

Universidade Federal do Ceará – UFC
Programa de Pós-Graduação em Engenharia Elétrica (PPGEE)
Disciplina: Reconhecimento de Padrões (TIP8311)

Trabalho Computacional III

Aluno: Luis Felipe Carneiro de Souza – 593034

Professor: Prof. Dr. Guilherme de Alencar Barreto

27 de fevereiro de 2026

Resumo

Este documento apresenta os resultados e discussões referentes ao trabalho computacional III da disciplina de Reconhecimento de Padrões. O trabalho aborda a implementação de quatro métodos algébricos distintos para o cálculo das matrizes de covariância global e específicas por classe, avaliando conjuntos de dados com 4 e 24 atributos (sensores). A acurácia matemática e a eficiência computacional das abordagens empíricas foram rigorosamente comparadas com a função nativa da linguagem Python por meio da norma da matriz de diferenças e de análises estatísticas de tempo de execução ao longo de 100 rodadas. Além disso, investigou-se a invertibilidade e a estabilidade numérica das matrizes resultantes utilizando o posto matricial e o número de condicionamento, aplicando-se técnicas de regularização de Tikhonov para viabilizar a inversão em cenários de singularidade ou instabilidade ocasionados pela alta dimensionalidade.. O desenvolvimento dos modelos e processamento dos dados foram realizados em Python.

1 Introdução

O presente trabalho aborda o problema de classificação de padrões em robótica móvel, utilizando um conjunto de dados composto por leituras de 24 sensores de ultrassom. O objetivo principal é classificar os movimentos do robô em quatro categorias distintas (Move-Forward, Sharp-Right-Turn, Slight-Left-Turn e Slight-Right-Turn) com base nas assinaturas sensoriais captadas.

Em problemas com alta dimensionalidade, fenômenos como a redundância de dados e a "maldição da dimensionalidade" podem degradar o desempenho dos modelos. Nesse contexto, este relatório avalia e compara o desempenho empírico de dois modelos de classificação: o Classificador Quadrático Gaussiano (CQG) e o classificador de Distância Mínima ao Protótipo (DMP). Adicionalmente, investiga-se o impacto da redução de dimensionalidade através da Análise de Componentes Principais (PCA) no desempenho preditivo e na estabilidade matemática dos classificadores desenvolvidos.

1.1 K-Means

O K-médias é um algoritmo clássico de aprendizado não-supervisionado cujo objetivo é particionar um conjunto de dados em K grupos (clusters) distintos. O algoritmo opera de forma iterativa, atribuindo cada amostra ao centroide mais próximo e, em seguida, recalculando a posição dos centroides com base na média das amostras de cada grupo. Matematicamente, o K-médias busca minimizar a soma das distâncias quadráticas intra-cluster (SSD - Sum of Squared Distances), dada pela Equação 1:

$$SSD = \sum_{j=1}^K \sum_{x \in S_j} \|x - \mu_j\|^2 \quad (1)$$

No qual:

- S_j representa o conjunto de amostras atribuídas ao cluster j
- μ_j é o vetor média (centroide) desse agrupamento

Neste trabalho, o K-médias não é utilizado como fim, mas sim como o motor de busca de protótipos para o classificador DMP.

1.2 Mínima Distância ao Protótipo

O Classificador de Distância Mínima ao Protótipo é uma abordagem supervisionada baseada em instâncias representativas. Em sua versão mais robusta, o modelo

contorna a limitação de representar classes com distribuições complexas utilizando múltiplos vetores de referência por classe. Para esse tipo de algoritmo baseado em protótipos, tem-se que o número total de protótipos (K) é maior que o número de classes (M). Consequentemente, o fato de diferentes protótipos dividirem uma mesma classe é uma característica comum nesta família de classificadores [1].

Durante o treinamento, é necessário definir a quantidade adequada de protótipos rotulados e suas corretas iniciações para reduzir erros de classificação. Para isso, o algoritmo K-médias (K-means) é aplicado individualmente aos dados de cada classe para varrer diferentes valores de K . Em seguida, técnicas de validação de agrupamentos (como os índices Davies-Bouldin, Dunn ou Silhouettes) são empregadas para selecionar o número ótimo de agrupamentos (K_{opt}) e suas posições iniciais confiáveis [1].

