On Separation of Signal Sources Using Kernel Estimates of Probability Densities

Oleg Michailovich[1] and Douglas Wiens[2]

[1] University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[2] University of Alberta, Edmonton, Alberta T6G 2G1, Canada

Abstract. The discussion in this paper revolves around the notion of *separation problems*. The latter can be thought of as a unifying concept which includes a variety of important problems in applied mathematics. Thus, for example, the problems of classification, clustering, image segmentation, and discriminant analysis can all be regarded as separation problems in which one is looking for a decision boundary to be used in order to separate a set of data points into a number of (homogeneous) subsets described by different conditional densities. Since, in this case, the decision boundary can be defined as a hyperplane, the related separation problems can be regarded as *geometric*. On the other hand, the problems of source separation, deconvolution, and independent component analysis represent another subgroup of separation problems which address the task of separating *algebraically* mixed signals. The main idea behind the present development is to show conceptually and experimentally that both geometric and algebraic separation problems are very intimately related, since there exists a general variational approach based on which one can recover either geometrically or algebraically mixed sources, while only little needs to be modified to go from one setting to another.

1 Introduction

Let $X = \{x_i \in \mathbb{R}^d, i = 1, \ldots, N\}$ be a set of N observations of a random variable $\mathcal{X}$ which is described by M conditional densities $\{p_k(x) \stackrel{def}{=} p(x \mid \mathcal{X} \in C_k)\}_{k=1}^{M}$, with C_k denoting a *class* to which a specific realization of $\mathcal{X}$ may belong. In other words, the set X can be viewed as a *mixture* of realizations of M random variables associated with different classes described by corresponding probability densities. In this case, the problem of classification (or, equivalently, *separation*) refers to the task of ascribing each observation x_i to the class C_k which it has *most likely* come from. The most challenging version of the above problem occurs in the case when the decision has to be made given the observed set X alone.

The setting considered above is standard for a variety of important problems in applied mathematics. Probably, the most famous examples here are unsupervised machine learning and data clustering [1,2]. Signal detection and image segmentation are among other important examples of the problems which could be embedded into the same separation framework [3,4]. It should be noted that, although a multitude of different approaches have been proposed previously to

M.E. Davies et al. (Eds.): ICA 2007, LNCS 4666, pp. 81–88, 2007.

address the above problems, most of them are similar at the conceptual level. Specifically, viewing the observations $\{x_i\}$ as points on either a linear or a non-linear manifold Ω, the methods search for such a partition of the latter so that the points falling at different subsets of Ω are most likely associated with different classes C_k. Moreover, the boundaries of the partition, which are commonly referred to as *decision boundaries*, are usually defined by means of geometric descriptors. The latter, for example, can be hyperplanes in machine learning [1] or active contours [4] in image segmentation. For this reason, we refer to the problems of this type as the problems of *geometric source separation* (GSS), in which case the data set X is considered to be a *geometric mixture* of unknown sources.

In parallel to the case of GSS, there exists an important family of problems concerned with separating sources that are mixed *algebraically* [5]. In a canonical setting, the problem of *algebraic source separation* (ASS) can be formulated as follows. Let $\mathbf{S}$ be a vector of M signals (sources) $[s_1(t), s_2(t), \ldots, s_M(t)]^T$, with $t = 1, \ldots, T$ being either a temporal or a spatial variable. Also, let $\mathbf{A} \in \mathbb{R}^{M\times M}$ be an unknown *mixing* matrix of full rank. Subsequently, the problem of blind source separation consists in recovering the sources given an observation of their *mixtures* $\mathbf{X} = [x_1(t), x_2(t), \ldots, x_M(t)]^T$ acquired according to[1]:

$$\mathbf{X} = \mathbf{A}\,\mathbf{S}. \tag{1}$$

Note that, in (1), neither the sources $\mathbf{S}$ nor the matrix $\mathbf{A}$ are known, and hence the above estimation problem is conventionally referred to as *blind*. Note that the problem of (algebraic) blind source separation constitutes a specific instance of *Independent Component Analysis*, which is a major theory encompassing a great number of applications [5]. Moreover, when $M = 1$ and $\mathbf{A}$ is defined to be a convolution operator, the resulting problem becomes the problem of blind deconvolution [6], which can also be inscribed in our framework of separation problems.

