Math. Program., Ser. A 96: 1–31 (2003)

Digital Object Identifier (DOI) 10.1007/s10107-003-0242-8

Tim Olson · Jong-Shi Pang* · Carey Priebe**

A likelihood-MPEC approach to target classification***

Received: October 26, 1998 / Accepted: June 11, 2001
Published online March 24, 2003 –

Abstract. In this paper we develop a method for classifying an unknown data vector as belonging to one of several classes. This method is based on the statistical methods of maximum likehood and borrowed strength estimation. We develop an MPEC procedure (for Mathematical Program with Equilibrium Constraints) for the classification of a multi-dimensional observation, using a finite set of observed training data as the inputs to a bilevel optimization problem. We present a penalty interior point method for solving the resulting MPEC and report numerical results for a multispectral minefield classification application. Related approaches based on conventional maximum likehood estimation and a bivariate normal mixture model, as well as alternative surrogate classification objective functions, are described.

1. Introduction

Classification is the problem of determining which one of several classes an unlabelled multi-dimensional observation belongs to. These problems occur in many different settings. Examples include the classification of white versus gray matter in the brain, the classification of masses in breast imaging as malignant or benign, and the classification of remotely sensed objects as mines or not mines.

One can develop classification rules based on a priori knowledge of the objects, or by utilizing a set of observed training data from the object classes. These different methodologies are generally referred to as model based and data-driven, respectively. Classification is an important subject in the vast area of data mining that has recently been evolving rapidly due to its importance in many diverse fields. In an excellent survey paper [2], Bradley, Fayyad, and Mangasarian have detailed the opportunities that optimizers can contribute to this fast growing topic of research.

T. Olson: Department of Mathematics, University of Florida, Gainesville, Florida 32611-8105, USA
e-mail: olson@math.ufl.edu

J.-S. Pang, C. Priebe: Department of Mathematical Sciences, Whiting School of Engineering, The Johns Hopkins University, Baltimore, Maryland 21218-2682, USA
e-mail: pang@mts.jhu.edu, cep@jhu.edu

Mathematics Subject Classification (2000): 62G07, 62P30, 90C26, 90C30, 90C33, 90C90

* The work of this author was based on research supported by the U.S. National Science Foundation under grant CCR-9624018.

** The work of this author was supported by the Office of Naval Research under grant N00014-95-1-0777.

*** The authors of this work were all partially supported by the Wright Patterson Air Force Base via Veda Contract F33615-94-D-1400. The first and third author were also supported by the National Science Foundation under grant DMS-9705220.

Along with its accompanying paper [10], this paper presents a novel methodology for target classification that employs both statistics and optimization methods. While the reference emphasizes the statistical aspects and practical application of the methodology, this paper, being addressed to the optimization community, emphasizes the constrained optimization problem and its solution. The optimization problem belongs to the class of mathematical programs with equilibrium constraints (MPECs) that has recently received a comprehensive treatment in [6]. As detailed in this reference, standard theory and methods from classical nonlinear programming are not appropriate to deal with this class of optimization problems; instead, we have employed the penalty interior point algorithm (PIPA) as the solution method for solving the MPEC arising from the target classification problem.

2. The statistical problem

In this section, we motivate the formulation of the classification problem as a constrained bi-level optimization problem, or MPEC. Before giving this motivation, we explain the notation used throughout the paper. The N-dimensional Euclidean space is denoted $\Re^N$. All vectors are column vectors; a superscript T denotes transposition of vectors and matrices. For a function $\phi(\boldsymbol{x}, \boldsymbol{y})$ of two arguments $(\boldsymbol{x}, \boldsymbol{y}) \in \Re^{N+M}$ and a subset U of $\Re^N$,

$$\operatorname{argmax} \{ \phi(\boldsymbol{x}, \boldsymbol{y}) \,:\, \boldsymbol{x} \in U \} \subset \Re^N$$

denotes the set of maximizers of the constrained maximization problem in the $\boldsymbol{x}$ variable, with $\boldsymbol{y}$ held fixed; thus, the above argmax set is dependent on the parameter $\boldsymbol{y}$. We also write $\nabla_{\boldsymbol{x}} \phi(\boldsymbol{x}, \boldsymbol{y}) \in \Re^N$ to denote the partial gradient vector of $\phi(\boldsymbol{x}, \boldsymbol{y})$ with respect to the $\boldsymbol{x}$ variable. For two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ of the same dimension, we write $\boldsymbol{a} \circ \boldsymbol{b}$ to denote the Hadamard product of $\boldsymbol{a}$ and $\boldsymbol{b}$; that is, $\boldsymbol{a} \circ \boldsymbol{b}$ is the vector whose components are equal to the products of the respective components of $\boldsymbol{a}$ and $\boldsymbol{b}$; moreover, we write $\boldsymbol{a} \perp \boldsymbol{b}$ to mean that $\boldsymbol{a}$ and $\boldsymbol{b}$ are perpendicular. For a vector $\boldsymbol{x} \in \Re^N$, we write $\operatorname{diag}(\boldsymbol{x})$ for the $N \times N$ diagonal matrix whose diagonal entries are the components of $\boldsymbol{x}$. The N-vector of all ones is denoted by $\mathbf{1}_N$; the identity matrix of order N is denoted by $\boldsymbol{I}_N$. Finally, for a vector $\boldsymbol{x}$ with nonzero components, let $\boldsymbol{x}^{-1}$ be the vector whose components are the reciprocals of the respective components of $\boldsymbol{x}$.

2.1. Background motivation

The most common performance metric for a classification system is the probability of correct classification (PCC). This performance metric is generally used only post-facto, i.e. after the construction of the classification rule in order to judge the success or failure of the algorithm. Our method, in contrast, will focus on the explicit optimization of this metric as the basis for constructing classification algorithms.

Many techniques exist to do some form of optimization in order to improve the performance of a classification system. An excellent overview of many of the methods for classification, or pattern recognition, can be found in [3]. Basic discriminant analysis

consists of fitting a model to the data, and utilizing this model to predict the classes of unknown data. Generally this is formulated in a least squares setting and therefore is seen as an optimization problem. Some types of neural networks are examples of such methods. These implement a particular model, and the algorithms are fit ("trained," in the neural network literature) to optimally differentiate between classes.

A recent method which has shown great promise is based on wavelets [12]. This method, like many methods which utilize optimization to improve classification rules, does not directly optimize the PCC performance of the system, but rather a secondary metric which is believed to be connected with the performance of the classification rule. Typical secondary cost functions are energy or entropy metrics. Our method, on the other hand, will be directly tied to optimizing the conditional PCC.

Our aim is to design a mechanism which automatically assigns an observation x to one of two classes, denoted by $i = I, II$. The classes are modelled as two unknown probability distribution F_I, F_{II} over the Euclidean space of possible observations x and the classification problem amounts to deciding from which of the two distributions an observation was drawn. Whilst misclassification cannot be completely ruled out, a good classification mechanism should have a small misclassification probability. We will focus in the sequel on classification mechanisms based on aggregate values $f(x)$. Any real-valued function f induces two distributions on the real line, corresponding to F_I and F_{II}. If we assume for the moment that the densities $\psi_{f,I}$, $\psi_{f,II}$ of these distributions are known then it is sensible to assign an observation x to class I if $\psi_{f,I}(f(x)) \geq \psi_{f,II}(f(x))$ and to class II otherwise. The quality of such a classification mechanism depends obviously on the chosen aggregation function f and we would wish to choose f from a suitable class of functions so that the misclassification probability is minimized. This approach is not readily implementable as it relies on the unknown densities $\psi_{f,i}$. We suggest to estimate these densities using a maximum likelihood approach based on a sample of vectors from each class, the training set, and the induced samples of aggregate values. In the simplest case where $\{\psi(\cdot;\theta) : \theta \in \Theta\}$ is a family of probability density functions on the real line and the observed training vectors x_k^i, $i = I, II$, are independent, we determine parameters $\theta^i \in \Theta$ which maximize

$$L\big(\theta | f\big(x_k^i\big)\big) = \prod_k \psi\big(f\big(x_k^i\big);\theta\big).$$

This leads naturally to a two-level optimization problem where on the lower level the maximum likelihood density estimates are computed for given f, whereas on the upper level f is chosen so that the estimated misclassification probability, based on the lower level density estimates, is minimized.

To understand the formulation of our method, we consider first a simple inner product of the observed training data vectors $\boldsymbol{x}_k^i$ from one of two classes against a fixed aggregation vector $\boldsymbol{f}$. This results in coefficients $\boldsymbol{c}_k^i = \boldsymbol{f}^T \boldsymbol{x}_k^i$. In order to determine the utility of the classifier, we must determine the extent to which the $\boldsymbol{c}_k^I$ differ from the $\boldsymbol{c}_k^{II}$. Thus it is necessary to estimate the class-conditional distributions of the $\boldsymbol{c}_k^i$. We wish to select the aggregation vector $\boldsymbol{f}$ which maximizes the separation between the distributions of the $\boldsymbol{c}_k^i$.

Fisher [4] is generally recognized as the first to formulate the above as an optimization problem. However, there are numerous approaches to addressing the optimization problem when the class-conditional distributions are unknown. We believe that this is partially due to the fact that in general no one directly optimizes the PCC performance for the system, but rather secondary metrics are utilized which may not be representative of the true performance of the system.

Our procedure starts with the training data and an initial aggregation vector $\boldsymbol{f}$. From the two classes of training data we obtain two estimates θ^i, for $i = I, II$. We would like to use these initial estimates of the class-conditional probability density functions to generate a cost function which will be equivalent to the conditional probability of correct classification. To do this, we need to understand the decision statistic. We utilize the simple likelihood ratio statistic. That is, the class whose density is largest for a given observation will be the class to which the observation is assigned. Therefore, the probability of false classification is minimized by minimizing the overlap of the class-conditional probability densities. This is equivalent to maximizing the L^1 difference between the distributions, i.e.

$$C(\boldsymbol{f}, \theta^I, \theta^{II}) = \int |\psi(x; \theta^I) - \psi(x; \theta^{II})|\, dx.$$

The difficulty in this optimization is that the above cost function, which is a function of the parameters θ^i, depends implicitly upon the aggregation vector $\boldsymbol{f}$ as well as the data.

2.2. *The statistical details*

In the outline above we have suppressed many of the statistical details. In particular, we did not address the parametric form of the class-conditional probability density functions $\psi(\cdot; \theta)$. While the basic ideas behind our method extend to arbitrary choices of distribution family, we will concentrate on mixture models (mixtures of normals) in this paper.

The choice of distribution family without a priori knowledge is generally a compromise between choosing a restrictive model for simplicity and choosing a more general model which requires significantly more data to accurately estimate the many parameters of the model. This quandry is generally referred to as the statistical "curse of dimensionality."

We concentrate on mixture models, which are simple sums of Gaussians. Mixture models are relatively robust; one can add additional terms if the model is too restrictive, and simple mixtures (with 3-5 components in the sum) are sufficient to model a rich variety of distributions. Specifically, let

$$\varphi(x; \mu, \sigma) \equiv \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad x \in (-\infty, \infty)$$

denote the density function of a normal random variable with arbitrary mean $\mu \in (-\infty, \infty)$ and positive standard deviation $\sigma \in (0, \infty)$. The Gaussian mixture model of

m normal density functions $\varphi(x;\, \mu_i, \sigma_i)$ for $i = 1, \ldots, m$ is given by

$$\psi(x;\, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) \;\equiv\; \sum_{i=1}^{m} \pi_i\, \varphi(x;\, \mu_i, \sigma_i), \qquad x \;\in\; (-\infty,\, \infty),$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are respectively the m-dimensional vectors of means μ_i and standard deviations σ_i, for $i = 1, \ldots, m$, and $\boldsymbol{\pi} \equiv (\pi_i) \in \Re^m$ is the vector of mixture coefficients that is an element of the unit simplex Δ_m in $\Re^m$:

$$\Delta_m \;\equiv\; \left\{ \boldsymbol{\pi} \;\in\; \Re^m_+ \;:\; \sum_{i=1}^{m} \pi_i \;=\; 1 \right\}.$$

Notice that for given $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$, $\psi(\cdot;\, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$ is a probability density function, though not necessarily normal.

Our statistical modeling and estimation approach consists of the well-known maximum likelihood density estimation of mixture models combined with the recent idea of borrowed strength for data analysis [8]. Borrowed strength is a dimensionality reduction technique which allows the optimization algorithm to choose a reduced dimensionality mixture model which is appropriate for the data to be analyzed. Specifically, borrowed strength utilizes all of the data in a data set to fix the means and variances of the mixture models, i.e. to fix $\boldsymbol{\sigma}, \boldsymbol{\mu}$. (The number of mixture components, m, is chosen via the methodology presented in [9].)

After fixing these means and variances we then want to separate the various classes. Specifically, we wish to classify an unknown target as one of two classes: I or II. Since we have fixed the variances and means at appropriate points for the entire data set, the only free parameters remaining in the mixture models are the mixing coefficients $\boldsymbol{\pi}$. Thus, the target classes are characterized by the m-dimensional mixing coefficients, generated from the maximum likelihood estimation:

$$\boldsymbol{\pi}^{\mathrm{I}} \;\equiv\; \left(\pi_i^{\mathrm{I}}\right)_{i=1}^{m} \quad \text{for class I target,}$$

$$\boldsymbol{\pi}^{\mathrm{II}} \;\equiv\; \left(\pi_j^{\mathrm{II}}\right)_{j=1}^{m} \text{ for class II target;}$$

These coefficients define an m-term Gaussian mixture model [5, 7] that is the basis of the overall classification procedure. This is accomplished by the introduction of a separation measure of the target classes whose maximization would be an appropriate surrogate to the minimization of the misclassification error. (As we shall see below, a direct minimization of the latter error is extremely difficult.) One simple (and somewhat naive) measure is a distance function $\theta_1(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}})$ of the mixing coefficients; for example, the squared Euclidean distance:

$$\theta_1(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}) \;\equiv\; \tfrac{1}{2} \sum_{i=1}^{m} \left[\left(\pi_i^{\mathrm{I}} - \pi_i^{\mathrm{II}}\right)^2 \right]. \tag{1}$$

Subsequently we will introduce another separation measure and compare the classification results based on these different measures. While these measures are not exactly

equivalent to the probability of correct classification, they can be shown to be closely tied to it. In the long run, we are interested in solving the L^1 optimization problem, which will be exactly equivalent to maximizing PCC.

