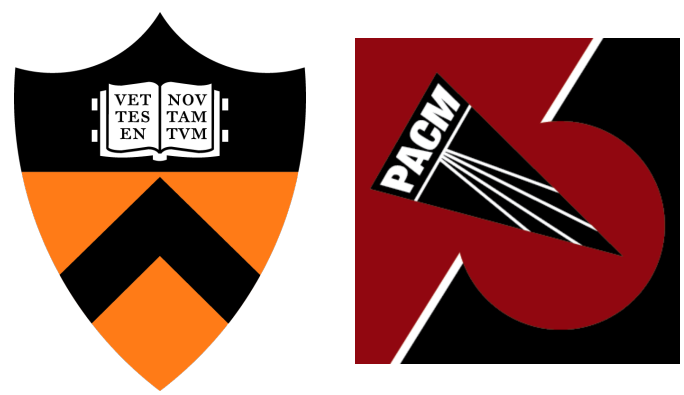




Generalized Method of Moments for Heteroscedastic Low-Rank Gaussian Mixtures

Diagonally-Weighted GMM (DGMM): scalable moment condition weighting

Liu Zhang¹, Oscar Mickelin², Sheng Xu¹, and Amit Singer^{1,3}

¹Program in Applied and Computational Mathematics (PACM), Princeton University

²Yau Mathematical Sciences Center, Tsinghua University

³Department of Mathematics, Princeton University



Model: heteroscedastic low-rank Gaussian mixture (GM)

► K -component GM:

$$\mathbf{y}_n \mid (h_n = j) \sim \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j),$$

$$\mathbb{P}(h_n = j) = \pi_j,$$

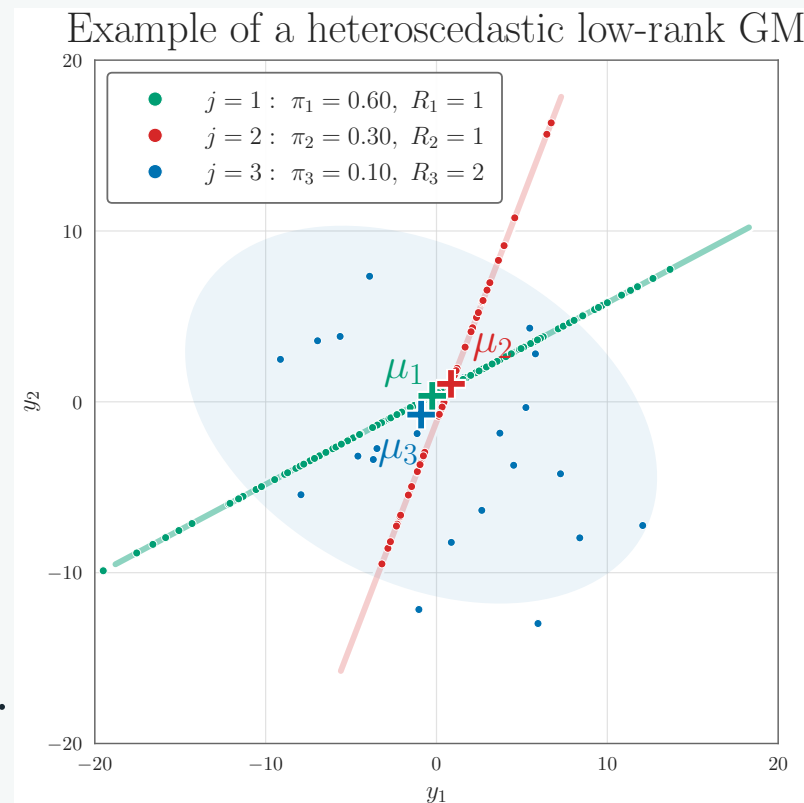
for $n = 1, \dots, N$, $j = 1, \dots, K$,

► Heteroscedastic:

component-specific covariances Σ_j ,

► Low-rank:

$$\Sigma_j = V_j V_j^\top, \quad \text{rank}(\Sigma_j) = R_j \leq R_{\max} \leq d.$$



Goal: to estimate $\boldsymbol{\theta}^* = (\pi_1^*, \dots, \pi_K^*; \boldsymbol{\mu}_1^*, \dots, \boldsymbol{\mu}_K^*; V_1^*, \dots, V_K^*) \in \Theta \subset \mathbb{R}^p$ from N observations, given K and $R_{\max}$.

