

Generalized Method of Moments (GMM) for Gaussian Mixture Modeling

Diagonally-weighted GMM
(DGMM) - scalability

Spectral-Gradient-Reweighted GMM
(SGR-GMM) - robustness

Liu Zhang¹

DGMM: joint work with Oscar Mickelin², Sheng Xu¹, and Amit Singer^{1,3}

SGR-GMM: joint work with Amit Singer^{1,3}

¹Program in Applied and Computational Mathematics, Princeton University

²Yau Mathematical Sciences Center, Tsinghua University

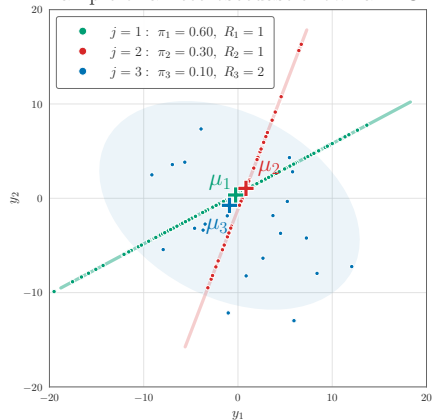
³Department of Mathematics, Princeton University

Gaussian mixture (GM) modeling

Heteroscedastic low-rank GM

- **K-component GM:** $y_n \mid (h_n = j) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_j, \Sigma_j)$, with $\mathbb{P}(h_n = j) = \pi_j$, $n = 1, \dots, N, j = 1, \dots, K$,

Example of a heteroscedastic low-rank GM



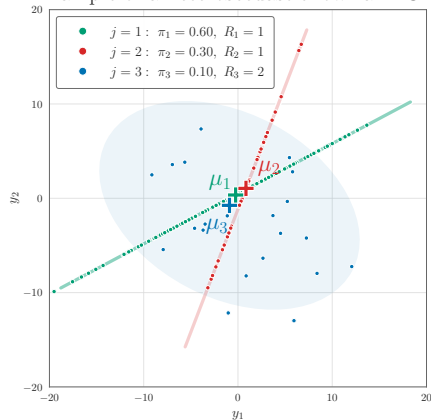
Zhang-Mickelin-Xu-Singer (2025), Model 1.1.

Gaussian mixture (GM) modeling

Heteroscedastic low-rank GM

- ▶ **K-component GM:** $\mathbf{y}_n \mid (h_n = j) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)$, with $\mathbb{P}(h_n = j) = \pi_j$, $n = 1, \dots, N, j = 1, \dots, K$,
- ▶ **Heteroscedastic:** component-specific Σ_j ,

Example of a heteroscedastic low-rank GM



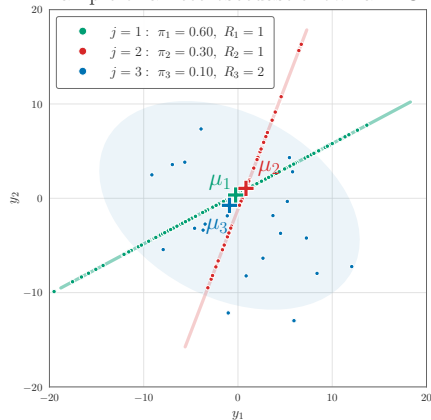
Zhang-Mickelin-Xu-Singer (2025), Model 1.1.

Gaussian mixture (GM) modeling

Heteroscedastic low-rank GM

- ▶ **K-component GM:** $\mathbf{y}_n \mid (h_n = j) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)$, with $\mathbb{P}(h_n = j) = \pi_j$, $n = 1, \dots, N, j = 1, \dots, K$,
- ▶ **Heteroscedastic:** component-specific Σ_j ,
- ▶ **Low-rank:** $\text{rank}(\Sigma_j) = R_j \leq R_{\max} \leq d, \Sigma_j = \mathbf{V}_j \mathbf{V}_j^\top$.

Example of a heteroscedastic low-rank GM



Zhang-Mickelin-Xu-Singer (2025), Model 1.1.

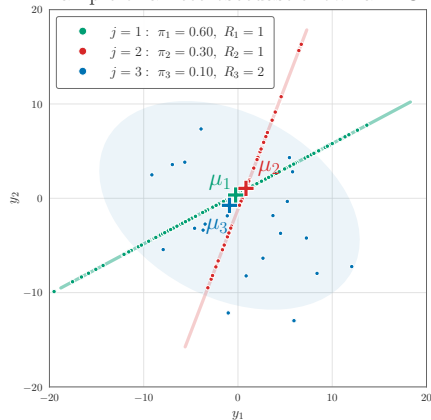
Gaussian mixture (GM) modeling

Heteroscedastic low-rank GM

- ▶ **K-component GM:** $y_n \mid (h_n = j) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_j, \Sigma_j)$, with $\mathbb{P}(h_n = j) = \pi_j$, $n = 1, \dots, N$, $j = 1, \dots, K$,
- ▶ **Heteroscedastic:** component-specific Σ_j ,
- ▶ **Low-rank:** $\text{rank}(\Sigma_j) = R_j \leq R_{\max} \leq d$, $\Sigma_j = V_j V_j^\top$.

Goal: to estimate $\theta^\star = (\pi_1^\star, \dots, \pi_K^\star; \mu_1^\star, \dots, \mu_K^\star; V_1^\star, \dots, V_K^\star)$ from N observations, given K and $R_{\max}$.

Example of a heteroscedastic low-rank GM



Zhang-Mickelin-Xu-Singer (2025), Model 1.1.

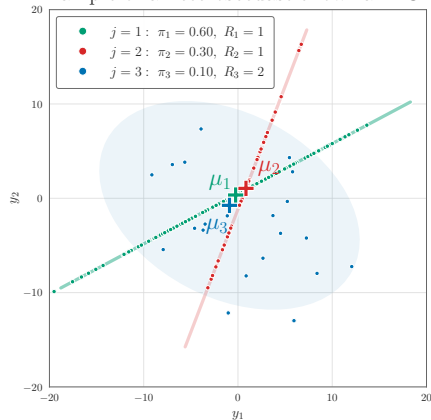
Gaussian mixture (GM) modeling

Heteroscedastic low-rank GM

- ▶ **K-component GM:** $\mathbf{y}_n \mid (h_n = j) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)$, with $\mathbb{P}(h_n = j) = \pi_j$, $n = 1, \dots, N, j = 1, \dots, K$,
- ▶ **Heteroscedastic:** component-specific Σ_j ,
- ▶ **Low-rank:** $\text{rank}(\Sigma_j) = R_j \leq R_{\max} \leq d, \Sigma_j = \mathbf{V}_j \mathbf{V}_j^\top$.

Weak-separation / low-SNR: $\|\Sigma_j\|_F \gg \|\boldsymbol{\mu}_j\|_2$.

Example of a heteroscedastic low-rank GM



Zhang-Mickelin-Xu-Singer (2025), Model 1.1.

Gaussian mixture (GM) modeling

Heteroscedastic low-rank GM

- ▶ **K-component GM:** $y_n \mid (h_n = j) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_j, \Sigma_j)$, with $\mathbb{P}(h_n = j) = \pi_j$, $n = 1, \dots, N$, $j = 1, \dots, K$,
- ▶ **Heteroscedastic:** component-specific Σ_j ,
- ▶ **Low-rank:** $\text{rank}(\Sigma_j) = R_j \leq R_{\max} \leq d$, $\Sigma_j = V_j V_j^\top$.

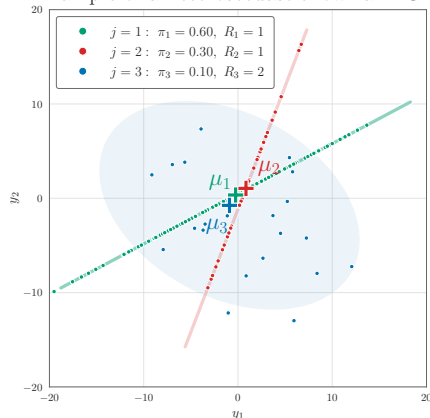
Weak-separation / low-SNR: $\|\Sigma_j\|_F \gg \|\mu_j\|_2$.