Inputs to the optimization problem consist of d types of training data points (observations):

$$\left\{ \boldsymbol{X}_1^{\mathrm{I}},\ \boldsymbol{X}_2^{\mathrm{I}},\ \cdots,\ \boldsymbol{X}_d^{\mathrm{I}} \right\} \text{ for class I target,}$$

$$\left\{ \boldsymbol{X}_1^{\mathrm{II}},\ \boldsymbol{X}_2^{\mathrm{II}},\ \cdots,\ \boldsymbol{X}_d^{\mathrm{II}} \right\} \text{ for class II target,}$$

with each $\boldsymbol{X}_\ell^{\mathrm{I}}$ and $\boldsymbol{X}_\ell^{\mathrm{II}}$ being vectors in the Euclidean spaces $\Re^{n_1}$ and $\Re^{n_2}$ respectively. For convenience, it would be useful to introduce the matrices $\boldsymbol{X}^{\mathrm{I,II}}$ whose columns are these vectors $\boldsymbol{X}_\ell^{\mathrm{I,II}}$; specifically,

$$\boldsymbol{X}^{\mathrm{I}} \equiv \left[\boldsymbol{X}_1^{\mathrm{I}}\ \ \boldsymbol{X}_2^{\mathrm{I}}\ \ \cdots\ \ \boldsymbol{X}_k^{\mathrm{I}} \right] \in \Re^{n_1 \times d} \quad \text{and} \quad \boldsymbol{X}^{\mathrm{II}} \equiv \left[\boldsymbol{X}_1^{\mathrm{II}}\ \ \boldsymbol{X}_2^{\mathrm{II}}\ \ \cdots\ \ \boldsymbol{X}_k^{\mathrm{II}} \right] \in \Re^{n_2 \times d}.$$

For each target class, the data types are combined via a certain aggregation function:

$$\boldsymbol{\chi}^{\mathrm{I}} : W \subset \Re^d \to \Re^{n_1} \text{ for class I target,}$$

$$\boldsymbol{\chi}^{\mathrm{II}} : W \subset \Re^d \to \Re^{n_2} \text{ for class II target,}$$

where W is the common set of admissible aggregation weights of the data. An example of such an aggregation function is the simple additive function:

$$\boldsymbol{\chi}^{\mathrm{I,II}}(\boldsymbol{w}) \equiv \sum_{\ell=1}^{d} w_\ell\, \boldsymbol{X}_\ell^{\mathrm{I,II}}, \quad \text{for } \boldsymbol{w} \equiv (w_\ell)_{\ell=1}^{d} \in \Re^d,$$

or more simply,

$$\boldsymbol{\chi}^{\mathrm{I}}(\boldsymbol{w}) = \boldsymbol{X}^{\mathrm{I}}\, \boldsymbol{w} \in \Re^{n_1} \quad \text{and} \quad \boldsymbol{\chi}^{\mathrm{II}}(\boldsymbol{w}) = \boldsymbol{X}^{\mathrm{II}}\, \boldsymbol{w} \in \Re^{n_2}. \tag{2}$$

An example of the admissible set W is the unit simplex in $\Re^d$. In general, we write

$$\chi_i^{\mathrm{I}}(\boldsymbol{w}) \quad i = 1 \ldots, n_1$$

$$\chi_j^{\mathrm{II}}(\boldsymbol{w}) \quad j = 1 \ldots, n_2$$

to denote the component functions of $\boldsymbol{\chi}^{\mathrm{I,II}}(\boldsymbol{w})$.

Based on the aggregated data (where the weight vector $\boldsymbol{w}$ has yet to be determined), the combined technique of borrowed strength and maximum likehood estimation is employed to compute target aggregation weights and model mixing coefficients in order to obtain the maximum separation of the two target classes with reference to the training data $\boldsymbol{X}^{\mathrm{I,II}}$.

The objective of the optimization problem is to seek the model means $\boldsymbol{\mu}$, standard deviations $\boldsymbol{\sigma}$, and the mixing coefficients $\boldsymbol{\pi}^{\mathrm{I,II}}$ so as to maximize the separation of the two density functions:

$$\psi(x; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{I}}) \quad \text{and} \quad \psi(x; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{II}}) \tag{3}$$

which are taken to be the density functions of the respective target classes. Specializing the function $C(\boldsymbol{f}, \theta^I, \theta^{II})$ to this context, we see that the separation is defined by the integral:

$$\int_{-\infty}^{\infty} \left| \psi(x; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{I}}) - \psi(x; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{II}}) \right| dx.$$

As in the general case, this integral is not easy to evaluate and it is not differentiable in the unknowns; we propose a surrogate objective function as a simplified but reasonable separation measure. Specifically:

$$\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \equiv \tfrac{1}{2} \int_{-\infty}^{\infty} \left(\psi(x; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{I}}) - \psi(x; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{II}}) \right)^2 dx. \tag{4}$$

The function $\theta_1(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}})$ is a plausible surrogate objective also, albeit perhaps oversimplified and therefore potentially less desirable than $\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma})$. The latter function can be shown to be a quadratic function of $\boldsymbol{\pi}^{\mathrm{I}} - \boldsymbol{\pi}^{\mathrm{II}}$ with a variable matrix that is a function of $(\boldsymbol{\mu}, \boldsymbol{\sigma})$. Thus, $\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ can be thought of as a distance function of $\boldsymbol{\pi}^{\mathrm{I}}$ and $\boldsymbol{\pi}^{\mathrm{II}}$, weighed by a transformation matrix that depends on the unknown model means and standard deviations.

To show that $\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ is of the mentioned class, we use the following formula that appears in [1]: for all scalars μ, μ', σ and σ' with the latter two positive,

$$\int_{-\infty}^{\infty} \varphi(x; \mu, \sigma)\, \varphi(x; \mu', \sigma')\, dx = \varphi(\mu; \mu', \tilde{\sigma}), \tag{5}$$

where

$$\tilde{\sigma} \equiv \sqrt{\sigma^2 + (\sigma')^2}.$$

Note that the function $\varphi(\cdot; \cdot, \sigma)$ is symmetric in its first two arguments for every fixed third component; that is, $\varphi(x; y, \sigma) = \varphi(y; x, \sigma)$ for all x and y and $\sigma > 0$.

For given vectors of means $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma}$, define the $m \times m$ symmetric matrix $\boldsymbol{Q}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ with entries: for $i, j = 1, \ldots, m$,

$$Q_{ij}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \equiv \int_{-\infty}^{\infty} \varphi(x; \mu_i, \sigma_i)\, \varphi(x; \mu_j, \sigma_j)\, dx = \varphi(\mu_i; \mu_j, \sigma_{ij}),$$

where $\sigma_{ij} \equiv \sqrt{\sigma_i^2 + \sigma_j^2}$. Note that for all $i = 1, \ldots, m$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$,

$$Q_{ii}(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{2\sqrt{\pi}\, \sigma_i};$$

thus the diagonal entries of $\boldsymbol{Q}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ are separable functions of $\boldsymbol{\sigma}$ and independent of $\boldsymbol{\mu}$. It is not difficult to show that

$$\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \tfrac{1}{2} \left(\boldsymbol{\pi}^{\mathrm{I}} - \boldsymbol{\pi}^{\mathrm{II}} \right)^T \boldsymbol{Q}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \left(\boldsymbol{\pi}^{\mathrm{I}} - \boldsymbol{\pi}^{\mathrm{II}} \right).$$

Since the function θ_2 is nonnegative valued, it follows that the matrix $\boldsymbol{Q}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ is positive semidefinite. In the Appendix, we give the partial derivatives of the function

$\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ with respect to its arguments. These derivatives are needed in the algorithm for solving the target classification problem.

For given vectors of means $\boldsymbol{\mu}$, standard deviations $\boldsymbol{\sigma}$, and mixing coefficients $\boldsymbol{\pi}$, the likelihood function of the aggregated data $\boldsymbol{\chi}^{\mathrm{I,II}}(\boldsymbol{w})$ is given by

$$L^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) \equiv \prod_{i=1}^{n_1} \psi\big(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}\big) \quad \text{for class I target}$$

$$L^{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) \equiv \prod_{j=1}^{n_2} \psi\big(\chi_j^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}\big) \text{ for class II target.}$$

2.3. Extending from two to k classes

We have presented the basic setup of the two-class classification problem in the previous discussion. In the rest of this paper, we concentrate on this two-class problem for two reasons. First, it is an important simplification of the k-class problem ($k \geq 2$). Second, recent research suggests that employing two-class classification, via pairwise comparison, is a preferred approach to k-class classification [14].

3. The bilevel optimization/MPEC formulation

We may now formally state a bilevel constrained optimization problem formulation for the classification of target classes I and II, based on the observed data of these classes $\boldsymbol{X}^{\mathrm{I}}$ and $\boldsymbol{X}^{\mathrm{II}}$. Let $\varepsilon > 0$ be a prescribed positive lower bound for the model standard deviations. The first-level decision variable is the vector of data aggregation weights $\boldsymbol{w}$ which are restricted by the admissible set W; the second-level decision variables consist of the model mixing coefficients $\boldsymbol{\pi}^{\mathrm{I,II}}$, auxiliary mixing coefficients $\boldsymbol{\pi}^0$, model means $\boldsymbol{\mu}$, and model standard deviations $\boldsymbol{\sigma}$. The overall model computes these variables in order to

$$\begin{aligned}
&\text{maximize } \theta(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\
&\text{subject to } \boldsymbol{w} \in W \\
&\qquad \boldsymbol{\pi}^{\mathrm{I}} \in \operatorname{argmax}\{ L^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{I}}) : \boldsymbol{\pi}^{\mathrm{I}} \in \Delta_m \} \\
&\qquad \boldsymbol{\pi}^{\mathrm{II}} \in \operatorname{argmax}\{ L^{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{II}}) : \boldsymbol{\pi}^{\mathrm{II}} \in \Delta_m \}
\end{aligned} \tag{6}$$

$$(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^0) \in \operatorname{argmax}\{L^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^0) L^{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^0) : \boldsymbol{\pi}^0 \in \Delta_m, \text{ and } \sigma_i \geq \varepsilon, \ \forall i\}, \tag{7}$$

where the first-level objective function θ is a (surrogate) separation measure of the two density functions (3). We are particularly interested in θ being one of the two functions: $\theta_1(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}})$ given by (1) and $\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ given by (4).

In general, the problem (6) is a bilevel optimization problem because the constraints involve several lower-level optimization subproblems each corresponding to a maximum likelihood estimation with certain variables held fixed. Specifically, the last maximiza-

tion (7) in the constraints takes as its input an unknown (but presumed fixed) vector of aggregation weights and computes the mean and standard deviation vectors as well as the Gaussian mixing coefficients via a joint likelihood maximization of the data of the two classes; the computed quantities are implicit functions of the input weights. Two of the three output quantities $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ along with the weights $\boldsymbol{w}$ from this maximization (7) are then fed as inputs into the individual likelihood maximization of the two classes. Thus, the maximization of $L^{\mathrm{I,II}}$ corresponds to the likelihood maximization of the individual target classes using the implicit mean and standard deviation functions computed from (7). The overall optimization problem (6) seeks a set of optimal weights and corresponding mixing coefficients, means and standard deviations in order to maximize the separation of the two target classes.

Bilevel optimization problems such as (6) are by no means easy to deal with. Nevertheless, substantial theory is known about these problems and iterative algorithms exist for computing stationary points of such problems [6]. In what follows, we present a reformulation of the problem (6) as an MPEC and an iterative method for computing a feasible tuple $(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ that satisfies the first-order stationarity condition of (6), more precisely, a B-stationary point (this terminology was coined in [13] to mean a stationary point of an MPEC that satisfies the first-order conditions presented in [6, Chap. 3]).

