractable EM algorithm for maximization of likelihood. Fortunately, we can derive a tractable EM algorithm for model (5), which is closely related to model (6).

Proposition 1. If ν goes to infinity, then RPMC is equivalent to SPPCA.

Proposition 2. If ν goes to infinity and $K = 0$, then RPMC is equivalent to PPCA.

Propositions 1 and 2 show that PPCA and SPPCA are the limiting cases of RPMC.

3.3 EM Algorithm for Parameter Estimation

Parameter estimation in latent variable models can be reduced to maximization of data likelihood with respect to all of the model's parameters. After observing n pairs of input and output, the likelihood of all of the observations $F = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, with iid assumption, is $p(F) = \prod_{i=1}^n p(\mathbf{x}_i, \mathbf{y}_i)$.

In the case of the RPMC model (5), the log-likelihood of the whole observation is defined as

$$\tilde{L} = \log \prod_{i=1}^{n} p(\mathbf{z}_i|\mathbf{t}_i, \theta_i, \boldsymbol{\mu}, \mathbf{W}, \boldsymbol{\Phi})p(\mathbf{t}_i|\theta_i)p(\theta_i|\nu), \tag{7}$$

where $\mathbf{z}_i = (\mathbf{x}_i \ \ \mathbf{y}_i)$.

Manipulating (7) and omitting the constant terms with respect to Ω gives

$$\begin{aligned} \tilde{L} = \frac{1}{4}\sum_{i=1}^{n} n \log|\boldsymbol{\Phi}| + \frac{n\nu}{2}\log\frac{\nu}{2} + \left(\frac{\nu}{2} - 1\right)\sum_{i=1}^{n}\log\theta_i \\ - \frac{1}{2}\sum_{i=1}^{n}\theta_i(\mathbf{z}_i - \mathbf{W}\mathbf{t}_i - \boldsymbol{\mu})^T\boldsymbol{\Phi}(\mathbf{z}_i - \mathbf{W}\mathbf{t}_i - \boldsymbol{\mu}) \\ - \frac{n}{2}\log|\boldsymbol{\Phi}| - n\log\Gamma\left(\frac{\nu}{2}\right) - \frac{\nu}{2}\sum_{i=1}^{n}\theta_i, \end{aligned}$$

where $|\cdot|$ calculates the determinant of the square matrix.

The absence of a closed-form solution to maximum likelihood of (7) contrasts with the linear Gaussian models. Fortunately, we can derive an EM algorithm that is applicable to the

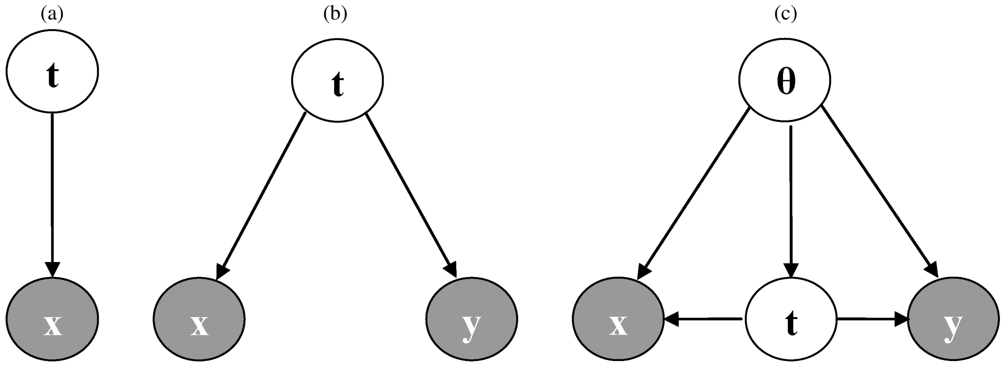

Figure 2. Graphical models of PPCA (a), SPPCA (b), and RPMC (c). The shaded nodes are observed variables; the arrows represent conditional dependencies between random variables.

model. The EM algorithm alternates between performing an expectation E-step and a maximization M-step. The parameters found in the M-step are then used to begin another E-step, and the process is repeated. In the E-step, we compute the expectation of (7), averaging over the latent variables $H = \{\theta_i, \mathbf{t}_i\}$, given the current estimate of the parameters $\Omega = \{\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\mu}, \nu\}$. In the M-step, we fix this expectation and maximize the complete-data likelihood with respect to the parameters. In this section we give only the updated equations; we provide the details of the derivation in Appendix B.

Statistics sufficient to update the parameters for the M-step are then given by

$$\langle\theta_i\rangle = \frac{M+K+\nu}{(\mathbf{z}_i-\boldsymbol{\mu})^T(\mathbf{W}\mathbf{W}^T+\boldsymbol{\Phi})^{-1}(\mathbf{z}_i-\boldsymbol{\mu})+\nu} \tag{8}$$

$$\langle\mathbf{t}_i\rangle = (\mathbf{I}_P+\mathbf{W}^T\boldsymbol{\Phi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{\Phi}^{-1}(\mathbf{z}_i-\boldsymbol{\mu}), \tag{9}$$

$$\langle\log\theta_i\rangle = \psi\left(\frac{M+K+\nu}{2}\right) - \log\left(\frac{(\mathbf{z}_i-\boldsymbol{\mu})^T(\mathbf{W}\mathbf{W}^T+\boldsymbol{\Phi})^{-1}(\mathbf{z}_i-\boldsymbol{\mu})+\nu}{2}\right),$$

and

$$\langle\theta_i\mathbf{t}_i\mathbf{t}_i^T\rangle = \langle\theta_i\rangle\langle\mathbf{t}_i\rangle\langle\mathbf{t}_i\rangle^T + (\mathbf{I}_P+\mathbf{W}^T\boldsymbol{\Phi}^{-1}\mathbf{W})^{-1},$$

where $\langle\cdot\rangle$ is the expectation operator and $\psi(\cdot)$ denotes the digamma function, that is, $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$.

In the M-step, we estimate parameters $\Omega = \{\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\mu}, \nu\}$ by maximizing the expected likelihood found on the E-step, that is,

$$\Omega^{j+1} = \arg\max_{\Omega} \Upsilon(\Omega|\Omega^j),$$

where $\Upsilon(\Omega|\Omega^j) = E[\tilde{L}|H, \Omega^j]$, denoting the conditional expectation of $\tilde{L}$ being taken with Ω in the conditional distribution of H fixed at Ω^j. Then we set $\frac{\partial\Upsilon(\Omega|\Omega^j)}{\partial\Omega} = 0$ to estimate Ω.