The main purpose of this paper is to show conceptually and experimentally that both GSS and ASS problems are intimately interrelated, since they can be solved using the same tool based on variational analysis [7]. To define this tool, let us first introduce an abstract, geometric separation operator $\varphi : X \mapsto \{S_k\}_{k=1}^M$ that "sorts" the points of X into M complementary and mutually exclusive subsets $\{S_k\}_{k=1}^M$ which represent estimates of the geometrically mixed sources. On the other hand, in the case of ASS, the separation operator is defined algebraically as a de-mixing matrix $\mathbf{W} \in \mathbb{R}^{M\times M}$ such that:

$$\mathbf{S} \simeq \mathbf{W}\,\mathbf{X}, \tag{2}$$

with $\mathbf{S}$ and $\mathbf{X}$ defined to be $\mathbf{S} = [s_1(t), \ldots, s_M(t)]^T$ and $\mathbf{X} = [x_1(t), \ldots, x_M(t)]^T$, respectively.

Additionally, let y_k be an estimate of either a geometric or an algebraic k-th source signal, computed via applying either φ or $\mathbf{W}$ to the data. This estimate

[1] Here and hereafter, the matrix $\mathbf{A}$ is assumed to be square which is merely a technical assumption which can be dropped; this is discussed in the sequel.

can be characterized by its *empirical* probability density function (*pdf*) which can be computed as given by:

$$\tilde{p}_k(z) = \frac{1}{N_k} \sum_{t=1}^{N_k} K(z - y_k(t)), \quad z \in \mathbb{R}^d, \tag{3}$$

where N_k is the size of the estimate (that is independent of k in the case of ASS). Note that (3) defines a *kernel based estimate* of the *pdf* of y_k when the *kernel* function $K(z)$ is normalized to have unit integral [8]. There exist a number of possibilities for choosing $K(z)$, among which the most frequent one is to define the kernel in the form of a Gaussian density function. Accordingly, this choice of $K(z)$ is used throughout the rest of this paper.

The core idea of the preset approach is quite intuitive and it is based on the assumption that the "overlap" between the informational contents of the estimated sources has to be minimal. To minimize this "overlap", we propose to find the optimal separation operator (viz. either φ or $\mathbf{W}$) as a minimizer of the cumulative Bhattacharyya coefficient between the empirical *pdf*s of the estimated sources, which is defined to be [9]:

$$B_M = \frac{2}{M(M-1)} \sum_{i<j} \int_{\mathbb{R}^d} \sqrt{\tilde{p}_i(z)\, \tilde{p}_j(z)}\, dz, \quad i, j = 1, \ldots, M. \tag{4}$$

It should be noted that, apart from the Bhattacharyya coefficient, a number of alternative metrics are available to assess the distance between the probability densities. Thus, for example, the Kullback-Leibler divergence was employed in [10] and [5] to solve the problems of image segmentation and blind source separation, respectively. However, for the reasons discussed below, we prefer using (4), since it results in comparatively more stable and reliable separation. To demonstrate how B_M can be used to unify the concept of separation, as it appears in both geometric and algebraic settings, we turn to some specific examples, among which the problem of image segmentation is chosen to be first.

2 Geometric Source Separation: Image Segmentation

In order to facilitate the discussion, we confine the derivations below to the case of two segmentation classes. In this case, the values of a vector-valued image $I(u) : \Omega \subseteq \mathbb{R}^2 \to \mathbb{R}^d$ are viewed as a geometric mixture of two sources, viz. the object of interest and its background. Consequently, the segmentation problem can be reformulated as the problem of partitioning the domain of definition Ω of $I(u)$ (with $u \in \Omega$) into two mutually exclusive and complementary subsets Ω_- and Ω_+. These subsets can be represented by their respective characteristic functions χ_- and χ_+, which can, in turn, be defined as $\chi_-(u) = \mathcal{H}(-\varphi(u))$ and $\chi_+(u) = \mathcal{H}(\varphi(u))$, with $\mathcal{H}$ standing for the Heaviside function.