Additive noise + strong ϵ -contamination:

$$\xi_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_\xi),$$
$$\check{y}_n = \begin{cases} y_n + \xi_n, & (\text{if } n \in \mathcal{I}_{\text{in}}) \\ a_n, & (\text{if } n \in \mathcal{I}_{\text{out}}) \end{cases},$$

with a partition $\{1, \dots, N\} = \mathcal{I}_{\text{in}} \sqcup \mathcal{I}_{\text{out}}, |\mathcal{I}_{\text{out}}| \leq \epsilon N$.

Example of a heteroscedastic low-rank GM



Zhang-Mickelin-Xu-Singer (2025), Model 1.1.

Method of Moments (MM)

Classical method of moments [Pearson, 1894]
GM parameter estimation by matching the moments.

Generalized method of moments [Hansen, 1982]
Asymptotically efficient when $W = cS^{-1}$.

- ▶ **Target GM parameter:** $\theta^\star = (\pi_1^\star, \dots, \pi_K^\star; \mu_1^\star, \dots, \mu_K^\star; V_1^\star, \dots, V_K^\star) \in \Theta \subset \mathbb{R}^p$.
- ▶ **Moment function:** for k -th model moment $\mathcal{M}^{(k)}(\theta) \in \mathbb{S}^k(\mathbb{R}^d)$ and $\mathbf{y} \in \mathbb{R}^d$, define

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

Method of Moments (MM)

Classical method of moments [Pearson, 1894]
GM parameter estimation by matching the moments.

Generalized method of moments [Hansen, 1982]
Asymptotically efficient when $W = cS^{-1}$.

- ▶ **Target GM parameter:** $\theta^\star = (\pi_1^\star, \dots, \pi_K^\star; \mu_1^\star, \dots, \mu_K^\star; V_1^\star, \dots, V_K^\star) \in \Theta \subset \mathbb{R}^p$.
- ▶ **Moment function:** for k -th model moment $\mathcal{M}^{(k)}(\theta) \in \mathbb{S}^k(\mathbb{R}^d)$ and $y \in \mathbb{R}^d$, define

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

- ▶ **Population moment conditions:** in expectation, model moments match data moments at $\theta^\star$:

$$\mathbb{E}[g(\theta^\star; y)] = \mathbf{0} \in \mathbb{R}^q.$$

- ▶ **Sample moment conditions:** $\hat{\theta}^{(\text{MM})}$ is obtained by solving for θ using the sample analogue:

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

Generalized Method of Moments (GMM)

Classical method of moments [Pearson, 1894]
GM parameter estimation by matching the moments.

Generalized method of moments [Hansen, 1982]
Asymptotically efficient when $W = cS^{-1}$.

A **GMM estimator** is obtained by solving a weighted least-squares problem:

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

Generalized Method of Moments (GMM)

Classical method of moments [Pearson, 1894]
GM parameter estimation by matching the moments.

Generalized method of moments [Hansen, 1982]
Asymptotically efficient when $W = cS^{-1}$.

A **GMM estimator** is obtained by solving a weighted least-squares problem:

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

- ▶ Reduces to $\widehat{\boldsymbol{\theta}}^{(\text{MM})}$ when $W = I$,
- ▶ [Hansen, 1982]: asymptotically efficient when $W = cS^{-1}$ for any $c > 0$ and

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

Generalized Method of Moments (GMM)

Classical method of moments [Pearson, 1894]
GM parameter estimation by matching the moments.

Generalized method of moments [Hansen, 1982]
Asymptotically efficient when $W = cS^{-1}$.

Multi-step GMM optimization is often used in practice to alternately obtain a consistent estimator $\widehat{S}^{[t]}$ and a consistent estimator $\widehat{\theta}^{[t]}$:

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

This terminates when $t = T$ or $\left\| \widehat{\theta}^{[t]} - \widehat{\theta}^{[t-1]} \right\| < \varepsilon_\theta$ with stopping criteria constants T, ε_θ .

Generalized Method of Moments (GMM)

► Multi-step GMM optimization:

$$\widehat{\boldsymbol{\theta}}^{[t]} = \arg \min_{\boldsymbol{\theta} \in \Theta} Q_N^{[t]}(\boldsymbol{\theta}) := \bar{\mathbf{g}}_N(\boldsymbol{\theta})^\top \widehat{\mathbf{W}}^{[t]} \bar{\mathbf{g}}_N(\boldsymbol{\theta}), \quad \widehat{\mathbf{W}}^{[t]} = \left(\widehat{\mathbf{S}}^{[t]} \right)^{-1} \in \mathbb{R}^{q \times q}, \quad q = O(d^L).$$

Computational bottlenecks

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

Generalized Method of Moments (GMM)

► Multi-step GMM optimization:

$$\widehat{\boldsymbol{\theta}}^{[t]} = \arg \min_{\boldsymbol{\theta} \in \Theta} Q_N^{[t]}(\boldsymbol{\theta}) := \bar{\mathbf{g}}_N(\boldsymbol{\theta})^\top \widehat{\mathbf{W}}^{[t]} \bar{\mathbf{g}}_N(\boldsymbol{\theta}), \quad \widehat{\mathbf{W}}^{[t]} = \left(\widehat{\mathbf{S}}^{[t]} \right)^{-1} \in \mathbb{R}^{q \times q}, \quad q = O(d^L).$$

Computational bottlenecks

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

Outlier sensitivity

- Sample moment conditions $\bar{\mathbf{g}}_N$: each observation has equal contribution.
- Effects of outliers get amplified.

Generalized Method of Moments (GMM)

► Multi-step GMM optimization:

$$\widehat{\theta}^{[t]} = \arg \min_{\theta \in \Theta} Q_N^{[t]}(\theta) := \bar{g}_N(\theta)^\top \widehat{W}^{[t]} \bar{g}_N(\theta), \quad \widehat{W}^{[t]} = \left(\widehat{S}^{[t]} \right)^{-1} \in \mathbb{R}^{q \times q}, \quad q = O(d^L).$$

Diagonally-weighted GMM (DGMM) [ZMXS25]

- Strategic weighting for moment conditions.
- Fast moment computation.
- Theoretical guarantees.

Outlier sensitivity

- Sample moment conditions $\bar{g}_N$: each observation has equal contribution.
- Effects of outliers get amplified.

Generalized Method of Moments (GMM)

► Multi-step GMM optimization:

$$\widehat{\theta}^{[t]} = \arg \min_{\theta \in \Theta} Q_N^{[t]}(\theta) := \bar{g}_N(\theta)^\top \widehat{W}^{[t]} \bar{g}_N(\theta), \quad \widehat{W}^{[t]} = \left(\widehat{S}^{[t]} \right)^{-1} \in \mathbb{R}^{q \times q}, \quad q = O(d^L).$$

Diagonally-weighted GMM (DGMM) [ZMXS25]

- Strategic weighting for moment conditions.
- Fast moment computation.
- Theoretical guarantees.

Spectral-Gradient-Reweighted GMM (SGR-GMM) [ZS26]

- Strategic weighting for per-observation moment-matching gradients.
- Robust DGMM specialization.
- Theoretical guarantees.