Therefore, the mean vector is updated by

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^n\langle\theta_i\rangle(\mathbf{z}_i-\mathbf{W}\langle\mathbf{t}_i\rangle)}{\sum_{i=1}^n\langle\theta_i\rangle}. \tag{10}$$

The factor loading matrices are updated by

$$\mathbf{W} = \left(\sum_{i=1}^n\langle\theta_i\rangle(\mathbf{z}_i-\boldsymbol{\mu})\langle\mathbf{t}_i\rangle^T\right)\left(\sum_{i=1}^n\langle\theta_i\mathbf{t}_i\mathbf{t}_i^T\rangle\right)^{-1}. \tag{11}$$

The noise levels are updated by

$$\sigma_{\mathbf{x}}^2 = \frac{1}{M\times n}\sum_{i=1}^n\bigl\{\mathrm{tr}\{\langle\theta_i\mathbf{t}_i\mathbf{t}_i^T\rangle\mathbf{W}_{\mathbf{x}}^T\mathbf{W}_{\mathbf{x}}\} + \langle\theta_i\rangle(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}})^T(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}}) - 2\langle\theta_i\rangle(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}})^T\mathbf{W}_{\mathbf{x}}\langle\mathbf{t}_i\rangle\bigr\}, \tag{12}$$

$$\sigma_{\mathbf{y}}^2 = \frac{1}{K\times n}\sum_{i=1}^n\bigl\{\mathrm{tr}\{\langle\theta_i\mathbf{t}_i\mathbf{t}_i^T\rangle\mathbf{W}_{\mathbf{y}}^T\mathbf{W}_{\mathbf{y}}\} + \langle\theta_i\rangle(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y}})^T(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y}}) - 2\langle\theta_i\rangle(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{y}})^T\mathbf{W}_{\mathbf{y}}\langle\mathbf{t}_i\rangle\bigr\},$$

where $\mathrm{tr}(\cdot)$ calculates the trace of the matrix.

Finally, the maximum likelihood solution of ν is calculated by solving

$$\frac{1}{n}\sum_{i=1}^n\{\langle\log\theta_i\rangle - \langle\theta_i\rangle\} + \log\left(\frac{\nu}{2}\right) + 1 - \psi\left(\frac{\nu}{2}\right) = 0. \tag{13}$$

In this EM algorithm, only the dimensionality P of the latent variable must be specified. One general way to select P is to maintain a balance between a good fit of the data and a reasonable number of parameters, which normally takes the form of a penalized likelihood function such as the Akaike information criterion (Akaike 1974), Bayes information criterion (Schwarz 1978), or ICOMP (Bozdogan 1988). We can readily adapt the EM algorithm presented here to this principle by just adding a penalty term that increases monotonically with the number of parameters. But there is no substitute for careful consideration in the context of each individual problem. In many real applications, we may have sufficient domain or prior knowledge to help guide our choice.

3.4 Outlier Detection

Outliers often are of primary interest. No matter the source of outliers, detecting them is an important task, because they imply that some form of action is necessary. Numerous terms, including "anomaly detection," have been used to denote outlier detection. The standard method for multivariate outlier detection is the Mahalanobis distance in conjunction with comparison with critical value of the chi-squared distribution (Rousseeuw and Van Zomeren 1990), which is based on a Gaussian assumption of the data. In this section we propose taking SPPCA and RPMC as the underlying probabilistic models for data generation, because both result in a low-dimensional Gaussian latent variable that can be measured by the chi-squared metrics.

3.4.1 Outlier Detection by SPPCA. Yu et al. (2006) showed that in SPPCA, the observed data, $\mathbf{z} = (\mathbf{x}\ \ \mathbf{y})$, are jointly Gaussian-distributed as

$$\mathbf{z} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Phi}),$$

where

$$\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_{\mathbf{x}}\\ \boldsymbol{\mu}_{\mathbf{y}}\end{pmatrix}, \qquad \boldsymbol{\Phi} = \begin{pmatrix}\sigma_{\mathbf{x}}^2\mathbf{I} & 0\\ 0 & \sigma_{\mathbf{y}}^2\mathbf{I}\end{pmatrix}, \qquad \mathbf{W} = \begin{pmatrix}\mathbf{W}_{\mathbf{x}}\\ \mathbf{W}_{\mathbf{y}}\end{pmatrix},$$

and $\boldsymbol{\mu}_{\mathbf{x}}$, $\boldsymbol{\mu}_{\mathbf{y}}$, $\sigma_{\mathbf{x}}$, $\sigma_{\mathbf{y}}$, $\mathbf{W}_{\mathbf{x}}$, and $\mathbf{W}_{\mathbf{y}}$ are as defined in the SPPCA model.

The chi-squared statistic for a new observation is given by

$$C = (\mathbf{z}-\boldsymbol{\mu})^T(\mathbf{W}\mathbf{W}^T+\boldsymbol{\Phi})^{-1}(\mathbf{z}-\boldsymbol{\mu}).$$

If the distribution of $\mathbf{z} \in \Re^d$ is multivariate Gaussian, then the statistic is approximately chi-squared distributed with d degrees of freedom (χ_d^2). Multivariate outliers can now be defined as observations having a large squared Mahalanobis distance, C. For this purpose, a quantile of the chi-squared distribution (e.g., the 95% quantile) could be considered.

But a recent rigorous mathematical analysis has validated a widely observed empirical fact, that a Gaussian distribution is often an accurate density model for low-dimensional data but very rarely for high-dimensional data (Dasgupta, Hsu, and Verma 2006). Therefore, instead of dealing with high-dimensional observed variable **z**, we monitor the low-dimensional latent variable **t** defined in (4), which is supposed to satisfy an isotropic Gaussian distribution. This replacement makes sense, because the outliers in the original data space generally are anomalous in the latent variable space. The corresponding chi-squared statistic is simply

$$C = \mathbf{t}^T\mathbf{t}. \tag{14}$$

3.4.2 Outlier Detection by RPMC. In SPPCA, the estimator for **t** may be substantially biased by the presence of outliers. Instead, we can use the robust parameters through RPMC in which the effect of outliers is downweighted. We still use the chi-squared statistic C defined in (14) as the criterion for measuring anomalies. The only difference from the procedure described in Section 3.4.1 is the use of RPMC to infer the latent variable **t**. We compare two estimators in case studies in the next section.

4. EXPERIMENTS

In this section we use various data sets—including artificial data, low-dimensional data with a single response, and high-dimensional data with multiple responses—to compare the performance of RPMC with that of the conventional multivariate calibration methods such as PLS, robust PLS (Hubert and Branden 2003), and SPPCA.

4.1 Simulation Study

In this section we report the results of a simulation study conducted to evaluate the robustness of RPMC. We show that in the absence of outliers, RPMC finds the same principal directions as SPPCA. Moreover, RPMC can reduce the effect of any outliers present in the data sets.

First, we consider a case without outliers. We generated 20 samples from each of two multivariate two-dimensional Gaussian distributions as $C_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $C_2 \sim \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\mu}_1 = [-6, 0]^T, \boldsymbol{\mu}_2 = [6, 0]^T$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{bmatrix} 3 & 1.5 \\ 1.5 & 3 \end{bmatrix}$. The data are shown in Figure 3.