Given a level-set function $\varphi(u)$, its *zero level set* $\{u \mid \varphi(u) \equiv 0, u \in \Omega\}$ is used to *implicitly* represent a curve – *active contour* – embedded into Ω. For the sake

of concreteness, we associate the subset Ω_- with the support of the object of interest, while Ω_+ is associated with the support of corresponding background. In this case, the objective of active-contour-based image segmentation is, given an initialization $\varphi_0(u)$, to construct a convergent sequence of level-set functions $\{\varphi_t(u)\}_{t>0}$ (with $\varphi_t(u)_{t=0} = \varphi_0(u)$) such that the zero level-set of $\varphi_\infty(u)$ coincides with the boundary of the object of interest.

The above sequence of level-set functions can be conveniently constructed using the variational framework. Specifically, the sequence can be defined by means of a *gradient flow* that minimizes the value of the cost functional (4). In the case of two segmentation classes, the optimal level set $\varphi^\star(u)$ is defined as:

$$\varphi^\star(u) = \arg\inf_{\varphi(u)} \{B_2(\varphi(u))\}, \tag{5}$$

where

$$B_2(\varphi(u)) = \int_{z\in\mathbb{R}^N} \sqrt{p_-(z \mid \varphi(u))\, p_+(z \mid \varphi(u))}\, dz. \tag{6}$$

with $p_-(z \mid \varphi(u))$ and $p_+(z \mid \varphi(u))$ being the kernel-based estimates of the *pdf*s of the class and background sources.

In order to contrive a numerical scheme for minimizing (5), its first variation should be computed first. The first variation of $B_2(\varphi(u))$ (with respect to $\varphi(u)$) can be shown to be given by:

$$\frac{\delta B_2(\varphi(u))}{\delta\varphi(u)} = \delta(\varphi(u))\, V(u), \tag{7}$$

where

$$V(u) = \frac{1}{2} B_2(\varphi(u))(A_-^{-1} - A_+^{-1}) + \frac{1}{2}\int_{z\in\mathbb{R}^d} K(z - I(u))\, L(z \mid \varphi(u))\, dz, \tag{8}$$

with

$$L(z \mid \varphi(u)) = \frac{1}{A_+}\sqrt{\frac{p_-(z \mid \varphi(u))}{p_+(z \mid \varphi(u))}} - \frac{1}{A_-}\sqrt{\frac{p_+(z \mid \varphi(u))}{p_-(z \mid \varphi(u))}}. \tag{9}$$

Note that, in the equations above, $\delta(\cdot)$ stands for the delta function, and A_- and A_+ are the areas of Ω_- and Ω_+ given by $\int_\Omega \chi_-(u)\, du$ and $\int_\Omega \chi_+(u)\, du$, respectively.

Finally, introducing an artificial time parameter t, the gradient flow of $\varphi(u)$ that minimizes (5) is given by:

$$\varphi_t(u) = -\frac{\delta B_2(\varphi(u))}{\delta\varphi(u)} = -\delta(\varphi(u))\, V(u), \tag{10}$$

where the subscript t denotes the corresponding partial derivative, and $V(u)$ is defined as given by (8).

It should be added that, in order to regularize the shape of the active contour, it is common to constrain its length and to replace the theoretical delta function

$\delta(\cdot)$ by its smoothed version $\bar{\delta}(\cdot)$. In this case, the final equation for the evolution of the active contour becomes:

$$\varphi_t(u) = \bar{\delta}(\varphi(u)) \left(\alpha\, \kappa(u) - V(u)\right), \tag{11}$$

where $\kappa(u)$ is the curvature of the active contour given by $\kappa(u) = -\text{div}\left\{\frac{\nabla\varphi(u)}{\|\nabla\varphi(u)\|}\right\}$ and $\alpha > 0$ is a user-defined regularization parameter. Note that, in the segmentation results reported in this paper, α was set to be equal to 1.