Weak separation / low signal-to-noise ratio (SNR): $\|\Sigma_j\|_F \gg \|\boldsymbol{\mu}_j\|_2$.

Generalized method of moments (GMM)

► **Moment function:** for k -th model moment $\mathcal{M}^{(k)}(\boldsymbol{\theta}) \in S^k(\mathbb{R}^d)$ and $\mathbf{y} \in \mathbb{R}^d$, define

$$g(\boldsymbol{\theta}; \mathbf{y}) := \left(\text{vec}(\mathcal{M}^{(1)}(\boldsymbol{\theta}) - \mathbf{y}); \dots; \text{vec}(\mathcal{M}^{(L)}(\boldsymbol{\theta}) - \mathbf{y}^{\otimes L}) \right)^\top \in \mathbb{R}^q,$$

$$q = d + \dots + d^L = \mathcal{O}(d^L).$$

► **Classical method of moments** [Pearson, 1894]: solving for $\boldsymbol{\theta}$ using the sample moment conditions

$$\bar{g}_N(\boldsymbol{\theta}) := \frac{1}{N} \sum_{n=1}^N g(\boldsymbol{\theta}; \mathbf{y}_n) = \mathbf{0} \in \mathbb{R}^q.$$

► **GMM estimator:** solving a weighted least-squares problem:

$$\hat{\boldsymbol{\theta}}^{(\text{GMM})} = \arg \min_{\boldsymbol{\theta} \in \Theta} Q_N(\boldsymbol{\theta}) := \bar{g}_N(\boldsymbol{\theta})^\top W \bar{g}_N(\boldsymbol{\theta}).$$

► [Hansen, 1982]: $\hat{\boldsymbol{\theta}}^{(\text{GMM})}$ is **asymptotically efficient** when $W = cS^{-1}$ for any $c > 0$ and

$$S = \lim_{N \rightarrow \infty} \text{Var} \left[\frac{1}{\sqrt{N}} \sum_{n=1}^N g(\boldsymbol{\theta}^*; \mathbf{y}_n) \right].$$

► **Multi-step GMM optimization:**

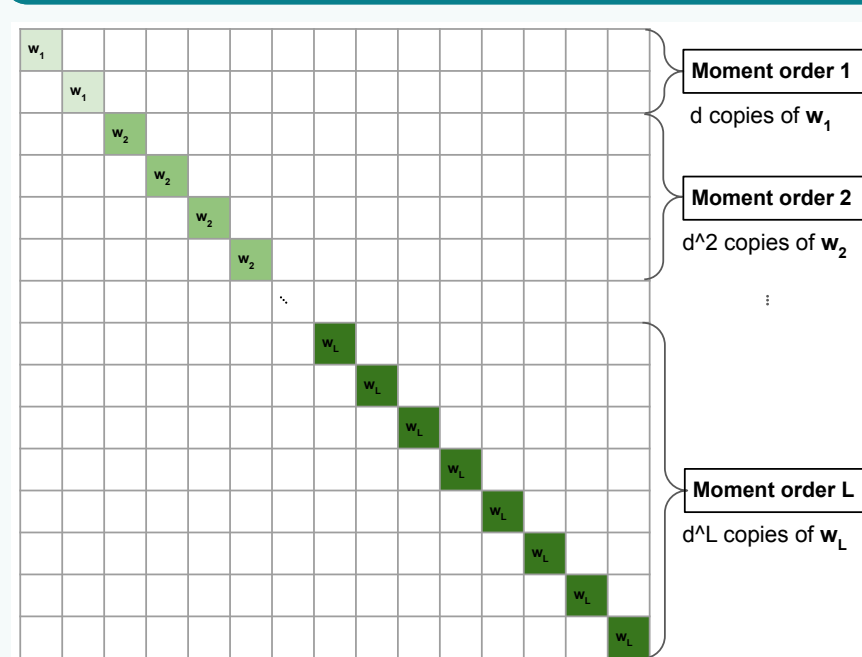
$$\hat{S}^{[t]} = \frac{1}{N} \sum_{n=1}^N g(\hat{\boldsymbol{\theta}}^{[t-1]}; \mathbf{y}_n) g(\hat{\boldsymbol{\theta}}^{[t-1]}; \mathbf{y}_n)^\top,$$

$$\hat{\boldsymbol{\theta}}^{[t]} = \arg \min_{\boldsymbol{\theta} \in \Theta} Q_N^{[t]}(\boldsymbol{\theta}) := \bar{g}_N(\boldsymbol{\theta})^\top \widehat{W}^{[t]} \bar{g}_N(\boldsymbol{\theta}), \quad \widehat{W}^{[t]} = \left(\hat{S}^{[t]} \right)^{-1}.$$

Computational bottlenecks:

- **Moment tensors:** $\mathcal{O}(d^L)$ entries.
- **Full GMM weighting matrix:** requires inverting an $\mathcal{O}(d^L)$ -by- $\mathcal{O}(d^L)$, often ill-conditioned covariance matrix.