We assigned a different binary output label for each class. We used SPPCA and RPMC for separate projections of the data into one-dimensional space (i.e., $P = 1$). In our experiments, to choose the initial values of the EM algorithm described in Section 3.3, we first used several different random initial estimates and chose the combination that produced the highest likelihood (Wang and Zhang 2006), which may increase the probability of hitting global maxima. (See Ueda and Nakano 1998; Elidan, Friedman, and Schuurmans 2002; Karciauskas et al. 2004 for more sophisticated strategies for the choice of initial values.) In this experiment, the initial values of the EM algorithm were $\boldsymbol{\Phi}_0 =$ identity matrix, $\mathbf{W}_0 = .3 \times \mathbf{1}, \boldsymbol{\mu}_0 =$ mean of observed $(\mathbf{x}, \mathbf{y})$, and $\nu_0 = 5$. Figure 3 includes the resulting principal directions. It clearly shows that in the absence of outliers, RPMC was able to recover the same principal direction as SPPCA.

In our second experiment, we added two outliers of $n_1 = [5, 3]^T$ and $n_2 = [10, 8]^T$ (indicated by circles in Fig. 4) into the data set. Figure 4 shows the principal directions found by SPPCA and RPMC. The two outliers had a significant impact on SPPCA. The principal direction found by SPPCA with outliers (black) was substantially skewed from the one found (gray) in the absence of outliers. In contrast, RPMC found approximately the same subspace as SPPCA found in the absence of outliers.

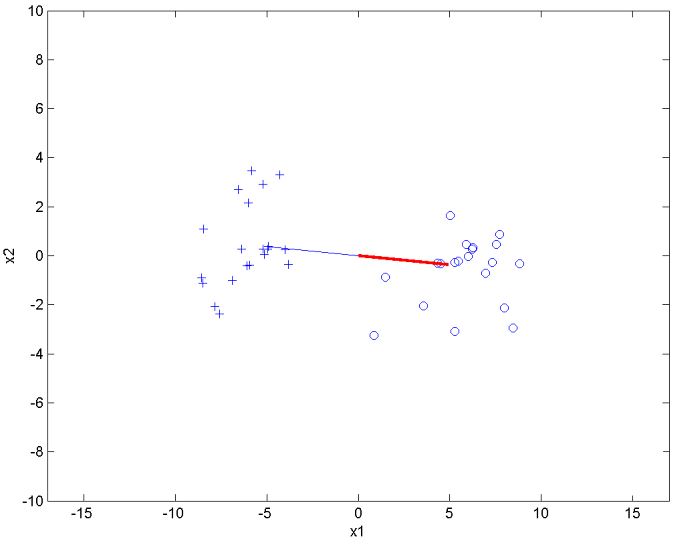

Figure 3. The first principal direction found in the absence of outliers by SPPCA and the one found by RPMC (+ Class 1; ○ Class 2; — SPPCA; — RPMC). The data are drawn from two different multivariate two-dimensional Gaussian distributions.

4.2 Case Study 1: Fish Data

Here we illustrate the RPMC approach on a low-dimensional example introduced by Naes (1985). This data set includes 45 observations of fish. The input variables consist of highly

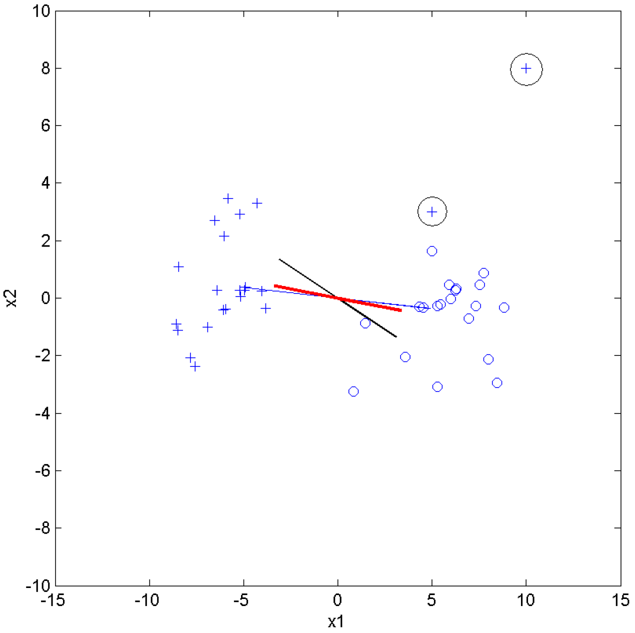

Figure 4. The first principal direction found by SPPCA and that found by RPMC in presence of two outliers, indicated by circles (+ Class 1; ○ Class 2; — SPPCA without outliers; — SPPCA with outliers; — RPMC with outliers). The data are drawn from two different multivariate two-dimensional Gaussian distributions.

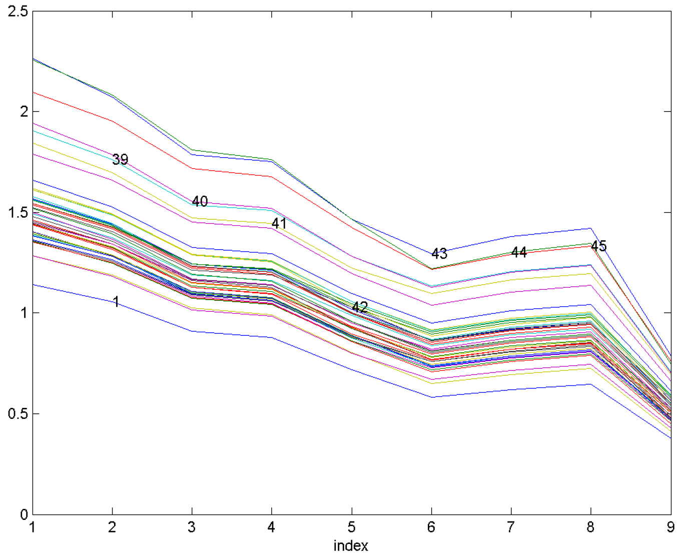

Figure 5. Forty-five input observations of the fish data set.

Table 1. MSEs for the fish validation data set with various P

P	Model	Calibration	Validation	Model	Calibration	Validation
3	PLS	1.9440	1.6293	SPPCA	2.3523	2.3436
	RPLS	3.0192	**.5055**	RPMC	2.2723	1.3543
4	PLS	1.6936	1.4706	SPPCA	1.7015	1.5373
	RPLS	2.5849	**.4604**	RPMC	2.1041	.8749
5	PLS	1.6617	1.3720	SPPCA	1.6841	1.4162
	RPLS	2.1337	**.6185**	RPMC	1.9586	.7270
6	PLS	1.3732	1.3763	SPPCA	1.6738	1.3905
	RPLS	2.5625	**.4997**	RPMC	1.9910	.7096
7	PLS	1.3459	1.3463	SPPCA	1.6622	1.3802
	RPLS	1.8145	.8092	RPMC	1.7540	**.7298**
8	PLS	1.3352	1.2085	SPPCA	1.5874	1.3340
	RPLS	9.3236	.9815	RPMC	1.6850	**.7701**

NOTE: The minimum error at each P is highlighted.

correlated spectra at nine wavelengths. The single output variable is the fat concentration of the fish. The goal of the analysis is to identify the relationship between the spectra and the fat concentration. The input variables are shown in Figure 5, in which observations 1 and 39–45 are highlighted. Naes (1985) reported that observations 39–45 are outliers. By observation, we can determine that each of these except observation 42 has a spectrum that apparently deviates from the majority.