3 Blind Separation of Algebraically Mixed Sources

It is surprising how little has to be done to modify the separation approach of the previous section to suit the ASS setting. Indeed, let $\mathbf{Y} = [y_1(t), y_2(t), \ldots, y_M(t)]^T$ be the matrix of estimated sources computed as $\mathbf{Y} = \mathbf{W}\,\mathbf{X}$. Additionally, let $\{p(z; \mathbf{w}_i) \stackrel{def}{=} \tilde{p}_i(z \mid \mathbf{W})\}_{i=1}^{M}$ (where $\mathbf{w}_i^T$ is the i^{th} row of $\mathbf{W}$) be the set of empirical densities computed as at (3) and that correspond to the source estimates in $\mathbf{Y}$. Consequently, the optimal separation matrix $\mathbf{W}^*$ can be found as:

$$\mathbf{W}^\star = \arg\inf_{\mathbf{W}}\{B_M(\mathbf{W})\}, \tag{12}$$

where

$$B_M(\mathbf{W}) = \frac{2}{M(M-1)} \int_{\mathbb{R}^d} \sum_{i<j} \sqrt{p(z; \mathbf{w}_i)\, p(z; \mathbf{w}_j)},\ i, j = 1, \ldots M. \tag{13}$$

It should be noted that intrinsic in blind (algebraic) source separation is the problem of permutation and normalization, as, using (2), the sources can only be recovered in an arbitrary order and up to arbitrary multiplication factors. While the order of the sources is rarely of importance, the normalization could become an issue, especially from the viewpoint of numerical minimization. To overcome this difficulty, it is common to *prewhiten* the mixtures X before they are passed into the computations. In this case, it can be easily shown that the optimal solution $\mathbf{W}^\star$ becomes a member of the orthogonal group $\mathbf{O}(M) = \{\mathbf{W} \in \mathbb{R}^{M\times M} \mid \mathbf{W}\mathbf{W}^T = \mathbf{I}\}$.

We solve this constrained minimization problem with the aid of Lagrange multipliers $\{\lambda_{\alpha\beta}\}_{\alpha\leq\beta}$. Consider the problem of minimizing

$$F\left(\mathbf{w}_1, ..., \mathbf{w}_M, \lambda\right) = B_M(\mathbf{W}) + \sum_{\alpha\leq\beta} \lambda_{\alpha\beta} \left(\mathbf{w}_\alpha^T \mathbf{w}_\beta - \delta_{\alpha\beta}\right), \tag{14}$$

where $\delta_{\alpha\beta}$ is Kronecker's delta. Solving the equations

$$\frac{\partial}{\partial \mathbf{w}_i} F\left(\mathbf{w}_1, ..., \mathbf{w}_M, \lambda\right) = \mathbf{0}^T\ (\in \mathbb{R}^{1\times M}), \tag{15}$$

together with $\mathbf{W}\mathbf{W}^T = \mathbf{I}$ (details available from authors) leads to the characterization of $\mathbf{W}^*$ as a fixed point of the function

$$G(\mathbf{W}) = (\mathbf{P}\mathbf{P}^T)^{-1/2}\mathbf{P}, \tag{16}$$

where $(\mathbf{P}\mathbf{P}^T)^{1/2}$ is a *symmetric* square root and $\mathbf{P} = \mathbf{P}(\mathbf{W})$ is defined as follows. Let $\dot{\mathbf{K}}(z)$ be the $N \times M$ matrix with $(i,j)^{th}$ element $K'(z - \mathbf{y}_j(i))$ and let $\mathbf{D}(z)$ be the diagonal matrix with diagonal elements

$$d_i(z) = \frac{\sum_{j=1,\ldots,M; j\neq i} \sqrt{p(z;\mathbf{w}_j)}}{\sqrt{p(z;\mathbf{w}_i)}}. \tag{17}$$

Then

$$\mathbf{P}^T = \frac{1}{NM\,(M-1)}\mathbf{X}\int_{\mathbb{R}^d} \dot{\mathbf{K}}(z)\,\mathbf{D}(z)\,dz. \tag{18}$$

We solve (16) by iteration:

1. Initialize $\mathbf{W}$, say $\mathbf{W}_{(0)} = \mathbf{I}_M$.
2. For $l = 0, 1, \ldots$ to convergence, compute $\mathbf{P}_{(l)}$ from (18), and update $\mathbf{W}_{(l)}$ to $\mathbf{W}_{(l+1)} = G(\mathbf{W}_{(l)}) = (\mathbf{P}_{(l)}\mathbf{P}_{(l)}^T)^{-1/2}\mathbf{P}_{(l)})$.