Na fase de inferência, a associação de um novo padrão de entrada $x(t)$ às classes predefinidas $\mathcal{C} = \{c_1, c_2, \dots, c_M\}$ é formulada a partir do critério de decisão de vizinho mais próximo. A amostra recebe o rótulo do protótipo que apresentar a menor distância, conforme a regra matemática apresentado em 2 [1]:

$$\text{Classe de } x(t) = \text{Classe de } m_c(t), \quad c = \arg \min_{i=1, \dots, K} d(x(t), m_i(t)) \quad (2)$$

No qual:

- $d(\cdot, \cdot)$ representa a medida de dissimilaridade
- c é o índice do protótipo mais próximo ao padrão de entrada entre os K protótipos disponíveis no modelo.

1.3 Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é uma técnica de projeção linear ortogonal utilizada para redução de dimensionalidade e extração de características. O PCA busca transformar o conjunto de atributos originais, possivelmente correlacionados, em um novo conjunto de variáveis descorrelacionadas (os componentes principais), ordenadas de forma decrescente pela variância que conseguem explicar. A transformação é baseada na Decomposição em Valores Singulares (SVD) da matriz de dados centrada. Ao selecionar apenas os primeiros q componentes, o PCA projeta os dados em um subespaço de menor dimensão, minimizando a perda de informação. Além da redução geométrica, pode-se aplicar o branqueamento (whitening) dividindo os componentes projetados por seus respectivos desvios padrões como apresentado em 3:

$$Z_{branc} = Z\Lambda^{-1/2} \quad (3)$$

No qual:

- Λ é a matriz diagonal dos autovalores.

Este processo esferiza os dados, o que é particularmente vantajoso para algoritmos baseados em distância Euclidiana, como o K-médias presente no classificador DMP.

2 Metodologia

2.1 Análise Inicial e Invertibilidade

O conjunto de dados é composto por instâncias distribuídas em 4 classes e descritas por um vetor de atributos originais no espaço \mathbb{R}^{24} . Verificou-se que as matrizes de covariância de algumas classes apresentavam determinante próximo a zero, indicando matrizes singulares (não invertíveis). Este comportamento é esperado em arranjos de múltiplos sensores de ultrassom, onde sensores adjacentes frequentemente registram leituras altamente correlacionadas, gerando multicolinearidade. Para garantir a estabilidade numérica na inversão das matrizes, aplicou-se um termo de regularização na diagonal principal de cada matriz de covariância C , tal que $C_{reg} = C + \lambda I$, adotando-se $\lambda = 10^{-6}$.

2.2 Classificadores Implementados

Classificador Quadrático Gaussiano (CQG): Modelou-se a distribuição de cada classe através de uma densidade Gaussiana multivariada. A inferência foi realizada minimizando a função discriminante quadrática para cada amostra x :

$$g_i(x) = (x - m_i)^T C_i^{-1} (x - m_i) + \ln |C_i| - 2 \ln P(\omega_i)$$

onde m_i é o vetor média, C_i é a matriz de covariância regularizada e $P(\omega_i)$ é a probabilidade a priori da classe i .

Classificador de Distância Mínima ao Protótipo (DMP): Implementou-se uma variação baseada em múltiplos protótipos, utilizando o algoritmo K-médias independentemente por classe. O número ótimo de protótipos K_{opt} foi determinado através da validação empírica minimizando o Índice de Davies-Bouldin. A classificação final de uma amostra de teste é dada pela classe associada ao protótipo mais próximo, utilizando a distância Euclidiana.

2.3 Redução de Dimensionalidade (PCA)

Para a Questão 2, implementou-se o algoritmo PCA via Decomposição em Valores Singulares (SVD). Os dados foram centrados na média e submetidos ao processo de branqueamento (*whitening*), garantindo que os componentes projetados tivessem variância unitária, de modo a otimizar a métrica de distância Euclidiana utilizada pelo DMP. O número de componentes q variou de 24 até 2 para construção da curva de variância explicada e avaliação de desempenho.

