Compressive Classification of Manifold-Structured Data via Bayesian Experiment Design

Phil Schniter

DARPA-KECoM 5/15/2012

Compressive Classification:

- In **compressive classification**, we are interested in the following problem:
  - Given:
    - = training data $\{\boldsymbol{u}_{cl}\}_{l=1}^{L}$ for each class $c \in \{1, \ldots, C\}$,
    - = prior class pmf $\{\pi_c\}_{c=1}^{C}$,
    - = noisy compressive linear measurements

$$\boldsymbol{y} = \boldsymbol{\Phi}\big(\boldsymbol{u} + \boldsymbol{w}\big) \in \mathbb{R}^M \quad \text{for a given noise prior } p(\boldsymbol{w}),$$

  - **Infer the class** c of the unknown signal $\boldsymbol{u} \in \mathbb{C}^N$.
- Note that we need to design both the kernel $\boldsymbol{\Phi}$ and the classifier.
- The noiseless version of this problem has been extensively studied in machine learning, where it is referred to as "**linear dimensionality reduction**" or "**supervised linear feature extraction**."

Optimal Linear Dimensionality Reduction:

- For a given classifier, the optimal dimensionality reducer $\boldsymbol{\Phi}$ is the one that minimizes the **classification error rate**. For most classifiers of interest, though, it is not possible to design this optimal $\boldsymbol{\Phi}$ in closed form.
- A popular alternative is the $\boldsymbol{\Phi}$ that maximizes **mutual information** (MI) between the observation vector $\boldsymbol{Y} \sim p(\boldsymbol{y}; \boldsymbol{\Phi})$ and class $C \sim p(c)$.
  - Why? Shannon MI yields upper (Hellman) and lower (Fano) **bounds** on classification error rate.
  - Unfortunately, the Shannon MI involves the log of a sum, making it difficult to evaluate:

$$I(\boldsymbol{Y}, C) = \sum_c \pi_c \int_{\boldsymbol{y}} p(\boldsymbol{y}|c) \Big[\log p(\boldsymbol{y}|c) - \log \Big(\sum_{c'} \pi_{c'}\, p(\boldsymbol{y}|c') \Big) \Big].$$

  - The **gradient** $\nabla_{\boldsymbol{\Phi}} I$, however, involves only conditional covariances [Palomar/Verdu 06], facilitating iterative maximization via gradient ascent.
  - Other options are to **approximate** the MI, to use other definitions of MI (e.g., Renyi's **quadratic** MI), or to use other heuristics like Fisher's **LDA**.

Bayes-Optimal Classification:

- Given the observation model $\boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{u} + \boldsymbol{w})$ and priors $\{p(\boldsymbol{u}|c), \pi_c\}_{c=1}^{C}$, the **class posterior** can be calculated as

$$\hat{\pi}_c \triangleq p(c|\boldsymbol{y}) = \frac{\pi_c \, p(\boldsymbol{y}|c)}{\sum_{c'} \pi_{c'} \, p(\boldsymbol{y}|c')} = \frac{\pi_c \int_{\boldsymbol{u}} p(\boldsymbol{y}|\boldsymbol{u}) \, p(\boldsymbol{u}|c)}{\sum_{c'} \pi_{c'} \int_{\boldsymbol{u}'} p(\boldsymbol{y}|\boldsymbol{u}') \, p(\boldsymbol{u}'|c')},$$

  from which the optimal (hard) decision is $\hat{c} = \arg\max_c \hat{\pi}_c$.

- Although not needed for classification, one could also compute the (class-conditional) **signal posterior** as

$$p(\boldsymbol{u}|c, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u}|c)}{\int_{\boldsymbol{u}'} p(\boldsymbol{y}|\boldsymbol{u}') p(\boldsymbol{u}'|c)},$$

  which can be useful for kernel design and adaptation.

