31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A Probabilistic Framework for Learning Robust Common Spatial Patterns

Wei Wu[1,2,3], Zhe Chen[1,3], Shangkai Gao[2], *Fellow*, *IEEE*, and Emery N. Brown[1,3], *Fellow*, *IEEE*

***Abstract*— Robustness in signal processing is crucial for the purpose of reliably interpreting physiological features from noisy data in biomedical applications. We present a robust algorithm based on the reformulation of a well-known spatial filtering and feature extraction algorithm named *Common Spatial Patterns* (CSP). We cast the problem of learning CSP into a probabilistic framework, which allows us to gain insights into the algorithm. To address the overfitting problem inherent in CSP, we propose an expectation-maximization (EM) algorithm for learning robust CSP using from a Student-t distribution. The efficacy of the proposed robust algorithm is validated with both simulated and real EEG data.**

I. Introduction

The Common Spatial Patterns (CSP) algorithm (also known as *Fukunaga-Koontz transform* in the machine learning field) was first proposed by Fukunaga and Koontz as an extension of Principal Component Analysis (PCA) for feature extraction [1], and since then it has been widely used in many fields, including digit and face recognition, target recognition, and identification of abnormal EEG patterns [2], [3], [4]. Notably, CSP has been successfully employed in brain-computer interfaces (BCIs) as a spatial filtering algorithm, as evidenced by recent international BCI competitions.

Given two classes of multivariable data, CSP aims to find a linearly transformed space where a good separability between the two classes can be attained. Mathematically, CSP is formulated as an optimization problem that maximizes the ratio of variance between one class of data and the other. On the face of it, CSP is viewed as a *discriminative* method for supervised learning without using probabilistic formulation. However, CSP is known to have three drawbacks. First, when the dimension of the observations (e.g., multi-channel EEG/MEG signals in BCIs) is high, CSP is prone to overfitting. Second, although CSP is a second-order statistics based algorithm, it is not robust to outliers in the data. Third, when CSP is used as a feature extraction method in BCI applications, the number of CSP components in the classification stage is often chosen in an ad-hoc manner. To our best knowledge, so far there has been no systematic analysis of CSP that deeply addresses the above issues.

The contribution of our paper is twofold. First, we provide a *generative* interpretation of CSP in that it can be equivalently derived as the maximum likelihood (ML) estimate of an underlying probability model. Mapping an existing algorithm to a probability model is desirable from both theoretical and practical viewpoints. From a statistical perspective, this allows us to examine and understand when the algorithm performs well or poorly. From a practical standpoint, associating a probability model with an algorithm leaves open the possibility of improving the algorithm by modifying the structure of the model. Second, we extend CSP to address the overfitting problem. Specifically, we use a Student-t distribution to model the data, thereby making the model more robust to outliers. A new expectation-maximization (EM) algorithm is developed for learning the robust CSP.

II. Methods

A. The Common Spatial Patterns Algorithm

Let us introduce the CSP algorithm in the context of EEG signal processing. Consider two classes of EEG signals $X^{(i)} \in \mathbb{R}^{C \times M_i} (i = 1, 2)$, where C and M_i denote the number of channels and sampled points, respectively. Without loss of generality, hereafter the signal in each channel is assumed to have zero mean. The spatial covariances for the two classes can then be computed as $\hat{R}^{(i)} = \frac{1}{M_i} X^{(i)} X^{(i)T} (i = 1, 2)$. The task of CSP is to find a linear transform by which the ratio of variance between the two classes can be maximized. Mathematically, this can be formulated as the following optimization problem

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \hat{R}^{(1)} \mathbf{w}}{\mathbf{w}^T \hat{R}^{(2)} \mathbf{w}} \qquad s.t. \quad ||\mathbf{w}|| = 1 \tag{1}$$

The solution can be obtained as the eigenvectors of the following generalized eigenvalue decomposition

$$\hat{R}^{(1)} W = \hat{R}^{(2)} W \Lambda \tag{2}$$

where Λ is a diagonal matrix with eigenvalues. Equivalently, the eigenvectors are given by joint diagonalization of the covariance matrices $\hat{R}^{(1)}$ and $\hat{R}^{(2)}$

$$W^T \hat{R}^{(i)} W = \Lambda^{(i)} (i = 1, 2) \tag{3}$$

B. A Generative View of CSP

We present a generative view of CSP, which casts the solution as an ML estimate from a probability model. The probability model is a mixture of two constrained factor analysis models [5], with each modeling one class of pattern:

$$\begin{aligned} X_k^{(i)} &= A Z_k^{(i)} + \Xi_k^{(i)} \\ Z_k^{(i)} &\sim \mathcal{N}(0, \Lambda^{(i)}), \Xi_k^{(i)} \sim \mathcal{N}(0, \Psi^{(i)}) (i = 1, 2) \end{aligned} \tag{4}$$

This work was supported by NIH Grants DP1-OD003646 and the National Natural Science Foundation of China under Grant 30630022.

