Riemannian Stochastic Optimization for Sufficient Dimension Reduction

Learning the predictive subspace of the data accurately and fast.

Thibault Pautrel[1] François Portier[2]

[1]L2S, CentraleSupélec, Université Paris-Saclay [2]ENSAI, CREST

Contributions

1. **Theory.** MAVE minimizers approximate the same Grassmannian target as OPG, via projected-space rather than ambient-space regression — unifying two major SDR families and motivating a Riemannian treatment.
2. **Reformulation.** MAVE recasts as a smooth $\mathcal{O}(d)$-invariant maximization on $\mathrm{St}(p,d)$ with closed-form Riemannian gradient.
3. **Algorithm & guarantees.** SMAVE uses mini-batch Riemannian SGD with sparse projected-space k-NN weights, reducing per-iteration cost. A simplified version enjoys almost-sure convergence and optimal $O(1/\sqrt{T})$ rate.
4. **Experiments.** SMAVE matches or beats RMAVE on all benchmarks at lower runtime, with the gap widening as p grows.

Sufficient Dimension Reduction (SDR)

Setup

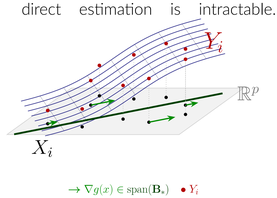
Given n pairs $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^p$ such that $Y = g(X) + \varepsilon$, $\mathbb{E}[\varepsilon \mid X] = 0$, $\mathrm{var}(\varepsilon \mid X) = \sigma^2(X)$.
We want to estimate $g(X) = \mathbb{E}[Y \mid X]$.

⚠ For large p, nonparametric rates decay as $n^{-4/(p+4)}$: direct estimation is intractable.

SDR seeks $\mathbf{B}_* \in \mathbb{R}^{p\times d}$, $d \ll p$, such that $g(X) = \mathbb{E}[Y \mid \mathbf{B}_*^\top X]$: estimation collapses from $\mathbb{R}^p$ to $\mathbb{R}^d$.
Goal: Minimize the population risk

$$D(\mathbf{B}) := \mathbb{E}\big[(Y - \mathbb{E}[Y \mid \mathbf{B}^\top X])^2\big] \tag{1}$$

over the *Stiefel manifold* $\mathrm{St}(p,d) = \{\mathbf{B} \in \mathbb{R}^{p\times d} : \mathbf{B}^\top\mathbf{B} = I_d\}$.
Since $D(\mathbf{B}) = D(\mathbf{BQ})$ for all $\mathbf{Q} \in \mathcal{O}(d)$, minimizers descend to the identifiable target $\mathrm{span}(\mathbf{B}_*) \in \mathrm{Gr}(p,d) \cong \mathrm{St}(p,d)/\mathcal{O}(d)$.



Outer Product of Gradients (OPG)

OPG [5] solves, for each anchor j, $(\hat{a}_j, \hat{b}_j) = \arg\min_{a,b\in\mathbb{R}^p} \sum_{i=1}^n \big(Y_i - a - b^\top(X_i - X_j)\big)^2 w_{ij}$, $w_{ij} \propto K_h(\|X_i - X_j\|)$, $h \propto n^{-1/(p+4)}$, then computes $\widehat{B}_{\mathrm{OPG}} = \text{top-}d\,\mathrm{eig}\big(\frac{1}{n}\sum_j \hat{b}_j\hat{b}_j^\top\big)$. Since $\hat{b}_j \to \nabla g(X_j)$, the aggregator targets $\mathcal{M} = \mathbb{E}[\nabla g(X)\nabla g(X)^\top]$. Under SDR, $\mathrm{range}(\mathcal{M}) = \mathrm{span}(\mathbf{B}_*)$, so $\widehat{B}_{\mathrm{OPG}} \to \mathrm{span}(\mathbf{B}_*)$.