To begin, it would be convenient for us to define the log-likelihood functions and compute their partial derivatives with respect to their arguments. For this purpose, we introduce some further notation: let

$$\boldsymbol{\varphi}(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) \equiv (\varphi(x; \mu_\ell, \sigma_\ell))_{\ell=1}^{m};$$

in terms of this vector function, we have

$$\psi(x; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \boldsymbol{\pi}^T \boldsymbol{\varphi}(x; \boldsymbol{\mu}, \boldsymbol{\sigma}).$$

In the Appendix, we introduce the notation and give the formulas for the first and second partial derivatives of the vector function $\boldsymbol{\varphi}(x; \boldsymbol{\mu}, \boldsymbol{\sigma})$ with respect to its arguments. For class I target, let

$$\mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) \equiv \log L^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{n_1} \log \psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$$

denote the logarithmic likelihood function; we have (see the Appendix for notation):

$$\nabla_{\boldsymbol{\pi}} \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{n_1} \left[\frac{1}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})} \boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}) \right], \tag{8}$$

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \mathrm{diag}(\boldsymbol{\pi}) \sum_{i=1}^{n_1} \left[\frac{1}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})} \boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}) \right], \tag{9}$$

$$\nabla_{\boldsymbol{\sigma}} \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \mathrm{diag}(\boldsymbol{\pi}) \sum_{i=1}^{n_1} \left[\frac{1}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})} \boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}) \right]. \tag{10}$$

Similar expressions can be derived for class II targets. Consider the following log-likelihood maximization problem in the variable $\boldsymbol{\pi}$ with $(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ held fixed:

$$\begin{aligned} &\text{maximize } \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) \\ &\text{subject to } \boldsymbol{\pi} \geq 0, \quad \sum_{i=1}^{m} \pi_i = 1. \end{aligned}$$

This is clearly equivalent to the first maximization constraint in (6). Introducing a Lagrange multiplier λ for the equality constraint, we may write the first-order optimality conditions for the above maximization problem as follows:

$$\begin{aligned} 0 \leq \boldsymbol{\pi} \perp -\nabla_{\boldsymbol{\pi}} \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) + \mathbf{1}_m \lambda \geq 0 \\ \sum_{i=1}^{m} \pi_i = 1, \quad \lambda \text{ unrestricted.} \end{aligned} \tag{11}$$

It is not difficult to see that

$$\boldsymbol{\pi}^T \nabla_{\boldsymbol{\pi}} \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = n_1.$$

Thus it follows that the conditions (11) are equivalent to

$$0 \leq \boldsymbol{\pi} \perp -\nabla_{\boldsymbol{\pi}} \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) + n_1 \mathbf{1}_m \geq 0.$$

Consider the following maximization problem in the variables $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$ for fixed $\boldsymbol{w}$ that is equivalent to (7):

$$\begin{aligned} &\text{maximize } \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) + \mathcal{L}^{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) \\ &\text{subject to } \boldsymbol{\pi} \in \Delta_m \text{ and } \boldsymbol{\sigma} \geq \varepsilon \mathbf{1}_m. \end{aligned}$$

The first-order optimality conditions for this problem are as follows:

$$\begin{gathered} 0 \leq \boldsymbol{\pi} \perp -\nabla_{\boldsymbol{\pi}} \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) - \nabla_{\boldsymbol{\pi}} \mathcal{L}^{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) + (n_1 + n_2) \mathbf{1}_m \geq 0. \\ \nabla_{\boldsymbol{\mu}} \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) + \nabla_{\boldsymbol{\mu}} \mathcal{L}^{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = 0 \\ 0 \leq \boldsymbol{\sigma} - \varepsilon \mathbf{1}_m \perp -\nabla_{\boldsymbol{\sigma}} \mathcal{L}^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) - \nabla_{\boldsymbol{\sigma}} \mathcal{L}^{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) \geq 0. \end{gathered}$$

Substituting the expressions (8), (9), and (10) and their analogs for the gradient vectors of $\mathcal{L}^{\mathrm{I,II}}$, we can now state the two-level optimization problem (6) as the following constrained optimization problem with nonlinear complementarity constraints in the variables $(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\pi}^{0}, \boldsymbol{\mu}, \boldsymbol{\sigma})$:

$$\begin{aligned} &\text{maximize} \quad \theta(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\ &\text{subject to} \quad \boldsymbol{w} \in W \\ &0 \leq \boldsymbol{\pi}^{\mathrm{I}} \perp -\sum_{i=1}^{n_1} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{I}})} \right] + n_1 \mathbf{1}_m \geq 0 \\ &0 \leq \boldsymbol{\pi}^{\mathrm{II}} \perp -\sum_{i=1}^{n_2} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{II}})} \right] + n_2 \mathbf{1}_m \geq 0 \end{aligned} \tag{12}$$

$$0 \le \boldsymbol{\pi}^0 \perp -\sum_{i=1}^{n_1}\left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}^0)}\right] - \sum_{i=1}^{n_2}\left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}^0)}\right] + (n_1+n_2)\,\mathbf{1}_m \ge 0$$

$$\mathrm{diag}(\boldsymbol{\pi}^0)\left\{\sum_{i=1}^{n_1}\left[\frac{\boldsymbol{d\varphi}_\mu(\chi_i^{\mathrm{I}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}^0)}\right] + \sum_{i=1}^{n_2}\left[\frac{\boldsymbol{d\varphi}_\mu(\chi_i^{\mathrm{II}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}^0)}\right]\right\} = 0$$

$$0 \le \boldsymbol{\sigma} - \varepsilon\,\mathbf{1}_m \perp -\mathrm{diag}(\boldsymbol{\pi}^0)\left\{\sum_{i=1}^{n_1}\left[\frac{\boldsymbol{d\varphi}_\sigma(\chi_i^{\mathrm{I}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}^0)}\right] + \sum_{i=1}^{n_2}\left[\frac{\boldsymbol{d\varphi}_\sigma(\chi_i^{\mathrm{II}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w});\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}^0)}\right]\right\} \ge 0.$$

The above constrained optimization problem is now in the form of an MPEC: $\boldsymbol{w}$ is the first-level decision variable and $(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\pi}^0, \boldsymbol{\mu}, \boldsymbol{\sigma})$ are the second-level decision variables that are (implicit) functions of $\boldsymbol{w}$; the constraints consist of the set W and a system of mixed complementarity conditions parametrized by $\boldsymbol{w}$. For a comprehensive treatment of MPECs, see [6]. Here we point out that in addition to all the computational issues associated with the complementarity constraints, the objective function θ, being a distance measure of $\boldsymbol{\pi}^{\mathrm{I}}$ and $\boldsymbol{\pi}^{\mathrm{II}}$ for fixed $(\boldsymbol{\mu}, \boldsymbol{\sigma})$, is typically a convex function of $(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}})$; cf. the functions θ_1 and θ_2. Nevertheless, as examplified by the latter function, θ can not be expected in realistic applications to be either convex or concave in its four arguments. Consequently, as a maximization problem of optimizing an objective function that is neither convex nor concave and subject to nonlinear disjunctive constraints, one can expect great difficulty for computing a global maximizer of the problem (12). Thus our computational goal is quite modest; namely, we wish to compute a stationary point of this problem which makes a satisfactory improvement based on a separation measure $\theta(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, starting from an initial value. As we shall see from the numerical results, this goal is easily achieved and in fact exceeded by our algorithm which we describe in Sect. 4.

Conventional maximum likelihood

The borrowed strength technique has certain statistical advantages in density function estimation; see [8]. Alternatively, one could apply conventional maximum likelihood estimation as the basis for target classification. The resulting MPEC is similar to (12) and so is the application of the PIPA described subsequently. In what follows we derive this alternative approach and presents the detailed MPEC formulation.

In the conventional maximum likelihood approach, we introduce two distinct sets of unknown means and standard deviations for the two target classes:

$$[\,(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}})\text{ for target I }] \quad \text{and} \quad [\,(\boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}})\text{ for target II }].$$

The objective $\theta(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}})$, to be maximized, is then a function of all these variables. The bilevel optimization problem arising from this approach is the following:

$$
\begin{aligned}
&\text{maximize } \theta(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}}) \\
&\text{subject to } \boldsymbol{w} \in W \\
&\qquad (\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}) \in \operatorname{argmax} \{ L^{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{I}}) : \boldsymbol{\pi}^{\mathrm{I}} \in \Delta_m \text{ and } \boldsymbol{\sigma}^{\mathrm{I}} \geq \varepsilon \mathbf{1}_m \} \\
&\text{and} \qquad (\boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}) \in \operatorname{argmax} \{ L^{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\pi}^{\mathrm{II}}) : \boldsymbol{\pi}^{\mathrm{II}} \in \Delta_m \text{ and } \boldsymbol{\sigma}^{\mathrm{II}} \geq \varepsilon \mathbf{1}_m \}.
\end{aligned} \tag{13}
$$

Similar to the derivation in Sect. 3, the above optimization problem has the following equivalent MPEC formulation:

$$
\begin{aligned}
&\text{maximize} \quad \theta(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}}) \\
&\text{subject to} \quad \boldsymbol{w} \in W \\
&0 \leq \boldsymbol{\pi}^{\mathrm{I}} \perp -\sum_{i=1}^{n_1} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{I}})} \right] + n_1 \mathbf{1}_m \geq 0 \\
&0 \leq \boldsymbol{\pi}^{\mathrm{II}} \perp -\sum_{i=1}^{n_2} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\pi}^{\mathrm{II}})} \right] + n_2 \mathbf{1}_m \geq 0 \\
&\operatorname{diag}(\boldsymbol{\pi}^{\mathrm{I}}) \sum_{i=1}^{n_1} \left[\frac{\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{I}})} \right] = 0 \\
&\operatorname{diag}(\boldsymbol{\pi}^{\mathrm{II}}) \sum_{i=1}^{n_2} \left[\frac{\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\pi}^{\mathrm{II}})} \right] = 0 \\
&0 \leq \boldsymbol{\sigma}^{\mathrm{I}} - \varepsilon \mathbf{1}_m \perp -\operatorname{diag}(\boldsymbol{\pi}^{\mathrm{I}}) \sum_{i=1}^{n_1} \left[\frac{\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{I}})} \right] \geq 0 \\
&0 \leq \boldsymbol{\sigma}^{\mathrm{II}} - \varepsilon \mathbf{1}_m \perp -\operatorname{diag}(\boldsymbol{\pi}^{\mathrm{II}}) \sum_{i=1}^{n_2} \left[\frac{\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\pi}^{\mathrm{II}})} \right] \geq 0.
\end{aligned} \tag{14}
$$

An objective function that is a generalization of (4) is given by

$$
\tilde{\theta}_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}}) \equiv \tfrac{1}{2} \int_{-\infty}^{\infty} \left(\psi(x; \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{I}}) - \psi(x; \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\pi}^{\mathrm{II}}) \right)^2 dx. \tag{15}
$$

Using the same formula (5) as before, we can show that

$$
\tilde{\theta}_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}}) = \tfrac{1}{2} \begin{pmatrix} \boldsymbol{\pi}^{\mathrm{I}} \\ \boldsymbol{\pi}^{\mathrm{II}} \end{pmatrix}^T \boldsymbol{Q}(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}}) \begin{pmatrix} \boldsymbol{\pi}^{\mathrm{I}} \\ \boldsymbol{\pi}^{\mathrm{II}} \end{pmatrix};
$$

here the $2m \times m$ symmetric positive semidefinite matrix $\boldsymbol{Q}(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}})$ is given by:

$$
\boldsymbol{Q}(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}}) \equiv \begin{bmatrix} Q_{\mathrm{I}}(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}) & -Q_{\mathrm{I,II}}(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}}) \\ -Q_{\mathrm{I,II}}(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}})^T & Q_{\mathrm{II}}(\boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}) \end{bmatrix},
$$

where

$$Q_\alpha(\boldsymbol{\mu}^\alpha, \boldsymbol{\sigma}^\alpha) \equiv \left(\varphi\big(\mu_i^\alpha, \mu_j^\alpha, \sigma_{ij}^\alpha\big) \right)_{i=1}^{m}, \quad \text{for } \alpha = I,\ II$$

and

$$Q_{\mathrm{I,II}}(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}}) \equiv \left(\varphi\big(\mu_i^{\mathrm{I}}, \mu_j^{\mathrm{II}}, \sigma_{ij}^{\mathrm{I,II}}\big) \right)_{i,j=1}^{m},$$

with

$$\sigma_{ij}^{\mathrm{I,II}} \equiv \sqrt{\big(\sigma_i^{\mathrm{I}}\big)^2 + \big(\sigma_j^{\mathrm{II}}\big)^2}\ .$$

Similar to (12), we can write the problem (14) in a compact form by introducing the vector functions $\boldsymbol{G}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}})$, $\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}})$, and $\boldsymbol{F}(\boldsymbol{w}, z, \boldsymbol{\lambda}, \boldsymbol{\mu}^{\mathrm{I,II}})$ as follows:

$$\boldsymbol{G}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}}) \equiv \begin{pmatrix} \boldsymbol{G}_{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{I}}) \\ \boldsymbol{G}_{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{II}}) \end{pmatrix} \tag{16}$$

$$\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}}) \equiv \begin{pmatrix} \boldsymbol{H}_{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{I}}) \\ \boldsymbol{H}_{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{II}}) \end{pmatrix} \tag{17}$$

and

$$\boldsymbol{F}(\boldsymbol{w}, z, \boldsymbol{\lambda}, \boldsymbol{\mu}^{\mathrm{I,II}}) \equiv \begin{pmatrix} z - \boldsymbol{G}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}}) \\ -\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}}) \end{pmatrix},$$

where

$$z \equiv \begin{pmatrix} \boldsymbol{x}^{\mathrm{I}} \\ \boldsymbol{x}^{\mathrm{II}} \\ \boldsymbol{y}^{\mathrm{I}} \\ \boldsymbol{y}^{\mathrm{II}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\lambda} \equiv \begin{pmatrix} \boldsymbol{\pi}^{\mathrm{I}} \\ \boldsymbol{\pi}^{\mathrm{II}} \\ \boldsymbol{\sigma}^{\mathrm{II}} \end{pmatrix},$$

$$\boldsymbol{G}_{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{I}}) \equiv \begin{pmatrix} -\sum_{i=1}^{n_1} \left[\dfrac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{I}})} \right] + n_1 \mathbf{1}_m \\ -\mathrm{diag}(\boldsymbol{\pi}^{\mathrm{I}}) \sum_{i=1}^{n_1} \dfrac{\boldsymbol{d\varphi}_\sigma(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{I}})} \end{pmatrix}$$

$$\boldsymbol{H}_{\mathrm{I}}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{I}}) \equiv -\mathrm{diag}(\boldsymbol{\pi}^{\mathrm{I}}) \sum_{i=1}^{n_1} \frac{\boldsymbol{d\varphi}_\mu(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{I}})};$$

and similar expressions exist for $\boldsymbol{G}_{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{II}})$ and $\boldsymbol{H}_{\mathrm{II}}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{II}})$. The compact formulation of (14) is:

$$\begin{aligned} &\text{maximize } \theta(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{II}}) \\ &\text{subject to } \boldsymbol{w} \in W \\ &\qquad\qquad \boldsymbol{F}(\boldsymbol{w}, z, \boldsymbol{\lambda}, \boldsymbol{\mu}^{\mathrm{I,II}}) = 0 \\ &\qquad\qquad z \circ (\boldsymbol{\lambda} - \boldsymbol{\lambda}_l) = 0 \\ &\qquad\qquad z \geq 0 \quad \text{and} \quad \boldsymbol{\lambda} \geq \boldsymbol{\lambda}_l, \end{aligned}$$

where

$$\boldsymbol{\lambda}_l \equiv \begin{pmatrix} 0 \\ 0 \\ \varepsilon\, \mathbf{1}_{2m} \end{pmatrix}.$$

From a computationally point of view, (12) and (14) are two different constrained optimization problems. They differ in several aspects. The former problem has $(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\pi}^{0}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \in \Re^{d+5m}$ as its variables; the latter problem has $(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}}) \in \Re^{d+6m}$ as its variables; thus (14) has m more variables than (12). Also, there is no obvious (mathematical) connection between the feasible regions of the two problems. The structure of the Jacobian matrix of the constraint functions $\boldsymbol{G}$ and $\boldsymbol{H}$ in the two problems is also different. For (12), the structure is given by the display (27); the associated linear algebraic computations within PIPA for solving (12) can take advantage of this structure; see Sect. 4. For (14), the Jacobian matrix of $\boldsymbol{G}$ and $\boldsymbol{H}$ are block diagonally structured, reflecting the separation of the variables $(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{I}})$ and $(\boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\sigma}^{\mathrm{II}}, \boldsymbol{\mu}^{\mathrm{II}})$ in these functions; see (16) and (17). In the implementation of PIPA, such block diagonal structure can also be put to use to simplify the linear algebraic computations.