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Department of Biomedical Engineering, Tsinghua University, Beijing 100084, China. [3]Neuroscience Statistics Research Lab, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA. (Email: weiwu@neurostat.mit.edu)

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

where $X_k^{(i)} (i = 1, 2)$ is the observation vector at the kth sample for class i. $Z^{(i)} \in \mathbb{R}^{S \times M_i}$ are the *factors*. The additive noise $\Xi^{(i)} \in \mathbb{R}^{C \times M_i}$ are the *specific factors*. $A \in \mathbb{R}^{C \times S}$ is the *factor loading* matrix (also known as the *mixing matrix* in the blind source separation literature) that contains spatial patterns, which is identical for the two classes. As a result, the factors and the mixing matrix are defined uniquely up to scaling and permutation indeterminancies, without the rotational indeterminancy as in the case of classical factor analysis. Matrices $\Lambda^{(1)}, \Lambda^{(2)}, \Psi^{(1)}, \Psi^{(2)}$ are all diagonal, implying that the factor variables are uncorrelated with each other, and that the observed variables are uncorrelated given the factors. The connection between the model (4) and CSP is revealed by the following theorem.

Theorem 1: The transformation matrix W in the CSP algorithm is equal to $\hat{A}^{-T}$, where $\hat{A}$ is the ML estimate of A in model (4) with two additional assumptions: (i) the additive noise vanishes to zero; (ii) A is a square matrix.

A sketchy proof of the theorem is presented in the appendix. In practice, the above two assumptions are hardly satisfied, which exactly make CSP suffer from overfitting. Another problem remains to be addressed is the sensitivity of CSP to outliers in the observed data, due to its underlying Gaussian assumption. These issues will be addressed below.

C. Robust CSP

The Student-t distribution is known to be able to be represented as the infinite mixture of Gaussian distributions that have the same mean but different variances controlled by a scaling variable [6]:

$$St(\boldsymbol{x}|\boldsymbol{\mu}, \Lambda, \nu) = \int_0^\infty \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, u^{-1}\Lambda)\text{Gam}(u|\frac{\nu}{2}, \frac{\nu}{2})du \tag{5}$$

where $St(\boldsymbol{x}|\boldsymbol{\mu}, \Lambda, \nu)$ denotes the multivariate Student-t distribution with mean $\boldsymbol{\mu}$, scale matrix Λ, and degrees of freedom (df) ν. $\text{Gam}(u|\frac{\nu}{2}, \frac{\nu}{2})$ denotes the Gamma distribution defined as $\text{Gam}(u|\alpha, \beta) = \frac{1}{\Gamma(\alpha)}\beta^\alpha u^{\alpha-1}e^{-\beta u}$.

The Student-t distribution provides a generalization of Gaussian distribution in that the additional parameter ν can be used to adjust the thickness of the tail in order to account for outliers. For each class, we model both the additive noise and the factors by Student-t distributions with the same df, as the marginal distribution of the observed variables in this case would still be a Student-t distribution.[1] Consequently, the robust probability model is given hierarchically by

$$\begin{aligned} &X_k^{(i)} = AZ_k^{(i)} + \Xi_k^{(i)} \\ &Z_k^{(i)} \sim \mathcal{N}(0, u_k^{(i)-1}\Lambda^{(i)}), \Xi_k^{(i)} \sim \mathcal{N}(0, u_k^{(i)-1}\Psi^{(i)}) \\ &u_k^{(i)} \sim \text{Gam}(\frac{\nu^{(i)}}{2}, \frac{\nu^{(i)}}{2}) \qquad (i = 1, 2) \end{aligned} \tag{6}$$

The ML estimates of the model parameters $\Theta = \{A, \Lambda^{(i)}, \Psi^{(i)}, \nu^{(i)}\}_{i=1}^2$ cannot be obtained in a closed form, we therefore resort to the EM algorithm [7] for iterative estimation. Due to space limit, we only give the final form of the EM algorithm by skipping step-by-step derivations.