3 Resultados

3.1 Desempenho no Espaço Original (Sem PCA)

Após 100 rodadas independentes de treinamento e teste (80% / 20%), o modelo CQG obteve uma acurácia global média de aproximadamente $68.18\% \pm 1.06\%$, o modelo DMP registrou $72.02\% \pm 1.66\%$. O desempenho por classes é apresentado na tabela 1

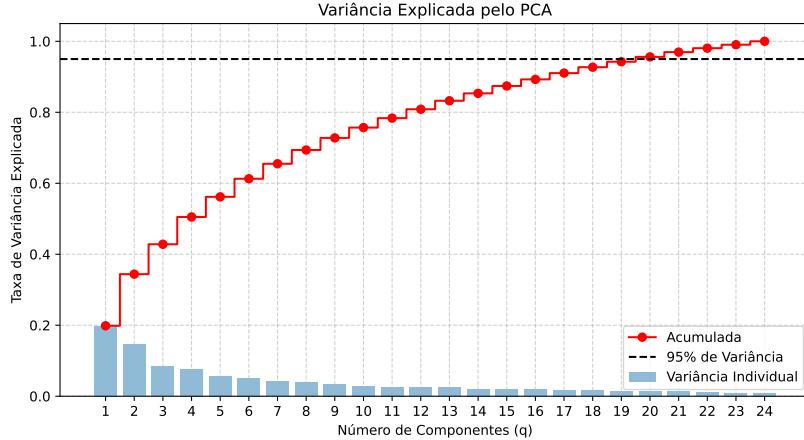
Tabela 1: Resultados de acurácia (Média \pm Desvio-Padrão) para o caso Base (Sem PCA).

Classificador	Global	Move-Forward	Sharp-Right-Turn	Slight-Left-Turn	Slight-Right-Turn
CQG	$68.18\% \pm 1.06\%$	$63.16\% \pm 1.95\%$	$62.64\% \pm 1.85\%$	$88.84\% \pm 3.01\%$	$87.63\% \pm 2.18\%$
DMP	$72.02\% \pm 1.66\%$	$66.48\% \pm 4.59\%$	$76.31\% \pm 4.08\%$	$93.44\% \pm 2.81\%$	$67.36\% \pm 11.56\%$

3.2 Impacto da Redução de Dimensionalidade

A Figura 1 apresenta a variância explicada em função das componentes do conjunto de dados.

Figura 1: Variância Explicada pelo PCA

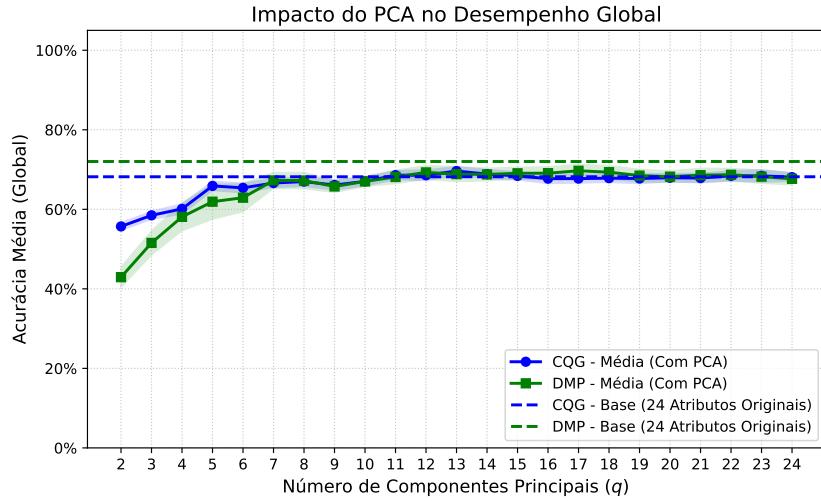


Fonte: Elaborado pelo autor

A Tabela 2 presente no apêndice A apresenta a acurácia global e para cada classe para a aplicação do PCA variando de 24 a 2 componentes.

A Figura 2 sintetiza os resultados da Tabela 2 de forma gráfica da aplicação do PCA no desempenho global.