Diagonally-weighted Generalized Method of Moments (DGMM)

- Multi-step GMM optimization:

$$\hat{\theta}^{[t]} = \arg \min_{\theta \in \Theta} Q_N^{[t]}(\theta) := \bar{g}_N(\theta)^\top \hat{W}^{[t]} \bar{g}_N(\theta), \quad \hat{W}^{[t]} = \left(\hat{S}^{[t]} \right)^{-1} \in \mathbb{R}^{q \times q}, \quad q = O(d^L).$$

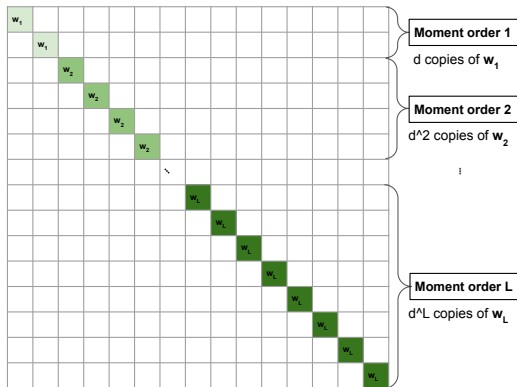
Diagonally-weighted GMM (DGMM) [ZMXS25]

- Strategic weighting for moment conditions.
- Fast moment computation.
- Theoretical guarantees.

Spectral-Gradient-Reweighted GMM (SGR-GMM) [ZS26]

- Strategic weighting for per-observation moment-matching gradients;
- Robust DGMM specialization;
- Theoretical guarantees.

DGMM: strategic weighting for moment conditions



DGMM weighting matrix.

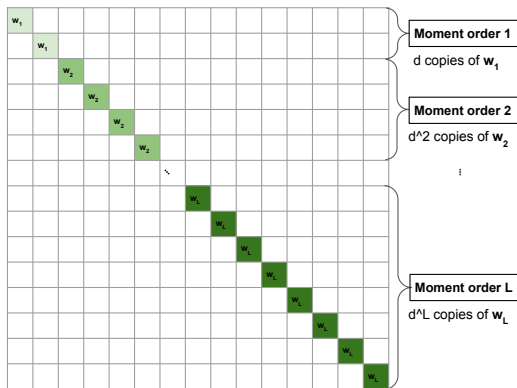
Idea. Optimal order-wise diagonal approximate inverse:

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

where $\mathcal{W}$ is the set of diagonal matrices of the form

$$\text{diag}(\underbrace{w_1, \dots, w_1}_{d \text{ copies}}, \dots, \underbrace{w_L, \dots, w_L}_{d^L \text{ copies}}).$$

DGMM: strategic weighting for moment conditions



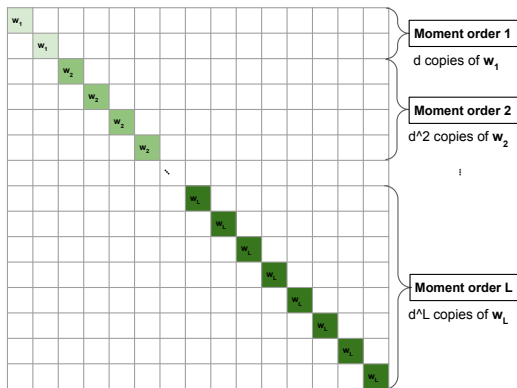
DGMM weighting matrix.

Idea. Optimal order-wise diagonal approximate inverse:

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

$$\hat{W}^{[t]} = \text{diag}(\underbrace{\hat{w}_1^{[t]}, \dots, \hat{w}_1^{[t]}}_{d \text{ copies}}, \dots, \underbrace{\hat{w}_L^{[t]}, \dots, \hat{w}_L^{[t]}}_{d^L \text{ copies}})$$

DGMM: strategic weighting for moment conditions



DGMM weighting matrix.

Idea. Optimal order-wise diagonal approximate inverse:

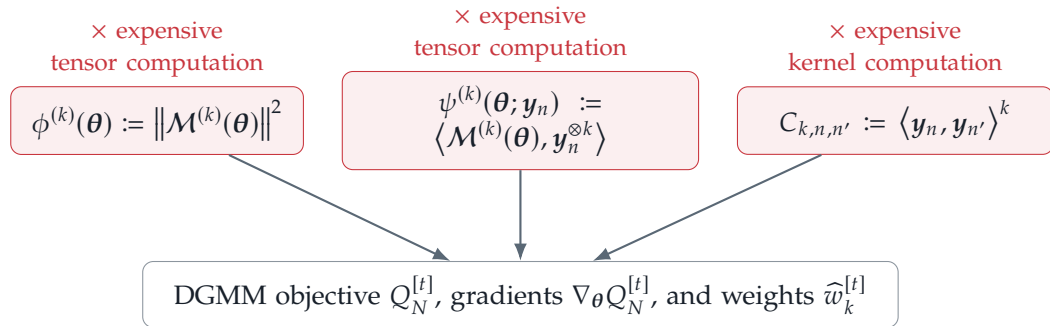
$$\text{closed form: } \hat{w}_k^{[t]} = \frac{\sum_{i \in \mathcal{I}_k} \hat{S}_{ii}^{[t]}}{\sum_{i \in \mathcal{I}_k} \sum_{j=1}^q \left(\hat{S}_{ij}^{[t]} \right)^2}, \quad k = 1, \dots, L.$$

$$\hat{W}^{[t]} = \text{diag}(\underbrace{\hat{w}_1^{[t]}, \dots, \hat{w}_1^{[t]}}_{d \text{ copies}}, \dots, \underbrace{\hat{w}_L^{[t]}, \dots, \hat{w}_L^{[t]}}_{d^L \text{ copies}})$$

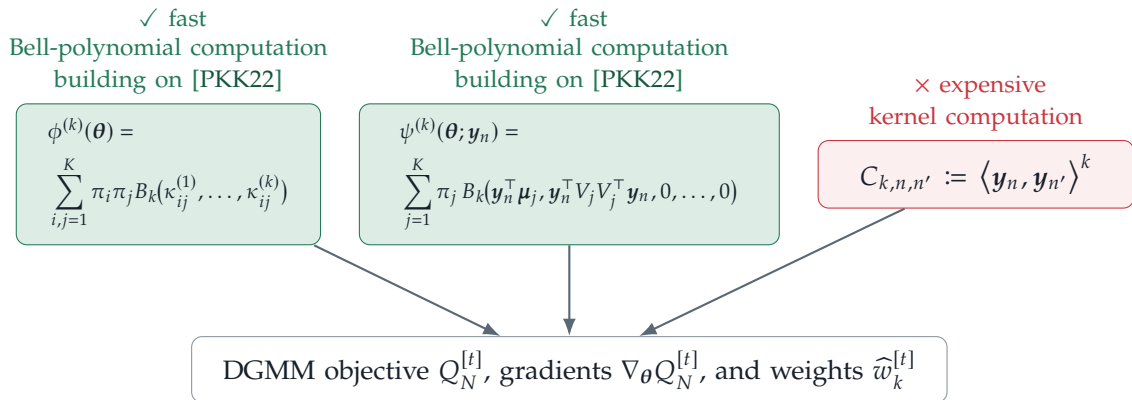
Advantages.

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

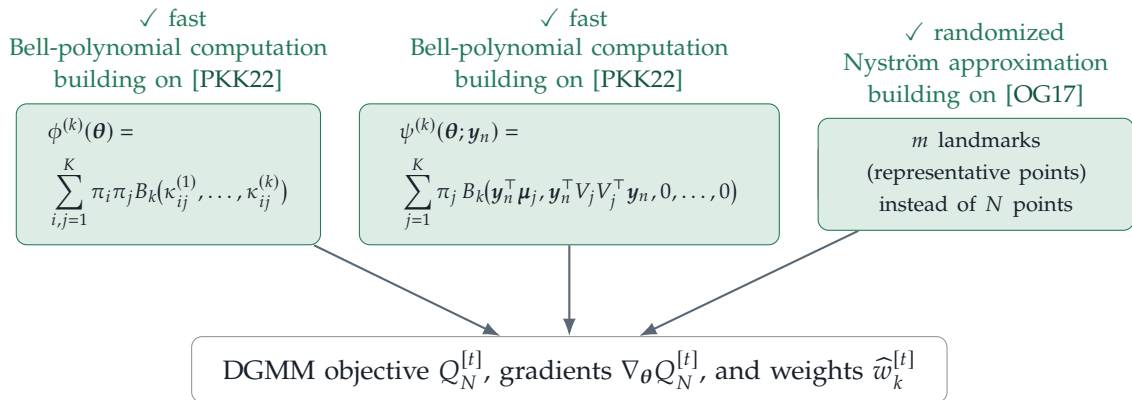
DGMM: fast moment computation



DGMM: fast moment computation



DGMM: fast moment computation



DGMM: theoretical guarantees

Statistical guarantees

Under standard global and local identification conditions,

► **Consistent:**

$$\widehat{\theta}^{(\text{DGMM})} \xrightarrow{p} \theta^\star$$

► **Asymptotically normal:**

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

► **Intermediate asymptotic efficiency:**

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

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

[Zhang-Mickelin-Xu-Singer, 2025], Thm. 4.1.