E-step: Calculating the posterior distribution of the hidden variables $\{Z^{(i)}, \boldsymbol{u}^{(i)}\}$ given $X^{(i)}$ $(i = 1, 2)$ yields

$$\begin{aligned} &p(Z^{(i)}, \boldsymbol{u}^{(i)}|X^{(i)}) = p(\boldsymbol{u}^{(i)}|X^{(i)})p(Z^{(i)}|\boldsymbol{u}^{(i)}, X^{(i)}) \\ &p(\boldsymbol{u}^{(i)}|X^{(i)}) = \prod_{k=1}^{M_i} \text{Gam}(u_k^{(i)}|\alpha_k^{(i)}, \beta_k^{(i)}), \alpha_k^{(i)} = \frac{M_i + \nu^{(i)}}{2} \\ &\beta_k^{(i)} = \frac{1}{2}[X_k^{(i)T}[A\Lambda^{(i)}A^T + \Psi^{(i)}]^{-1}X_k^{(i)} + \nu^{(i)}] \\ &p(Z^{(i)}|\boldsymbol{u}^{(i)}, X^{(i)}) = \prod_{k=1}^{M_i} \mathcal{N}(Z_k^{(i)}|\boldsymbol{\mu}_k^{(i)}, \Sigma_k^{(i)}) \\ &\Sigma_k^{(i)} = u_k^{(i)-1}[A^T\Psi^{(i)-1}A + \Lambda^{(i)-1}]^{-1} \\ &\boldsymbol{\mu}_k^{(i)} = u_k^{(i)}\Sigma_k^{(i)}A^T\Psi^{(i)-1}X_k^{(i)} \end{aligned}$$

M-step: Maximizing the expectation of the complete log-likelihood with respect to the hidden variables yields

$$\begin{aligned} &\Lambda^{(i)} = \frac{1}{M_i}\sum_{k=1}^{M_i} \mathbb{E}[u_k^{(i)}]\text{diag}\{C_k^{(i)}\} \\ &\Psi^{(i)} = \frac{1}{M_i}\sum_{k=1}^{M_i} \mathbb{E}[u_k^{(i)}] \times \\ &\qquad \text{diag}\{X_k^{(i)}X_k^{(i)T} - X_k^{(i)}\boldsymbol{\mu}_k^{(i)T}A^T - A\boldsymbol{\mu}_k^{(i)}X_k^{(i)T} + AC_k^{(i)}A^T\} \\ &\boldsymbol{a}_j = (\sum_{i=1}^2\sum_{k=1}^{M_i} \mathbb{E}[u_k^{(i)}]\psi_j^{(i)-1}X_{jk}^{(i)}\boldsymbol{\mu}_k^{(i)})(\sum_{i=1}^2\sum_{k=1}^{M_i} \mathbb{E}[u_k^{(i)}]\psi_j^{(i)-1}C_k^{(i)})^{-1} \end{aligned}$$

where $C_k^{(i)} = \Sigma_k^{(i)} + \boldsymbol{\mu}_k^{(i)}\boldsymbol{\mu}_k^{(i)T}$, $\boldsymbol{a}_j$ is the jth row of A, $\psi_j^{(i)}$ is the jth diagonal element of $\Psi^{(i)}$, $X_{jk}^{(i)}$ is the jth element of $X_k^{(i)}$, $\text{diag}\{\mathcal{S}\}$ denotes a diagonal matrix with diagonal entries taken from the main diagonal of matrix $\mathcal{S}$. Note that A and Φ^i are coupled in the M-step iterations, hence must be solved alternately. In addition, $\nu^{(i)}$ is obtained by solving the following nonlinear equation

$$1 + \ln(\frac{\nu^{(i)}}{2}) - \digamma(\frac{\nu^{(i)}}{2}) + \frac{1}{M_i}\sum_{k=1}^{M_i}\left(\mathbb{E}[\ln(u_k^{(i)})] - \mathbb{E}[u_k^{(i)}]\right) = 0$$

where $\mathbb{E}[u_k^{(i)}] = \frac{\alpha_k^{(i)}}{\beta_k^{(i)}}, \mathbb{E}[\ln(u_k^{(i)})] = \digamma(\alpha_k^{(i)}) - \ln(\beta_k^{(i)})$, and $\digamma$ denotes the digamma function.