4 Results

4.1 Image Segmentation

The image of Lizard shown in Subplot A of Fig. 1 is considered to be relatively hard to segment due to the multimodality of the *pdf* related to the object class. Moreover, the intensity distributions of the object and background classes of the image are very similar, which makes it impossible to segment the image based on gray-level information alone. To overcome this difficulty, the input image $I(u)$ was defined to be the bivariate image of the partial derivatives of Lizard, which are shown in Subplots B and C of the figure.

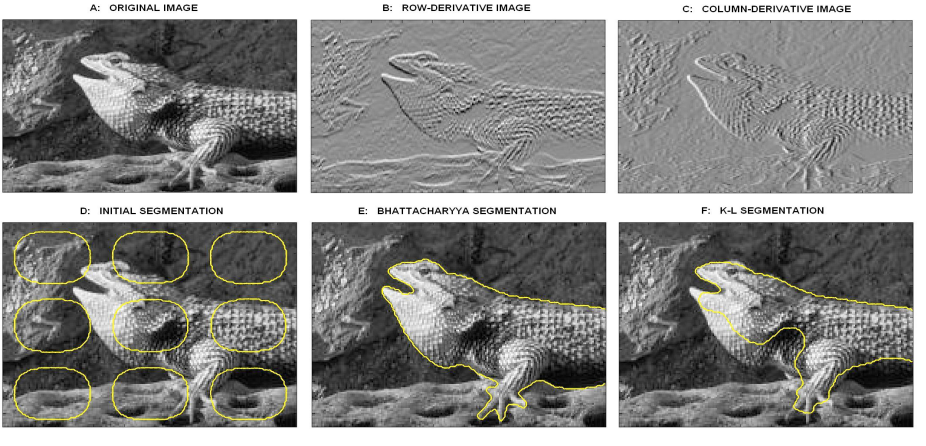

Fig. 1. (*Subplot A*) Original image of Lizard; (*Subplot B*) Row-derivative of the image; (*Subplot C*) Column-derivative of the image; (*Subplot D*) Initial segmentation; (*Subplot E*) Separation by the Bhattacharyya flow; (*Subplot F*) Separation by the K-L flow

The initial segmentation of Lizard and its segmentation obtained using the proposed method are shown in Subplots D and E of Fig. 1, respectively. For the sake of comparison, we have also segmented the image of Lizard using the active contour that maximized the Kullback-Leibler (K-L) divergence between the empirical *pdf*'s of the object and background classes. The resulting segmentation is shown in Subplot F of Fig.1. It is obvious that the proposed approach (i.e., the one that exploits the Bhattacharyya metric) is the best performer here.

It is worthwhile noting that the relatively worse performance of the image segmentation using the K-L divergence seems to be stemming from the properties of the functions involved in its definition, viz. of the logarithm. In particular, the latter is known to be very sensitive to variations of its argument in vicinity of relatively small values of the latter. Moreover, the logarithm is undefined at zero, which makes computing the K-L gradient flow prone to the errors caused by inaccuracies in estimating the *tails* of probability densities. On the other hand, the square root is a well-defined function in vicinity of zero. Moreover, for relatively small values of its argument, the variability of the square root is considerably smaller than that of the logarithm. As a result, the Bhattacharyya flow is much less susceptible to the influence of the inaccuracies mentioned above.