4. A solution method and its implementation

This section presents a computational method for solving the optimization problem (12). In addition to the algorithms described in [6, Chap. 6], there are recent advances on algorithms for solving an MPEC. In what follows, we describe a penalty interior point algorithm (PIPA) that was designed to treat optimization problems of this kind. This is the algorithm that we have implemented in our computational study whose results we report in Sect. 5.

We introduce several vector functions for the constraints of the problem (12):

$$\begin{aligned}
&\boldsymbol{G}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\pi}^{0}, \boldsymbol{\sigma}, \boldsymbol{\mu}) \equiv \begin{pmatrix} \boldsymbol{G}_I(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\sigma}, \boldsymbol{\mu}) \\ \boldsymbol{G}_{II}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\sigma}, \boldsymbol{\mu}) \\ \boldsymbol{G}_0(\boldsymbol{w}, \boldsymbol{\pi}^{0}, \boldsymbol{\sigma}, \boldsymbol{\mu}) \\ \boldsymbol{G}_\sigma(\boldsymbol{w}, \boldsymbol{\pi}^{0}, \boldsymbol{\sigma}, \boldsymbol{\mu}) \end{pmatrix} \\
&\equiv \begin{pmatrix} -\sum_{i=1}^{n_1} \left[\dfrac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{I}})} \right] + n_1\, \mathbf{1}_m \\ -\sum_{i=1}^{n_2} \left[\dfrac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{\mathrm{II}})} \right] + n_2\, \mathbf{1}_m \\ -\sum_{i=1}^{n_1} \left[\dfrac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{0})} \right] - \sum_{i=1}^{n_2} \left[\dfrac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{0})} \right] + (n_1 + n_2)\, \mathbf{1}_m \\ -\mathrm{diag}(\boldsymbol{\pi}^{0}) \left\{ \sum_{i=1}^{n_1} \left[\dfrac{d\boldsymbol{\varphi}_\sigma(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{0})} \right] + \sum_{i=1}^{n_2} \left[\dfrac{d\boldsymbol{\varphi}_\sigma(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{0})} \right] \right\} \end{pmatrix}
\end{aligned} \tag{18}$$

$$\begin{aligned}
&\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{\pi}^{0}, \boldsymbol{\sigma}, \boldsymbol{\mu}) \\
&\equiv -\mathrm{diag}(\boldsymbol{\pi}^{0}) \left\{ \sum_{i=1}^{n_1} \left[\dfrac{d\boldsymbol{\varphi}_\mu(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{0})} \right] + \sum_{i=1}^{n_2} \left[\dfrac{d\boldsymbol{\varphi}_\mu(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}); \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}^{0})} \right] \right\}
\end{aligned} \tag{19}$$

and

$$\boldsymbol{F}(\boldsymbol{w}, \boldsymbol{x}^{\mathrm{I}}, \boldsymbol{x}^{\mathrm{II}}, \boldsymbol{x}^{0}, \boldsymbol{y}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\pi}^{0}, \boldsymbol{\sigma}, \boldsymbol{\mu}) \equiv \begin{pmatrix} \begin{pmatrix} \boldsymbol{x}^{\mathrm{I}} \\ \boldsymbol{x}^{\mathrm{II}} \\ \boldsymbol{x}^{0} \\ \boldsymbol{y} \end{pmatrix} - \boldsymbol{G}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\pi}^{0}, \boldsymbol{\sigma}, \boldsymbol{\mu}) \\ -\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{\pi}^{0}, \boldsymbol{\sigma}, \boldsymbol{\mu}) \end{pmatrix}. \quad (20)$$

We can then rewrite (12) as

$$\begin{aligned} &\text{maximize } \theta(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\ &\text{subject to } \boldsymbol{w} \in W \\ &\qquad \boldsymbol{F}(\boldsymbol{w}, z, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0 \\ &\qquad z \circ (\boldsymbol{\lambda} - \boldsymbol{\lambda}_l) = 0 \end{aligned}$$

and

$$0 \leq z \equiv \begin{pmatrix} \boldsymbol{x}^{\mathrm{I}} \\ \boldsymbol{x}^{\mathrm{II}} \\ \boldsymbol{x}^{0} \\ \boldsymbol{y} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\lambda} \equiv \begin{pmatrix} \boldsymbol{\pi}^{\mathrm{I}} \\ \boldsymbol{\pi}^{\mathrm{II}} \\ \boldsymbol{\pi}^{0} \\ \boldsymbol{\sigma} \end{pmatrix} \geq \boldsymbol{\lambda}_l \equiv \begin{pmatrix} 0 \\ 0 \\ 0 \\ \varepsilon \mathbf{1}_m \end{pmatrix},$$

which is in the special form of an MPEC to which PIPA can be directly applied. A general iteration (labeled ν) of PIPA is as follows. Let c, ρ, $\alpha_{\nu-1}$, and κ_ν be given scalars satisfying

$$c > 0, \quad \rho \in (0, 1), \quad \alpha_{\nu-1} > 1, \quad \text{and} \quad \kappa_\nu \in (0, 1];$$

Let Q_ν be a symmetric positive definite matrix of order d, and

$$(\boldsymbol{w}^\nu, z^\nu, \boldsymbol{\lambda}^\nu, \boldsymbol{\mu}^\nu) \quad (21)$$

be an iterate satisfying

- (feasibility of weights) $\boldsymbol{w}^\nu \in W$;
- (positivity of complementary variables) $(z^\nu, \boldsymbol{\lambda}^\nu - \boldsymbol{\lambda}_l) > 0$;
- (centrality condition) $z^\nu \circ (\boldsymbol{\lambda}^\nu - \boldsymbol{\lambda}_l) \geq \rho\, g_\nu \mathbf{1}_{4m}$, where

$$g_\nu \equiv \frac{(z^\nu)^T (\boldsymbol{\lambda}^\nu - \boldsymbol{\lambda}_l)}{4m}$$

is the average gap of the complementary pairs.

Let $\boldsymbol{d}\boldsymbol{F}^\nu$ denote the Jacobian matrix of $\boldsymbol{F}$ evaluated at the tuple (21). Also write

$$\boldsymbol{F}^\nu \equiv \boldsymbol{F}(\boldsymbol{w}^\nu, z^\nu, \boldsymbol{\lambda}^\nu, \boldsymbol{\mu}^\nu), \quad \boldsymbol{G}^\nu \equiv \boldsymbol{G}(\boldsymbol{w}^\nu, \boldsymbol{\lambda}^\nu, \boldsymbol{\mu}^\nu), \quad \boldsymbol{H}^\nu \equiv \boldsymbol{H}(\boldsymbol{w}^\nu, \boldsymbol{\pi}^{0,\nu}, \boldsymbol{\sigma}^\nu, \boldsymbol{\mu}^\nu),$$

and

$$\begin{pmatrix} d\theta^{\mathrm{I},\nu} \\ d\theta^{\mathrm{II},\nu} \\ d\theta^{\mu,\nu} \\ d\theta^{\sigma,\nu} \end{pmatrix} \equiv \nabla\theta(\pi^{\mathrm{I},\nu}, \pi^{\mathrm{II},\nu}, \mu^{\nu}, \sigma^{\nu}),$$

with the latter being the gradient vector of θ evaluated at $(\pi^{\mathrm{I},\nu}, \pi^{\mathrm{II},\nu}, \mu^{\nu}, \sigma^{\nu})$.

We compute a search direction $(dw, dz, d\lambda, d\mu)$ by solving the convex quadratic program: for an arbitrary constant $c > 0$,

$$\begin{aligned} \text{minimize } & -(d\theta^{\mathrm{I},\nu})^T d\pi^{\mathrm{I}} - (d\theta^{\mathrm{II},\nu})^T d\pi^{\mathrm{II}} - (d\theta^{\mu,\nu})^T d\mu \\ & \qquad - (d\theta^{\sigma,\nu})^T d\sigma + \tfrac{1}{2}\, dw^T\, Q_\nu\, dw \\ \text{subject to } & w^\nu + dw \in W \\ & \|dw\|_\infty \le c\, \left(\|F^\nu\| + (z^\nu)^T (\lambda^\nu - \lambda_l) \right) \\ & dF^\nu \begin{pmatrix} dw \\ dz \\ d\lambda \\ d\mu \end{pmatrix} = -F^\nu \\ & \mathrm{diag}(z^\nu)\, d\lambda + \mathrm{diag}(\lambda^\nu - \lambda_l)\, dz = -z^\nu \circ (\lambda^\nu - \lambda_l) + \kappa_\nu\, g_\nu\, \mathbf{1}_{4m}. \end{aligned} \tag{22}$$

By using the last two equations in the constraints to eliminate the triple $(dz, d\lambda, d\mu)$, the above quadratic program can be equivalently reformulated in the variable dw only. Indeed, writing dF^ν in partitioned form:

$$dF^\nu = \left[\begin{array}{cccc} dF^\nu_w & dF^\nu_z & dF^\nu_\lambda & dF^\nu_\mu \end{array} \right],$$

where the subscripts denote the partial Jacobian submatrices of dF^ν with respect to the respective arguments, we can rewrite the two equations in question as:

$$\begin{aligned} & \begin{bmatrix} dF^\nu_w \\ 0 \end{bmatrix} dw + \begin{bmatrix} dF^\nu_z & dF^\nu_\lambda & dF^\nu_\mu \\ \mathrm{diag}(\lambda^\nu - \lambda_l) & \mathrm{diag}(z^\nu) & 0 \end{bmatrix} \begin{pmatrix} dz \\ d\lambda \\ d\mu \end{pmatrix} \\ & \qquad = \begin{pmatrix} -F^\nu \\ -z^\nu \circ (\lambda^\nu - \lambda_l) + \kappa_\nu\, g_\nu\, \mathbf{1}_{4m} \end{pmatrix}. \end{aligned} \tag{23}$$

By the special form (20) of the function F, we may solve for $(dz, d\lambda, d\mu)$ in terms of dw. For this purpose, we need to introduce some further notation. Let dG^ν_λ denote the partial Jacobian matrix of G with respect to the variable λ evaluated at $(w^\nu, \lambda^\nu, \mu^\nu)$; similarly, let dG^ν_w and dG^ν_μ denote respectively the partial Jacobian matrix of G with respect to w and μ evaluated at $(w^\nu, \lambda^\nu, \mu^\nu)$; similar notation is used for the partial Jacobian matrices of H with respect to its arguments. We also write $\lambda^{-\nu} \equiv (\lambda^\nu - \lambda_l)^{-1}$.

With the above notation, the equation (23) takes the form:

$$\begin{bmatrix} \mathrm{diag}(z^\nu \circ \lambda^{-\nu}) + d\boldsymbol{G}_\lambda^\nu & d\boldsymbol{G}_\mu^\nu \\ d\boldsymbol{H}_\lambda^\nu & d\boldsymbol{H}_\mu^\nu \end{bmatrix} \begin{pmatrix} d\lambda \\ d\mu \end{pmatrix} = -\begin{pmatrix} \boldsymbol{G}^\nu - \kappa_\nu\, g_\nu\, \lambda^{-\nu} \\ \boldsymbol{H}^\nu \end{pmatrix} - \begin{bmatrix} d\boldsymbol{G}_w^\nu \\ d\boldsymbol{H}_w^\nu \end{bmatrix} d\boldsymbol{w}, \tag{24}$$

and

$$d\boldsymbol{z} \;=\; -\mathrm{diag}(z^\nu \circ \lambda^{-\nu})\, d\lambda - z^\nu + \kappa_\nu\, g_\nu\, \lambda^{-\nu}.$$

From (24), we may obtain $d\boldsymbol{\pi}^{\mathrm{I,II}}$ as an affine function of $d\boldsymbol{w}$; the directional quadratic subprogram (22) therefore is equivalent to:

$$\begin{aligned} &\text{minimize } q(d\boldsymbol{w}) \\ &\text{subject to } \boldsymbol{w}^\nu + d\boldsymbol{w} \;\in\; W \\ &\qquad\qquad \|\, d\boldsymbol{w}\|_\infty \;\le\; c\, \left(\|\, \boldsymbol{F}^\nu \,\| + (\, z^\nu\,)^T (\, \lambda^\nu - \lambda_{\,l}\,) \right), \end{aligned} \tag{25}$$

where

$$\begin{aligned} q(d\boldsymbol{w}) \equiv -(d\boldsymbol{\theta}^{\,\mathrm{I},\nu})^T d\boldsymbol{\pi}^{\mathrm{I}}(d\boldsymbol{w}) - (d\boldsymbol{\theta}^{\,\mathrm{II},\nu})^T d\boldsymbol{\pi}^{\mathrm{II}}(d\boldsymbol{w}) - (d\boldsymbol{\theta}^{\,\mu,\nu})^T d\boldsymbol{\mu}(d\boldsymbol{w}) \\ - (d\boldsymbol{\theta}^{\,\sigma,\nu})^T d\boldsymbol{\sigma}(d\boldsymbol{w}) + \tfrac{1}{2}\, d\boldsymbol{w}^T\, Q_\nu\, d\boldsymbol{w}. \end{aligned}$$

Clearly (25) has $d\boldsymbol{w}$ as its only variable.