We refer to the above EM algorithm for learning the probabilistic model (6) the *robust CSP* algorithm. In M-step, we see that the posterior mean of the scaling variable $\mathbb{E}[u_k^{(i)}]$ acts as a weighting coefficient for each data point. The weighting coefficients are optimized so that the effect of outliers will be suppressed on the parameter estimation. Instead of taking binary values of 0 or 1, the continous-valued weighting coefficient reflects the confidence of excluding a data point as an outlier: the smaller the weighting coefficient, the more likely a data point is an outlier.

III. Results

A. Results on Synthetic Data

First, we compare the performance of the robust CSP algorithm and the standard CSP algorithm on the reconstruction accuracy of the mixing matrix via Monte Carlo simulations.

[1] The primary objective is to use the Student-t distribution to model the observed data that may lessen the unfavorable influence of the outliers in the data.

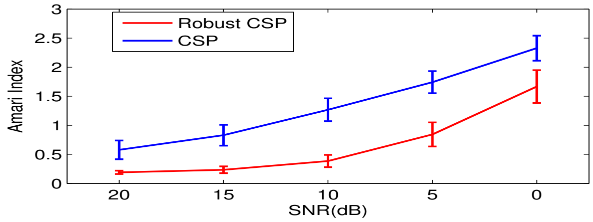

Fig. 1. Comparison of Amari indices between CSP and robust CSP at different signal-to-noise ratios (SNR).

In each run, two sets of 10 mutually uncorrelated factors are generated, with each set corresponding to one class. Each factor comprises $1,000$ data points that are independently and identically Gaussian distributed with zero mean. The standard deviations of the 10 factors for one class are in descending order from 10 to 1, while for the other class the standard deviations are ascending from 1 to 10. A 20×10 mixing matrix is also randomly generated, with each entry uniformly distributed within $[0, 1]$. Additive white Gaussian noise is simulated with varying SNR $20 \sim 0$ dB.

The noisy mixture signals are presented to both algorithms. For robust CSP, the dimension of the factors is assumed to be known (The issue of learning the factor dimension is discussed in Section IV). Since CSP specifically addresses the square mixing case (Theorem 1), we select the 10 columns that correspond to the 5 largest and the 5 smallest eigenvalues to form the estimated mixing matrix $\hat{A}$. The Amari index is used as a measure of the closeness of $\hat{A}$ and the true mixing matrix A, which is invariant to permutation and scaling of the columns of A and $\hat{A}$:

$$d(\hat{A}, A) = \frac{1}{2S}\left[\sum_{i=1}^{S}\frac{\sum_{j=1}^{S}|b_{ij}|}{\max_j |b_{ij}|} + \sum_{j=1}^{S}\frac{\sum_{i=1}^{S}|b_{ij}|}{\max_i |b_{ij}|} - 2S\right]$$

where $b_{ij} = ((A^T A)^{-1} A^T \hat{A})_{ij}$. The Amari indices averaged over 50 runs are plotted in Fig. 1. With no surprise, the index increases with increasing SNR for both methods. However, the robust CSP outperforms CSP in that the Amari index substantially reduces under the same SNR. Next, we test the performance of robust CSP when data are contaminated with outliers. The basic setting of the simulation is similar to the former case, except that: (i) for demonstration purpose, the simulation consists of only one Monte Carlo run; (ii) The mixing matrix of size 45×6 is generated from a real head model, with one column representing the spatial pattern of a cortical patch over the foot representation area, which is the pattern of interest (POI). The ratio of standard deviations between the first and the second class for the corresponding factor is set to 10:1. (iii) The SNR is fixed to 20 dB. The additive noise is a mixture of two Gaussians: one has a standard deviation of 1, while the other one has a standard deviation of 500 to simulate the impulsive noise. We consider two cases when the percentage of the impulse noise in the data are 1% and 5%, respectively. For comparison we visualize the spatial patterns (i.e., columns in the mixing matrix) associated with the POI that are computed by CSP and robust CSP. The spatial pattern is selected to be the column in A that has the maximum ratio of variance $\Lambda_{jj}^{(1)}/\Lambda_{jj}^{(2)}$. It was found in Fig. 2 that robust CSP in both cases performs fairly well even when the percentage of the impulse noise reaches 5%, while CSP fails in both cases in revealing the true spatial pattern.