DGMM: strategic weighting on moment conditions



Idea. Optimal order-wise diagonal approximate inverse:

$$\widehat{W}^{[t]} = \arg \min_{W \in \mathcal{W}} \left\| W \widehat{S}^{[t]} - I \right\|_F^2,$$

$$\mathcal{W} = \left\{ \text{diag} \left(\underbrace{w_1, \dots, w_1}_{d \text{ copies}}, \dots, \underbrace{w_L, \dots, w_L}_{d^L \text{ copies}} \right) \right\}.$$

$$\text{closed form: } \widehat{w}_k^{[t]} = \frac{\sum_{i \in \mathcal{I}_k} \widehat{S}_{ii}^{[t]}}{\sum_{i \in \mathcal{I}_k} \sum_{j=1}^q \left(\widehat{S}_{ij}^{[t]} \right)^2}, \quad k = 1, \dots, L.$$

Advantages.

- down-weight noisier moment orders
- avoid the unstable inverse $\left(\widehat{S}^{[t]} \right)^{-1}$
- reduce weight-estimation noise by only using L weights

Acknowledgements and references

Acknowledgements:

The DGMM work is supported in part by AFOSR FA9550-20-1-0266 and FA9550-23-1-0249, the Simons Foundation Math+X Investigator Award, NSF DMS 2009753, and NIH/NIGMS R01GM136780-01. We would like to thank Afonso S. Bandeira, Tamir Bendory, Charles Fefferman, Marc A. Gilles, Anya Katsevich, Joe Kileel, João M. Pereira, and Liza Rebrova for helpful discussions. In particular, we thank Marc A. Gilles for suggesting the randomly pivoted Cholesky algorithm to speed up the code and João M. Pereira for helpful suggestions on numerical optimization.

References:

- [Pearson, 1894]
- [Hansen, 1982]
- [OG17]

- K. Pearson. “Contributions to the mathematical theory of evolution”. In: *Philosophical Transactions of the Royal Society of London. A* 185 (1894), pp. 71–110.
- L. P. Hansen. “Large Sample Properties of Generalized Method of Moments Estimators”. In: *Econometrica* 50.4 (1982), pp. 1029–1054.
- D. Oglic and T. Gärtner. “Nyström method with kernel k -means++ samples as landmarks”. In: *International Conference on Machine Learning*. PMLR, 2017, pp. 2652–2660.

- J. M. Pereira, J. Kileel, and T. G. Kolda. “Tensor Moments of Gaussian Mixture Models: Theory and applications”. In: *arXiv preprint arXiv:2202.06930* (2022).
- L. Zhang, O. Mickelin, S. Xu, and A. Singer. “Diagonally-Weighted Generalized Method of Moments Estimation for Gaussian Mixture Modeling”. In: *arXiv preprint arXiv:2507.20459* (2025).

DGMM: fast moment computation

DGMM objective, gradients, weights are obtained without explicitly computing or storing the $\mathcal{O}(d^L)$ moment tensors or the $N \times N$ kernel Gram matrices.

1. Moment-cumulant-Bell-polynomial identities (building on [PKK22])

$$\phi^{(k)}(\boldsymbol{\theta}) := \left\| \mathcal{M}^{(k)}(\boldsymbol{\theta}) \right\|^2 = \sum_{i,j=1}^K \pi_i \pi_j B_k(\kappa_{ij}^{(1)}, \dots, \kappa_{ij}^{(k)}),$$

$$\psi^{(k)}(\boldsymbol{\theta}; \mathbf{y}_n) := \left\langle \mathcal{M}^{(k)}(\boldsymbol{\theta}), \mathbf{y}_n^{\otimes k} \right\rangle$$

$$= \sum_{j=1}^K \pi_j B_k(\mathbf{y}_n^T \boldsymbol{\mu}_j, \mathbf{y}_n^T V_j V_j^T \mathbf{y}_n, 0, \dots, 0).$$