Having obtained the search direction as described above, we compute a positive step size as follows. Let us denote

$$\begin{pmatrix} \boldsymbol{w}^\nu(\tau) \\ z^\nu(\tau) \\ \lambda^\nu(\tau) \\ \boldsymbol{\mu}^\nu(\tau) \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{w}^\nu \\ z^\nu \\ \lambda^\nu \\ \boldsymbol{\mu}^\nu \end{pmatrix} + \tau \begin{pmatrix} d\boldsymbol{w} \\ d\boldsymbol{z} \\ d\lambda \\ d\boldsymbol{\mu} \end{pmatrix}, \qquad \tau \;\in\; [0,\,1].$$

Notice that $\boldsymbol{w}^\nu(\tau) \in W$ for all $\tau \in [0, 1]$. Define the penalized objective function with penalty parameter α:

$$P_\alpha(\boldsymbol{w}, z, \lambda, \boldsymbol{\mu}) \;\equiv\; -\theta(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) + \alpha \left(\|\boldsymbol{F}(\boldsymbol{w}, z, \lambda, \boldsymbol{\mu})\|^2 + z^T(\lambda - \lambda_{\,l}) \right), \quad \alpha > 0.$$

Penalty update rule. Let $\alpha_\nu \equiv \alpha_{\nu-1}^{p_\nu}$, where $p_\nu \ge 1$ is the smallest integer $p \ge 1$ such that

$$\begin{aligned} &- (d\boldsymbol{\theta}^{\,\mathrm{I},\nu})^T d\boldsymbol{\pi}^{\mathrm{I}} - (d\boldsymbol{\theta}^{\,\mathrm{II},\nu})^T d\boldsymbol{\pi}^{\mathrm{II}} - (d\boldsymbol{\theta}^{\,\mu,\nu})^T d\boldsymbol{\mu} - (d\boldsymbol{\theta}^{\,\sigma,\nu})^T d\boldsymbol{\sigma} \\ -\alpha_{\nu-1}^p \left[2\|\boldsymbol{F}^\nu\|^2 + (1-\kappa_\nu)(z^\nu)^T(\lambda^\nu - \lambda_{\,l}) \right] &< -\left[\|\boldsymbol{F}^\nu\|^2 + (z^\nu)^T(\lambda^\nu - \lambda_{\,l}) \right]. \end{aligned}$$

We seek a step size $\tau_\nu \in (0, 1]$ such that the new iterate

$$\begin{pmatrix} \boldsymbol{w}^{\nu+1} \\ z^{\nu+1} \\ \lambda^{\nu+1} \\ \boldsymbol{\mu}^{\nu+1} \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{w}^\nu(\tau_\nu) \\ z^\nu(\tau_\nu) \\ \lambda^\nu(\tau_\nu) \\ \boldsymbol{\mu}^\nu(\tau_\nu) \end{pmatrix}$$

satisfies the positivity condition

$$(z^{\nu+1}, \lambda^{\nu+1} - \lambda_l) > 0$$

and the centrality condition

$$z^{\nu+1} \circ (\lambda^{\nu+1} - \lambda_l) \geq \rho\, g_{\nu+1}\, \mathbf{1}_{4m},$$

where

$$g_{\nu+1} \equiv \frac{(z^{\nu+1})^T(\lambda^{\nu+1} - \lambda_l)}{4m};$$

moreover $P_{\alpha_\nu}(w^{\nu+1}, z^{\nu+1}, \lambda^{\nu+1}, \mu^{\nu+1})$ will be sufficiently smaller than $P_{\alpha_\nu}(w^{\nu}, z^{\nu}, \lambda^{\nu}, \mu^{\nu})$. The determination of τ_ν is as follows. Let

$$g_\nu(\tau) \equiv \frac{z^\nu(\tau)^T(\lambda^\nu(\tau) - \lambda_l)}{4m}.$$

Compute the scalar

$$\bar{\tau}_\nu \equiv \sup\Bigg\{\tau \in (0,1] : \min_{1\leq i\leq m}\Big(\min_{k=0,I,II}\big[\pi_i^{k,\nu}(\tau)x_i^{k,\nu}(\tau) - \rho\, g_\nu(\tau)\big], \\ y_i^\nu(\tau)\sigma_i^\nu(\tau) - \rho\, g_\nu(\tau)\Big) \geq 0\Bigg\};$$

this can be accomplished very easily because each term in the above expression is a quadratic function of τ. Next perform the

Armijo inexact line search: for a given backtracking factor $\delta \in (0, 1)$ and a constant $\gamma \in (0, 1)$, let

$$\tau_\nu \equiv 0.9999\, \delta^{\ell'_\nu}\, \bar{\tau}_\nu,$$

where ℓ'_ν is the smallest nonnegative integer ℓ such that with

$$\tau \equiv 0.9999\, \delta^{\ell}\, \bar{\tau}_\nu,$$

we have

$$\begin{aligned} P_{\alpha_\nu}(w^\nu(\tau), z^\nu(\tau), \lambda^\nu(\tau), \mu^\nu(\tau)) - P_{\alpha_\nu}(w^\nu, z^\nu, \lambda^\nu, \mu^\nu) \\ \leq -\gamma\,\tau\,\big[(dw^{\mathrm{I},\nu})^T dx^{\mathrm{I}} + (dw^{\mathrm{II},\nu})^T dx^{\mathrm{II}} + (dw^{\mu,\nu})^T d\mu + (dw^{\sigma,\nu})^T d\sigma \\ + \alpha_\nu\,\big(2\,\|\,F^\nu\,\|^2 + (1-\kappa_\nu)\,(z^\nu)^T(\lambda^\nu - \lambda_l)\big)\big]. \end{aligned}$$

This completes the description of a general iteration of PIPA. A termination rule for the algorithm is:

$$\|\,F^\nu\,\| + \|\,\min(z^\nu, \lambda^\nu - \lambda_\ell)\,\| + \|\,dw\,\| \leq \text{tolerance}. \tag{26}$$

Solving the quadratic program (25)

We now give some more computationally details for solving the quadratic subprogram (25) numerically. For this purpose, we discuss how to obtain the functions $d\boldsymbol{\pi}^{\mathrm{I,II}}(d\boldsymbol{w})$, $d\boldsymbol{\mu}(d\boldsymbol{w})$, and $d\boldsymbol{\sigma}(d\boldsymbol{w})$ from the equation (24). See the Appendix for further notation. We write

$$\boldsymbol{\sigma}^{-\nu} \equiv \left(\boldsymbol{\sigma}^{\nu} - \varepsilon \mathbf{1}_m\right)^{-1}.$$

Introducing the positive diagonal matrices:

$$\boldsymbol{D}^{\mathrm{I},\nu} \equiv \operatorname{diag}\left(\boldsymbol{x}^{\mathrm{I},\nu} \circ \boldsymbol{\pi}^{\mathrm{I},-\nu}\right), \quad \boldsymbol{D}^{\mathrm{II},\nu} \equiv \operatorname{diag}\left(\boldsymbol{x}^{\mathrm{II},\nu} \circ \boldsymbol{\pi}^{\mathrm{II},-\nu}\right),$$

$$\boldsymbol{D}^{0,\nu} \equiv \operatorname{diag}\left(\boldsymbol{x}^{0,\nu} \circ \boldsymbol{\pi}^{0,-\nu}\right), \quad \boldsymbol{D}^{s,\nu} \equiv \operatorname{diag}\left(\boldsymbol{y}^{\nu} \circ \boldsymbol{\sigma}^{-\nu}\right),$$

the equation (24) can be written in the following block form:

$$\begin{bmatrix} d\boldsymbol{G}_{\mathrm{I}}^{\nu} + \boldsymbol{D}^{\mathrm{I},\nu} & 0 & 0 & d\boldsymbol{G}_{\mathrm{I},\sigma}^{\nu} & d\boldsymbol{G}_{\mathrm{I},\mu}^{\nu} \\ 0 & d\boldsymbol{G}_{\mathrm{II}}^{\nu} + \boldsymbol{D}^{\mathrm{II},\nu} & 0 & d\boldsymbol{G}_{\mathrm{II},\sigma}^{\nu} & d\boldsymbol{G}_{\mathrm{II},\mu}^{\nu} \\ 0 & 0 & d\boldsymbol{G}_{0}^{\nu} + \boldsymbol{D}^{0,\nu} & d\boldsymbol{G}_{0,\sigma}^{\nu} & d\boldsymbol{G}_{0,\mu}^{\nu} \\ 0 & 0 & d\boldsymbol{G}_{\sigma,0}^{\nu} & d\boldsymbol{G}_{\sigma}^{\nu} + \boldsymbol{D}^{s,\nu} & d\boldsymbol{G}_{\sigma,\mu}^{\nu} \\ 0 & 0 & d\boldsymbol{H}_{0}^{\nu} & d\boldsymbol{H}_{\sigma}^{\nu} & d\boldsymbol{H}_{\mu}^{\nu} \end{bmatrix} \begin{pmatrix} d\boldsymbol{\pi}^{\mathrm{I}} \\ d\boldsymbol{\pi}^{\mathrm{II}} \\ d\boldsymbol{\pi}^{0} \\ d\boldsymbol{\sigma} \\ d\boldsymbol{\mu} \end{pmatrix}$$

$$- \begin{pmatrix} \boldsymbol{G}_{\mathrm{I}}^{\nu} - \kappa_{\nu}\, g_{\nu}\, \boldsymbol{\pi}^{\mathrm{I},-\nu} \\ \boldsymbol{G}_{\mathrm{II}}^{\nu} - \kappa_{\nu}\, g_{\nu}\, \boldsymbol{\pi}^{\mathrm{II},-\nu} \\ \boldsymbol{G}_{0}^{\nu} - \kappa_{\nu}\, g_{\nu}\, \boldsymbol{\pi}^{0,-\nu} \\ \boldsymbol{G}_{\sigma}^{\nu} - \kappa_{\nu}\, g_{\nu}\, \boldsymbol{\sigma}^{-\nu} \\ \boldsymbol{H}^{\nu} \end{pmatrix} - \begin{bmatrix} d\boldsymbol{G}_{\mathrm{I},w}^{\nu} \\ d\boldsymbol{G}_{\mathrm{II},w}^{\nu} \\ d\boldsymbol{G}_{0,w}^{\nu} \\ d\boldsymbol{G}_{\sigma,w}^{\nu} \\ d\boldsymbol{H}_{w}^{\nu} \end{bmatrix} d\boldsymbol{w}. \tag{27}$$

The solution of the latter equation can be accomplished in two steps. First solve for $(d\boldsymbol{\pi}^{0}, d\boldsymbol{\sigma}, d\boldsymbol{\mu})$ in the equation:

$$\begin{bmatrix} d\boldsymbol{G}_{0}^{\nu} + \boldsymbol{D}^{0,\nu} & d\boldsymbol{G}_{0,\sigma}^{\nu} & d\boldsymbol{G}_{0,\mu}^{\nu} \\ d\boldsymbol{G}_{\sigma,0}^{\nu} & d\boldsymbol{G}_{\sigma}^{\nu} + \boldsymbol{D}^{s,\nu} & d\boldsymbol{G}_{\sigma,\mu}^{\nu} \\ d\boldsymbol{H}_{0}^{\nu} & d\boldsymbol{H}_{\sigma}^{\nu} & d\boldsymbol{H}_{\mu}^{\nu} \end{bmatrix} \begin{pmatrix} d\boldsymbol{\pi}^{0} \\ d\boldsymbol{\sigma} \\ d\boldsymbol{\mu} \end{pmatrix} = - \begin{pmatrix} \boldsymbol{G}_{0}^{\nu} - \kappa_{\nu}\, g_{\nu}\, \boldsymbol{\pi}^{0,-\nu} \\ \boldsymbol{G}_{\sigma}^{\nu} - \kappa_{\nu}\, g_{\nu}\, \boldsymbol{\sigma}^{-\nu} \\ \boldsymbol{H}^{\nu} \end{pmatrix} - \begin{bmatrix} d\boldsymbol{G}_{0,w}^{\nu} \\ d\boldsymbol{G}_{\sigma,w}^{\nu} \\ d\boldsymbol{H}_{w}^{\nu} \end{bmatrix} d\boldsymbol{w}$$

and then back substitute to obtain $d\boldsymbol{\pi}^{\mathrm{I}}$ and = $d\boldsymbol{\pi}^{\mathrm{II}}$ from the equations:

$$\left(d\boldsymbol{G}_{\mathrm{I}}^{\nu} + \boldsymbol{D}^{\mathrm{I},\nu}\right) d\boldsymbol{\pi}^{\mathrm{I}} = -\left(d\boldsymbol{G}_{\mathrm{I},\sigma}^{\nu} d\boldsymbol{\sigma} + d\boldsymbol{G}_{\mathrm{I},\mu}^{\nu} d\boldsymbol{\mu} + \boldsymbol{G}_{\mathrm{I}}^{\nu} - \kappa_{\nu} g_{\nu} \boldsymbol{\pi}^{\mathrm{I},-\nu} + d\boldsymbol{G}_{\mathrm{I},w}^{\nu} d\boldsymbol{w}\right),$$
$$\left(d\boldsymbol{G}_{\mathrm{II}}^{\nu} + \boldsymbol{D}^{\mathrm{II},\nu}\right) d\boldsymbol{\pi}^{\mathrm{II}} = -\left(d\boldsymbol{G}_{\mathrm{II},\sigma}^{\nu} d\boldsymbol{\sigma} + d\boldsymbol{G}_{\mathrm{II},\mu}^{\nu} d\boldsymbol{\mu} + \boldsymbol{G}_{\mathrm{II}}^{\nu} - \kappa_{\nu} g_{\nu} \boldsymbol{\pi}^{\mathrm{II},-\nu} + d\boldsymbol{G}_{\mathrm{II},w}^{\nu} d\boldsymbol{w}\right).$$

Let us denote

$$\mathbf{\Gamma}^{\nu} \equiv \begin{bmatrix} d\boldsymbol{G}_{\sigma}^{\nu} + \boldsymbol{D}^{s,\nu} & d\boldsymbol{G}_{\sigma,\mu}^{\nu} \\ d\boldsymbol{H}_{\sigma}^{\nu} & d\boldsymbol{H}_{\mu}^{\nu} \end{bmatrix} - \begin{bmatrix} d\boldsymbol{G}_{\sigma,0}^{\nu} \\ d\boldsymbol{H}_{0}^{\nu} \end{bmatrix} \left(d\boldsymbol{G}_{0}^{\nu} + \boldsymbol{D}^{0,\nu} \right)^{-1} \begin{bmatrix} d\boldsymbol{G}_{0,\sigma}^{\nu} & d\boldsymbol{G}_{0,\mu}^{\nu} \end{bmatrix}.$$