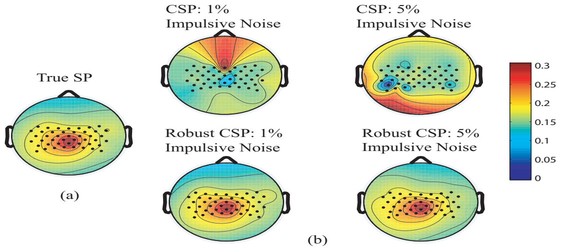

Fig. 2. (a) True spatial pattern (SP) of the POI; (b) Spatial patterns computed from CSP and robust CSP under different noise conditions.

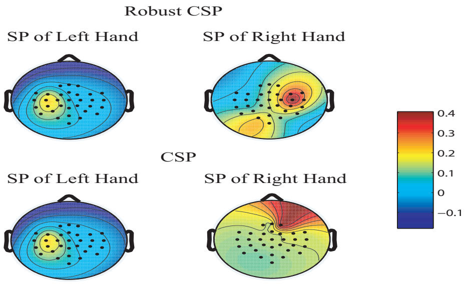

Fig. 3. Spatial patterns (SP) for left- and right-hand motor imageries computed by CSP and robust CSP from real EEG signals.

B. Results on Real EEG Data

We also demonstrate the efficacy of the robust CSP algorithm in tackling the outliers in real EEG recordings. Due to space limit, we only present the result on one dataset for the purpose of illustration. The EEG data here were recorded from a healthy female subject participating in a real-time BCI experiment, in which the task is to control the vertical movement (upward or downward) of a cursor on the screen via imagination of her left or right hand movement. The 32-channel EEG recordings (sampling rate 256 Hz) consisting of 20 trials (10 trials per class) in a single session is used for present analysis. Each trial lasts 2 s, during which the subject was performing the motor imagery task. Each class of EEG signals are band-pass filtered between 8 Hz and 30 Hz before being presented to CSP and robust CSP for analysis.

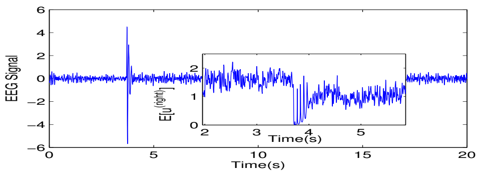

Fig. 4. The waveform of one EEG channel that was contaminated by strong outliers around 4 s. The inset figure highlights the weighting coefficients over the time period when outliers occur.

For robust CSP, the number of factors is determined by cross-validation.

The spatial pattern associated with left-hand motor imagery is selected to be the column in A that has the maximum ratio of variance $\Lambda_{jj}^{(left)}/\Lambda_{jj}^{(right)}$. Similar criterion is used for the right-hand spatial pattern. The resultant spatial patterns are shown in Fig. 3. The difference between the results of robust CSP and CSP lies primarily in the spatial pattern of right-hand motor imagery: in robust CSP the spatial pattern focuses on the true left-hand representation area (as expected from physiology [8]); whereas the spatial pattern found by CSP incorrectly centers on the frontal region. To understand why CSP fails to find the correct right-hand spatial pattern in this case, we closely inspected the EEG signal. It was found that during certain period ($3.6 \sim 4$ s) when the task was right-hand motor imagery, one channel of the recorded EEG signal was contaminated by outliers, as shown in Fig. 4. By observing the weighting coefficients $\mathbb{E}[u^{(i)}]$, we see that these outliers actually have negligible influence on the performance of robust CSP. For this subject, the BCI classification accuracies (using Fisher discriminant analysis on a separate test set of 20 trials) yield 85% for robust CSP and 65% for CSP, indicating a significant improvement in performance.

IV. Discussion and Future Work

The proposed probability model can be viewed as a generative counterpart of the discriminative model in [11]. However, the performance of the algorithm therein highly depends on the preprocessing of the data. For example, since *mu-rhythms* recorded by channels covering the motor cortex in scalp surface are typically rather weak in the case of motor imagery, the performance in [11] would degrade significantly without data prewhitening. By contrast, the ML solution of our probabilistic generative model is unaffected by the scaling of the coordinates in the data.

Our paper also serves as a basis for exploring various extensions. First, the proposed robust generative model can be extended to the state-space form to account for the temporal dynamics of data, as similarly derived in [9], [10] for the CSP counterpart. Second, the ML framework enables us to use unlabelled data for augmenting the labelled training data to perform semi-supervised learning in classification. Third, the extension of the current model to the multi-class case is straightforward. Fourth, an important question unanswered in this paper is the automatic determination of the number of factors in the model. A promising approach is Bayesian inference, in which hyperparameters can be introduced to control the number of factors in learning [12]. All the above directions are currently under investigation. We also plan to employ the algorithm to explore the differences in spatial patterns of brain activities at different stages of general anesthesia.