The total data set was divided into two parts: a calibration set and a validation set. The calibration set was used to train the predictive models of PLS, RPLS, SPPCA, and RPMC, and the validation set was used to test the predictive performance of these models. Because our interest lies in determining the impact of outliers on model development, we included the outliers in the calibration set; specifically, observations 1–15 and 31–45 were taken for the calibration set, and the others were used for validation purposes. We then used between three and eight components of P in conjunction with PLS, RPLS, SPPCA, and RPMC. Table 1 contains the mean squared errors (MSEs) for the calibration and validation sets. The best prediction at each P is highlighted. Among the six total cases, RPLS generated four best predictions, and RPMC obtained two best results. The overall best prediction on the validation set was achieved by RPLS at $P = 4$. An interesting observation is that RPLS was the best when P was relatively small, and RPMC outperformed RPLS as P increased. This indicates that RPMC may be particularly suitable to high-dimensional data, where a large P is usually required. It also is noteworthy that in many cases the test error is even lower than the training error. This can be explained by the presence of outliers in the training data set and their absence in the test data set.

To identify the outliers, we applied the method proposed in Section 3.4. The data $(\mathbf{x}, \mathbf{y})$ were assumed to be generated by a latent variable model. Any observation with the Mahalanobis distance (chi-squared statistic) larger than the cutoff value $\chi^2_{P,.95}$ was considered an outlier. In this experiment, all of the observations were included in model development. Figure 6(a) shows the chi-squared statistic for each observation, as well as the 95% quantile threshold obtained when the SPPCA model was assumed; in contrast, Figure 6(b) shows the results obtained by the RPMC model. The plots clearly show that RPMC identified observations 1 and 39–45 as outliers and that SPPCA produced too many misclassifications. Hubert and Branden (2003) reported that the RPLS approach did not detect observation 42 as an outlier and also misclassified observation 12 as an outlier. To the best of our knowledge, the performance of RPMC in detecting outliers in this data set is the most accurate of any technique used to date.

4.3 Case Study 2: Biscuit Dough Data

In this experiment we applied RPMC to the well-known high-dimensional biscuit dough data set with multiple responses (Osborne, Fearn, Miller, and Douglas 1984). The study aims to predict the constituents of biscuit dough based on the spectral characteristics of the dough as measured using near-infrared (NIR) spectroscopy. The preprocessing step suggested by Hubert, Rousseeuw, and Verboven (2002) resulted in a data set of NIR spectra in 600 dimensions (input). Figures 7(a) and 7(b) show the spectra signals of the input and the output. The data set originally contained four output variables—the concentrations of fat, flour, sucrose, and water—contained in 40 observations of the biscuit dough. As suggested by Hubert et al. (2002), we removed the output concentration of fat, because it had a higher variance and was not highly correlated with the other output variables. Although observation 23 is a known outlier, we still included it in the model development.

We used observations 1–35 as a calibration set and the others as a validation set. We then performed PLS, RPLS, SPPCA, and RPMC on the data with $P = 3$, 4, and 5. (As suggested in Hubert and Branden 2003, $P = 3$ is sufficient for PLS.) Table 2 summarizes the MSE results. Of the nine prediction tasks, RPMC and RPLS performed the best in four cases. Overall, RPMC had the minimum MSE on two responses (flour and water) and PLS had the minimum MSE on one response (sucrose). These results indicate that even when the data have few outliers, RPMC remains superior to traditional methods. As a wider family of Gaussian distributions, t-distributions seem more flexible in adjusting themselves to real-life data that do not exactly fit the Gaussian assumptions in most cases.

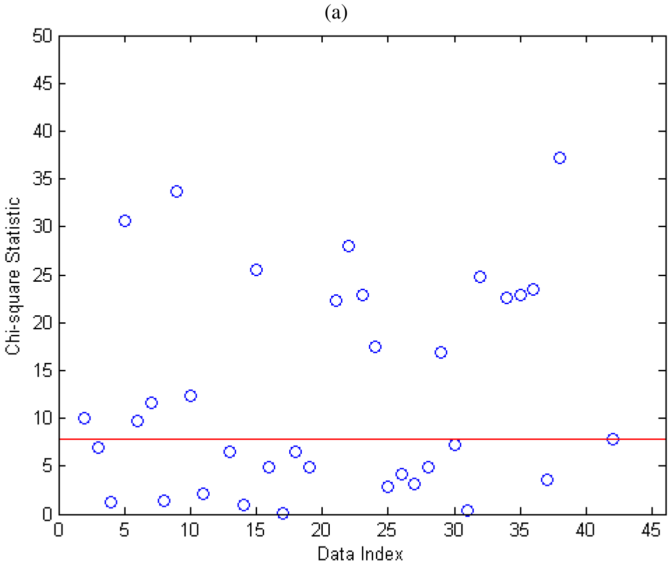

Figure 6. Outlier detection results for the fish data set obtained by monitoring the latent variable from (a) SPPCA and (b) RPMC. The circles represent the chi-squared statistic of the observations; the line indicates the threshold of the 95% quantile of the chi-squared distribution with 3 degrees of freedom.

We applied the proposed outlier detection procedure to the data; Figure 8 graphs the results. In Figure 8(b), which shows the results obtained with RPMC, observation 23 stands out as a clear outlier, with a large chi-squared statistic (>20). This result is consistent with our previous knowledge. In contrast, SPPCA produced a false alarm on observation 7, as illustrated in Figure 8(a). We also applied RPMC separately to each output dimension (i.e., flour, sucrose, and water); Figure 9 gives the outlier detection plots. We can see that some false alarms were generated based on the 95% chi-squared statistic, although observation 23 still was identified as an outlier. Therefore, the true outlier is likely to be associated with the three outputs as a whole, not with any single response. This result also illustrates the strength of RPMC, which has a principled way of dealing with multiple response variables.

