Performance Limits on the Classification of Kronecker-structured Models

Ishan Jindal and Matthew Nokleby
Electrical and Computer Engineering
Wayne State University
Motivation: Subspace models

- (Union of) subspace models useful for signal processing/machine learning problems
- Represent signals via (overcomplete) dictionary:

\[ y = Ax + z \]

[Basri and Jacob, “Lambertian Reflectance Functions and Linear Subspaces”, 2003]
Motivation: Subspace models

- (Union of) subspace models useful for signal processing/machine learning problems
- Represent signals via (overcomplete) dictionary:
  \[ y = Ax + z \]
- Offers a compact representation
- Efficient algorithms for classification, clustering, and dictionary learning
- Good understanding of fundamental performance limits

[Zhang and Li, “Discriminative K-SVD for dictionary learning…” 2010]
[Heckel et al., “Dimensionality-reduced subspace clustering” 2015]
[MN, Rodrigues, Calderbank, "Discrimination on the Grassmann Manifold…” 2015]
Motivation: Kronecker-structured models

• Matrix-valued subspace model:

\[ Y = AXB^T + Z \]

• Constituent dictionaries for rows and columns
• Accounts for multidimensional data structure
• Useful for video inpainting, tomographic image reconstruction

[Tsiligkaridis and Hero "Covariance estimation … via Kronecker product expansion" 2013]
[Ha"we et al., “Separable dictionary learning” 2013]
[Shakeri et al., “Minimax lower bounds for Kronecker-structured dictionary learning” 2016]
Motivation: Kronecker-structured models

- Matrix-valued subspace model:
  \[ Y = AXB^T + Z \]

- Constituent dictionaries for rows and columns
- Accounts for multidimensional data structure
- Useful for video inpainting, tomographic image reconstruction

[Tsiligkaridis and Hero "Covariance estimation ... via Kronecker product expansion” 2013]
[Hawe et al., “Separable dictionary learning” 2013]
[Shakeri et al., “Minimax lower bounds for Kronecker-structured dictionary learning” 2016]
Motivation: Kronecker-structured models

- Matrix-valued subspace model:
  \[ Y = AXB^T + Z \]
- Constituent dictionaries for rows and columns
- Accounts for multidimensional data structure
- Useful for video inpainting, tomographic image reconstruction

[Tsiligkaridis and Hero "Covariance estimation … via Kronecker product expansion" 2013]
[Hawe et al., “Separable dictionary learning" 2013]
[Shakeri et al., “Minimax lower bounds for Kronecker-structured dictionary learning” 2016]
Motivation: Kronecker-structured models

• Matrix-valued subspace model:

\[ Y = AXB^T + Z \]

• Constituent dictionaries for rows and columns
• Accounts for multidimensional data structure
• Useful for video inpainting, tomographic image reconstruction

[Tsiligkaridis and Hero “Covariance estimation … via Kronecker product expansion” 2013]
[Hawe et al., “Separable dictionary learning” 2013]
[Shakeri et al., “Minimax lower bounds for Kronecker-structured dictionary learning” 2016]
Motivation: Kronecker-structured models

- Matrix-valued subspace model:
  \[ Y = AXB^T + Z \]

- Constituent dictionaries for rows and columns
- Accounts for multidimensional data structure
- Useful for video inpainting, tomographic image reconstruction

- Even more compact representation
- Efficient dictionary learning methods

- What is the **classification performance**?

  [Tsiligkaridis and Hero "Covariance estimation … via Kronecker product expansion" 2013]
  [Hawe et al., “Separable dictionary learning” 2013]
  [Shakeri et al., “Minimax lower bounds for Kronecker-structured dictionary learning” 2016]
Contributions

• Information-theoretic analysis of the **classification performance** of K-S models

Performance metrics:
• Diversity order: How fast does the classification error rate decay as the SNR approaches infinity?

Main results:
• Exact derivation of the diversity order
• Find the diversity gap with unstructured subspace models

Not in this talk:
• (Classification) capacity: How many K-S subspaces can one discern in the limit of large signal dimension?
Contributions

• Information-theoretic analysis of the **classification performance** of K-S models

Performance metrics:
• Diversity order: How fast does the classification error rate decay as the SNR approaches infinity?

Main results:
• Exact derivation of the diversity order
• Find the diversity gap with unstructured subspace models

Not in this talk:
• (Classification) capacity: How many K-S subspaces can one discern in the limit of large signal dimension?