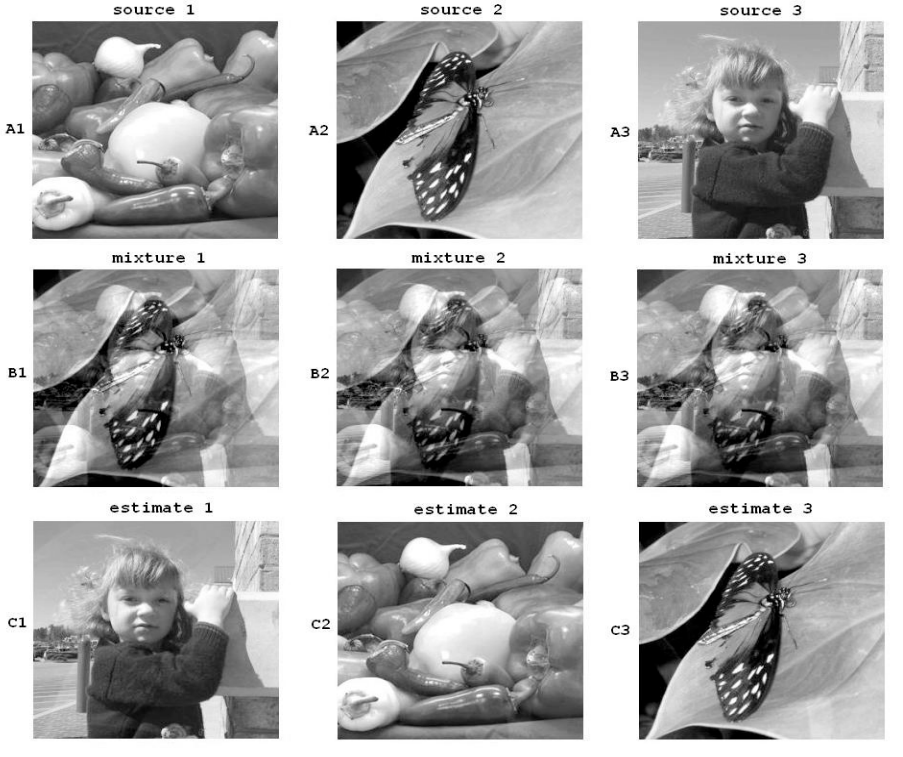

Fig. 2. (*Subplots A1-A3*) Original image sources; (*Subplot B1-B3*) Corresponding mixtures; (*Subplot C1-C3*) Estimated sources

4.2 Blind Source Separation

Subplots A1-A3 of Fig. 2 show the original source images which have been used to test the performance of the proposed separation methodology. The corresponding mixtures obtained using a random mixing matrix $\mathbf{A}$ are shown in Subplots B1-B3 of the same figure, whereas Subplots C1-C3 of Fig.2 show the source images estimated by applying 50 iterations of the fixed point algorithm described in Section 3. One can see that the algorithm results in virtually perfect reconstruction of the image sources. For this case, the average interference-to-signal ratio (ISR) was found to be equal to 0.0024, while minimizing the mutual information between the estimated sources resulted in ISR equal to 0.036.

5 Conclusions

The present study has demonstrated the applicability and practicability of the method for separating different components of a data signal based on the notion of a distance between probability distributions. The latter was defined by means of the Bhattacharyya coefficient which was shown to be advantageous over the K-L divergence (and, hence, over the related criterion of mutual information) in practical settings, in which class-conditional densities have to be estimated in a non-paramentric manner. Additionally, the versatility of the proposed criterion was demonstrated via its application to the problems of blind separation of both geometrically and algebraically mixed sources. Thus, from a certain perspective, the proposed method can be seen as unifying for the problems of both classes.

References

1. Duda, R., Hart, R., Stork, D.: Pattern Recognition. Willey, New York (2001)
2. Han, J., Kamber, M.: Data mining: Concepts and techniques. Morgan Kaufmann, San Francisco (2001)
3. Tuzlukov, V.: Signal detection theory. Springer, Heidelberg (2001)
4. Sapiro, G.: Geometric partial differential equations and image analysis. Cambridge University Press, Cambridge (2001)
5. Hyvarinen, A., Karhunen, J., Oja, E.: Independent component analysis. John Wiley and Sons, Chichester (2001)
6. Haykin, S.: Blind deconvolution. Prentice Hall, Englewood Cliffs (1994)
7. Gelfand, I., Fomin, S., Silverman, R.: Calculus of variations. Prentice-Hall, Englewood Cliffs (1975)
8. Silverman, B.: Density estimation for statistics and data analysis. CRC Press, Boca Raton (1986)
9. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc. 78, 99–109 (1943)
10. Kim, L., Fisher, J., Yezzi, A., Cetin, M., Willsky, A.: A nonparametric statistical method for image segmentation using information theory and curve evolution IEEE Proc. Image Processing 4(10), 1486–1502 (2005)

Some omitted details re

"On Geometric and Algebraic Separation of Signal Sources using Kernel Estimates of Probability Densities",

by Oleg Michailovich and Douglas Wiens.