5. CONCLUSIONS

In this article we report a RPMC model constructed based on Student t-distributions. RPMC aims to use a set of latent variables to identify the relationship between input and output. We proved that RPMC is a wider family of latent variable models that encompasses PPCA and SPPCA as limiting cases. The t-distribution diminishes the impact of outliers on model development. Consequently, RPMC produces parameter estimators that can be used with greater confidence. Using these more accurate estimators, we presented an outlier detection approach based on a chi-squared statistic. We also derived a tractable EM algorithm for parameter estimation in RPMC. The approach works well on artificial data as well as on well-known public data sets. Another potential advantage of RPMC is its ability to

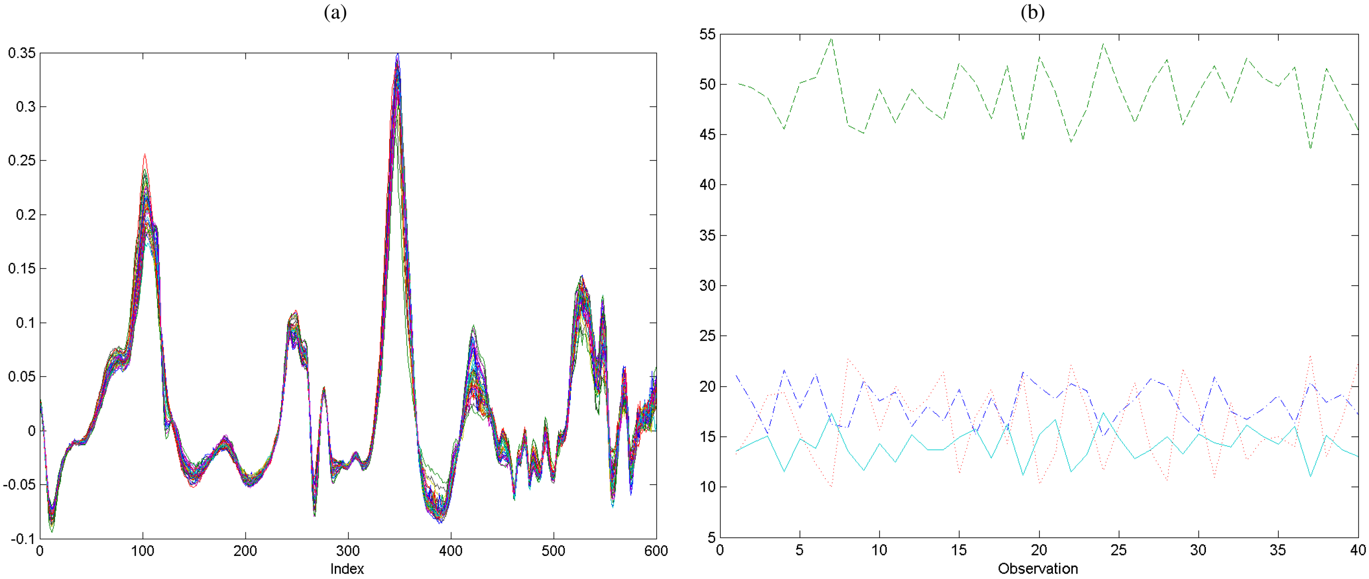

Figure 7. The input (a) and output (b) observations of the biscuit dough data set (-·-, fat; ---, flour; ·····, sucrose; —, water).

Table 2. MSEs for the biscuit dough validation data set

		Calibration			Validation		
P		Flour	Sucrose	Water	Flour	Sucrose	Water
3	PLS	1.2949	2.2783	.2365	.2425	.4565	.0474
	SPPCA	1.5160	3.4759	.2372	.2543	.5636	.0419
	RPLS	1.6442	2.5534	.3295	.1982	.4416	**.0216**
	RPMC	1.6138	2.7112	.4347	**.1941**	**.4208**	.0243
4	PLS	1.1060	1.7750	.2142	.2093	**.3779**	.0413
	SPPCA	1.3320	2.3054	.2373	.2443	.4949	.0425
	RPLS	1.7406	2.7325	.3244	**.1998**	.4997	**.0200**
	RPMC	1.9319	3.9956	.3968	.2072	.4832	.0208
5	PLS	.4264	.6509	.1573	.3964	.6360	.0508
	SPPCA	1.3058	2.2923	.2355	.4170	.6501	.0507
	RPLS	1.6261	2.5038	.2809	.2140	**.4817**	.0278
	RPMC	2.3670	3.4659	.5019	**.1463**	.4935	**.0073**

NOTE: The minimum error at each *P* is highlighted.

interpret data. As a class of latent variable models, factor analysis has been heavily used in psychometrics to identify underlying causes or factors by explaining latent variables. As we have shown, RPMC is also a latent variable model, and in some applications its latent variables may involve domain-specific implications.

Our future work includes building nonlinear versions of RPMC. A possible approach to this is to use the kernel trick by constructing a dual form of RPMC (Scholkopf and Smola 2002; Fukumizu, Bach, and Jordan 2004). Another approach might be mixture models of RPMC. The probabilistic formulation of RPMC very likely will lead to a well-defined mixture model whose parameters can be determined using an EM algorithm, as in the formulation of PPCA mixture models (Tipping and Bishop 1999a).

ACKNOWLEDGMENTS

This research was supported by National Science Foundation grant CMMI-0644830. The authors thank Dr. Shipeng Yu and Dr. Cédric Archambeau for sharing their codes, Dr. Mia Hubert for providing the fish data and the biscuit dough data, and the anonymous reviewers for their constructive comments.

APPENDIX A: PROOFS

Proof of Theorem 1

Here we sketch the proof only of the marginal distribution over $\mathbf{t}$, because the marginal distributions of $\mathbf{x}$ and $\mathbf{y}$ (conditional on $\mathbf{t}$) can be derived in a similar fashion. Write

$$J_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \equiv (\mathbf{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu}).$$

By noting a useful definition of the gamma function (Liu and Rubin 1995; Khan and Dellaert 2004),

$$\int_0^\infty e^{-(\alpha\theta)^\beta}\theta^\tau\,d\theta = \Gamma\left(\frac{\tau+1}{\beta}\right)\Big/(\beta\alpha^{\tau+1}),$$

we can obtain that

$$p(\mathbf{t}) = \int_0^\infty p(\mathbf{t}|\theta)p(\theta)\,d\theta$$

$$\propto \int_0^\infty \theta^{P/2}e^{-(\theta/2)J_{\mathbf{0},\mathbf{I}_P}}\theta^{\nu/2-1}e^{-\nu\theta/2}\,d\theta$$

$$\propto \int_0^\infty \theta^{(\nu+P)/2-1}e^{-\theta/2(J_{\mathbf{0},\mathbf{I}_P}+\nu)}\,d\theta$$

$$\propto \left(\frac{J_{\mathbf{0},\mathbf{I}_P}}{\nu}+1\right)^{-(\nu+P)/2},$$

which has the form of a Student *t*-distribution with ν degrees of freedom, with the normalized term

$$(\nu\pi)^{-P/2}\Gamma\left(\frac{\nu+P}{2}\right)\Big/\Gamma\left(\frac{\nu}{2}\right).$$