Minimum Average Variance Estimation (MAVE)

(1) is intractable since $\mathbb{E}[Y \mid \mathbf{B}^\top X]$ is unknown. [5] replaces it by a local linear first-order Taylor expansion in a ball of radius u around each x:

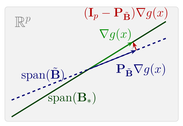
$$D_{(x,u)}(\mathbf{B}) := \min_{a,\mathbf{b}} \mathbb{E}_{(x,u)}\Big[\Big(Y - a - \mathbf{b}^\top\mathbf{B}^\top(X - x)\Big)^2\Big], \qquad a \approx g(x), \quad \mathbf{b} \approx \nabla g(x).$$

Global gradient alignment

Under smoothness and local covariance lower bound assumptions, any minimizer $\bar{\mathbf{B}}$ of $D_u(\mathbf{B}) = \int D_{(x,u)}(\mathbf{B})\,\mathbb{P}(\mathrm{d}x)$ satisfies

$$\mathrm{tr}\big((\mathbf{I}_p - \mathbf{P}_{\bar{\mathbf{B}}})\mathcal{M}\big) \lesssim u^2,$$

where $\mathcal{M} := \mathbb{E}[\nabla g(X)\nabla g(X)^\top]$ with $\mathrm{range}(\mathcal{M}) = \mathrm{span}(\mathbf{B}_*)$,
$\mathbf{P}_{\bar{\mathbf{B}}} = \bar{\mathbf{B}}\bar{\mathbf{B}}^\top$ is the orthogonal projector on $\mathrm{span}(\bar{\mathbf{B}})$.



Empirical MAVE

Replace $\int D_{(x,u)}\,\mathbb{P}(\mathrm{d}x)$ by n anchors with weights w_{ij}:

$$\min_{\mathbf{B}\in\mathrm{St}(p,d)} \sum_j \min_{a_j,\mathbf{b}_j} \sum_i \big(Y_i - a_j - \mathbf{b}_j{}^\top\mathbf{B}^\top(X_i - X_j)\big)^2 w_{ij}. \qquad \hat{a}_j \approx g(\mathbf{B}^\top X_j), \quad \hat{b}_j \approx \nabla g(X_j) \tag{2}$$

Refined MAVE (RMAVE)

RMAVE [5] is a deterministic iterative refinement of MAVE: T iterations, localized in the projected space $\mathbb{R}^d$.
Initialization $\hat{\mathbf{B}}_0 = \hat{\mathbf{B}}_{\mathrm{OPG}}$
Localization $w_{ij} \propto K_h\big(\|\hat{\mathbf{B}}_t^\top(X_i - X_j)\|\big)$, $h \propto n^{-1/(d+4)}$
Update coordinate descent over the d columns of $\mathbf{B}$, weights recomputed at every outer iteration.

The SMAVE Algorithm

SMAVE (**S**tochastic **MAVE**) recasts the empirical MAVE criterion (2) as a smooth *maximization* on $\mathrm{St}(p,d)$:

$$\max_{\mathbf{B}\in\mathrm{St}(p,d)} \left\{ F(\mathbf{B}) = \frac{1}{n}\sum_{j=1}^n (\mu^{(j)})^\top\mathbf{B}\big(\mathbf{B}^\top G^{(j)}\mathbf{B}\big)^{-1}\mathbf{B}^\top\mu^{(j)} \right\}.$$

with *local Gram matrices* and *moment vectors* defined as

$$G^{(j)} = \sum_i w_{ij}\,\tilde{X}_i^{(j)}(\tilde{X}_i^{(j)})^\top \in \mathbb{R}^{p\times p}, \qquad \mu^{(j)} = \sum_i w_{ij}\,\tilde{X}_i^{(j)}\tilde{Y}_i^{(j)} \in \mathbb{R}^p.$$

For any $\mathbf{Q} \in \mathcal{O}(d)$, $F(\mathbf{BQ}) = F(\mathbf{B})$, so F is well-defined on $\mathrm{Gr}(p,d)$.
Its Riemannian gradient is known in **closed form** and equal to the Euclidean gradient.