Figura 2: Impacto do PCA no Desempenho Global

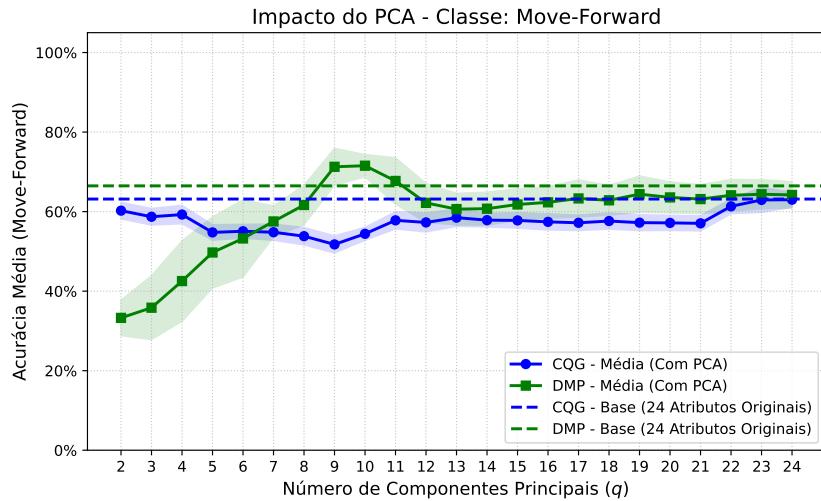


Fonte: Elaborado pelo autor

Observa-se desempenho aceitável no uso de uma quantidade de componentes superiores a 7 componentes.

A Figura 3 apresenta o desempenho da acurácia da classe *Move-Forward* com a aplicação do PCA.

Figura 3: Impacto do PCA - Classe: Move-Forward



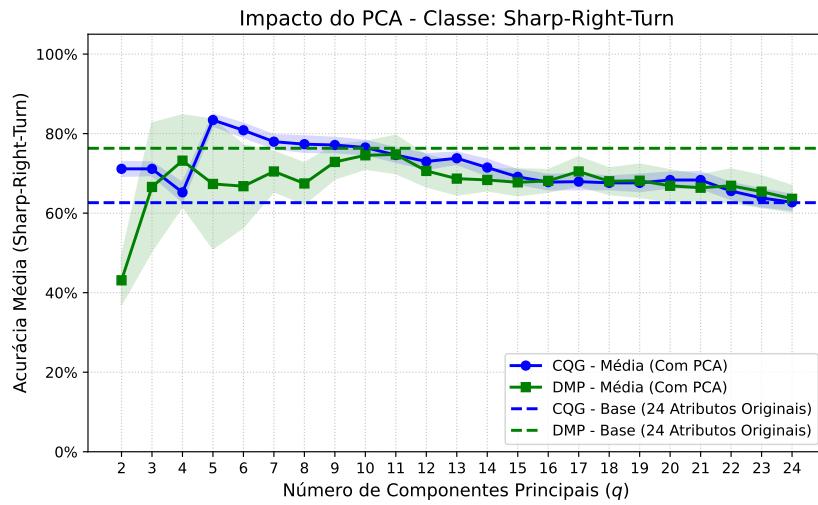
Fonte: Elaborado pelo autor

Pontuamos:

- **No Espaço Original (Sem PCA):** O DMP (66.48%) apresenta uma leve vantagem sobre o CQG (63.16%). É uma classe de dificuldade intermediária/alta para ambos os classificadores.

- **Impacto do PCA:** Ao aplicar o PCA, o desempenho do CQG cai levemente e se estabiliza na faixa dos 57% a 59% na maior parte das reduções. O DMP também sofre uma queda para a casa dos 60% a 64%.
- **Redução Extrema ($q = 2$ e $q = 3$):** O CQG mostra-se muito mais robusto na redução extrema, mantendo cerca de 60% com $q = 2$, enquanto o DMP desaba para preocupantes 33.28%