Sequential Classification with an Adaptive Kernel:

- Now suppose that there is a high cost per measurement, so that we want to **minimize the number of measurements** M.
- We can do this by **sequentially** taking a small numbers of measurements M_t at round $t = 1, 2, 3, \ldots$. The observation model then becomes

$$\underbrace{\begin{bmatrix} \underline{\boldsymbol{y}}_{t-1} \\ \boldsymbol{y}_t \end{bmatrix}}_{\underline{\boldsymbol{y}}_t} = \underbrace{\begin{bmatrix} \underline{\boldsymbol{\Phi}}_{t-1} \\ \boldsymbol{\Phi}_t \end{bmatrix}}_{\underline{\boldsymbol{\Phi}}_t} \boldsymbol{u} + \underbrace{\begin{bmatrix} \underline{\boldsymbol{w}}_{t-1} \\ \boldsymbol{\Phi}_t \boldsymbol{w}_t \end{bmatrix}}_{\underline{\boldsymbol{w}}_t \in \mathbb{R}^{\underline{M}_t}} \begin{matrix} \in \mathbb{R}^{\underline{M}_{t-1}} \\ \in \mathbb{R}^{M_t} \end{matrix},$$

  where underbars are used to denote **cumulative** quantities.
- Sequential classification is then performed as follows: For $t = 1, 2, 3, \ldots$
  1. Design $\boldsymbol{\Phi}_t$ based on step-t **priors** $p(\boldsymbol{u}|\underline{\boldsymbol{y}}_{t-1})$ and $p(c|\underline{\boldsymbol{y}}_{t-1})$.
  2. Take take M_t new **measurements** $\boldsymbol{y}_t$.
  3. Compute **posteriors** on class $\hat{\pi}_c \triangleq p(c|\underline{\boldsymbol{y}}_t)$ and signal $p(\boldsymbol{u}|\underline{\boldsymbol{y}}_{t-1})$.
  4. Repeat with $t = t + 1$.

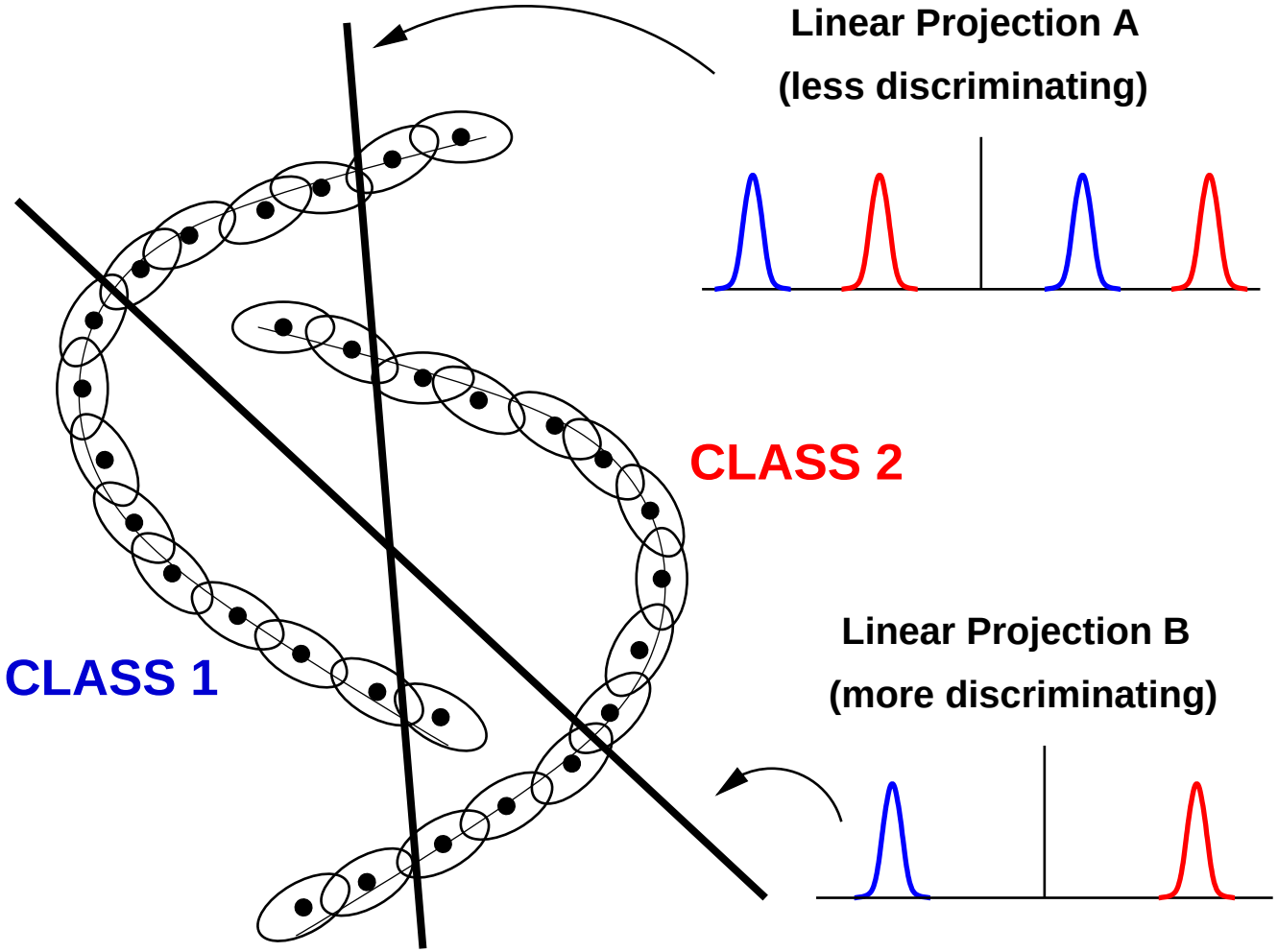
Manifold Structured Priors:

- Motivations:
  - In many high-dimensional problems (e.g., radar), the signal $\boldsymbol{u} \in \mathbb{C}^N$ lives on a low-dimensional **manifold**. The prior $p(\boldsymbol{u}|c)$ should reflect this.
  - We are seldom given an analytical description of the manifold, but rather labeled training data $\{\boldsymbol{u}_{cl}\}_{l=1}^L$, from which to **learn** the prior.
  - The form of the prior should facilitate kernel adaptation.
- These considerations motivate the Gaussian mixture model (GMM) prior

$$p(\boldsymbol{u}|c) = \sum_{q=1}^{Q} \hat{\lambda}_{cq} \mathcal{N}(\boldsymbol{u}; \hat{\boldsymbol{u}}_{cq}; \hat{\boldsymbol{\Sigma}}_{cq} + \hat{\sigma}_u^2 \boldsymbol{I}) \text{ with } \textbf{low rank } \hat{\boldsymbol{\Sigma}}_{cq}.$$

  - When $\{\hat{\lambda}_{cq}, \hat{\boldsymbol{u}}_{cq}, \hat{\boldsymbol{\Sigma}}_{cq}, \hat{\sigma}_u^2\}$ are fitted to training data using EM or similar, it's called **mixture of factor analyzers** [Ghahramani/Hinton 96].
  - When $\hat{\boldsymbol{u}}_{cq} = \boldsymbol{u}_{cq}$ (training) and $\hat{\lambda}_{cq} = \frac{1}{Q}$, and when $\{\hat{\boldsymbol{\Sigma}}_{cq}, \hat{\sigma}_u^2\}$ are fitted using local averages, it's called **manifold Parzen** [Vincent & Bengio 03].

Illustration of Manifold Model and Linear Projections:



Fast GMM Fitting via Bilinear Approximate Message Passing:

- According to the GMM,

$$\boldsymbol{u}|c = \sum_q \left[\hat{\boldsymbol{u}}_{cq},\, \hat{\boldsymbol{V}}_{cq}\right] \begin{bmatrix} 1 \\ \boldsymbol{z}_{cq} \end{bmatrix} b_{cq} + \boldsymbol{w} \quad \text{for} \quad \begin{cases} \boldsymbol{z}_{cq} \sim \mathcal{N}(0, \boldsymbol{I}) \\ b_{cq} \in \{0,1\} \sim \mathcal{B}(\lambda_{cq}) \\ \boldsymbol{w} \sim \mathcal{N}(0, \hat{\sigma}_u^2 \boldsymbol{I}). \end{cases}$$

  where low-rank $\hat{\boldsymbol{\Sigma}}_{cq} = \hat{\boldsymbol{V}}_{cq} \hat{\boldsymbol{V}}_{cq}^H$.

- Thus, for class c,

$$\boldsymbol{u} = \underbrace{\left[\hat{\boldsymbol{u}}_{c1}, \hat{\boldsymbol{V}}_{c1} \;\middle|\; \hat{\boldsymbol{u}}_{c2}, \hat{\boldsymbol{V}}_{c2} \;\middle|\; \cdots \;\middle|\; \hat{\boldsymbol{u}}_{cQ}, \hat{\boldsymbol{V}}_{cQ}\right]}_{\triangleq\, \boldsymbol{D}_c} \boldsymbol{z} + \boldsymbol{w}$$

  for **block-sparse** Bernoulli-Gaussian $\boldsymbol{z}$.