[**MN**, Rodrigues, Calderbank, "Discrimination on the Grassmann Manifold…" 2015]
Problem Definition: Signal model

• Signal belongs to one of \( L \) classes
• Signal model for class \( 1 \leq l \leq L \):

\[
Y = A_l X B_l^T + Z
\]

- \( Y \in \mathbb{R}^{m_1 \times m_2} \): signal of interest
- \( A_l \in \mathbb{R}^{m_1 \times n_1} \): column subspace basis
- \( B_l \in \mathbb{R}^{m_2 \times n_2} \): row subspace basis
- \( X \in \mathbb{R}^{n_1 \times n_2} \): coefficients
- \( Z \in \mathbb{R}^{m_1 \times m_2} \): model noise
- \( n_1 < m_1, \quad n_2 < m_2 \)

• Signal **columns** live near an \( n_1 \)-dimensional subspace, signal **rows** live near an \( n_2 \)-dimensional subspace
Problem Definition: Signal model

- Useful to work with vectorized signal model:

\[ y = (B_l \otimes A_l)x + z, \]

\[ y = \text{vec}(Y) \in \mathbb{R}^M \quad z = \text{vec}(Z) \in \mathbb{R}^M \quad x = \text{vec}(X) \in \mathbb{R}^N \]

- Signal lives near an \( n_1 n_2 \)-dimensional subspace with a Kronecker-structured dictionary

- Classification is over a **restricted set** of subspace models
Problem Definition: Classification model

• Let coefficients and i.i.d. noise be Gaussian:
  \[ x \sim \mathcal{N}(0, I), \quad z \sim \mathcal{N}(0, \sigma^2 \cdot I) \]

• Noise variance quantifies deviation from K-S model

• Class-conditional distribution is Gaussian:
  \[ p(y|A_l, B_l) = \mathcal{N}(0, (B_l \otimes A_l)(B_l \otimes A_l)^T + \sigma^2 \cdot I) \]

• Equivalent to classification of a Gaussian mixture model with matrix normal components
Problem Definition: Classification model

• Let coefficients and i.i.d. noise be Gaussian:
  \[ x \sim \mathcal{N}(0, I), \quad z \sim \mathcal{N}(0, \sigma^2 \cdot I) \]

• Noise variance quantifies deviation from K-S model

• Class-conditional distribution is Gaussian:
  \[ p(y|A_l, B_l) = \mathcal{N}(0, (B_l \otimes A_l)(B_l \otimes A_l)^T + \sigma^2 \cdot I) \]

• Equivalent to classification of a Gaussian mixture model with matrix normal components

• Classifier estimates the class label \( l \) from \( y \) (e.g. by ML rule)

• Want to understand the probability of error:
  \[ P_e = \frac{1}{L} \sum_{l=1}^{L} \Pr(\hat{l} \neq l | y \sim p(y|A_l, B_l)) \]
• How fast does the error probability decay in the model noise?
• Fix $L$ and bases $A_l, B_l$ for each class

**Definition:** The **diversity order** is the exponent:

$$d = \lim_{\sigma^2 \to 0} \left( \frac{\log(P_e)}{\frac{1}{2} \log(1/\sigma^2)} \right)$$
Diversity Order

- How fast does the error probability decay in the model noise?
- Fix $L$ and bases $\mathbf{A}_l, \mathbf{B}_l$ for each class

**Definition:** The **diversity order** is the exponent:

$$d = \lim_{\sigma^2 \to 0} - \frac{\log(P_e)}{\frac{1}{2} \log(1/\sigma^2)}$$

- Isolates the impact of subspace model on performance
- For standard subspace models ($m_2=n_2=1$):
  $$d = \min\{n_1, m_1 - n_1\}$$
- Equivalent K-S subspace "conjecture":
  $$d = \min\{n_1n_2, m_1m_2 - n_1n_2\}$$

[MN, Rodrigues, Calderbank, "Discrimination on the Grassmann Manifold…" 2015]
Lemma [Bhattacharyya Bound]: For two zero-mean Gaussian classes with covariances $\Sigma_1$ and $\Sigma_2$ and equal priors, the pairwise maximum-likelihood classification error is bounded by:

$$P_e \leq \frac{1}{2} \left( \frac{\left| \frac{\Sigma_1 + \Sigma_2}{2} \right|}{\left| \Sigma_1 \right|^{1/2} \left| \Sigma_2 \right|^{1/2}} \right)^{-1/2}$$
Lemma [Bhattacharyya Bound]: For two zero-mean Gaussian classes with covariances $\Sigma_1$ and $\Sigma_2$ and equal priors, the pairwise maximum-likelihood classification error is bounded by:

$$P_e \leq \frac{1}{2} \left( \frac{\left| \frac{\Sigma_1 + \Sigma_2}{2} \right|}{\left| \Sigma_1 \right|^{1/2} \left| \Sigma_2 \right|^{1/2}} \right)^{-1/2}$$

- Tight w.r.t the exponent of the pairwise error as $P_e \rightarrow 0$
- Union bound is exponentially tight, too
- Diversity order: worst pairwise exponent of this bound

[Kailath, “The divergence and Bhattacharyya distance measures in signal selection” 1967]
Diversity Order: Bhattacharyya bound

- Recall the class-conditional distributions
  \[ p(y|A_l, B_l) = \mathcal{N}(0, (B_l \otimes A_l)(B_l \otimes A_l)^T + \sigma^2 \cdot I) \]

- Define Kronecker-structured dictionaries for each class:
  \[ D_l = B_l \otimes A_l \in \mathbb{R}^{m_1 m_2 \times n_1 n_2} \]
Diversity Order: Bhattacharyya bound

- Recall the class-conditional distributions
  \[ p(y|A_l, B_l) = \mathcal{N}(0, (B_l \otimes A_l)(B_l \otimes A_l)^T + \sigma^2 \cdot I) \]