DGMM: theoretical guarantees

Statistical guarantees

Under standard global and local identification conditions,

► **Consistent:**

$$\widehat{\theta}^{(\text{DGMM})} \xrightarrow{P} \theta^\star$$

► **Asymptotically normal:**

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

► **Intermediate asymptotic efficiency:**

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

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

Zhang-Mickelin-Xu-Singer (2025), Thm. 4.1.

Computational guarantees

► **Overall computational complexity of DGMM:**

$$O(L^2 K^2 + L K^2 d R_{\max}^2 + NLK + NKdR_{\max} + NLdm).$$

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

► **MM:** $O(Nd^{L+1}KR_{\max})$

► **GMM:** $O(Nd^{2L} + d^{3L} + Nd^{L+1}KR_{\max} + d^{2L+1}KR_{\max})$

Zhang-Mickelin-Xu-Singer (2025), Thm. 4.11.

DGMM numerical results

Covariance error Err_Σ across four cases

lower is better

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

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

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

DGMM improves parameter estimation error across the four cases.

DGMM: numerical results

Runtime across four cases

lower is better

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 baseline uses Bell-polynomial moment computation with identical weights.

DGMM improves runtime across the four cases.

Spectral-Gradient-Reweighted GMM (SGR-GMM)

- Multi-step GMM optimization:

$$\hat{\theta}^{[t]} = \arg \min_{\theta \in \Theta} Q_N^{[t]}(\theta) := \bar{g}_N(\theta)^\top \hat{W}^{[t]} \bar{g}_N(\theta), \quad \hat{W}^{[t]} = \left(\hat{S}^{[t]} \right)^{-1} \in \mathbb{R}^{q \times q}, \quad q = O(d^L).$$

Diagonally-weighted GMM (DGMM) [ZMXS25]

- Strategic weighting for moment conditions.
- Fast moment computation.
- Theoretical guarantees.

Spectral-Gradient-Reweighted GMM (SGR-GMM) [ZS26]

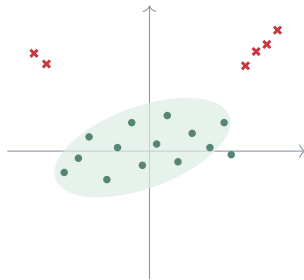
- Strategic weighting for per-observation moment-matching gradients.
- Robust DGMM specialization.
- Theoretical guarantees.

SGR-GMM: motivations

Strong ϵ -contamination

$$\check{y}_n = \begin{cases} y_n, & (\text{if } n \in \mathcal{I}_{\text{in}}) \\ a_n, & (\text{if } n \in \mathcal{I}_{\text{out}}) \end{cases},$$

with a partition $\{1, \dots, N\} = \mathcal{I}_{\text{in}} \sqcup \mathcal{I}_{\text{out}}, |\mathcal{I}_{\text{out}}| \leq \epsilon N$.



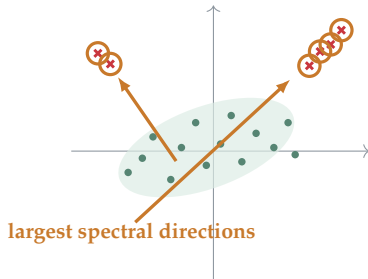
SGR-GMM: motivations

Strong ϵ -contamination

$$\check{y}_n = \begin{cases} y_n, & (\text{if } n \in \mathcal{I}_{\text{in}}) \\ a_n, & (\text{if } n \in \mathcal{I}_{\text{out}}) \end{cases},$$

with a partition $\{1, \dots, N\} = \mathcal{I}_{\text{in}} \sqcup \mathcal{I}_{\text{out}}, |\mathcal{I}_{\text{out}}| \leq \epsilon N$.

Spectral principle (e.g., [LRV16; DKKLMS19; DKKLSS19; DM22; DK23]) for moment-matching gradients: if a subset of outliers substantially shifts the gradient mean, it must create a large spectral direction in the gradient covariance.

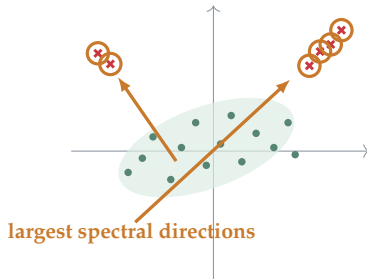


SGR-GMM: motivations

Spectral principle (e.g., [LRV16; DKKLMS19; DKKLSS19; DM22; DK23]) for moment-matching gradients: if a subset of outliers substantially shifts the gradient mean, it must create a large spectral direction in the gradient covariance.

The feasibility problem is nonconvex in w :
to find $w \in \Delta_{N,\varepsilon}$, such that

$$\left\| \sum_n w_n \left(\check{g}_n^{(k)} - \overline{\check{g}_w^{(k)}} \right) \left(\check{g}_n^{(k)} - \overline{\check{g}_w^{(k)}} \right)^\top \right\|_{\text{op}} \leq C_{\text{stop},k}, \quad \overline{\check{g}_w^{(k)}} = \sum_n w_n \check{g}_n^{(k)}.$$

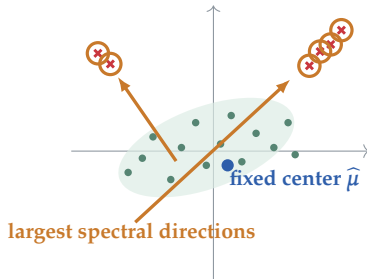


SGR-GMM: motivations

Spectral principle (e.g., [LRV16; DKKLMS19; DKKLSS19; DM22; DK23]) for moment-matching gradients: if a subset of outliers substantially shifts the gradient mean, it must create a large spectral direction in the gradient covariance.

For arbitrary fixed-center $\hat{\mu}$, the feasibility problem is convex in w : to find $w \in \Delta_{N,\varepsilon}$, such that

$$\gamma(w; \hat{\mu}) = \left\| \sum_n w_n \left(\check{g}_n^{(k)} - \hat{\mu} \right) \left(\check{g}_n^{(k)} - \hat{\mu} \right)^\top \right\|_{\text{op}} \leq C_{\text{stop},k}.$$



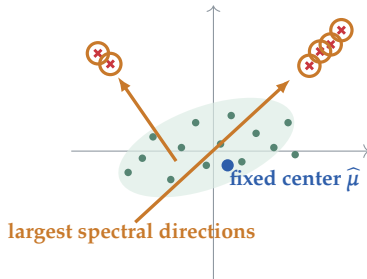
SGR-GMM: motivations

Spectral principle (e.g., [LRV16; DKKLMS19; DKKLSS19; DM22; DK23]) for moment-matching gradients: if a subset of outliers substantially shifts the gradient mean, it must create a large spectral direction in the gradient covariance.