Figura 4: Impacto do PCA - Classe: Sharp-Right-Turn

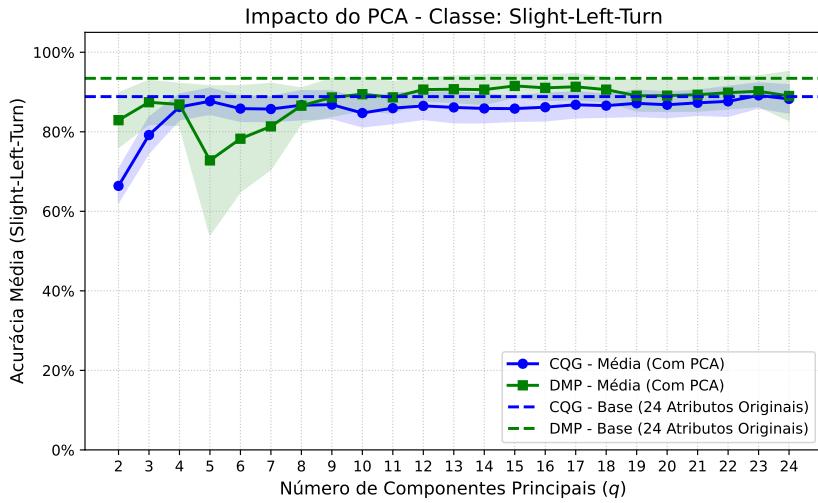


Fonte: Elaborado pelo autor

Pontuamos:

- **No Espaço Original (Sem PCA):** O DMP domina completamente aqui (76.31%) contra um desempenho modesto do CQG (62.64%).
- **Impacto do PCA:** Este é o caso mais interessante para o CQG. Conforme você reduz os componentes com o PCA, o CQG melhora significativamente, saltando de 62% para impressionantes 83.42% em $q = 5$. O DMP se mantém instável, oscilando entre 63% e 74%.

Figura 5: Impacto do PCA - Classe: Slight-Left-Turn

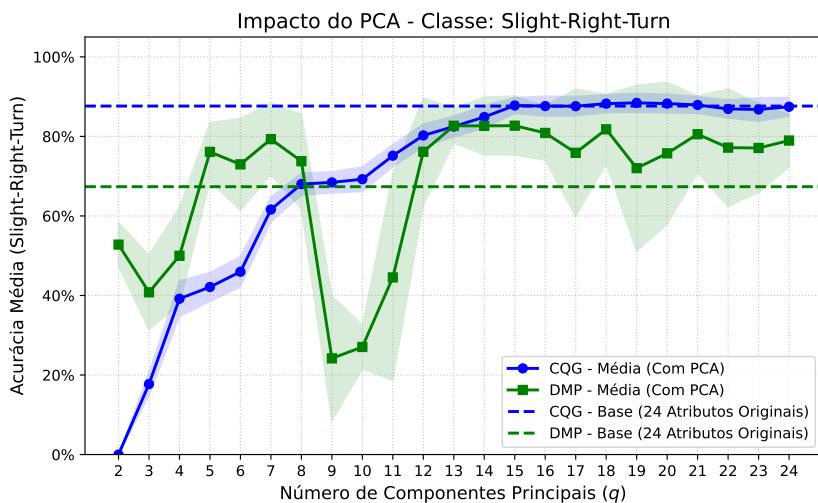


Fonte: Elaborado pelo autor

Pontuamos:

- **No Espaço Original (Sem PCA):** É a classe "mais fácil" do conjunto de dados. O DMP alcança excelentes 93.44% e o CQG atinge 88.84%.
- **Impacto do PCA:** A acurácia se mantém altíssima e extremamente estável para ambos os modelos, mesmo descartando mais da metade dos sensores. O CQG fica na faixa de 86-89% até $q = 4$, e o DMP na faixa de 89-91% até $q = 12$.

Figura 6: Impacto do PCA - Classe: Slight-Right-Turn



Fonte: Elaborado pelo autor

Pontuamos:

- **No Espaço Original (Sem PCA):** O cenário se inverte completamente em relação à curva fechada à direita. Aqui, o CQG tem um desempenho excelente (87.63%), enquanto o DMP sofre bastante (67.36%) e apresenta uma variância altíssima (desvio padrão de 11.56%).
- **Impacto do PCA e o "Colapso":** O CQG se mantém firme na casa dos 87% até $q = 12$. A partir de $q = 11$, o desempenho começa a cair. O dado mais drástico de toda a sua tabela ocorre aqui: em $q = 2$, o CQG atinge 0.00% de acurácia. O DMP também apresenta oscilações violentas, caindo para 27% em $q = 10$ e voltando para 79% em $q = 7$.