- Define Kronecker-structured dictionaries for each class:
  \[ D_l = B_l \otimes A_l \in \mathbb{R}^{m_1 m_2 \times n_1 n_2} \]

- Pairwise error probability between classes i and j is bounded:
  \[ P_e(D_i, D_j) \leq \frac{1}{2} \left( \frac{|D_i D_i^T + D_j D_j^T + 2\sigma^2 I|}{|D_i D_i^T + \sigma^2 I|^\frac{1}{2} |D_j D_j^T + \sigma^2 I|^\frac{1}{2}} \right)^{-\frac{1}{2}} \]

- Simplifying, we obtain, for almost every class pair:
  \[ P_e(D_i, D_j) = 2^{\frac{n_1 n_2 - 2}{2}} \left( 1 + \frac{\lambda_{ij} r^*_{ij}}{\sigma^2} \right)^{-\frac{r^*_{ij} - n_1 n_2}{2}} \]

  \( r^*_{ij} : \) rank of \( D_i D_i^T + D_j D_j^T \)

  \( \lambda_{ij} r^*_{ij} \): smallest nonzero eigenvalue of \( D_i D_i^T + D_j D_j^T \)
Theorem: For any K-S classification problem, the diversity order is

\[ d(a) = r^* - n_1 n_2 \]

where

\[ r^* = \min_{i \neq j} \text{rank}(D_i D_i^T + D_j D_j^T) \]
Theorem: For any K-S classification problem, the diversity order is

\[ d(a) = r^* - n_1 n_2 \]

where

\[ r^* = \min_{i \neq j} \text{rank}(D_i D_i^T + D_j D_j^T) \]

- Almost every dictionary \( D_l \) has rank \( n_1 n_2 \)
- In general: \( \text{rank}(D_1 + D_2) = \text{rank}(D_1) + \text{rank}(D_2) - \dim(\mathcal{R}(D_1) \cap \mathcal{R}(D_2)) \)
Lemma: Let the intersection of row and column range spaces be
\[
\dim[\mathcal{R}(A_i) \cap \mathcal{R}(A_j)] = x
\]
\[
\dim[\mathcal{R}(B_i) \cap \mathcal{R}(B_j)] = y
\]
Then,
\[
\dim[\mathcal{R}(A_i \otimes B_i) \cap \mathcal{R}(A_j \otimes B_j)] = xy
\]

• Intuition: If row or column subspaces are disjoint, entire K-S subspaces are disjoint
Lemma: Let the intersection of row and column range spaces be
\[ \dim[\mathcal{R}(A_i) \cap \mathcal{R}(A_j)] = x \]
\[ \dim[\mathcal{R}(B_i) \cap \mathcal{R}(B_j)] = y \]
Then,
\[ \dim[\mathcal{R}(A_i \otimes B_i) \cap \mathcal{R}(A_j \otimes B_j)] = xy \]

Corollary: For almost every pair of KS dictionaries,
\[ \text{rank}(D_i D_i^T + D_j D_j^T) = 2n_1 n_2 - [2n_1 - m_1]^+ [2n_2 - m_2]^+ \]

- Intuition: If row or column subspaces are disjoint, entire K-S subspaces are disjoint
**Main Result: Diversity Order**

**Corollary:** For almost any K-S classification problem, the diversity order is

\[ d = n_1 n_2 - [2n_1 - m_1]^+ [2n_2 - m_2]^+ \]

- How does this compare to the unstructured “conjecture”?

\[ d = \min\{n_1 n_2, m_1 m_2 - n_1 n_2\} = n_1 n_2 - [2n_1 n_2 - m_1 m_2]^+ \]

- When subspace dimensions are small, diversity is the same
- With larger dimension, the K-S diversity order is smaller than that of unstructured subspaces
- Gap is bilinear in dimensions => quadratic in overall dimension

[MN, Rodrigues, Calderbank, "Discrimination on the Grassmann Manifold..." 2015]
Numerical results: Synthetic data

- Synthetic data: L=2 K-S bases drawn randomly, ML classification performance plotted vs. SNR
- Validates numerically the diversity results
Numerical Results: YaleB

• Learned K-S dictionaries from 10 samples per class of the YaleB face dataset
• Compare to: standard subspace classification, classification from SIFT features
• Much more compact representation, competitive classification accuracy

<table>
<thead>
<tr>
<th></th>
<th>Subspaces</th>
<th>SIFT[1]</th>
<th>K-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Accuracy (%)</td>
<td>79.36</td>
<td>84.3</td>
<td>84</td>
</tr>
<tr>
<td>Number of parameters</td>
<td>10240</td>
<td>5040</td>
<td>512</td>
</tr>
</tbody>
</table>

[Ramirez et al. “Classification and clustering via dictionary learning…” 2010]
Conclusion

Summary:
• Studied the classification performance of Kronecker-structured subspace models
• Derived the exact diversity order
• Found the error rate is higher than of unstructured subspace models

Future work:
• Discriminative K-S dictionary/subspace learning
• Good feature extraction for K-S data
• Higher order tensor models