Rewritten as a convex-concave game:

$$\begin{aligned}\gamma(w; \hat{\mu}) &= \left\| \sum_n w_n \left(\check{g}_n^{(k)} - \hat{\mu} \right) \left(\check{g}_n^{(k)} - \hat{\mu} \right)^\top \right\|_{\text{op}} \\ &= \max_{\rho \in \mathcal{D}_p} \sum_n w_n \left(\check{g}_n^{(k)} - \hat{\mu} \right)^\top \rho \left(\check{g}_n^{(k)} - \hat{\mu} \right).\end{aligned}$$

→ classical entropy-regularized multiplicative-weights (MW) - matrix multiplicative-weights (MMW) algorithm [AHK12].



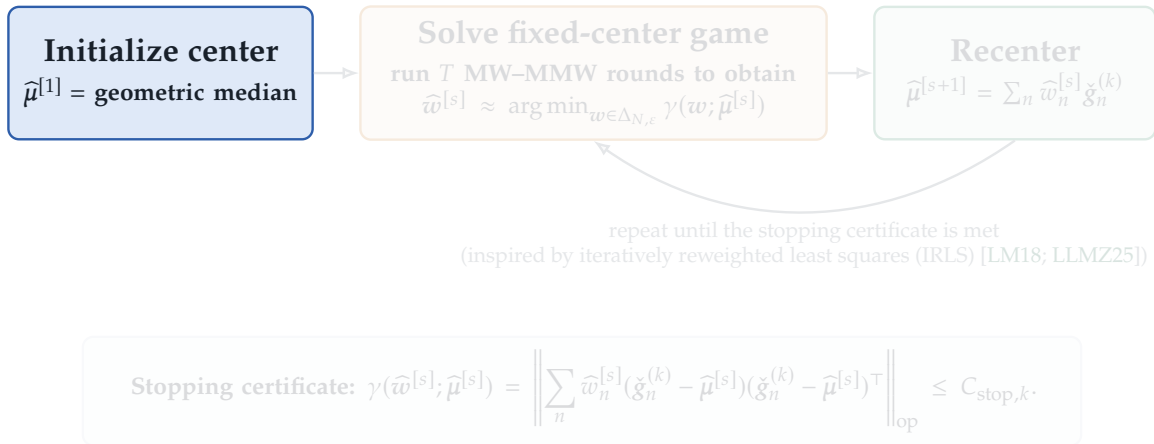
Density-matrix player:

expose the spectral directions with the largest weighted energy.

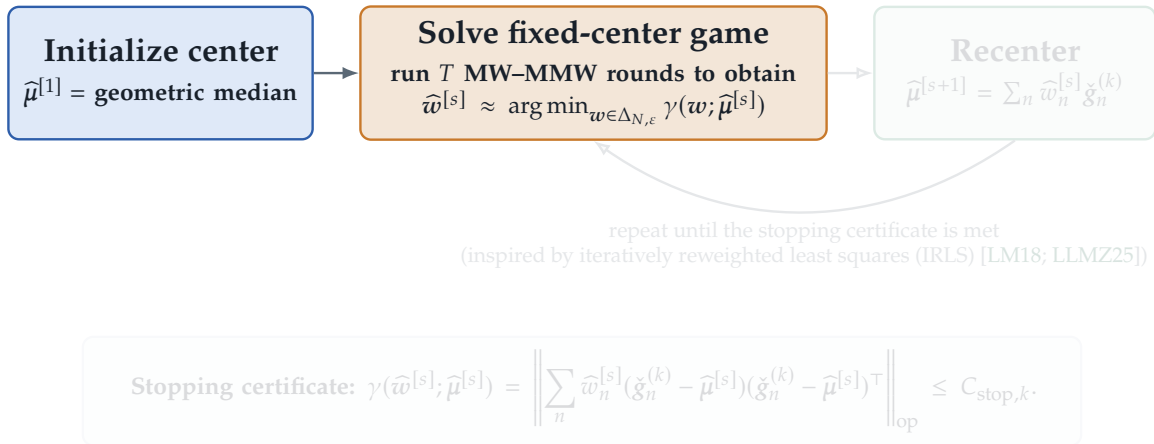
Gradient-weight player:

suppress gradients most responsible for the exposed directions.