Design choices

- Mini-batches $J_t \subset [n]$, $|J_t| = m \ll n$, give the stochastic Riemannian gradient

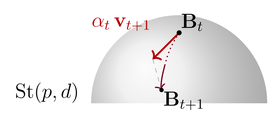
$$\mathbf{g}_t = \frac{2}{m}\sum_{j\in J_t}\big(\mu^{(j)} - G^{(j)}\mathbf{B}\,u^{(j)}\big)(u^{(j)})^\top. \tag{3}$$

  where $u^{(j)} := (\mathbf{B}^\top G^{(j)}\mathbf{B})^{-1}\mathbf{B}^\top\mu^{(j)}$.
- Sparse k-NN weights in the *projected* space $\mathbb{R}^d$: letting $\mathcal{N}_k(\mathbf{B}, X_j)$ be the k nearest neighbours of X_j under $\|\mathbf{B}^\top(\cdot - X_j)\|$, with $k \ll n$

$$w_{ij} = \tfrac{1}{k}\mathbb{1}_{X_i\in\mathcal{N}_k(\mathbf{B},X_j)}, \quad k = \lceil n^{4/(d+4)}\rceil.$$

- Geometry-aware update with normalized gradient, momentum and QR retraction keeping the iterate on $\mathrm{St}(p,d)$ [1] with $\alpha_t = \frac{\alpha_0}{1+\gamma t}$:

$$\boxed{\mathbf{B}_{t+1} = R^{\mathrm{QR}}_{\mathbf{B}_t}(\alpha_t\,\mathbf{v}_{t+1})}, \quad \mathbf{v}_{t+1} = \beta\mathbf{v}_t + \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|_F} \tag{4}$$



SMAVE algorithm

Input iterations T, refresh τ, step-size α_0, decay γ, momentum β
Init $\mathbf{B}_0 \in \mathrm{St}(p,d)$ at random; build k-NN index on $\{\mathbf{B}_0^\top X_i\}_{i=1}^n$
For $t = 0, \dots, T-1$ **do**
  Draw $J_t \subset \{1,\dots,n\}$, $|J_t| = m$, w.o.r.
  Compute $\mathbf{g}_t$ via (3); update $\mathbf{B}_{t+1}$ via (4)
  Every τ steps: rebuild k-NN index on $\{\mathbf{B}_{t+1}^\top X_i\}$
Return $\mathbf{B}_T$

Computational complexity

	Loc.	Weights	Gram	Iters
OPG	$\mathbb{R}^p$	$O(n^2p)$	$O(n^2p^2)+O(np^3)$	1
RMAVE	$\mathbb{R}^d$	$O(n^2d)$	$O(n^2p^2)\times T$	T
SMAVE	$\mathbb{R}^d$	$O(n\log n)$	$O(mkp^2)\times T$	T

OPG and RMAVE are non-stochastic.

Theoretical Guarantees

We analyze a simplified regime (fixed neighborhoods, no momentum, no gradient normalization) with the ε-regularized objective (i.e. $G_\varepsilon^{(j)} := G^{(j)} + \varepsilon I_p$).

- **Almost-sure convergence** (from [3]) for step sizes $\alpha_t = \alpha_0/(1+\gamma t)$, almost surely: $F_\varepsilon(\mathbf{B}_t)$ converges to a finite limit; $\mathrm{grad}\,F_\varepsilon(\mathbf{B}_t) \to 0$. Every accumulation point of $\{\mathbf{B}_t\}$ is a critical point of F_ε.
- **Non-asymptotic rate** for any constant step $\alpha > 0$ small enough, and $T \geq 1$, there exists constants $L_\varepsilon, \sigma_\varepsilon^2, \Delta_\varepsilon$ such that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\mathrm{grad}\,F_\varepsilon(\mathbf{B}_t)\|_F^2\big] \;\leq\; \frac{2\Delta_\varepsilon}{\alpha T} + L_\varepsilon\,\alpha\,\sigma_\varepsilon^2,$$

  Tuning $\alpha \propto 1/\sqrt{T}$ gives $O(1/\sqrt{T})$, matching the **optimal rate** for non-convex stochastic first-order methods [4, 2].