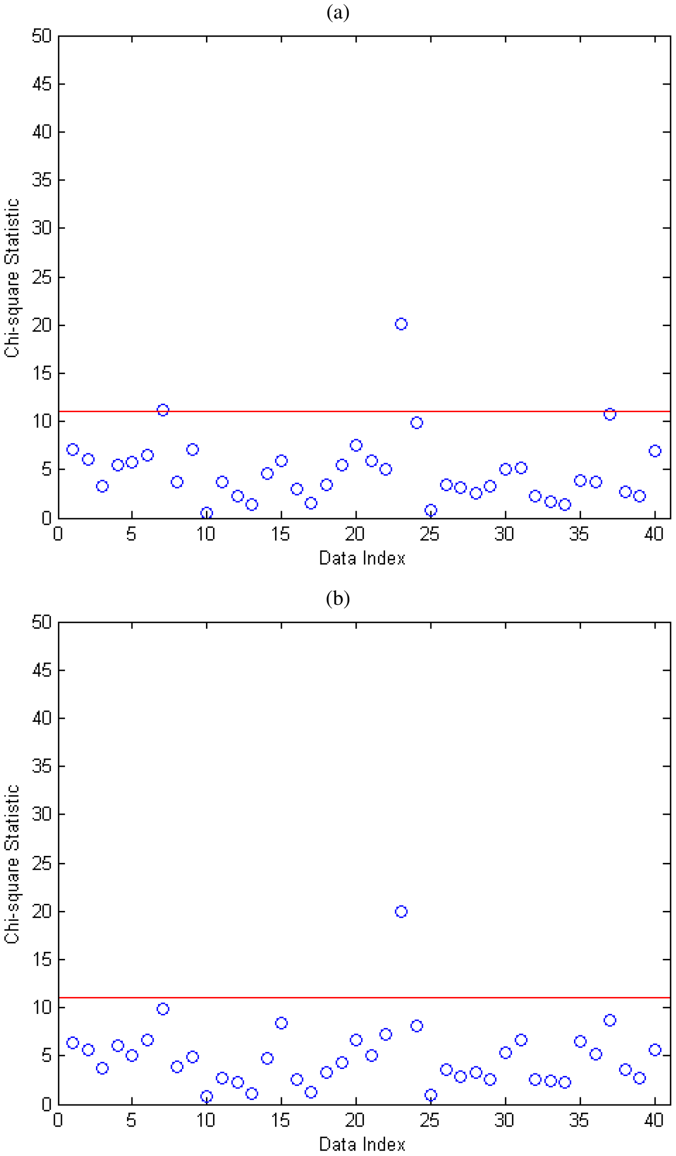

Figure 8. Outlier detection results from monitoring the latent variable from SPPCA (a) and RPMC (b) for the biscuit dough data set with three outputs. The circles represent the chi-squared statistic of the observations; the line indicates the threshold of the 95% quantile of the chi-squared distribution with 5 degrees of freedom.

Proof of Proposition 1

By Theorem 1, we obtain the equivalent model (6) of RPMC. The Student *t*-distribution approaches the Gaussian distribution

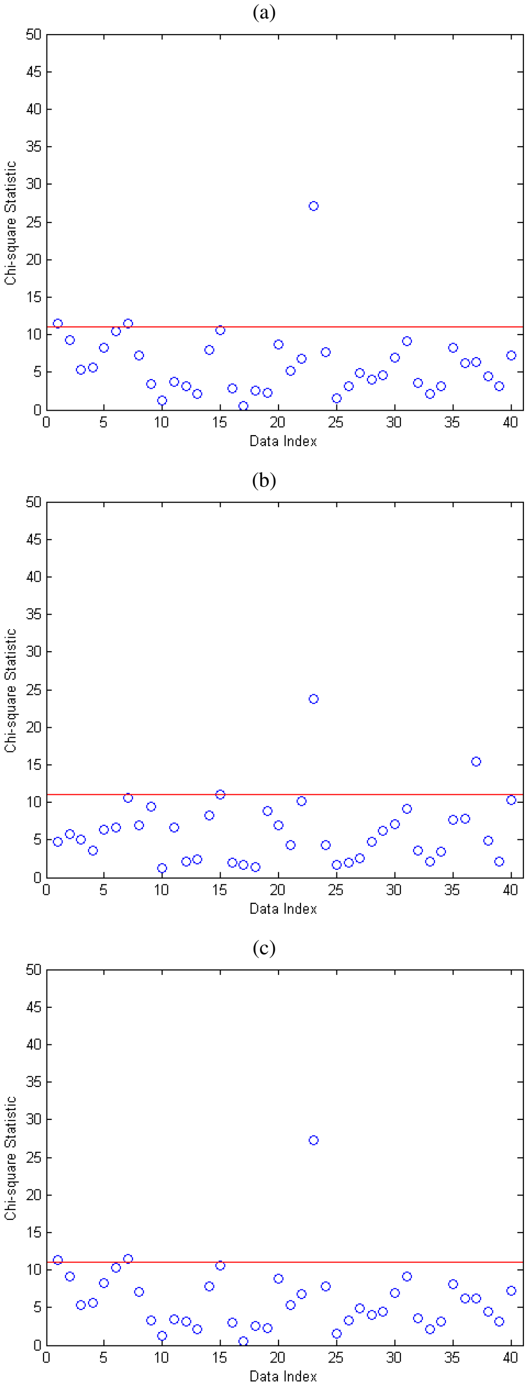

Figure 9. Outlier detection results by monitoring the latent variable from RPMC for the biscuit dough data set with single output flour (a), sucrose (b), and water (c). The circles represent the chi-squared statistic of the observations; the line indicates the threshold of the 95% quantile of the chi-squared distribution with 5 degrees of freedom.

as $\nu \to \infty$. It follows that (6) degrades to the SPPCA model as defined in (3).

Proof of Proposition 2

According to Proposition 1, the RPMC model degrades to SPPCA as $\nu \to \infty$. In addition, if $K = 0$, then the SPPCA model is unsupervised and then degrades to PPCA (Yu et al. 2006). It follows that RPMC is equivalent to PPCA as defined in (1) when the two conditions are satisfied.

APPENDIX B: EM ALGORITHM FOR RPMC

a. When both input and output $\mathbf{z}_i = (\mathbf{x}_i\ \mathbf{y}_i)$ are observed, the posterior distribution of the latent variable θ_i can be obtained as

$$\begin{aligned}p(\theta_i|\mathbf{z}_i) &\propto p(\mathbf{z}_i|\theta_i)p(\theta_i)\\ &= G\left(\theta_i \middle| \frac{M+K+\nu}{2},\right.\\ &\qquad\left.\frac{(\mathbf{z}_i-\boldsymbol{\mu})^T(\mathbf{W}\mathbf{W}^T+\boldsymbol{\Phi})^{-1}(\mathbf{z}_i-\boldsymbol{\mu})+\nu}{2}\right).\end{aligned}$$