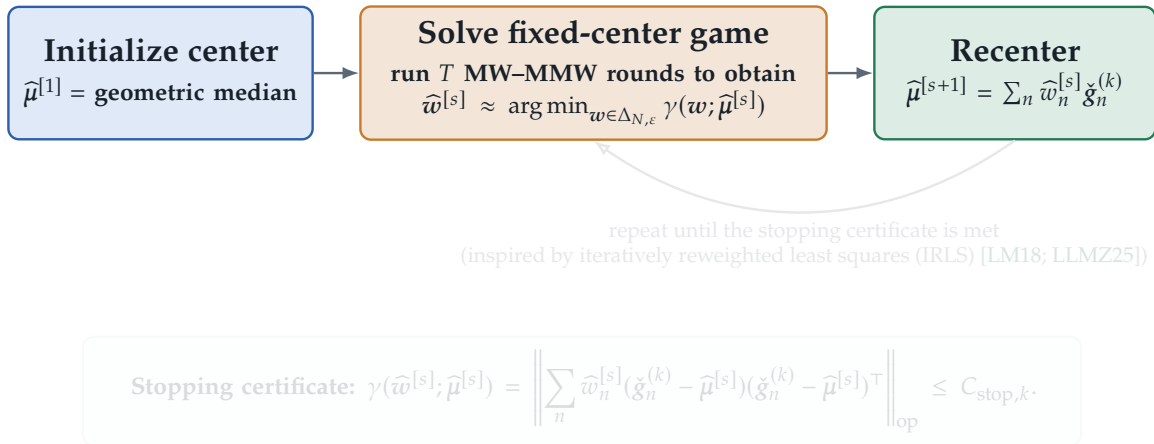
SGR primitive: solve, recenter, certify



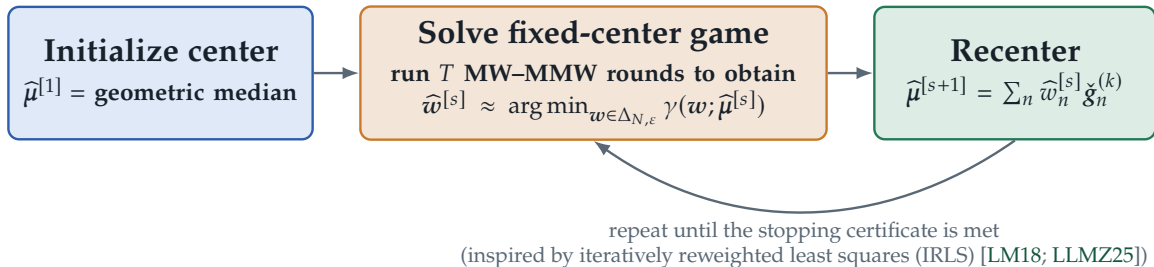
SGR primitive: solve, recenter, certify



SGR primitive: solve, recenter, certify

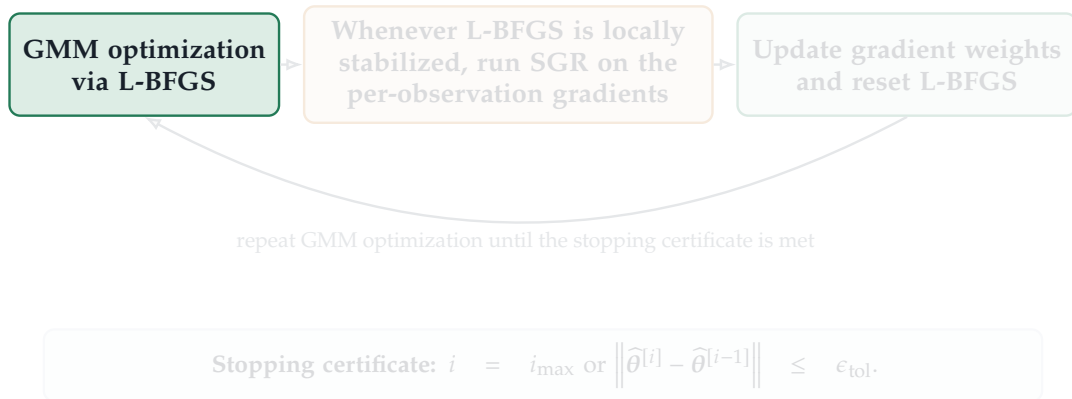


SGR primitive: solve, recenter, certify

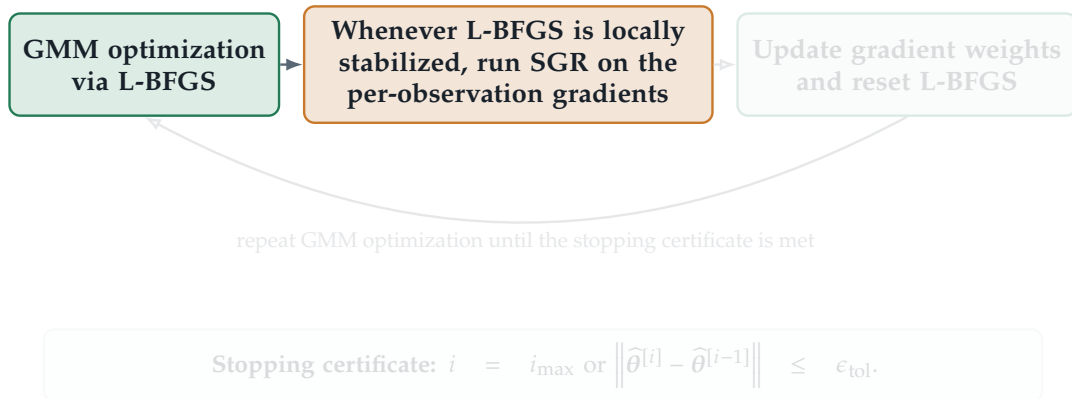


Stopping certificate: $\gamma(\hat{w}^{[s]}; \hat{\mu}^{[s]}) = \left\| \sum_n \hat{w}_n^{[s]} (\check{g}_n^{(k)} - \hat{\mu}^{[s]}) (\check{g}_n^{(k)} - \hat{\mu}^{[s]})^\top \right\|_{\text{op}} \leq C_{\text{stop},k}.$

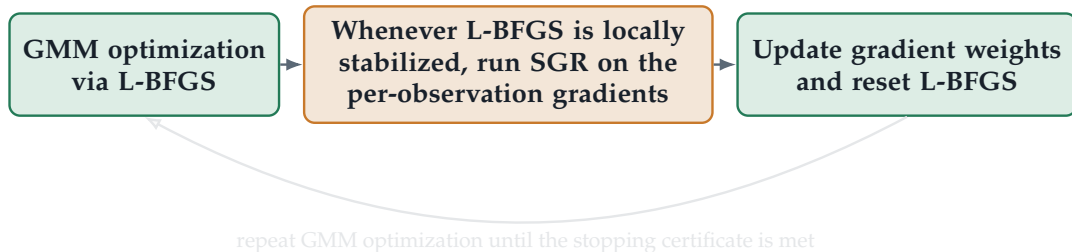
SGR-GMM: robustify, reweight, certify



SGR-GMM: robustify, reweight, certify

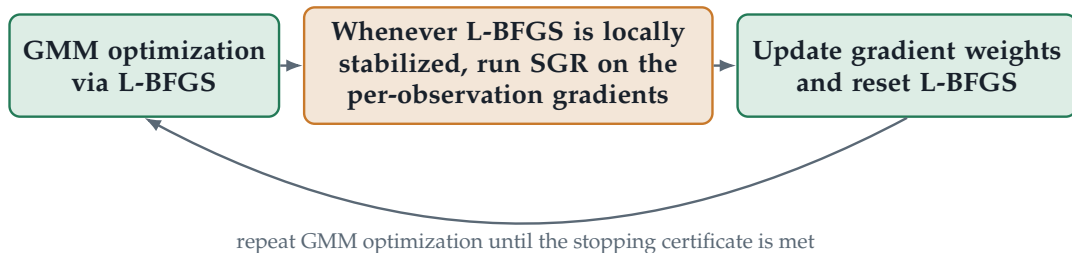


SGR-GMM: robustify, reweight, certify



Stopping certificate: $i = i_{\max}$ or $\|\hat{\theta}^{[i]} - \hat{\theta}^{[i-1]}\| \leq \epsilon_{\text{tol}}$.

SGR-GMM: robustify, reweight, certify



Stopping certificate: $i = i_{\max}$ or $\left\| \hat{\theta}^{[i]} - \hat{\theta}^{[i-1]} \right\| \leq \epsilon_{\text{tol}}$.

SGR-GMM: theoretical guarantees

1. Fixed-center game regret bound

$$\gamma(\widehat{\mathbf{w}}^{[s]}; \widehat{\boldsymbol{\mu}}^{[s]}) - \text{OPT}(\widehat{\boldsymbol{\mu}}^{[s]}) \leq 4v \left(\sqrt{\frac{\log(1/(1-\varepsilon))}{T}} + \sqrt{\frac{\log p}{T}} \right)$$

Zhang-Singer (2026), Thm. 3.14.

SGR-GMM: theoretical guarantees

1. Fixed-center game regret bound

$$\gamma(\widehat{w}^{[s]}; \widehat{\mu}^{[s]}) - \text{OPT}(\widehat{\mu}^{[s]}) \leq 4v \left(\sqrt{\frac{\log(1/(1-\varepsilon))}{T}} + \sqrt{\frac{\log p}{T}} \right)$$

Zhang-Singer (2026), Thm. 3.14.

2. Fixed-center updates convergence

$$e^{[s+1]} \leq \alpha_\varepsilon e^{[s]} + R_{\varepsilon, T}, \quad \alpha_\varepsilon = \sqrt{\frac{\varepsilon}{1-2\varepsilon}} < 1 \quad \text{for } 0 \leq \varepsilon < \frac{1}{3}$$

Zhang-Singer (2026), Thm. 3.23.

SGR-GMM: theoretical guarantees

1. Fixed-center game regret bound

$$\gamma(\widehat{w}^{[s]}; \widehat{\mu}^{[s]}) - \text{OPT}(\widehat{\mu}^{[s]}) \leq 4v \left(\sqrt{\frac{\log(1/(1-\varepsilon))}{T}} + \sqrt{\frac{\log p}{T}} \right)$$

Zhang-Singer (2026), Thm. 3.14.

2. Fixed-center updates convergence

$$e^{[s+1]} \leq \alpha_\varepsilon e^{[s]} + R_{\varepsilon, T}, \quad \alpha_\varepsilon = \sqrt{\frac{\varepsilon}{1-2\varepsilon}} < 1 \quad \text{for } 0 \leq \varepsilon < \frac{1}{3}$$

Zhang-Singer (2026), Thm. 3.23.

3. Local finite-sample parameter error

$$\left\| \widehat{\theta}^{(\text{SGR-GMM})} - \theta^\star \right\|_2 \leq \frac{2}{\lambda^\star} \left[\sum_{k=1}^L a_k \left(\delta_{\mu, k}(\zeta) + \alpha_\varepsilon \sqrt{C_k} \right) + \delta_{\text{opt}} \right]$$

Zhang-Singer (2026), Thm. 3.33.

SGR-GMM: RobustDGMM specialization

Heteroscedastic
low-rank GM



Additive noise +
strong ϵ -contamination

$$\xi_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_\xi),$$
$$\check{y}_n = \begin{cases} y_n + \xi_n, & (\text{if } n \in \mathcal{I}_{\text{in}}) \\ a_n, & (\text{if } n \in \mathcal{I}_{\text{out}}) \end{cases},$$

with a partition $\{1, \dots, N\} = \mathcal{I}_{\text{in}} \sqcup \mathcal{I}_{\text{out}}, |\mathcal{I}_{\text{out}}| \leq \epsilon N$.