- Building the matrix $\boldsymbol{U}_c$ column-wise from the class-c training vectors:

$$\boldsymbol{U}_c = \boldsymbol{D}_c \boldsymbol{Z}_c + \boldsymbol{W}_c,$$

  where $\{\boldsymbol{D}_c, \boldsymbol{Z}_c\}$ can be learned with **BiG-AMP** [Schniter/Cevher/Parker 12].

GMM Posterior Update:

Given the GMM signal prior

$$p(\boldsymbol{u}|c) = \textstyle\sum_{q=1}^{Q} \hat{\lambda}_{cq} \mathcal{N}(\boldsymbol{u}; \hat{\boldsymbol{u}}_{cq}; \hat{\boldsymbol{\Sigma}}_{cq} + \hat{\sigma}_u^2 \boldsymbol{I}) \text{ with } \textbf{low rank } \hat{\boldsymbol{\Sigma}}_{cq}.$$

and the step-t measurement model

$$\underline{\boldsymbol{y}}_t = \underline{\boldsymbol{\Phi}}_t \boldsymbol{u} + \underline{\boldsymbol{w}}_t \text{ for } \underline{\boldsymbol{w}}_t \sim \mathcal{N}(0, \underline{\boldsymbol{R}}_t),$$

we get the step-t class posterior

$$\hat{\pi}_{c,t} \triangleq p(c|\underline{\boldsymbol{y}}_t) \propto \pi_c \textstyle\sum_{q=1}^{Q} \hat{\lambda}_{cq} \mathcal{L}_{cq,t}$$

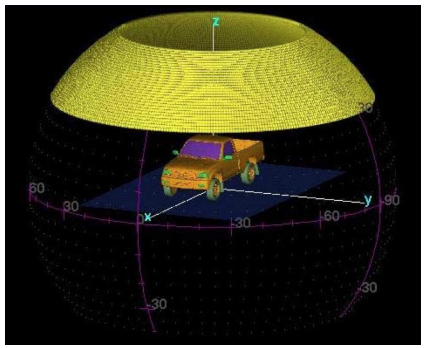
$$\mathcal{L}_{cq,t} \triangleq (\underline{\boldsymbol{y}}_t - \underline{\boldsymbol{\Phi}}_t \hat{\boldsymbol{u}}_{cq})^{\mathsf{H}} \big(\underline{\boldsymbol{\Phi}}_t (\hat{\boldsymbol{\Sigma}}_{cq} + \hat{\sigma}_v^2 \boldsymbol{I}) \underline{\boldsymbol{\Phi}}_t^H + \underline{\boldsymbol{R}}_t\big)^{-1} (\underline{\boldsymbol{y}}_t - \underline{\boldsymbol{\Phi}}_t \hat{\boldsymbol{u}}_{cq})$$

and the step-t signal posterior

$$p(\boldsymbol{u}|c, \underline{\boldsymbol{y}}_t) = \textstyle\sum_q \hat{\lambda}_{cq,t} \mathcal{N}(\boldsymbol{u}; \hat{\boldsymbol{u}}_{cq,t}; \hat{\boldsymbol{\Sigma}}_{cq,t} + \hat{\sigma}_u^2 \boldsymbol{I})$$

The low-rank and sequential nature of the problem yields **fast updates** for the posterior parameters $\{\hat{\lambda}_{cq,t}, \hat{\boldsymbol{u}}_{cq,t}, \hat{\boldsymbol{\Sigma}}_{cq,t}\}$, where again $\hat{\boldsymbol{\Sigma}}_{cq,t}$ is **low rank**.

Experiment — Vehicle Classification from Radar Returns:



Goal: Classify a noisy cluttered radar return as one of 3 vehicles (with unknown azimuth).

Given: $N = 128$-dimensional training vectors $\{\boldsymbol{u}_{cl}\}_{l=1}^{5760}$ for each class $c \in \{1, 2, 3\}$.