The nonsingularity of this matrix is essential to the overall performance of PIPA. In terms of $\mathbf{\Gamma}^{\nu}$, we have

$$\begin{bmatrix} d\boldsymbol{\sigma} \\ d\boldsymbol{\mu} \end{bmatrix} = (\mathbf{\Gamma}^{\nu})^{-1} \left\{ \begin{bmatrix} d\boldsymbol{G}_{\sigma,0}^{\nu} \\ d\boldsymbol{H}_{0}^{\nu} \end{bmatrix} \left(d\boldsymbol{G}_{0}^{\nu} + \boldsymbol{D}^{0,\nu} \right)^{-1} \left(\boldsymbol{G}_{0}^{\nu} - \kappa_{\nu} g_{\nu} \boldsymbol{\pi}^{0,-\nu} \right) \right.$$
$$\left. - \begin{pmatrix} \boldsymbol{G}_{\sigma}^{\nu} - \kappa_{\nu} g_{\nu} \boldsymbol{\sigma}^{-\nu} \\ \boldsymbol{H}^{\nu} \end{pmatrix} + \left(\begin{bmatrix} d\boldsymbol{G}_{\sigma,0}^{\nu} \\ d\boldsymbol{H}_{0}^{\nu} \end{bmatrix} \left(d\boldsymbol{G}_{0}^{\nu} + \boldsymbol{D}^{0,\nu} \right)^{-1} d\boldsymbol{G}_{0,w}^{\nu} - \begin{bmatrix} d\boldsymbol{G}_{\sigma,w}^{\nu} \\ d\boldsymbol{H}_{w}^{\nu} \end{bmatrix} \right) d\boldsymbol{w} \right\}.$$

Consequently, we have

$$-(d\boldsymbol{\theta}^{\mathrm{I},\nu})^{T} d\boldsymbol{\pi}^{\mathrm{I}} = \text{ constant } + (\boldsymbol{c}^{\mathrm{I},\nu})^{T} d\boldsymbol{w}$$

where

$$(\boldsymbol{c}^{\mathrm{I},\nu})^{T} \equiv (d\boldsymbol{\theta}^{\mathrm{I},\nu})^{T} \left(d\boldsymbol{G}_{\mathrm{I}}^{\nu} + \boldsymbol{D}^{\mathrm{I},\nu} \right)^{-1} \left\{ d\boldsymbol{G}_{\mathrm{I},w}^{\nu} + \begin{bmatrix} d\boldsymbol{G}_{\mathrm{I},\sigma}^{\nu} & d\boldsymbol{G}_{\mathrm{I},\mu}^{\nu} \end{bmatrix} (\mathbf{\Gamma}^{\nu})^{-1} \right.$$
$$\left. \left(\begin{bmatrix} d\boldsymbol{G}_{\sigma,0}^{\nu} \\ d\boldsymbol{H}_{0}^{\nu} \end{bmatrix} \left(d\boldsymbol{G}_{0}^{\nu} + \boldsymbol{D}^{0,\nu} \right)^{-1} d\boldsymbol{G}_{0,w}^{\nu} - \begin{bmatrix} d\boldsymbol{G}_{\sigma,w}^{\nu} \\ d\boldsymbol{H}_{w}^{\nu} \end{bmatrix} \right) \right\}.$$

A similar expression for $-(d\boldsymbol{\theta}^{\mathrm{II},\nu})^{T} d\boldsymbol{\pi}^{\mathrm{II}}$ can be derived. Moreover, we have

$$-(d\boldsymbol{\theta}^{\mu,\nu})^{T} d\boldsymbol{\sigma} - (d\boldsymbol{\theta}^{\sigma,\nu})^{T} d\boldsymbol{\mu} = \text{ constant } - (\boldsymbol{c}^{\mu\sigma,\nu})^{T} d\boldsymbol{w}$$

where

$$(\boldsymbol{c}^{\mu\sigma,\nu})^{T} \equiv$$
$$\begin{bmatrix} (d\boldsymbol{\theta}^{\mu,\nu})^{T} & (d\boldsymbol{\theta}^{\sigma,\nu})^{T} \end{bmatrix} (\mathbf{\Gamma}^{\nu})^{-1} \left(\begin{bmatrix} d\boldsymbol{G}_{\sigma,0}^{\nu} \\ d\boldsymbol{H}_{0}^{\nu} \end{bmatrix} \left(d\boldsymbol{G}_{0}^{\nu} + \boldsymbol{D}^{0,\nu} \right)^{-1} d\boldsymbol{G}_{0,w}^{\nu} - \begin{bmatrix} d\boldsymbol{G}_{\sigma,w}^{\nu} \\ d\boldsymbol{H}_{w}^{\nu} \end{bmatrix} \right).$$

Consequently, the quadratic subprogram (24) can now be written as

$$\begin{aligned} &\text{minimize } (\boldsymbol{c}^{\mathrm{I},\nu} + \boldsymbol{c}^{\mathrm{II},\nu} + \boldsymbol{c}^{\mu\sigma,\nu})^{T} d\boldsymbol{w} + \tfrac{1}{2} d\boldsymbol{w}^{T} Q_{\nu} d\boldsymbol{w} \\ &\text{subject to } \boldsymbol{w}^{\nu} + d\boldsymbol{w} \in W \\ &\qquad \| d\boldsymbol{w} \|_{\infty} \leq c \left(\| \boldsymbol{F}^{\nu} \| + (z^{\nu})^{T} (\boldsymbol{\lambda}^{\nu} - \boldsymbol{\lambda}_{l}) \right). \end{aligned} \tag{28}$$

In the special case where W is the unit simplex and Q_{ν} is chosen to be a positive diagonal matrix (such as a positive multiple of the identity matrix), the latter quadratic program can be solved very effectively by a specialized algorithm.

An implicit programming approach

As an alternative to PIPA, a piecewise programming approach can also be used for solving the bilevel optimization problem (6). Specifically, this approach, IMPA, is based on an equivalent implicit programming formulation of (6) as a one-level nonsmooth optimization problem in the first-level variable $\boldsymbol{w}$ alone. The cornerstone of the IMPA is the following:

Implicit program postulate: The lower-level optimization problem (7) has an optimal solution

$$(\boldsymbol{\mu}(\boldsymbol{w}), \boldsymbol{\sigma}(\boldsymbol{w}), \boldsymbol{\pi}^0(\boldsymbol{w}))$$

that is a B(ouligand)-differentiable function of the first-level variable $\boldsymbol{w} \in W$; similarly, the two log-likehood maximization problems: for $\alpha = \text{I}, \text{II}$,

$$\begin{aligned} &\text{maximize } \mathcal{L}^{\alpha}(\boldsymbol{w}, \boldsymbol{\mu}(\boldsymbol{w}), \boldsymbol{\sigma}(\boldsymbol{w}), \boldsymbol{\pi}) \\ &\text{subject to } \boldsymbol{\pi} \geq 0, \quad \sum_{i=1}^{m} \pi_i = 1, \end{aligned} \tag{29}$$

have optimal solutions $\boldsymbol{\pi}^{\text{I}}(\boldsymbol{w})$ and $\boldsymbol{\pi}^{\text{II}}(\boldsymbol{w})$ that are B-differentiable in $\boldsymbol{w}$.

It should be noted that for a given triple $(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, the function $\mathcal{L}^{\text{I,II}}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \cdot)$ is concave in the last argument, as confirmed by the positive semidefiniteness of the two Jacobian matrices (30) and (31); thus the maximization problem (29) is not expected to be difficult for each fixed $\boldsymbol{w}$.

Under the above postulate, the bilevel optimization problem (6) may now be formulated in the following implicit form:

$$\begin{aligned} &\text{maximize } \hat{\theta}(\boldsymbol{w}) \equiv \theta(\boldsymbol{\pi}^{\text{I}}(\boldsymbol{w}), \boldsymbol{\pi}^{\text{II}}(\boldsymbol{w}), \boldsymbol{\mu}(\boldsymbol{w}), \boldsymbol{\sigma}(\boldsymbol{w})) \\ &\text{subject to } \boldsymbol{w} \in W, \end{aligned}$$

where the (implicitly defined) objective function $\boldsymbol{w}$ is a B-differentiable function of the single variable $\boldsymbol{w}$. Based on the latter formulation, iterative algorithms can be developed for computing a stationarity point of the problem (6).

5. A numerical application

In this section we report the results of a computational experiment with the application of PIPA for solving the MPECs arising from the borrowed strength and conventional statistical models in a multispectral minefield application. Specifically, we consider data collected under the Coastal Battlefield Reconnaissance and Analysis (COBRA) Program. The COBRA remote sensing data are taken from a passive multispectral video sensor with six different spectral bands aboard an unmanned aerial vehicle and are made available by NSWC Coastal Systems Station, Dahlgren Division, Panama City, Florida, to aid in the analysis of algorithms and approaches. A point pattern map provided with

the data contains candidate detections, both false positives and true mines, used in this experiment. See [10] for more details.

Before giving details of the results, we emphasize the fact that the optimization problems being solved are highly nonlinear and nonconvex; the complementarity constraints are a salient feature of these problems that make them particularly challenging. Since PIPA is a local search method and we make no claim of its ability to obtain a global optimum, our goal is therefore to demonstrate the effectiveness of the algorithm for obtaining a satisfactory (as opposed to best) target classification.

Our experience confirms that this class of MPECs is indeed not easy to be solved computationally. A straightforward implementation of PIPA yielded undesirable results; careful tuning of certain parameters in PIPA were essential. The following are two successful strategies in adjusting these parameters. Details are contained in the computer codes which are available upon request.

(a) In each directional quadratic subprogram (25), the following modified bound on the size of the search direction $\boldsymbol{dw}$ was employed:

$$\| \boldsymbol{dw} \|_\infty \leq 2 \min \left(\| \boldsymbol{F}^\nu \|_2 + (z^\nu)^T (\boldsymbol{\lambda}^\nu - \boldsymbol{\lambda}_l), \text{ rad} \right),$$

where "rad" was a constant initially set equal to 10^{-3} and was adjusted according to

$$\text{rad} = \max(10^4, \text{ rad} * .1)$$

when the search step became too small. The matrix Q_ν was set equal to the identity matrix.

(b) The centering parameter κ_ν has a substantial effect on the convergence of the algorithm. We set this parameter using an adaptive scheme that depends on the step size $\tau_{\nu-1}$ computed at the previous iteration.

We coded up PIPA as described in Sect. 4 in MATLAB. The "qp" function within the MATLAB Optimization Toolbox was employed to solve the directional quadratic subprograms (28). The first-level feasible set W was the unit simplex in $\Re^d$. The training data were given by the 12 x 6 matrix $\boldsymbol{X}^{\text{I}}$ and the 27 x 6 matrix $\boldsymbol{X}^{\text{II}}$; thus $d = 6$. An initial weight vector

$$\boldsymbol{w}^0 = \tfrac{1}{6} \mathbf{1}_6$$

was used. With this vector, the EM algorithm commonly used in statistical estimation was employed to generate an initial triple under a very crude stopping rule (the pre-processing step).

This triple was used as the starting iterate for an interior-point method (which is essentially PIPA with all the first-level elements removed) to compute an initial feasible solution to the MPEC (12); that is, we first solved a state problem under a tight termination tolerance (the initialization step). Note: the state problem is itself is a nonconvex constrained optimization problem. We recorded the objective value calculated at the computed feasible solution. This solution was then slightly perturbed to ensure the positivity of the complementary variables as required by PIPA; the perturbed vector then became the initial iterate used by PIPA (the main calculation). We terminated PIPA using the termination rule (26) where tolerance = 1.e-7. We recorded the objective

value at termination of PIPA and computed the ratio of this value with that of the initial feasible solution.

In the multispectral minefield application to which the PIPA was applied, there are six spectral bands of data ($d = 6$); class I data are obtained from 12 true mine observations ($n_1 = 12$) and class II data from 27 false mine observations ($n_2 = 27$). We have run our MATLAB code on a 2-term, 3-term, and 4-term borrowed strength model using the θ_2 objective function (thus $m = 2, 3, 4$) as well as a 2-term conventional model. Recall that the total number of MPEC variables is equal to $d + 5m$ in the borrow strength model and $d + 6m$ in the conventional model. Thus the largest MPEC that we have solved has 26 variables. Although this number is very small, the high nonlinearity of the defining functions and the nonconvexity of the constraints and objective function make the overall MPEC a difficult problem. The results of the runs are summarized as follows.

In all cases, the pre-processing step was uneventful, with the number of iterations ranging from low 20 to over 200. In the main calculation, the number of PIPA iterations in the successful runs ranges from the low teens to about 60. Invariably, the ratio of objective values was greater than unity, meaning that a better feasible solution was obtained by PIPA. Specifically, in the 2-term borrowed strength model, the best improvement in objective values was a little higher than 8%. In the 3-term borrowed strength model, this improvement reached higher than 40%. In the 2-term conventional model, we obtained a 45% improvement in the objective values. The most impressive result occurred in a run with the 4-term borrowed strength model; PIPA computed a feasible solution that was 3.5 times better in objective value than that obtained at the initialization step. This solution was then employed in a statistical test to calculate the probability of misclassification error. The estimated probability of error using the PIPA solution is by far the smallest among all known probabilities reported in the literature for this set of land mine data. For more details, see the accompanying paper [10].

The conclusion of this experiment is that although it has not been an easy experience with solving this class of MPECs, PIPA is nevertheless successful in obtaining the best result on a practical application having to do with land mine classification where the data originated from a realistic source.

Notes added in proof: Recent work by Sven Leiffer at Argonne National Laboratory has shown that there is an outstanding convergence issue with PIPA. Dr. Leiffer's discovery could explain the difficulty we have had in our computational experiments with the use of PIPA in this application.