A análise da variância explicada demonstrou que grande parte da variação dos dados pode ser retida com um subconjunto reduzido de componentes. Ao executar os classificadores no espaço transformado pelo PCA, observou-se que a redução de $q = 24$ para a faixa de $q \approx 12$ componentes resultou em um leve ganho de desempenho computacional e estabilidade, alcançando um pico de acurácia global do CQG próximo a 69.64%. Este fenômeno indica que os últimos componentes principais descartados representavam primordialmente ruídos de sensoriamento.

Entudo, uma redução extrema mostrou-se agressiva. Ao reduzir o espaço para $q \leq 4$, a acurácia sofreu decaimento agudo. De forma notável, no cenário com $q = 2$, a capacidade preditiva do CQG para a classe *Slight-Right-Turn* desabou para 0.00%, evidenciando que em espaços de dimensionalidade muito baixa, as assinaturas desta classe sofrem total sobreposição com as demais, tornando o problema linearmente inseparável.

4 Conclusão

Os experimentos realizados validam a hipótese de que o sensoriamento em massa (24 sensores) introduz redundância linear significativa, o que pode ser tratado eficazmente pela regularização das matrizes e reduzido através do PCA. O Classificador Quadrático Gaussiano demonstrou ser o modelo mais robusto para a topologia deste conjunto de dados geométricos. Conclui-se que a aplicação do PCA não apenas evidenciou a dimensionalidade intrínseca do problema (em torno de 12 a 14 dimensões úteis), como também auxiliou na remoção de ruídos em estágios moderados de redução, sem incorrer em perda de eficiência preditiva.

5 Códigos e Repositório

Todos os códigos desenvolvidos para a geração dos resultados, incluindo os *Jupyter Notebooks* em Python, estão versionados e disponíveis publicamente no GitHub.

O repositório pode ser acessado através do seguinte link:

<https://github.com/LuisFelipeCSouza/reconhecimento-de-padroes/tree/main/trabalho-3>

Referências

- [1] D. Sousa, C. Cavalcante, G. Barreto, and C. Medeiros, “Uma análise heurística baseada em protótipos para o monitoramento de estados de motores de indução,” in *Anais do 14º Congresso Brasileiro de Inteligência Computacional* (B. J. T. Fernandes and A. P. Júnior, eds.), (Curitiba, PR), pp. 1–6, ABRICOM, 2019.
- [2] G. de Alencar Barreto, “Introdução à classificação de padrões.” Notas de aula, 2025.
Contato: gbarreto@ufc.br.

A Desempenho dos classificadores mediante aplicação do PCA

Tabela 2: Resultados de acurácia (Média \pm Desvio-Padrão) em função da redução do PCA ($q = 24$ a $q = 2$).