Measurement kernel design (offline):

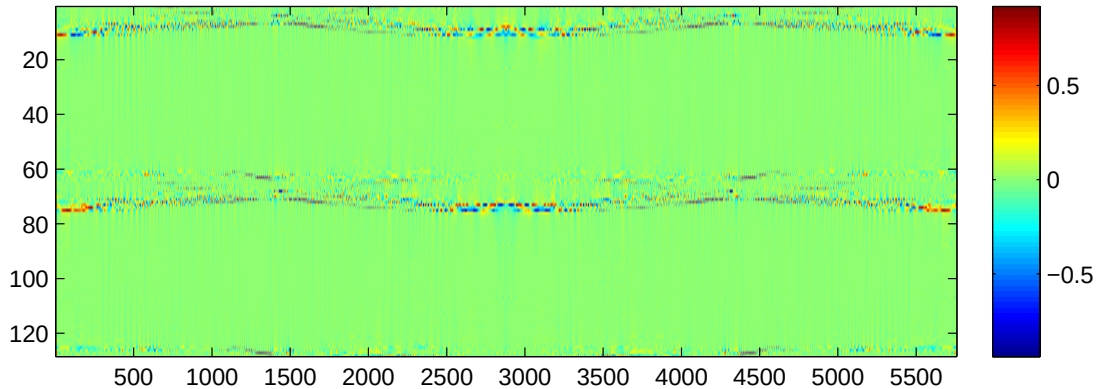
1. Fit a manifold-Parzen prior to training data using rank-1 covariances $\hat{\boldsymbol{\Sigma}}_{cq}$.
2. Design a $2 \times N$ kernel via Fisher's linear discriminant analysis (LDA).
   (One eigen-decomp — *much* cheaper than gradient of mutual information!)

Classification and measurement kernel adaptation (online): For $t = 1, 2, 3, \ldots$

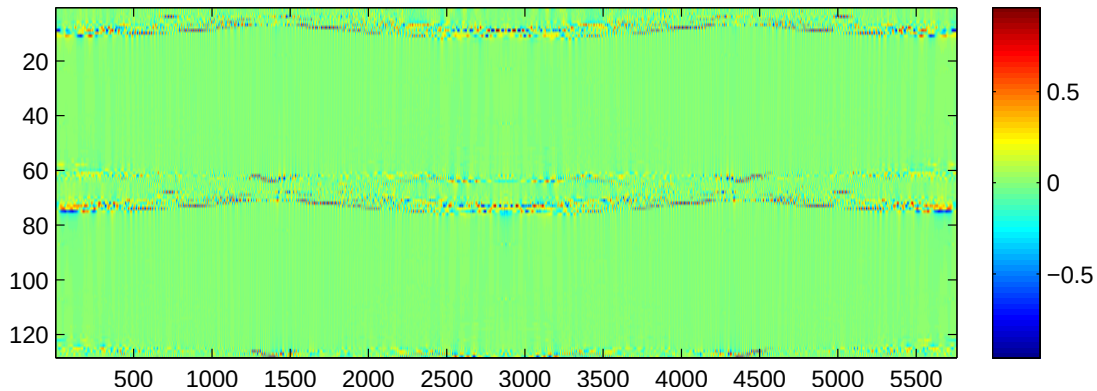
1. Take 2 measurements: $\boldsymbol{y}_t = \boldsymbol{\Phi}_t\big(\boldsymbol{u} + \boldsymbol{v} + \boldsymbol{w}_t\big) \in \mathbb{R}^2$
2. Compute posterior class pmf and posterior GMM parameters.
3. Design next $2 \times N$ kernel from GMM posteriors using LDA.
4. Repeat with $t = t + 1$.

Training Vectors:

Camry:



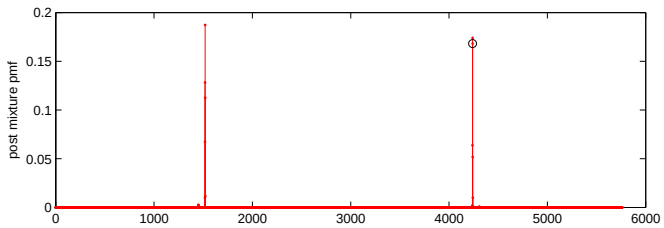
Maxima:

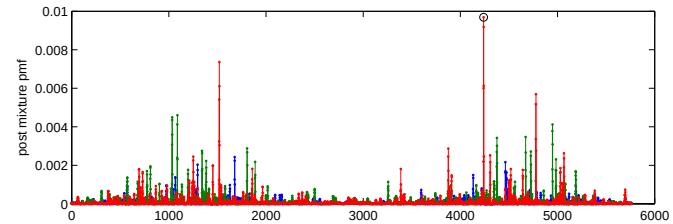


Preliminary Results on Vehicle Classification:

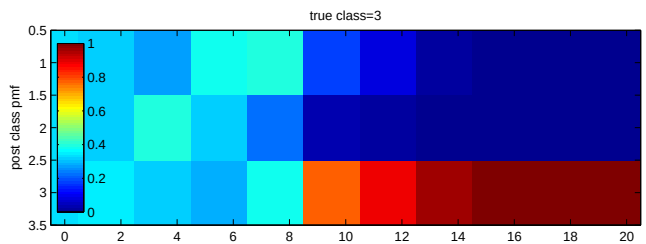
A typical problem realization (for $\frac{\textbf{signal}}{\textbf{clutter+noise}} = \textbf{0 dB}$) looks like this:

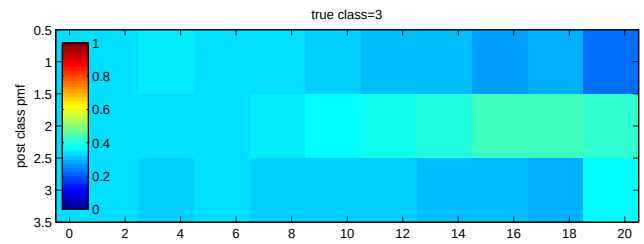
final posterior mixture pmf:





posterior class pmf versus time:



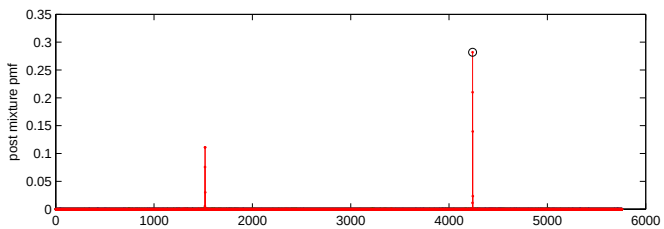


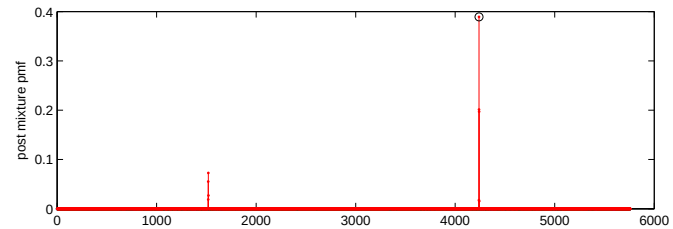
The LDA-based scheme gives correct classification after **10** measurements, while the iid-random scheme is still **uncertain after 20** measurements.

Preliminary Results on Vehicle Classification:

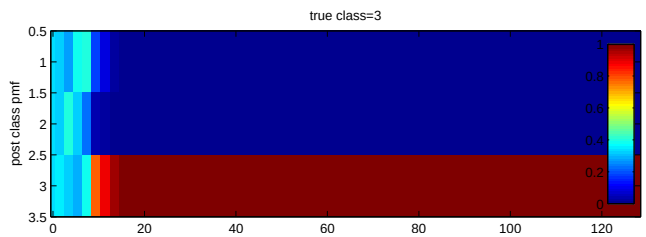
A typical problem realization (for $\frac{\textbf{signal}}{\textbf{clutter+noise}} = \mathbf{0\ dB}$) looks like this:

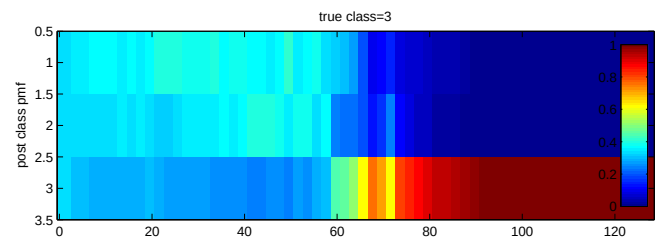
final posterior mixture pmf:





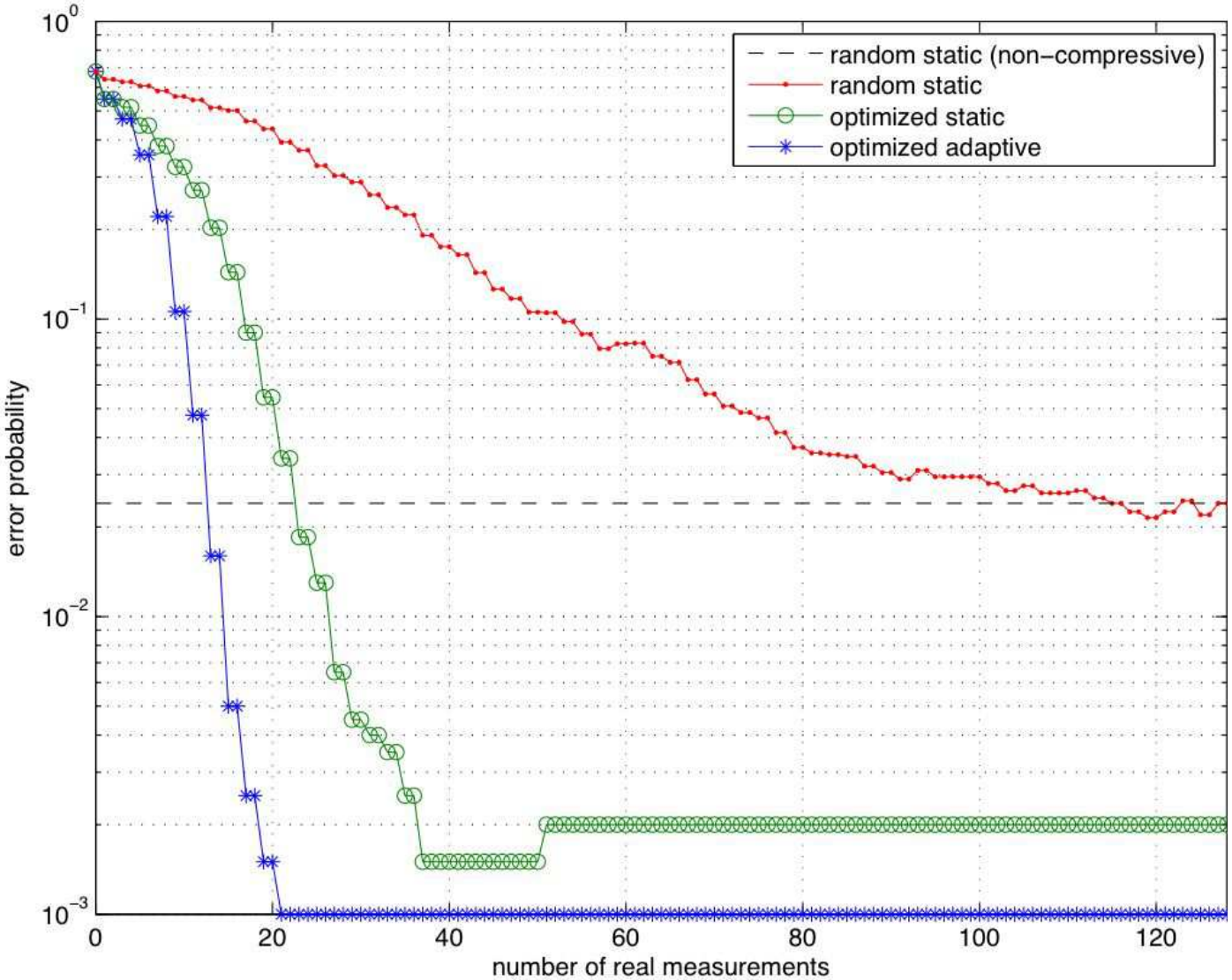
posterior class pmf versus time:





The LDA-based scheme gives correct classification after **10** measurements, whereas the iid-random scheme requires **75** measurements.

Average Classification Error Probability vs # of Measurements:



Summary:

- We considered the problem of **designing** and **adapting** linear measurement kernels to maximize **classification** performance from **very few** measurements.
- In particular, we proposed a novel **sequential classification** scheme using ideas from
  - **information-theoretic linear dimensionality reduction**,
  - **Bayes-optimal classification**, and
  - **manifold-structured signal priors**.
- We applied our approach to **vehicle classification from radar returns**.
  - Unknown azimuth, so signal lives on a 1-dimensional **manifold** in $\mathbb{R}^{128}$.
  - The returns were **extremely cluttered and noisy**: 0 dB!
  - Our scheme was able to classify correctly after **10** measurements, whereas iid random measurements required **75**.
  - LDA was used for kernel adaptation: no gradient search $\Rightarrow$ much **faster**!