6. An extension: bivariate normal density

We introduce an extension of the basic statistical method that has led to the bilevel optimization problems (6) and (13). Specifically, in addition to the aggregation function $\chi^{\mathrm{I,II}}$, we consider a second pair of such functions:

$$\Xi^{\mathrm{I}} : W \subset \Re^d \to \Re^{n_1} \text{ for class I target,}$$
$$\Xi^{\mathrm{II}} : W \subset \Re^d \to \Re^{n_2} \text{ for class II target.}$$

For instance, these can again be be additive functions just like $\boldsymbol{\chi}^{\mathrm{I,II}}$. Let

$$\phi(\boldsymbol{x};\boldsymbol{m},\boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{n/2}\sqrt{\det\boldsymbol{\Sigma}}}\exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})\right), \quad \boldsymbol{x}\in\Re^2,$$

denote the bivariate normal density function with $\boldsymbol{m}$ being the 2-dimensional vector of unknown model means and $\boldsymbol{\Sigma}$ being a 2×2 positive definite unknown model covariance matrix. Based on this bivariate normal density function ϕ, the density functions of the two targets are then taken to be:

$$\sum_{i=1}^{m}\pi_i^{\mathrm{I}}\,\phi(\boldsymbol{x};\boldsymbol{m}^{\mathrm{I},i},\boldsymbol{\Sigma}^{\mathrm{I},i}) \quad \text{for target I,}$$

$$\sum_{i=1}^{m}\pi_i^{\mathrm{II}}\,\phi(\boldsymbol{x};\boldsymbol{m}^{\mathrm{II},i},\boldsymbol{\Sigma}^{\mathrm{II},i}) \quad \text{for target II;}$$

again, we wish to separate these density functions by using a separation function such as

$$\tfrac{1}{2}\int_{\Re^2}\left[\sum_{i=1}^{m}\pi_i^{\mathrm{I}}\,\phi(\boldsymbol{x};\boldsymbol{m}^{\mathrm{I},i},\boldsymbol{\Sigma}^{\mathrm{I},i})-\pi_i^{\mathrm{II}}\,\phi(\boldsymbol{x};\boldsymbol{m}^{\mathrm{II},i},\boldsymbol{\Sigma}^{\mathrm{II},i})\right]^2 d\boldsymbol{x}.$$

Using either conventional or borrowed strength maximum likelihood estimation, one can formulate an MPEC whereby one seeks a vector of suitably constrained weights $\boldsymbol{w}$, the vectors of model means $\boldsymbol{m}^{\mathrm{I},i}$ and $\boldsymbol{m}^{\mathrm{II},i}$, the vectors of model covariance matrices $\boldsymbol{\Sigma}^{\mathrm{I},i}$ and $\boldsymbol{\Sigma}^{\mathrm{II},i}$ and mixture coefficients $\pi_i^{\mathrm{I,II}}$ so as to maximize a separation measure subject to maximum likelihood constraints. The detailed formulation of such an MPEC is omitted. Further extension to a multivariate normal density model is also possible.

A major difference between the MPEC resulting from a bivariate (and more generally, a multivariate) normal density model and that derived from the univariate normal density model is that instead of the unknown standard deviations (which are scalars) in the latter model, we encounter covariance matrices that are part of the unknowns of the MPEC; these matrices are restricted to be positive definite. Therefore, the maximum likelihood constraints themselves belong to the class of semidefinite programs; this aspect adds another layer of computational complication to the overall optimization problem. We expect great challenges in solving these highly sophisticated MPECs.

7. Conclusion

In this paper, we have presented the mathematical and algorithmic details of the MPEC approach for target classification. Several statistical methods are described that lead to various MPEC models. The general computational difficulties of these MPEC models are noted. The overall methodology is successfully demonstrated on a realistic classification application. Our conclusion with this research is that the MPEC methodology is promising for target classification but further research is needed to fine tune the methodology in order to enhance its effectiveness in practical applications of this kind.

Acknowledgements. The authors are grateful to a referee for a suggestion that has clarified the statistical model.

Appendix: derivative formulas

This appendix has three parts. The first part gives the explicit expressions for the partial derivatives of the scalar function $\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ (given by (4)) with respect to its arguments. The second part gives the various partial derivatives of the vector function $\boldsymbol{\varphi}(x; \boldsymbol{\mu}; \boldsymbol{\sigma})$ and the third part gives the partial Jacobian matrices for the functions $\boldsymbol{G}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\pi}^{0}, \boldsymbol{\sigma}, \boldsymbol{\mu})$ and $\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\pi}^{0}, \boldsymbol{\sigma}, \boldsymbol{\mu})$ (given by (18) and (19)) that define the MPEC (12) of the target classification problem. Although these derivatives are not difficult to calculate, they are nevertheless cumbersome and yet their explicit forms are essential for the efficient implementation of the PIPA for solving the MPEC. Thus we document the detailed formulas for these derivatives. Analogous derivative formulas can be obtained for the functions $\tilde{\theta}_2(\pi^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}})$, $\boldsymbol{G}(\boldsymbol{w}, \pi^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}})$, and $\boldsymbol{H}(\boldsymbol{w}, \pi^{\mathrm{I,II}}, \boldsymbol{\mu}^{\mathrm{I,II}}, \boldsymbol{\sigma}^{\mathrm{I,II}})$ (given by (15), (16), and (17), respectively); the latter formulas are omitted.

The function $\theta_2(\boldsymbol{\pi}^{I}, \boldsymbol{\pi}^{II}, \boldsymbol{\mu}, \boldsymbol{\sigma})$

Let $\nabla_{\mathrm{I,II}}\theta_2$ denote these partial derivatives with respect to $\boldsymbol{\pi}^{\mathrm{I}}$ and $\boldsymbol{\pi}^{\mathrm{II}}$. We have

$$\nabla_{\mathrm{I}}\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \boldsymbol{Q}(\boldsymbol{\mu}, \boldsymbol{\sigma})\left(\boldsymbol{\pi}^{\mathrm{I}} - \boldsymbol{\pi}^{\mathrm{II}}\right) = -\nabla_{\mathrm{II}}\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}).$$

Similarly, letting $\nabla_{\mu}\theta_2$ and $\nabla_{\sigma}\theta_2$ denote the partial derivatives of θ_2 with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, we have

$$\begin{aligned}
\nabla_{\mu}\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) &= \left(\boldsymbol{\pi}^{\mathrm{I}} - \boldsymbol{\pi}^{\mathrm{II}}\right) \circ \boldsymbol{M}(\boldsymbol{\mu}, \boldsymbol{\sigma})\left(\boldsymbol{\pi}^{\mathrm{I}} - \boldsymbol{\pi}^{\mathrm{II}}\right) \\
\nabla_{\sigma}\theta_2(\boldsymbol{\pi}^{\mathrm{I}}, \boldsymbol{\pi}^{\mathrm{II}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) &= \left(\boldsymbol{\pi}^{\mathrm{I}} - \boldsymbol{\pi}^{\mathrm{II}}\right) \circ \boldsymbol{S}(\boldsymbol{\mu}, \boldsymbol{\sigma})\left(\boldsymbol{\pi}^{\mathrm{I}} - \boldsymbol{\pi}^{\mathrm{II}}\right),
\end{aligned}$$

where $\boldsymbol{M}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ and $\boldsymbol{S}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ are $m \times m$ matrices with entries given respectively by

$$m_{ij}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \equiv \frac{\mu_j - \mu_i}{\sigma_i^2 + \sigma_j^2}\,\varphi(\mu_i; \mu_j, \sigma_{ij})$$

$$s_{ij}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \equiv \begin{cases} -\dfrac{1}{4\sqrt{\pi}\,\sigma_i^2}, & \text{if } i = j \\ \dfrac{\sigma_i}{\sigma_i^2 + \sigma_j^2}\left[-1 + \dfrac{(\mu_j - \mu_i)^2}{\sigma_i^2 + \sigma_j^2},\right]\varphi(\mu_i; \mu_j, \sigma_{ij}) & \text{if } i \neq j; \end{cases}$$

recall $\sigma_{ij} = \sqrt{\sigma_i^2 + \sigma_j^2}$. Notice that the matrix $\boldsymbol{M}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ is skew-symmetric; thus, in particular, it has zero diagonals.

The function $\boldsymbol{\varphi}(x;\boldsymbol{\mu};\boldsymbol{\sigma})$

We have the first partial derivatives:

$$\boldsymbol{d\varphi}_x(x;\boldsymbol{\mu},\boldsymbol{\sigma}) \equiv \left(\frac{\partial\varphi(x;\mu_\ell,\sigma_\ell)}{\partial x}\right)_{\ell=1}^{m},$$

$$\boldsymbol{d\varphi}_\mu(x;\boldsymbol{\mu},\boldsymbol{\sigma}) \equiv \left(\frac{\partial\varphi(x;\mu_\ell,\sigma_\ell)}{\partial\mu_\ell}\right)_{\ell=1}^{m}, \quad \boldsymbol{d\varphi}_\sigma(x;\boldsymbol{\mu},\boldsymbol{\sigma}) \equiv \left(\frac{\partial\varphi(x;\mu_\ell,\sigma_\ell)}{\partial\sigma_\ell}\right)_{\ell=1}^{m};$$

and for the second partial derivatives:

$$\boldsymbol{d^2\varphi}_\mu(x;\boldsymbol{\mu},\boldsymbol{\sigma}) \equiv \left(\frac{\partial^2\varphi(x;\mu_\ell,\sigma_\ell)}{\partial\mu_\ell^2}\right)_{\ell=1}^{m},$$

$$\boldsymbol{d^2\varphi}_\sigma(x;\boldsymbol{\mu},\boldsymbol{\sigma}) \equiv \left(\frac{\partial^2\varphi(x;\mu_\ell,\sigma_\ell)}{\partial\sigma_\ell^2}\right)_{\ell=1}^{m},$$

$$\boldsymbol{d^2\varphi}_{\mu,x}(x;\boldsymbol{\mu},\boldsymbol{\sigma}) = \boldsymbol{d^2\varphi}_{x,\mu}(x;\boldsymbol{\mu},\boldsymbol{\sigma}) \equiv \left(\frac{\partial^2\varphi(x;\mu_\ell,\sigma_\ell)}{\partial\mu_\ell\,\partial x}\right)_{\ell=1}^{m},$$

$$\boldsymbol{d^2\varphi}_{\sigma,x}(x;\boldsymbol{\mu},\boldsymbol{\sigma}) = \boldsymbol{d^2\varphi}_{x,\sigma}(x;\boldsymbol{\mu},\boldsymbol{\sigma}) \equiv \left(\frac{\partial^2\varphi(x;\mu_\ell,\sigma_\ell)}{\partial\sigma_\ell\,\partial x}\right)_{\ell=1}^{m},$$

and

$$\boldsymbol{d^2\varphi}_{\mu,\sigma}(x;\boldsymbol{\mu},\boldsymbol{\sigma}) = \boldsymbol{d^2\varphi}_{\sigma,\mu}(x;\boldsymbol{\mu},\boldsymbol{\sigma}) \equiv \left(\frac{\partial^2\varphi(x;\mu_\ell,\sigma_\ell)}{\partial\mu_\ell\,\partial\sigma_\ell}\right)_{\ell=1}^{m}.$$

We note the following explicit formulas:

$$-\frac{\partial\varphi(x;\mu,\sigma)}{\partial x} = \frac{\partial\varphi(x;\mu,\sigma)}{\partial\mu} = \frac{x-\mu}{\sigma^2}\,\varphi(x;\mu,\sigma),$$

$$\frac{\partial\varphi(x;\mu,\sigma)}{\partial\sigma} = \left(\frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}\right)\varphi(x;\mu,\sigma),$$

$$\frac{\partial^2\varphi(x;\mu,\sigma)}{\partial\mu^2} = \left[\left(\frac{x-\mu}{\sigma^2}\right)^2 - \frac{1}{\sigma^2}\right]\varphi(x;\mu,\sigma) = \frac{1}{\sigma}\,\frac{\partial\varphi(x;\mu,\sigma)}{\partial\sigma},$$

$$\frac{\partial^2\varphi(x;\mu,\sigma)}{\partial\sigma^2} = \left[\frac{(x-\mu)^4}{\sigma^6} - \frac{5\,(x-\mu)^2}{\sigma^4} + \frac{2}{\sigma^2}\right]\varphi(x;\mu,\sigma),$$

$$\frac{\partial^2\varphi(x;\mu,\sigma)}{\partial\mu\,\partial x} = \frac{\partial^2\varphi(x;\mu,\sigma)}{\partial x\,\partial\mu} = -\frac{\partial^2\varphi(x;\mu,\sigma)}{\partial\mu^2},$$

$$\frac{\partial^2\varphi(x;\mu,\sigma)}{\partial\sigma\,\partial x} = \frac{\partial^2\varphi(x;\mu,\sigma)}{\partial x\,\partial\sigma} = \left[-\frac{(x-\mu)^3}{\sigma^5} + \frac{3\,(x-\mu)}{\sigma^3}\right]\varphi(x;\mu,\sigma),$$

and

$$\frac{\partial^2\varphi(x;\mu,\sigma)}{\partial\sigma\,\partial\mu} = \frac{\partial^2\varphi(x;\mu,\sigma)}{\partial\mu\,\partial\sigma} = \left[\frac{(x-\mu)^3}{\sigma^5} - \frac{3\,(x-\mu)}{\sigma^3}\right]\varphi(x;\mu,\sigma).$$

The functions $\boldsymbol{G}(\boldsymbol{w}, \boldsymbol{\pi}^I, \boldsymbol{\pi}^{II}, \boldsymbol{\pi}^0, \boldsymbol{\sigma}, \boldsymbol{\mu})$ *and* $\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{\pi}^I, \boldsymbol{\pi}^{II}, \boldsymbol{\pi}^0, \boldsymbol{\sigma}, \boldsymbol{\mu})$