Cenário	Classificador	Global	Move-Forward	Sharp-Right-Turn	Slight-Left-Turn	Slight-Right-Turn
PCA ($q = 24$)	CQG	68.08% \pm 1.06%	62.99% \pm 1.89%	62.73% \pm 1.97%	88.26% \pm 3.54%	87.46% \pm 2.35%
PCA ($q = 24$)	DMP	67.67% \pm 1.53%	64.20% \pm 3.34%	63.61% \pm 3.32%	89.00% \pm 6.16%	78.96% \pm 6.62%
PCA ($q = 23$)	CQG	68.42% \pm 1.54%	62.93% \pm 3.14%	63.83% \pm 2.24%	89.15% \pm 3.13%	86.77% \pm 2.95%
PCA ($q = 23$)	DMP	68.21% \pm 1.63%	64.37% \pm 3.66%	65.41% \pm 4.01%	90.19% \pm 3.81%	77.09% \pm 11.23%
PCA ($q = 22$)	CQG	68.41% \pm 1.23%	61.33% \pm 1.85%	65.56% \pm 2.15%	87.65% \pm 3.76%	86.91% \pm 2.28%
PCA ($q = 22$)	DMP	68.69% \pm 1.58%	64.11% \pm 3.98%	66.89% \pm 4.25%	89.83% \pm 4.07%	77.18% \pm 14.77%
PCA ($q = 21$)	CQG	67.86% \pm 1.19%	57.05% \pm 1.90%	68.33% \pm 2.09%	87.28% \pm 3.12%	87.91% \pm 2.10%
PCA ($q = 21$)	DMP	68.58% \pm 1.72%	63.15% \pm 3.63%	66.40% \pm 3.21%	89.31% \pm 4.13%	80.56% \pm 9.63%
PCA ($q = 20$)	CQG	67.94% \pm 0.97%	57.17% \pm 1.89%	68.33% \pm 1.96%	86.79% \pm 3.21%	88.26% \pm 2.33%
PCA ($q = 20$)	DMP	68.21% \pm 1.42%	63.57% \pm 3.89%	66.88% \pm 4.01%	89.10% \pm 3.96%	75.74% \pm 17.86%
PCA ($q = 19$)	CQG	67.74% \pm 1.28%	57.23% \pm 2.00%	67.61% \pm 2.16%	87.16% \pm 3.31%	88.44% \pm 2.40%
PCA ($q = 19$)	DMP	68.48% \pm 1.58%	64.39% \pm 4.56%	68.16% \pm 4.17%	89.05% \pm 3.82%	72.02% \pm 20.85%
PCA ($q = 18$)	CQG	67.83% \pm 1.03%	57.63% \pm 2.03%	67.62% \pm 1.68%	86.57% \pm 2.88%	88.25% \pm 2.26%
PCA ($q = 18$)	DMP	69.34% \pm 1.64%	62.84% \pm 3.72%	68.02% \pm 3.33%	90.60% \pm 3.01%	81.79% \pm 8.71%
PCA ($q = 17$)	CQG	67.72% \pm 1.10%	57.19% \pm 1.95%	67.93% \pm 2.00%	86.79% \pm 3.29%	87.60% \pm 2.43%
PCA ($q = 17$)	DMP	69.69% \pm 1.57%	63.31% \pm 4.69%	70.49% \pm 3.66%	91.32% \pm 3.26%	75.88% \pm 16.01%
PCA ($q = 16$)	CQG	67.71% \pm 1.22%	57.42% \pm 2.04%	67.80% \pm 1.92%	86.18% \pm 3.43%	87.63% \pm 2.49%
PCA ($q = 16$)	DMP	69.09% \pm 1.59%	62.34% \pm 3.93%	68.12% \pm 2.82%	91.05% \pm 3.15%	80.87% \pm 6.85%
PCA ($q = 15$)	CQG	68.40% \pm 0.99%	57.78% \pm 1.96%	69.13% \pm 1.82%	85.82% \pm 3.17%	87.77% \pm 2.01%
PCA ($q = 15$)	DMP	69.06% \pm 1.50%	61.80% \pm 3.96%	67.75% \pm 3.37%	91.54% \pm 2.86%	82.70% \pm 7.39%
PCA ($q = 14$)	CQG	68.85% \pm 1.15%	57.85% \pm 1.72%	71.47% \pm 2.05%	85.86% \pm 3.59%	84.88% \pm 2.83%
PCA ($q = 14$)	DMP	68.77% \pm 1.55%	60.73% \pm 4.14%	68.36% \pm 2.72%	90.60% \pm 3.73%	82.65% \pm 7.28%
PCA ($q = 13$)	CQG	69.63% \pm 1.10%	58.48% \pm 2.18%	73.81% \pm 1.54%	86.13% \pm 3.85%	82.47% \pm 2.63%