This can be readily obtained using the fact that the gamma distribution is conjugate to the exponential family. It follows formula (8) by noting that for the gamma distribution,

$$G(\theta|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \qquad \text{and} \qquad \langle\theta\rangle = \frac{\alpha}{\beta}.$$

b. By Bayes's theorem, the posterior distribution of the latent vector $\mathbf{t}_i$ is given by

$$\begin{aligned}p(\mathbf{t}_i|\mathbf{z}_i,\theta_i) &\propto p(\mathbf{z}_i|\mathbf{t}_i,\theta_i)p(\mathbf{t}_i|\theta_i)\\ &= N\big(\mathbf{t}_i|(\mathbf{W}^T\boldsymbol{\Phi}^{-1}\mathbf{W}+\mathbf{I}_P)^{-1}\mathbf{W}^T\boldsymbol{\Phi}^{-1}(\mathbf{z}_i-\boldsymbol{\mu}),\\ &\qquad \theta_i^{-1}(\mathbf{W}^T\boldsymbol{\Phi}^{-1}\mathbf{W}+\mathbf{I}_P)\big).\end{aligned}$$

This readily follows formula (9).

c. Setting $\frac{\partial\Upsilon(\Omega|\Omega^j)}{\partial\boldsymbol{\mu}} = 0$, we readily obtain formula (10).

d. Given the formulas of the trace derivatives (Golub and Van Loan 1996; Khan and Dellaert 2004),

$$\frac{\partial\,\mathrm{tr}(\mathbf{X}^T\mathbf{C}\mathbf{X}\mathbf{D})}{\partial\mathbf{X}} = 2\mathbf{C}\mathbf{X}\mathbf{D},$$

we can obtain

$$\begin{aligned}\frac{\partial\Upsilon(\Omega|\Omega^j)}{\partial\mathbf{W}} &= \sum_{i=1}^{n}\left[\frac{\partial}{\partial\mathbf{W}}\left\langle\theta_i\mathbf{t}_i^T\mathbf{W}^T\boldsymbol{\Phi}^{-1}(\mathbf{z}_i-\boldsymbol{\mu})\right.\right.\\ &\qquad\left.\left.-\frac{1}{2}\theta_i\,\mathrm{tr}(\mathbf{W}^T\boldsymbol{\Phi}^{-1}\mathbf{W}\mathbf{t}_i\mathbf{t}_i^T)\right\rangle\right]\\ &= \sum_{i=1}^{n}\left[\boldsymbol{\Phi}^{-1}(\mathbf{z}_i-\boldsymbol{\mu})\langle\theta_i\mathbf{t}_i\rangle^T - \boldsymbol{\Phi}^{-1}\mathbf{W}\langle\theta_i\mathbf{t}_i\mathbf{t}_i^T\rangle\right].\end{aligned}$$

By setting $\frac{\partial\Upsilon(\Omega|\Omega^j)}{\partial\mathbf{W}} = 0$, it follows formula (11).

e. Using the formula

$$\begin{aligned}&(\mathbf{z}_i-\mathbf{W}\mathbf{t}_i-\boldsymbol{\mu})^T\boldsymbol{\Phi}(\mathbf{z}_i-\mathbf{W}\mathbf{t}_i-\boldsymbol{\mu})\\ &\quad = \mathrm{tr}\big(\boldsymbol{\Phi}^{-1}(\mathbf{z}_i-\boldsymbol{\mu})(\mathbf{z}_i-\boldsymbol{\mu})^T\big) + \mathrm{tr}(\mathbf{W}^T\boldsymbol{\Phi}^{-1}\mathbf{W}\mathbf{t}_i\mathbf{t}_i^T)\\ &\qquad - 2\mathbf{t}_i^T\mathbf{W}^T\boldsymbol{\Phi}^{-1}(\mathbf{z}_i-\boldsymbol{\mu})\end{aligned}$$

and the formulas of the trace derivatives

$$\frac{\partial\,\mathrm{tr}(\mathbf{X}^{-1}\mathbf{C})}{\partial\mathbf{X}} = -\mathbf{X}^{-1}\mathbf{C}^T\mathbf{X}^{-1} \qquad \text{and}$$

$$\frac{\partial\,\mathrm{tr}(\mathbf{C}^T\mathbf{X}^{-1}\mathbf{D})}{\partial\mathbf{X}} = -\mathbf{X}^{-1}\mathbf{C}\mathbf{D}^T\mathbf{X}^{-1},$$

we have

$$\begin{aligned}
&\frac{\partial \Upsilon(\Omega|\Omega^j)}{\partial \boldsymbol{\Phi}} \\
&= \sum_{i=1}^{n} \biggl[\frac{\partial}{\partial \boldsymbol{\Phi}} \biggl\langle -\frac{1}{2}\log|\boldsymbol{\Phi}| - \frac{1}{2}\theta_i \mathbf{tr}\bigl(\boldsymbol{\Phi}^{-1}(\mathbf{z}_i - \boldsymbol{\mu})(\mathbf{z}_i - \boldsymbol{\mu})^T\bigr) \\
&\qquad + \theta_i \mathbf{t}_i^T \mathbf{W}^T \boldsymbol{\Phi}^{-1}(\mathbf{z}_i - \boldsymbol{\mu}) - \frac{1}{2}\operatorname{tr}(\mathbf{W}^T \boldsymbol{\Phi}^{-1} \mathbf{W} \theta_i \mathbf{t}_i \mathbf{t}_i^T) \biggr\rangle \biggr] \\
&= \sum_{i=1}^{n} \bigl[-\boldsymbol{\Phi}^{-1} + \langle \theta_i \rangle \boldsymbol{\Phi}^{-1}(\mathbf{z}_i - \boldsymbol{\mu})(\mathbf{z}_i - \boldsymbol{\mu})^T \boldsymbol{\Phi}^{-1} \\
&\qquad - 2\boldsymbol{\Phi}^{-1}\mathbf{W}\langle \theta_i \mathbf{t}_i \rangle (\mathbf{z}_i - \boldsymbol{\mu})^T \boldsymbol{\Phi}^{-1} \\
&\qquad + \boldsymbol{\Phi}^{-1}\mathbf{W}\langle \theta_i \mathbf{t}_i \mathbf{t}_i^T \rangle \mathbf{W}^T \boldsymbol{\Phi}^{-1} \bigr].
\end{aligned}$$

Simplifying this and setting $\frac{\partial \Upsilon(\Omega|\Omega^j)}{\partial \boldsymbol{\Phi}} = 0$, we can obtain formula (12).

f.

$$\begin{aligned}
\nu^{j+1} = \arg\max_{\nu} \biggl[&\frac{n\nu}{2}\log\frac{\nu}{2} + \left(\frac{\nu}{2} - 1\right)\sum_{i=1}^{n}\langle \log\theta_i \rangle \\
&- \frac{\nu}{2}\sum_{i=1}^{n}\langle \theta_i \rangle - n\log\Gamma\left(\frac{\nu}{2}\right) \biggr].
\end{aligned}$$

By setting $\frac{\partial \Upsilon(\Omega|\Omega^j)}{\partial \nu} = 0$, we need only solve the one-dimensional nonlinear equation (13).

[Received February 2007. Revised June 2007.]