Let $\mathbf{y}_i^T$ and $\mathbf{w}_i^T$ be the i^{th} rows of $\mathbf{Y}_{M\times N}$ and $\mathbf{W}_{M\times M}$ respectively. Then $\mathbf{y}_i^T = \mathbf{w}_i^T\mathbf{X}$ for $i = 1, ..., M$. Subject to

$$\mathbf{W}\mathbf{W}^T = \mathbf{I}, \tag{1}$$

we aim to minimize

$$B_M(\mathbf{W}) = \frac{2}{M(M-1)} \sum_{i<j} \int \sqrt{p(z;\mathbf{w}_i)p(z;\mathbf{w}_j)}dz,$$

where

$$p(z;\mathbf{w}_i) \stackrel{def}{=} p_i(z|\mathbf{W}) = \frac{1}{N}\sum_{k=1}^{N} K(z - y_{ik}).$$

For this, we first introduce Lagrange multipliers $\{\lambda_{\alpha\beta}\}_{\alpha\leq\beta}$ and consider the problem of minimizing

$$F(\mathbf{w}_1, ..., \mathbf{w}_M, \boldsymbol{\lambda}) = B_M(\mathbf{W}) + \sum_{\alpha\leq\beta} \lambda_{\alpha\beta}\left(\mathbf{w}_\alpha^T\mathbf{w}_\beta - \delta_{\alpha\beta}\right).$$

Here $\delta_{\alpha\beta}$ ($= 1$ if $\alpha = \beta$, $= 0$ otherwise) is Kronecker's delta. This leads to the equations

$$\frac{\partial}{\partial\mathbf{w}_i} F(\mathbf{w}_1, ..., \mathbf{w}_M, \boldsymbol{\lambda}) = \mathbf{0}^T : 1\times M\ , \tag{2}$$

together with (1). Equations (2) can be evaluated as follows. First define

$$\begin{aligned} \dot{\mathbf{p}}^T(z;\mathbf{w}_i) &= \frac{\partial p(z;\mathbf{w}_i)}{\partial\mathbf{w}_i} : 1\times M, \\ \mathbf{k}_i^T(z) &= \left(K'(z - y_{i1}), \cdots, K'(z - y_{iN})\right) : 1\times N, \end{aligned}$$

and denote by $\mathbf{x}_k$ the k^{th} column of $\mathbf{X}$. Then since $y_{ik} = \mathbf{w}_i^T\mathbf{x}_k$ we obtain

$$\dot{\mathbf{p}}^T(z;\mathbf{w}_i) = \frac{\partial}{\partial\mathbf{w}_i}\frac{1}{N}\sum_{k=1}^{N} K\left(z - \mathbf{w}_i^T\mathbf{x}_k\right) = \frac{-1}{N}\sum_{k=1}^{N} K'\left(z - \mathbf{w}_i^T\mathbf{x}_k\right)\mathbf{x}_k^T,$$

whence

$$\dot{\mathbf{p}}(z;\mathbf{w}_i) = -\mathbf{X}\mathbf{k}_i(z)/N.$$

Now write $\sum_{(i)}$ for $\sum_{j=1,...,M;j\neq i}$. Then with $\lambda_{\beta\alpha}$ defined as $\lambda_{\alpha\beta}$ when $\alpha > \beta$ we have