PCA ($q = 13$)	DMP	68.84% \pm 1.87%	60.63% \pm 3.97%	68.68% \pm 4.11%	90.70% \pm 3.30%	82.67% \pm 4.37%
PCA ($q = 12$)	CQG	68.56% \pm 1.32%	57.28% \pm 2.39%	72.99% \pm 1.93%	86.50% \pm 3.38%	80.25% \pm 2.85%
PCA ($q = 12$)	DMP	69.28% \pm 1.68%	62.19% \pm 4.91%	70.63% \pm 3.99%	90.61% \pm 2.83%	76.15% \pm 13.56%
PCA ($q = 11$)	CQG	68.58% \pm 1.13%	57.81% \pm 2.16%	74.62% \pm 1.83%	85.94% \pm 3.85%	75.15% \pm 2.80%
PCA ($q = 11$)	DMP	68.13% \pm 1.66%	67.71% \pm 5.85%	74.76% \pm 4.80%	88.65% \pm 3.94%	44.54% \pm 25.88%
PCA ($q = 10$)	CQG	67.02% \pm 1.10%	54.42% \pm 1.61%	76.51% \pm 1.77%	84.72% \pm 3.48%	69.24% \pm 3.10%
PCA ($q = 10$)	DMP	67.02% \pm 1.25%	71.54% \pm 2.81%	74.55% \pm 3.47%	89.45% \pm 4.16%	27.05% \pm 5.51%
PCA ($q = 9$)	CQG	66.11% \pm 1.26%	51.75% \pm 2.17%	77.13% \pm 1.92%	86.82% \pm 3.45%	68.44% \pm 2.65%
PCA ($q = 9$)	DMP	65.74% \pm 1.47%	71.27% \pm 4.64%	72.88% \pm 4.29%	88.64% \pm 4.73%	24.19% \pm 15.73%
PCA ($q = 8$)	CQG	66.97% \pm 1.17%	53.82% \pm 2.19%	77.35% \pm 2.05%	86.68% \pm 3.64%	68.01% \pm 2.78%
PCA ($q = 8$)	DMP	67.23% \pm 2.03%	61.66% \pm 4.94%	67.46% \pm 5.18%	86.57% \pm 4.43%	73.77% \pm 11.92%
PCA ($q = 7$)	CQG	66.61% \pm 1.11%	54.82% \pm 2.02%	77.97% \pm 1.73%	85.72% \pm 3.16%	61.61% \pm 3.16%
PCA ($q = 7$)	DMP	67.26% \pm 1.99%	57.53% \pm 3.91%	70.49% \pm 5.07%	81.35% \pm 10.80%	79.32% \pm 9.07%
PCA ($q = 6$)	CQG	65.39% \pm 1.23%	55.06% \pm 1.75%	80.84% \pm 1.79%	85.82% \pm 3.18%	45.97% \pm 3.81%
PCA ($q = 6$)	DMP	62.92% \pm 3.48%	53.25% \pm 9.70%	66.78% \pm 10.35%	78.25% \pm 13.29%	72.97% \pm 11.56%
PCA ($q = 5$)	CQG	65.89% \pm 1.14%	54.77% \pm 1.93%	83.42% \pm 1.49%	87.68% \pm 3.26%	42.13% \pm 3.65%
PCA ($q = 5$)	DMP	61.92% \pm 4.39%	49.72% \pm 8.94%	67.35% \pm 16.28%	72.80% \pm 18.71%	76.10% \pm 7.38%
PCA ($q = 4$)	CQG	60.11% \pm 1.25%	59.26% \pm 2.31%	65.23% \pm 2.43%	86.24% \pm 3.23%	39.19% \pm 4.48%
PCA ($q = 4$)	DMP	58.10% \pm 3.47%	42.53% \pm 10.12%	73.20% \pm 11.49%	86.87% \pm 5.27%	49.98% \pm 12.25%
PCA ($q = 3$)	CQG	58.49% \pm 0.99%	58.71% \pm 2.08%	71.14% \pm 1.68%	79.18% \pm 4.51%	17.72% \pm 2.96%
PCA ($q = 3$)	DMP	51.55% \pm 2.99%	35.83% \pm 8.03%	66.59% \pm 16.11%	87.43% \pm 5.50%	40.80% \pm 9.38%
PCA ($q = 2$)	CQG	55.69% \pm 0.96%	60.24% \pm 2.11%	71.14% \pm 1.86%	66.39% \pm 4.13%	0.00% \pm 0.00%
PCA ($q = 2$)	DMP	42.93% \pm 2.52%	33.28% \pm 4.45%	43.11% \pm 6.09%	82.92% \pm 6.83%	52.79% \pm 5.53%