REFERENCES

Akaike, H. (1974), "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Archambeau, C. (2005), "Probabilistic Models in Noisy Environments and Their Application to a Visual Prosthesis for the Blind," doctoral dissertation, Catholic University of Louvain, Belgium.

Archambeau, C., Delannay, N., and Verleysen, M. (2006), "Robust Probabilistic Projections," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 33–40.

Bach, F. R., and Jordan, M. I. (2005), "A Probabilistic Interpretation of Canonical Correlation Analysis," Technical Report 688, University of California, Berkeley, Dept. of Statistics.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by Supervised Principal Components," *Journal of the American Statistical Association*, 101, 119–137.

Bellman, R. (1961), *Adaptive Control Processes*, Princeton, NJ: Princeton University Press.

Bozdogan, H. (1988), "ICOMP: A New Model-Selection Criterion," in *Proceedings of the Conference of the International Federation of Classification Societies*, pp. 599–608.

Chatterjee, S., and Price, B. (1977), *Regression Analysis by Examples*, New York: Wiley.

Dasgupta, S., Hsu, D. J., and Verma, N. (2006), "A Concentration Theorem for Projections," in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*.

Davies, P. (1987), "Asymptotic Behavior of S–Estimators of Multivariate Location and Dispersion Matrices," *The Annals of Statistics*, 15, 1269–1292.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Donoho, D. (1982), "Breakdown Properties of Multivariate Location Estimators," Ph.D. qualifying paper, Harvard University, Dept. of Statistics.

Elidan, M. N., Friedman, N., and Schuurmans, D. (2002), "Data Perturbation for Escaping Local Maxima in Learning," in *Proceedings of Eighteenth National Conference on Artificial Intelligence*, pp. 132–139.

Everitt, B. S. (1984), *An Introduction to Latent Variable Models*, London: Chapman & Hall.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004), "Dimensionality Reduction for Supervised Learning With Reproducing Kernel Hilbert Spaces," *Journal of Machine Learning Research*, 5, 73–99.

Ghoshdastider, B., and Schafer, J. L. (2003), "Outlier Detection and Editing Procedures for Continuous Multivariate Data," Technical Report 0307, Princeton University, Office of Population Research.

Golub, G. H., and Van Loan, C. F. (1996), *Matrix Computations*, Baltimore: John Hopkins University Press.

Hotelling, H. (1936), "Relations Between Two Sets of Variables," *Biometrika*, 28, 321–377.

Huber, P. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73–101.

Hubert, M., and Branden, V. (2003), "Robust Methods for Partial Least Squares Regression," *Journal of Chemometrics*, 17, 537–549.

Hubert, M., Rousseeuw, P., and Verboven, S. (2002), "A Fast Method for Robust Principal Components With Applications to Chemometrics," *Chemometrics and Intelligent Laboratory Systems*, 60, 101–111.

Karciauskas, G., Kocka, T., Jensen, F. V., Larranaga, P., and Lozano, J. A. (2004), "Learning of Latent Class Models by Splitting and Merging Components," in *Proceedings of Second Workshop on Probabilistic Graphical Models, Leiden*, pp. 137–144.

Khan, Z., and Dellaert, F. (2004), "Robust Generative Subspace Modeling: The Subspace *t* Distribution," Technical Report GIT-GVU-04-11, Georgia Institute of Technology, GVU Center.

Lange, K., Little, R. J. A., and Taylor, J. M. G. (1989), "Robust Statistical Modeling Using the *t* Distribution," *Journal of the American Statistical Association*, 84, 881–896.

Liu, C., and Rubin, D. B. (1995), "ML Estimation of the *t* Distribution Using EM and Its Extensions, ECM and ECME," *Statistica Sinica*, 5, 19–39.

Lopuhaa, H. (1989), "On the Relation Between S–Estimators and M–Estimators of Multivariate Location and Covariance," *The Annals of Statistics*, 17, 1662–1683.

Moller, S., Frese, J., and Bro, R. (2005), "Robust Methods for Multivariate Data Analysis," *Journal of Chemometrics*, 19, 549–563.

Naes, T. (1985), "Multivariate Calibration When the Error Covariance Matrix Is Structured," *Technometrics*, 27, 301–311.

Osborne, B., Fearn, T., Miller, A., and Douglas, S. (1984), "Application of Near-Infrared Reflectance Spectroscopy to the Compositional Analysis of Biscuits and Biscuit Dough," *Journal of the Science of Food and Agriculture*, 35, 99–105.

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, 2, 559–572.

Peel, D., and McLachlan, G. J. (2000), "Robust Mixture Modeling Using the *t*-Distribution," *Statistics and Computing*, 10, 339–348.

Rousseeuw, P. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P., and Van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–651.

Roweis, S., and Ghahramani, Z. (1999), "A Unifying Review of Linear Gaussian Models," *Neural Computation*, 11, 305–345.

Scholkopf, B., and Smola, A. J. (2002), *Learning With Kernels*, Cambridge, MA: MIT Press.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Spearman, C. (1904), "Proof and Measurement of Association Between Two Things," *American Journal of Psychology*, 15, 72–101.

Stahel, W. (1981), "Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators," doctoral dissertation, ETH, Zurich.

Svensen, M., and Bishop, C. M. (2004), "Robust Bayesian Mixture Modeling," *Neurocomputing*, 64, 235–252.

Thomas, E., and Ge, N. (2000), "Development of Robust Multivariate Calibration Models," *Technometrics*, 42, 168–177.

Tipping, M. E., and Bishop, C. M. (1999a), "Mixtures of Probabilistic Principal Component Analyzers," *Neural Computation*, 11, 443–482.

——— (1999b), "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society*, Ser. B, 61, 611–622.

Torre, F., and Black, M. J. (2001), "Robust Principal Component Analysis for Computer Vision," in *Proceedings of the International Conference on Computer Vision*, pp. 362–369.

Tukey, J. W. (1962), "The Future of Data Analysis," *Annals of Mathematical Statistics*, 33, 1–67.

Ueda, N., and Nakano, R. (1998), "Deterministic Annealing EM Algorithm," *Neural Networks*, 11, 271–282.

Wang, Y., and Zhang, N. L. (2006), "Severity of Local Maxima for the EM Algorithm: Experiences With Hierarchical Latent Class Models," in *Proceedings of the Third European Workshop on Probabilistic Graphical Model*, pp. 301–308.

Wold, H. (1966), “Estimation of Principal Components and Related Models by Iterative Least Squares,” in *Multivariate Analysis*, ed. P. R. Krishnaiah, New York: Academic Press, pp. 391–420.

Wold, S., Berglund, A., and Kettaneh, N. (2002), “New and Old Trend in Chemometrics: How to Deal With the Increasing Data Volumes in R&D&P. With Examples From Pharmaceutical Research and Process Modeling,” *Journal of Chemometrics*, 16, 377–386.

Yu, S., Yu, K., Tresp, V., Kriegel, P., and Wu, M. (2006), “Supervised Probabilistic Principal Component Analysis,” in *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press New York, pp. 464–473.