- **Model term** $\phi^{(k)}(\boldsymbol{\theta}), \nabla \phi^{(k)}(\boldsymbol{\theta})$: $\mathcal{O}(k^2 K^2 + k K^2 d R_{\max}^2)$.
- **Model-data cross term** $\psi^{(k)}(\boldsymbol{\theta}; \mathbf{y}_n), \nabla \psi^{(k)}(\boldsymbol{\theta}; \mathbf{y}_n)$: $\mathcal{O}(k K + K d R_{\max})$.

2. Randomized Nyström low-rank approximation (building on [OG17])

$$\sum_{n'} H_{nn'}^{(k)}, \quad H_{nn'}^{(k)} := \langle \mathbf{y}_n^{\otimes k}, \mathbf{y}_{n'}^{\otimes k} \rangle = \langle \mathbf{y}_n, \mathbf{y}_{n'} \rangle^k.$$

- **Data term:** the inner-product sum above (used in the weight computation) takes $\mathcal{O}(N d m)$, with $m \ll N$ k -means++ landmarks.
- Under polynomial spectral decay of the kernel ($\lambda_i(H^{(k)}) \in \mathcal{O}(i^{-a}), a > 1$) and clustering potential condition, the expected error is

$$\mathbb{E} \left[\left\| H^{(k)} - \widetilde{H}_m^{(k)} \right\|_F \right] \in \mathcal{O} \left(\ln(m) \sqrt{N - m} \left(m^{-a+1/2} \right) \right).$$

DGMM: theoretical guarantees

Statistical guarantees ([ZMXS25, Thm. 4.1])

Under standard global and local identification conditions,

► **Consistent:**

$$\widehat{\boldsymbol{\theta}}^{(\text{DGMM})} \xrightarrow{p} \boldsymbol{\theta}^*.$$

► **Asymptotically normal:**

$$\sqrt{N} \left(\widehat{\boldsymbol{\theta}}^{(\text{DGMM})} - \boldsymbol{\theta}^* \right) \xrightarrow{d} \mathcal{N}(0, V^{(\text{DGMM})}).$$

► **Intermediate asymptotic efficiency:**

$$V^{(\text{DGMM})} = A^{-1} B A^{-1},$$

$$A = \sum_{k=1}^L w_k G_k^\top G_k, \quad B = \sum_{k,k'=1}^L w_k w_{k'} G_k^\top S_{kk'} G_{k'}.$$

Computational guarantees ([ZMXS25, Thm. 4.11])

► **Overall computational complexity of DGMM:**

$$\mathcal{O}(L^2 K^2 + L K^2 d R_{\max}^2 + N L K + N K d R_{\max} + N L d m)$$

for constant-sized Gaussian mixtures:
fixed $K, L, R_{\max}, m \implies$ **linear in N and d**

► **MM:** $\mathcal{O}(N d^{L+1} K R_{\max})$

► **GMM:** $\mathcal{O}(N d^{2L} + d^{3L} + N d^{L+1} K R_{\max} + d^{2L+1} K R_{\max})$

DGMM numerical results

Covariance estimation error across four cases

Regime	MM	GMM	DGMM
$d = 10$, same rank	0.01244	0.05239	0.00581
$d = 100$, same rank	7.4772*	NA [†]	0.01556
$d = 10$, diff. ranks	0.13736	0.33879	0.01986
$d = 100$, diff. ranks	0.08657*	NA [†]	0.00744

Runtime across four cases

Regime	MM	GMM	DGMM
$d = 10$, same rank	28.3s $\pm$ 2.71s	11m43s $\pm$ 12.1s	14.3s $\pm$ 0.089s
$d = 100$, same rank	17m29s $\pm$ 24.9s*	NA [†]	2m51s $\pm$ 2.31s
$d = 10$, diff. ranks	2m29s $\pm$ 1.66s	7m39s $\pm$ 10.7s	42.2s $\pm$ 0.396s
$d = 100$, diff. ranks	17m41s $\pm$ 2.4s*	NA [†]	6m24s $\pm$ 11.9s

[†] Standard MM and GMM are computationally infeasible in the two $d = 100$ cases;

* The reported $d = 100$ MM uses Bell-polynomial moment computation with identical weights.