The analysis extends to any localization scheme (kernel, fixed-radius, similarity graphs) producing an objective of the same form.

Convergence of full SMAVE (periodic k-NN refresh + heavy-ball momentum) remains open.

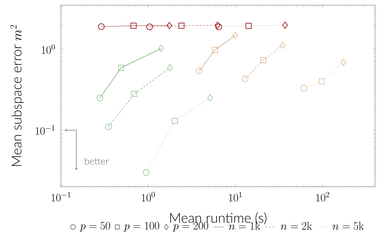
Synthetic experiments: accuracy meets speed

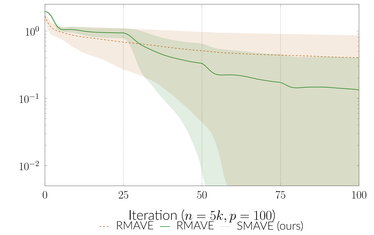
SDR subspace of dimension $d = 2$,

$$Y = g(\mathbf{B}_*^\top X) + \xi, \qquad X \sim \mathcal{N}(0,\mathbf{\Sigma}), \xi \sim \mathcal{N}(0, 0.5^2).$$

$\mathbf{B}_* \in \mathrm{St}(p,d)$ sparse (20% nonzero).
Sparsity reflects high-dimensional settings where only a subset of covariates drives the response.
Metric: squared subspace distance $m^2(\widehat{\mathbf{B}}, \mathbf{B}_*) := \big\|(I_p - \mathbf{B}_*\mathbf{B}_*^\top)\widehat{\mathbf{B}}\big\|_F^2$.

Link g	Polynomial, Sinusoidal, Exponential, Interaction,Rational
Covariance $\mathbf{\Sigma}$	$\mathbf{I}_p$; AR(1): $\Sigma_{ij} = 0.5^{\lvert i-j\rvert}$
Sample size n	1000, 2000, 5000
Dimension p	50, 100, 200
Replications	10 per configuration





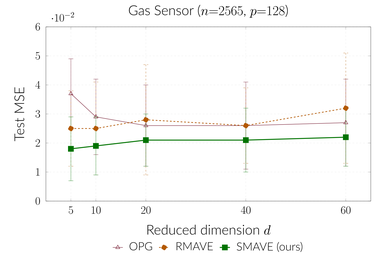
34× faster than RMAVE at n=5000, p=200 (5 s vs. 174 s)

3× lower error than RMAVE at n=5000, p=200 (0.25 vs. 0.69)

9× wins all 9 moderate-to-high-p settings simultaneously (accuracy *and* speed)

Real-world data: four datasets

Since $\mathbf{B}_*$ is unknown, we project $X \to Z = X\widehat{\mathbf{B}}$, fit a k-NN regressor on Z, and evaluate test MSE. The dimension d^* is selected per method by CV. MAVE methods select parsimonious subspaces.



Dataset	Method	MSE	d^*	Time (s)
Wine	OPG	0.52	8	1.4
n=6497	RMAVE	**0.50**	8	600.2
p=11	SMAVE	0.51	8	0.2
Seoul	OPG	0.27	8	2.8
n=8760	RMAVE	**0.17**	4	327.3
p=14	SMAVE	0.17	4	0.3
Pumadyn	OPG	0.17	2	4.4
n=8192	RMAVE	**0.09**	2	55.4
p=32	SMAVE	0.13	4	0.5
Gas	OPG	0.03	20	1.9
n=2565	RMAVE	0.03	5	35.1
p=128	SMAVE	**0.02**	5	0.2

bold: best; underline: second best

Take-home

SMAVE recovers the predictive directions of your data as accurately as the best classical method, with convergence guarantees, and up to orders of magnitude faster — and it scales when ambient-space methods break.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.
[2] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.
[3] Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.
[4] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
[5] Yingcun Xia, Howell Tong, W. K. Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B*, 64(3):363–410, 2002.

thibault.pautrel@centralesupelec.fr