References

[1] F. Fukunaga and W. Koontz, "Applications of the Karhunen-Loève expansion to feature selection and ordering," *IEEE Trans. Comput.*, vol. 19, pp. 311-318, 1970.

[2] S. Zhang and T. Sim, "Discriminant subspace analysis: a Fukunaga-Koontz approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, pp. 1732-1745, 2007.

[3] X. Huo, "A statistical analysis of Fukunaga-Koontz transform," *IEEE Signal Proc. Letters*, vol. 11, pp. 123-126, 2004.

[4] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial patterns underlying population differences in the background EEG," *Brain Topo.*, vol. 2, pp. 275-284, 1990.

[5] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis* (3rd ed.), Wiley 2003.

[6] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t-distribution," *J. Ame. Stat. Assoc.*, vol. 84, pp. 881-896, 1989.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. B*, vol. 39, pp. 1-38, 1977.

[8] G. Pfurtscheller and A. Aranibar, "Event-related cortical desynchronization detected by power measurements of scalp EEG," *Electroencephalogr. Clin. Neurophysiol.*, vol. 42, pp. 817-826, 1977.

[9] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, "Spatio-spectral filters for improved classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, pp. 1541-1548, 2005.

[10] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 53, pp. 2274-2281, 2006.

[11] R. Tomioka and K. Aihara, "Classifying matrices with a spectral regularization," in *Proc. 24th Int. Conf. Machine Learning*, ACM Press, pp. 895-902, 2007.

[12] H. Attias, "A variational Bayesian framework for graphical models," in *Adv. Neural Info. Proc. Syst. 12*, pp. 209-215, 2000.

Appendix: Proof of Theorem 1

In the noiseless and square mixing matrix setup, the log-likelihood of the observed variables is

$$
\begin{aligned}
L=&\sum_{i=1}^{2}\sum_{k=1}^{M_i} p(X_k^{(i)}|A,\Lambda^{(1)},\Lambda^{(2)})\\
=&-\sum_{i=1}^{2}\frac{M_i}{2}[C\ln(2\pi)+\ln|R^{(i)}|+\mathrm{Tr}((R^{(i)})^{-1}\hat{R}^{(i)})]\\
=&-\sum_{i=1}^{2}\frac{M_i}{2}[\mathrm{Tr}(R^{(i)})^{-1}\hat{R}^{(i)})-\ln|(R^{(i)})^{-1}\hat{R}^{(i)}|-C]+\mathrm{Const}\\
=&-\sum_{i=1}^{2}[D_{\mathrm{KL}}(\hat{R}^{(i)}\|R^{(i)})]+\mathrm{Const}
\end{aligned}
$$

where $R^{(i)} = A\Lambda^{(i)}A^T$, and $D_{\mathrm{KL}}(\mathcal{S}_1\|\mathcal{S}_2)$ denotes the Kullback-Leibler (KL) divergence between two Gaussian distributions with covariance matrices $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. The last equality follows from the definition of KL divergence.

Because KL divergence is invariant to invertible linear transformation and the Pythagorean decomposition holds as the involved distributions are all Gaussian, the log-likelihood is rewritten as

$$
\begin{aligned}
L &= -\sum_{i=1}^{2} D_{\mathrm{KL}}(A^{-1}\hat{R}^{(i)}A^{-T}\|\Lambda^{(i)})+\mathrm{Const}\\
&= -\sum_{i=1}^{2}\Big[D_{\mathrm{KL}}(A^{-1}\hat{R}^{(i)}A^{-T}\|\mathrm{diag}(A^{-1}\hat{R}^{(i)}A^{-T}))\\
&\qquad +D_{\mathrm{KL}}(\mathrm{diag}\{A^{-1}\hat{R}^{(i)}A^{-T}\}\|\Lambda^{(i)})\Big]+\mathrm{Const}
\end{aligned}
$$

$\Lambda^{(1)}$ and $\Lambda^{(2)}$ are fully parameterized diagonal matrices, thus regardless of A the second KL divergence in the bracket can always be made exactly to zero. The first KL divergence will be zero if and only if $A^{-1}\hat{R}^{(i)}A^{T}$ is a diagonal matrix. In other words, the log-likelihood is maximized if and only if $\hat{R}^{(1)}$ and $\hat{R}^{(2)}$ are jointly diagonalized by A^{-T}. ■