SGR-GMM: RobustDGMM specialization

Heteroscedastic
low-rank GM



Additive noise +
strong ϵ -contamination

$$\xi_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_\xi),$$
$$\check{y}_n = \begin{cases} y_n + \xi_n, & (\text{if } n \in \mathcal{I}_{\text{in}}) \\ a_n, & (\text{if } n \in \mathcal{I}_{\text{out}}) \end{cases},$$

with a partition $\{1, \dots, N\} = \mathcal{I}_{\text{in}} \sqcup \mathcal{I}_{\text{out}}, |\mathcal{I}_{\text{out}}| \leq \epsilon N$.

DGMM weights: $\widehat{w}_1^{[t]}, \dots, \widehat{w}_L^{[t]} \in \mathbb{R}$



SGR-GMM weights: $\widehat{w}^{(k)} = \left(\widehat{w}_1^{(k)}, \dots, \widehat{w}_N^{(k)} \right) \in \Delta_{N, \epsilon}$



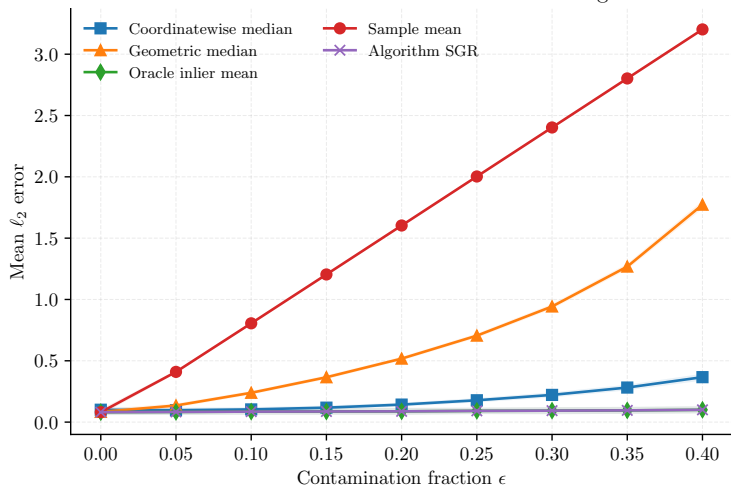
Robust DGMM gradient:

$$\nabla_{\theta} Q_N^{[t]}(\theta) = \sum_{k=1}^L \widehat{w}_k^{[t]} \left(\nabla_{\theta} \check{\phi}^{(k)}(\theta; \xi) - 2 \sum_{n=1}^N \widehat{w}_n^{(k)} \nabla_{\theta} \check{\psi}^{(k)}(\theta; \xi, \check{y}_n) \right)$$

$\check{\phi}$ and $\check{\psi}$ incorporate Σ_ξ through the Gaussian cumulants.

SGR-GMM: numerical results

Estimation error for different estimators under increasing contamination



Isolated SGR experiment

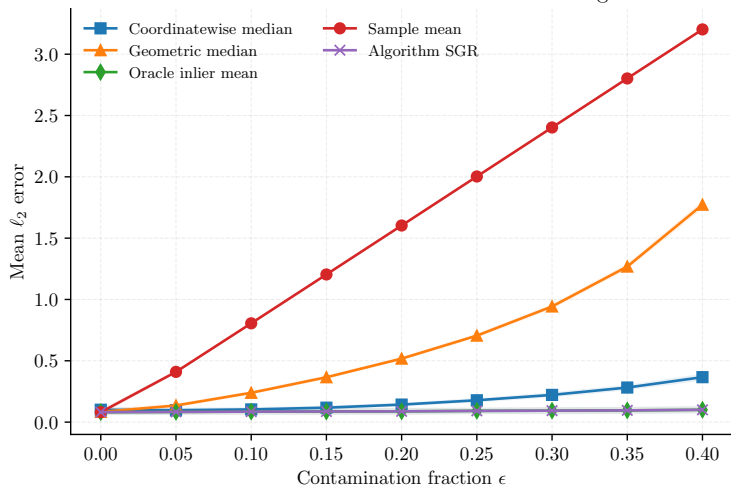
- ▶ $p = 10$,
- ▶ $N = 600$,
- ▶ directional Gaussian replacement outliers.

Near-oracle weighting

Even with increasing ϵ , the SGR error differs from the oracle inlier-mean error by at most 7.24×10^{-5} .

SGR-GMM: numerical results

Estimation error for different estimators under increasing contamination



Isolated SGR experiment

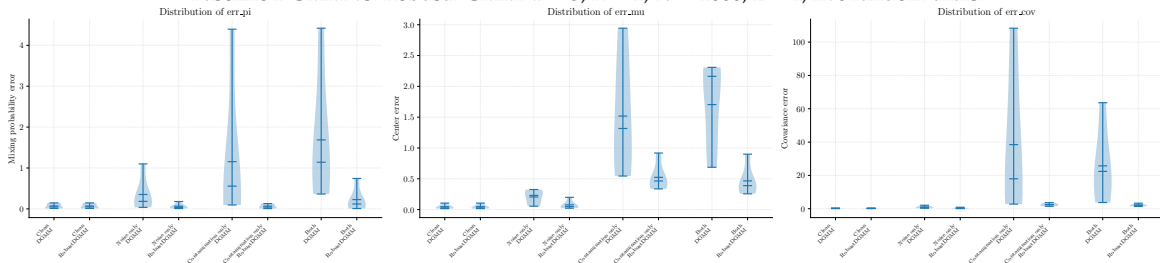
- ▶ $p = 10$,
- ▶ $N = 600$,
- ▶ directional Gaussian replacement outliers.

Near-oracle weighting

Even with increasing ϵ , the SGR error differs from the oracle inlier-mean error by at most 7.24×10^{-5} .

SGR-GMM: numerical results

Baseline DGMM vs. RobustDGMM: $d = 5, K = 2, N = 1000, L = 4$, five random trials.



Clean GM
Numerically
identical.

GM + additive noise
RobustDGMM shows
slight improvements
over baseline DGMM.

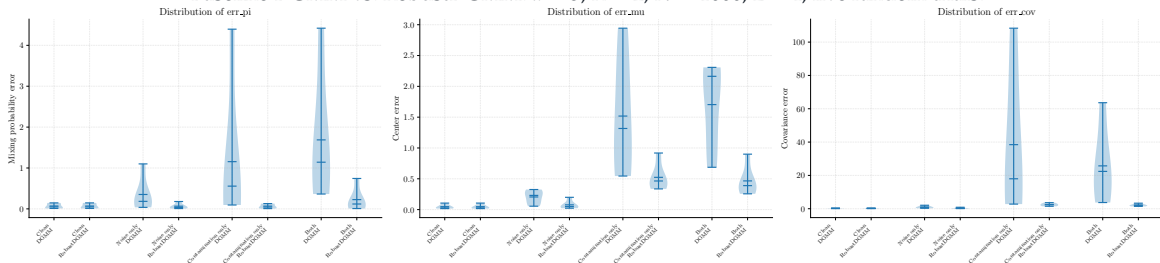
GM + contamination
baseline DGMM error
vs. RobustDGMM error:
 $17.1 \times (\text{Err}_{\pi})$, $2.9 \times (\text{Err}_{\mu})$, $15.6 \times (\text{Err}_{\Sigma})$.

GM + additive noise
+ contamination
baseline DGMM error
vs. RobustDGMM error:
 $6.2 \times (\text{Err}_{\pi})$, $3.5 \times (\text{Err}_{\mu})$, $12.6 \times (\text{Err}_{\Sigma})$.

Zhang-Singer (2026), Section 5.2.2, Table 1 and Fig. 5.

SGR-GMM: numerical results

Baseline DGMM vs. RobustDGMM: $d = 5, K = 2, N = 1000, L = 4$, five random trials.



Clean GM
Numerically
identical.

GM + additive noise
RobustDGMM shows
slight improvements
over baseline DGMM.