Finally, we give the explicit forms of the matrices $d\boldsymbol{G}_{\lambda}^{\nu}$, $d\boldsymbol{G}_{\mu}^{\nu}$, $d\boldsymbol{G}_{w}^{\nu}$, $d\boldsymbol{H}_{\lambda}^{\nu}$, and $d\boldsymbol{H}_{\mu}^{\nu}$. In the formulas given below, we assume that the functions $\boldsymbol{\chi}^{\mathrm{I,II}}$ are additive as expressed by (2). From their definitions (18) and (19), we have

$$d\boldsymbol{G}_{\lambda}^{\nu} = \begin{pmatrix} d\boldsymbol{G}_{I}^{\nu} & 0 & 0 & d\boldsymbol{G}_{I,\sigma}^{\nu} \\ 0 & d\boldsymbol{G}_{II}^{\nu} & 0 & d\boldsymbol{G}_{II,\sigma}^{\nu} \\ 0 & 0 & d\boldsymbol{G}_{0}^{\nu} & d\boldsymbol{G}_{0,\sigma}^{\nu} \\ 0 & 0 & d\boldsymbol{G}_{\sigma,0}^{\nu} & d\boldsymbol{G}_{\sigma}^{\nu} \end{pmatrix}, \quad d\boldsymbol{G}_{\mu}^{\nu} = \begin{pmatrix} d\boldsymbol{G}_{\mathrm{I},\mu}^{\nu} \\ d\boldsymbol{G}_{\mathrm{II},\mu}^{\nu} \\ d\boldsymbol{G}_{0,\mu}^{\nu} \\ d\boldsymbol{G}_{\sigma,\mu}^{\nu} \end{pmatrix}, \quad d\boldsymbol{G}_{w}^{\nu} = \begin{pmatrix} d\boldsymbol{G}_{\mathrm{I},w}^{\nu} \\ d\boldsymbol{G}_{\mathrm{II},w}^{\nu} \\ d\boldsymbol{G}_{0,w}^{\nu} \\ d\boldsymbol{G}_{\sigma,w}^{\nu} \end{pmatrix}$$

$$d\boldsymbol{H}_{\lambda}^{\nu} = \begin{pmatrix} 0 & 0 & d\boldsymbol{H}_{0}^{\nu} & d\boldsymbol{H}_{\sigma}^{\nu} \end{pmatrix},$$

where the various partial Jacobian matrices are given as follows:

$$d\boldsymbol{G}_{I}^{\nu} \equiv \sum_{i=1}^{n_1} \frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})^T}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{I},\nu})\,)^2}, \tag{30}$$

$$d\boldsymbol{G}_{II}^{\nu} \equiv \sum_{i=1}^{n_2} \frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{II},\nu})\,)^2}, \tag{31}$$

$$d\boldsymbol{G}_{I,\sigma}^{\nu} \equiv \sum_{i=1}^{n_1} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \left(\mathrm{diag}(\boldsymbol{\pi}^{\mathrm{I},\nu})\, d\boldsymbol{\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{I},\nu})\,)^2} - \frac{\mathrm{diag}(d\boldsymbol{\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{I},\nu})} \right],$$

$$d\boldsymbol{G}_{II,\sigma}^{\nu} \equiv \sum_{i=1}^{n_2} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \left(\mathrm{diag}(\boldsymbol{\pi}^{\mathrm{II},\nu})\, d\boldsymbol{\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{II},\nu})\,)^2} - \frac{\mathrm{diag}(d\boldsymbol{\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{II},\nu})} \right],$$

$$d\boldsymbol{G}_{0}^{\nu} \equiv \sum_{i=1}^{n_1} \frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})^T}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} + \sum_{i=1}^{n_2} \frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2},$$

$$d\boldsymbol{G}_{0,\sigma}^{\nu} \equiv \sum_{i=1}^{n_1} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \left(\mathrm{diag}(\boldsymbol{\pi}^{0,\nu})\, d\boldsymbol{\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} - \frac{\mathrm{diag}(d\boldsymbol{\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \right] +$$

$$\sum_{i=1}^{n_2} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \left(\mathrm{diag}(\boldsymbol{\pi}^{0,\nu})\, d\boldsymbol{\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} - \frac{\mathrm{diag}(d\boldsymbol{\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu}); \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \right],$$

$$\boldsymbol{dG}^{\nu}_{\sigma,0} \equiv \sum_{i=1}^{n_1} \left[\frac{\mathrm{diag}(\,\boldsymbol{\pi}^{0,\nu}\,)\,\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\mu})^T}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})\,)^2} - \frac{\mathrm{diag}(\,\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})} \right] + \sum_{i=1}^{n_2} \left[\frac{\mathrm{diag}(\,\boldsymbol{\pi}^{0,\nu}\,)\,\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})\,)^2} - \frac{\mathrm{diag}(\,\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})} \right],$$

$$\boldsymbol{dG}^{\nu}_{\sigma} \equiv \mathrm{diag}(\boldsymbol{\pi}^{0,\nu}) \left\{ \sum_{i=1}^{n_1} \left[\frac{\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,\left(\mathrm{diag}(\boldsymbol{\pi}^{0,\nu})\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})\,)^2} - \frac{\mathrm{diag}(\,\boldsymbol{d^2\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})} \right] + \sum_{i=1}^{n_2} \left[\frac{\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,\left(\mathrm{diag}(\boldsymbol{\pi}^{0,\nu})\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})\,)^2} - \frac{\mathrm{diag}(\boldsymbol{d^2\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})} \right] \right\},$$

$$\boldsymbol{dG}^{\nu}_{I,\,\mu} \equiv \sum_{i=1}^{n_1} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,\left(\mathrm{diag}(\,\boldsymbol{\pi}^{\mathrm{I},\nu}\,)\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{\mathrm{I},\nu})\,)^2} - \frac{\mathrm{diag}(\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{\mathrm{I},\nu})} \right],$$

$$\boldsymbol{dG}^{\nu}_{II,\,\mu} \equiv \sum_{i=1}^{n_2} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,\left(\mathrm{diag}(\,\boldsymbol{\pi}^{\mathrm{II},\nu}\,)\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{\mathrm{II},\nu})\,)^2} - \frac{\mathrm{diag}(\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{\mathrm{II},\nu})} \right],$$

$$\boldsymbol{dG}^{\nu}_{0,\,\mu} \equiv \sum_{i=1}^{n_1} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,\left(\mathrm{diag}(\,\boldsymbol{\pi}^{0,\nu}\,)\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})\,)^2} - \frac{\mathrm{diag}(\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})} \right] + \sum_{i=1}^{n_2} \left[\frac{\boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu})\,\left(\mathrm{diag}(\,\boldsymbol{\pi}^{0,\nu}\,)\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})\,)^2} - \frac{\mathrm{diag}(\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\,\boldsymbol{\mu}^{\nu},\boldsymbol{\sigma}^{\nu},\boldsymbol{\pi}^{0,\nu})} \right],$$

$$
\begin{aligned}
\boldsymbol{dG}^{\nu}_{\sigma,\mu} \equiv \operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \Bigg\{ \sum_{i=1}^{n_1} \Bigg[& \frac{\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \left(\operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^{T}}{(\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu}))^2} \\
& - \frac{\operatorname{diag}(\boldsymbol{d^2\varphi}_{\sigma,\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] + \\
& \sum_{i=1}^{n_2} \Bigg[\frac{\boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \left(\operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^{T}}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} \\
& - \frac{\operatorname{diag}(\boldsymbol{d^2\varphi}_{\sigma,\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] \Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{dG}^{\nu}_{I,w} \equiv \sum_{i=1}^{n_1} \Bigg[& \frac{(\boldsymbol{\pi}^{\mathrm{I},\nu})^{T} \boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{I},\nu})\,)^2} \boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \\
& - \frac{\boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{I},\nu})} \Bigg] \boldsymbol{X}^{\mathrm{I}}_{i\cdot},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{dG}^{\nu}_{II,w} \equiv \sum_{i=1}^{n_2} \Bigg[& \frac{(\boldsymbol{\pi}^{\mathrm{II},\nu})^{T} \boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{II},\nu})\,)^2} \boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \\
& - \frac{\boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{\mathrm{II},\nu})} \Bigg] \boldsymbol{X}^{\mathrm{II}}_{i\cdot},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{dG}^{\nu}_{0,w} \equiv \sum_{i=1}^{n_1} \Bigg[& \frac{(\boldsymbol{\pi}^{0,\nu})^{T} \boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} \boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \\
& - \frac{\boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] \boldsymbol{X}^{\mathrm{I}}_{i\cdot} + \\
& \sum_{i=1}^{n_2} \Bigg[\frac{(\boldsymbol{\pi}^{0,\nu})^{T} \boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} \boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \\
& - \frac{\boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] \boldsymbol{X}^{\mathrm{II}}_{i\cdot},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{dG}^{\nu}_{\sigma,w} \equiv \operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \Bigg\{ \sum_{i=1}^{n_1} \Bigg[& \frac{(\boldsymbol{\pi}^{0,\nu})^{T} \boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} \boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \\
& - \frac{\boldsymbol{d^2\varphi}_{\sigma,x}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] \boldsymbol{X}^{\mathrm{I}}_{i\cdot} + \\
& \sum_{i=1}^{n_2} \Bigg[\frac{(\boldsymbol{\pi}^{0,\nu})^{T} \boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} \boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \\
& - \frac{\boldsymbol{d^2\varphi}_{\sigma,x}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] \boldsymbol{X}^{\mathrm{II}}_{i\cdot} \Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{dH}_0^{\nu} \equiv \sum_{i=1}^{n_1} \Bigg[& \frac{\operatorname{diag}(\boldsymbol{\pi}^{0,\nu})\, \boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \boldsymbol{\varphi}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\mu})^T}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} \\
& - \frac{\operatorname{diag}(\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] + \\
\sum_{i=1}^{n_2} \Bigg[& \frac{\operatorname{diag}(\boldsymbol{\pi}^{0,\nu})\, \boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\, \boldsymbol{\varphi}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\mu})^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} \\
& - \frac{\operatorname{diag}(\,\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg],
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{dH}_{\sigma}^{\nu} \equiv \operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \Bigg\{ \sum_{i=1}^{n_1} \Bigg[& \frac{\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \left(\operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^T}{(\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu}))^2} \\
& - \frac{\operatorname{diag}(\,\boldsymbol{d^2\varphi}_{\mu,\sigma}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\,)}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] + \\
\sum_{i=1}^{n_2} \Bigg[& \frac{\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \left(\operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \boldsymbol{d\varphi}_{\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} \\
& - \frac{\operatorname{diag}(\boldsymbol{d^2\varphi}_{\mu,\sigma}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] \Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{dH}_{\mu}^{\nu} \equiv \operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \Bigg\{ \sum_{i=1}^{n_1} \Bigg[& \frac{\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \left(\operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^T}{(\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu}))^2} \\
& - \frac{\operatorname{diag}(\boldsymbol{d^2\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] + \\
\sum_{i=1}^{n_2} \Bigg[& \frac{\boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \left(\operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})\right)^T}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2} \\
& - \frac{\operatorname{diag}(\boldsymbol{d^2\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}))}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] \Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{dH}_{w}^{\nu} \equiv \operatorname{diag}(\boldsymbol{\pi}^{0,\nu}) \Bigg\{ \sum_{i=1}^{n_1} \Bigg[& \frac{(\boldsymbol{\pi}^{0,\nu})^T \boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{(\,\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2}\, \boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \\
& - \frac{\boldsymbol{d^2\varphi}_{\mu,x}(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{\psi(\chi_i^{\mathrm{I}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] \boldsymbol{X}_{i\cdot}^{\mathrm{I}} + \\
\sum_{i=1}^{n_2} \Bigg[& \frac{(\boldsymbol{\pi}^{0,\nu})^T \boldsymbol{d\varphi}_{x}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{(\,\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})\,)^2}\, \boldsymbol{d\varphi}_{\mu}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}) \\
& - \frac{\boldsymbol{d^2\varphi}_{\mu,x}(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu})}{\psi(\chi_i^{\mathrm{II}}(\boldsymbol{w}^{\nu});\, \boldsymbol{\mu}^{\nu}, \boldsymbol{\sigma}^{\nu}, \boldsymbol{\pi}^{0,\nu})} \Bigg] \boldsymbol{X}_{i\cdot}^{\mathrm{II}} \Bigg\}.
\end{aligned}
$$

References

1. Aldershof, B., Marron, J.S., Park, B.U., Wand, M.P. (1991): Facts about the Gaussian probability density function. Unpublished manuscript
2. Bradley, P.S., Fayyad, U.M., Mangasarian, O.L. (1999): Data mining: overview and optimization opportunities. INFORMS Journal on Computing **11**, 217–238
3. Devroye, L., Gyorfi, L., Lugosi, G. (1996): A Probabilistic Theory of Pattern Recognition. Springer, New York
4. Fisher, A.R. (1936): The use of multiple measurements in taxonomic problems. Annals of Eugenics **7**, 179–188
5. Lindsay, B.G. (1995): Mixture Models: Theory, Geometry and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5, Institute of Mathematical Sciences, Hayward
6. Luo, Z.Q., Pang, J.S., Ralph, D. (1996): Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge
7. McLachlan, G.J., Basford, K.E. (1988): Mixture Models. Marcel Dekker, New York
8. Priebe, C.E. (1996): Nonhomogeneity analysis using borrowed strength. Journal of the American Statistical Association **91**, 1497–1503.
9. Priebe, C.E., Marchette, D.J. (2000): Alternating kernel and mixture density estimates. Computational Statistics & Data Analysis **35**, 43–65
10. Priebe, C.E., Pang, J.S., Olson, T., Olson, T. (1998): Likelihood constrained bilevel optimization of classification performance. Manuscript, Department of Mathematical Sciences, The Johns Hopkins University, Maryland
11. Ripley, B.D. (1997): Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge
12. Saito, N., Coifman, R.R. (1994): Local discriminant bases. SPIE Proc. **2303**, 2–14
13. Scheel, H., Scholtes, S. (2000): Mathematical programs with equilibrium constraints: stationarity, optimality, and sensitivity. Mathematics of Operations Research **25**, 1–22
14. Hastie, T., Tibsharani, R. (1998): Classification by pairwise coupling. Annals of Statistics **26**, 451–471