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}_i} F(\mathbf{w}_1, ..., \mathbf{w}_M, \boldsymbol{\lambda}) &= \frac{\partial}{\partial \mathbf{w}_i} \left\{ \frac{2}{M(M-1)} \sum_{\alpha<\beta} \int \sqrt{p(z;\mathbf{w}_\alpha) p(z;\mathbf{w}_\beta)} dz + \sum_{\alpha\leq\beta} \lambda_{\alpha\beta} \left(\mathbf{w}_\alpha^T \mathbf{w}_\beta - \delta_{\alpha\beta} \right) \right\} \\
&= \frac{\partial}{\partial \mathbf{w}_i} \left\{ \begin{array}{c} \frac{2}{M(M-1)} \sum_{(i)} \int \sqrt{p(z;\mathbf{w}_i) p(z;\mathbf{w}_j)} dz + \sum_{j=1}^{M} \lambda_{ij} \left(\mathbf{w}_i^T \mathbf{w}_j - \delta_{ij} \right) \\ + \text{ terms not involving } \mathbf{w}_i \end{array} \right\} \\
&= \frac{1}{M(M-1)} \sum_{(i)} \int \sqrt{\frac{p(z;\mathbf{w}_j)}{p(z;\mathbf{w}_i)}} \dot{\mathbf{p}}^T(z;\mathbf{w}_i) dz + \sum_{(i)} \lambda_{ij} \frac{\partial \mathbf{w}_i^T \mathbf{w}_j}{\partial \mathbf{w}_i} + \lambda_{ii} \frac{\partial \mathbf{w}_i^T \mathbf{w}_i}{\partial \mathbf{w}_i} \\
&= \frac{-1}{NM(M-1)} \sum_{(i)} \int \sqrt{\frac{p(z;\mathbf{w}_j)}{p(z;\mathbf{w}_i)}} \left(\mathbf{X}\mathbf{k}_i(z) \right)^T dz + \sum_{(i)} \lambda_{ij} \mathbf{w}_j^T + 2\lambda_{ii} \mathbf{w}_i^T \\
&= -\mathbf{a}_i^T + \sum_{(i)} \lambda_{ij} \mathbf{w}_j^T + 2\lambda_{ii} \mathbf{w}_i^T, \qquad (3)
\end{aligned}$$

where

$$\mathbf{a}_i = \frac{1}{NM(M-1)} \sum_{(i)} \int \sqrt{\frac{p(z;\mathbf{w}_j)}{p(z;\mathbf{w}_i)}} \mathbf{X}\mathbf{k}_i(z) dz : M \times 1. \qquad (4)$$

Let $\boldsymbol{\Lambda}$ be the triangular matrix with elements λ_{ij} $(i \leq j)$ on and above the main diagonal. Let $\boldsymbol{\Lambda}_S$ be the symmetric matrix $\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T$ and let $\mathbf{P}$ be the $M \times M$ matrix with rows $\mathbf{a}_i^T$. Then the equations (3) become

$$\mathbf{P} = \boldsymbol{\Lambda}_S \mathbf{W}.$$

The solution to these equations, together with (1), is easily seen to be

$$\begin{aligned}
\boldsymbol{\Lambda}_S &= \left(\mathbf{P}\mathbf{P}^T \right)^{1/2}, \\
\mathbf{W} &= \boldsymbol{\Lambda}_S^{-1} \mathbf{P}^T = \left(\mathbf{P}\mathbf{P}^T \right)^{-1/2} \mathbf{P},
\end{aligned}$$

where $\left(\mathbf{P}\mathbf{P}^T \right)^{1/2}$ is a *symmetric* square root.

We have shown that the solution $\mathbf{W}^*$ being sought is a fixed point of the function

$$G(\mathbf{W}) = \left(\mathbf{P}\mathbf{P}^T \right)^{-1/2} \mathbf{P}.$$

To write $\mathbf{P}$, hence $G(\mathbf{W})$, in a more compact form, let $\dot{\mathbf{K}}(z)$ be the $M \times M$ matrix with columns $\mathbf{k}_j(z)$, i.e. with $(i,j)^{th}$ element $K'\left(z - \mathbf{w}_j^T \mathbf{x}_i \right)$ and let $\mathbf{D}(z)$ be the diagonal matrix with diagonal elements

$$d_i(z) = \frac{\sum_{(i)} \sqrt{p(z;\mathbf{w}_j)}}{\sqrt{p(z;\mathbf{w}_i)}}.$$

Then from (4) we obtain

$$\mathbf{P}^T = \frac{1}{NM(M-1)} \mathbf{X} \int \dot{\mathbf{K}}(z) \mathbf{D}(z) dz. \qquad (5)$$