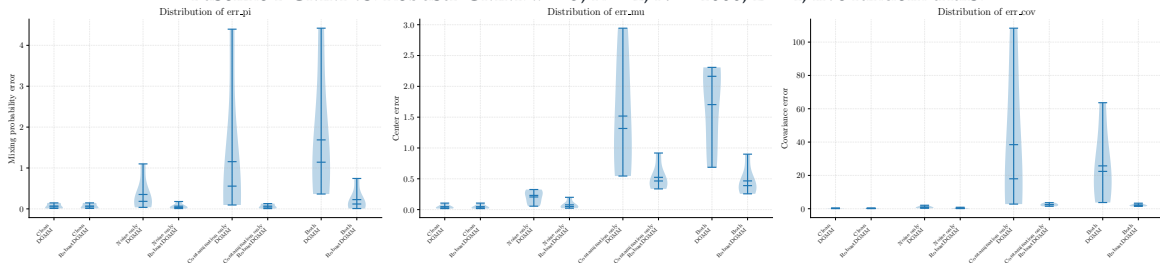
GM + contamination
baseline DGMM error
vs. RobustDGMM error:
 $17.1 \times (\text{Err}_\pi)$, $2.9 \times (\text{Err}_\mu)$, $15.6 \times (\text{Err}_\Sigma)$.

**GM + additive noise
+ contamination**
baseline DGMM error
vs. RobustDGMM error:
 $6.2 \times (\text{Err}_\pi)$, $3.5 \times (\text{Err}_\mu)$, $12.6 \times (\text{Err}_\Sigma)$.

Zhang-Singer (2026), Section 5.2.2, Table 1 and Fig. 5.

SGR-GMM: numerical results

Baseline DGMM vs. RobustDGMM: $d = 5, K = 2, N = 1000, L = 4$, five random trials.



Clean GM
Numerically
identical.

GM + additive noise
RobustDGMM shows
slight improvements
over baseline DGMM.

GM + contamination
baseline DGMM error
vs. RobustDGMM error:
 $17.1 \times (\text{Err}_\pi)$, $2.9 \times (\text{Err}_\mu)$, $15.6 \times (\text{Err}_\Sigma)$.

**GM + additive noise
+ contamination**
baseline DGMM error
vs. RobustDGMM error:
 $6.2 \times (\text{Err}_\pi)$, $3.5 \times (\text{Err}_\mu)$, $12.6 \times (\text{Err}_\Sigma)$.

Zhang-Singer (2026), Section 5.2.2, Table 1 and Fig. 5.

Takeaways

Diagonally-weighted GMM (DGMM) [ZMXS25]

- ▶ Strategic weighting for moment conditions.
- ▶ Fast moment computation.
- ▶ Theoretical guarantees.

[ZMXS25]

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).

[ZS26]

L. Zhang and A. Singer. “Robust Moment-Based Estimation via Spectral Gradient Reweighting”. In: *arXiv preprint arXiv:2605.27718* (2026).

Spectral-Gradient-Reweighted GMM (SGR-GMM) [ZS26]

- ▶ Strategic weighting for per-observation moment-matching gradients.
- ▶ Robust DGMM specialization.
- ▶ Theoretical guarantees.



DGMM

DGMM: scalable moment condition weighting
SGR-GMM: robust per-observation gradient weighting



SGR-GMM

Future directions

- ▶ apply DGMM to other parameter estimation problems, especially in low-SNR regimes,
- ▶ apply SGR-GMM to other GMM procedures, e.g., Instrumental Variables (IV) estimation, Z-estimation, ...
- ▶ additional theoretical questions: sample complexity, moment identifiability, ...

[ZMXS25]

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).

[ZS26]

L. Zhang and A. Singer. “Robust Moment-Based Estimation via Spectral Gradient Reweighting”. In: *arXiv preprint arXiv:2605.27718* (2026).



DGMM

DGMM: scalable moment condition weighting
SGR-GMM: robust per-observation gradient weighting



SGR-GMM

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.
- ▶ The SGR-GMM work is supported in part by AFOSR FA9550-23-1-0249, DARPA HR0011-25-3-E002, NSF DMS 2510039, and the Simons Foundation Math+X Investigator Award.
- ▶ Registration and travel support for this presentation was provided in part by the Society for Industrial and Applied Mathematics Travel Award and Workshop Celebrating Diversity (WCD) Reimbursement.

[ZMXS25]

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).

[ZS26]

L. Zhang and A. Singer. “Robust Moment-Based Estimation via Spectral Gradient Reweighting”. In: *arXiv preprint arXiv:2605.27718* (2026).



DGMM

DGMM: scalable moment condition weighting
SGR-GMM: robust per-observation gradient weighting



SGR-GMM

Thank you.

Questions and comments are welcome.

Contact: lz1619@princeton.edu.

[ZMXS25]

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).

[ZS26]

L. Zhang and A. Singer. “Robust Moment-Based Estimation via Spectral Gradient Reweighting”. In: *arXiv preprint arXiv:2605.27718* (2026).



DGMM

DGMM: scalable moment condition weighting
SGR-GMM: robust per-observation gradient weighting



SGR-GMM

References I

- [AHK12] S. Arora, E. Hazan, and S. Kale. “The multiplicative weights update method: a meta-algorithm and applications”. In: *Theory Comput.* 8 (2012), pp. 121–164. ISSN: 1557-2862. DOI: [10.4086/toc.2012.v008a006](https://doi.org/10.4086/toc.2012.v008a006).
- [DK23] I. Diakonikolas and D. M. Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.
- [DKKLMS19] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. “Robust estimators in high-dimensions without the computational intractability”. In: *SIAM Journal on Computing* 48.2 (2019), pp. 742–864.
- [DKKLSS19] I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart. “SEVER: A robust meta-algorithm for stochastic optimization”. In: *International Conference on Machine Learning*. PMLR, 2019, pp. 1596–1606.
- [DM22] A. S. Dalalyan and A. Minasyan. “All-in-one robust estimator of the Gaussian mean”. In: *Ann. Statist.* 50.2 (2022), pp. 1193–1219. ISSN: 0090-5364,2168-8966. DOI: [10.1214/21-aos2145](https://doi.org/10.1214/21-aos2145).
- [DS96] J. E. Dennis Jr. and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Vol. 16. Classics in Applied Mathematics. Corrected reprint of the 1983 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996, pp. xvi+378. ISBN: 0-89871-364-1. DOI: [10.1137/1.9781611971200](https://doi.org/10.1137/1.9781611971200).

References II

- [Hansen, 1982] L. P. Hansen. “Large Sample Properties of Generalized Method of Moments Estimators”. In: *Econometrica* 50.4 (1982), pp. 1029–1054.
- [LLMZ25] G. Lerman, K. Li, T. Maunu, and T. Zhang. “Global Convergence of Iteratively Reweighted Least Squares for Robust Subspace Recovery”. In: *arXiv preprint arXiv:2506.20533* (2025).
- [LM18] G. Lerman and T. Maunu. “Fast, robust and non-convex subspace recovery”. In: *Inf. Inference* 7.2 (2018), pp. 277–336. ISSN: 2049-8764,2049-8772. DOI: [10.1093/imaiai/iax012](https://doi.org/10.1093/imaiai/iax012).
- [LRV16] K. A. Lai, A. B. Rao, and S. Vempala. “Agnostic estimation of mean and covariance”. In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2016, pp. 665–674.
- [OG17] 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.
- [Pearson, 1894] K. Pearson. “Contributions to the mathematical theory of evolution”. In: *Philosophical Transactions of the Royal Society of London. A* 185 (1894), pp. 71–110.
- [PKK22] 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).
- [ZMXS25] 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).
- [ZS26] L. Zhang and A. Singer. “Robust Moment-Based Estimation via Spectral Gradient Reweighting”. In: *arXiv preprint arXiv:2605.27718* (2026).