



Paper Type: Original Article

Heterogeneous EEG Feature Fusion: Brain Storm Optimization with Covariance-Aware Mutation

Astrid L. Nielsen*

Department of Applied Mathematics, Technical University of Denmark, Lyngby, Denmark; alni@dtu.dk.

Citation:

Received: 28 March 2026

Revised: 01 May 2026

Accepted: 27 July 2026

Nielsen, A. L. (2026). Heterogeneous EEG feature fusion: Brain storm optimization with covariance-aware mutation. *Metaheuristic Algorithms with Applications*, 3(2), 141-167.

Abstract

Decoding accuracy in Electroencephalographic (EEG) brain–computer interfaces is governed by how heterogeneous feature descriptors are fused: time-domain, frequency-domain, and time–frequency blocks are acquired with markedly different dimensionalities, scales, and correlation structures. Selecting and weighting these blocks, while the per-block regularization that controls the lower-level classifier fit is tuned in response, constitutes a bilevel optimization problem, with a leader fixing a fusion configuration and a follower returning regularized estimates. This paper develops HEF-Brain Storm Optimization (BSO), a bilevel optimizer whose swarm dynamics follow the BSO paradigm of idea clustering and idea generation, enhanced by a covariance-aware mutation operator defined on Symmetric Positive-Definite (SPD) matrices estimated from the heterogeneous feature blocks. Cluster-level covariance estimates are stabilized by shrinkage, safeguarded by eigenvalue clipping in $[10^{-4}, 10^3]$, and coordinated with the lower level through warm-started follower solves governed by a 60/40 evaluation-budget split. A supermartingale-based convergence argument and an amortized complexity of $O(Nd + N \log k + d^3/5)$ per iteration are supplied. The experimental program evaluates ten SMD bilevel instances over 30 independent runs against nine rivals, BLEAQ among them, and contrasts the learned fusion with the EEGNet, ShallowConvNet, and DeepConvNet decoders under leave-one-subject-out transfer on an eleven-subject motor-imagery cohort. HEF-BSO attains the smallest median optimality gap on seven of ten instances, with a suite-wide median of 4.4×10^{-6} , and Pratt-adjusted Wilcoxon tests with Holm correction consolidate the margin. A hierarchical Bayesian model with subject-level partial pooling yields a posterior mean leave-one-subject-out balanced accuracy of 75.5%, against 74.1% for EEGNet, with posterior probabilities of superiority of 0.90 against EEGNet and above 0.995 against the remaining decoders. Perturbation and coupling-stress protocols hold degradations between two-fifths and one-half of the strongest rival's and bilevel infeasibility at 7.8% under a starved thirty-percent follower budget.

Keywords: Brain storm optimization, Bilevel optimization, Electroencephalographic feature fusion, Covariance-aware mutation, Positive-definite matrices, Robustness analysis.

1 | Introduction

A Brain–Computer Interface (BCI) must translate a weak, artifact-contaminated Electroencephalographic (EEG) signal into a reliable control command, and two decades of research have located the bottleneck not in the sensing hardware but in the quality of the extracted representations [1], [2]. A conventional motor-imagery measurement chain band-passes the acquired signal, derives spatial filters, concatenates the resulting

Corresponding Author: alni@dtu.dk

<https://doi.org/10.48313/maa.vi.82>

Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

descriptors, and hands the fused vector to a classifier [3]. Because the discriminative structure of the EEG is strongly subject-specific and drifts across sessions, the choice of which descriptors to extract, how to rescale them, and how to weight them against one another becomes a genuine calibration problem rather than a fixed preprocessing convention. The feature-fusion task accordingly supplies the specification for the algorithm itself: the measurement side fixes the structural requirements, the optimizer satisfies them.

Modern EEG analysis rests on three heterogeneous descriptor families whose mismatched geometries are an everyday reality in measurement. Time-domain statistics such as Hjorth activity, mobility, and complexity capture waveform shape at low computational cost [4]. Frequency-domain descriptors track event-related desynchronization and synchronization in physiologically motivated bands [5], while spatial-filtering methods such as Common Spatial Patterns (CSP) and their filter-bank extensions exploit covariance contrasts between imagery conditions [6], [7]. When trials are scarce relative to dimension, regularized and Riemannian variants refine these estimators [8], [9]. Dimensionality, statistical distribution, and dynamic range separate the three families: Raw variances, logarithmic band powers, and CSP log-variances share no common scale, and their pairwise correlations are strongly block-structured. Under such mismatch, naive concatenation is dominated by whichever block carries the largest variance, so principled fusion through per-block weighting, whitening, and selection becomes a measurement requirement. Compact convolutional decoders have since become the principal alternative to handcrafted fusion, and their subject-transfer behavior fixes the bar that any optimized fusion pipeline must clear [10], [11].

The fusion task is, moreover, hierarchically coupled, and the coupling is what makes it hard. Once a candidate fusion configuration fixes the weighting and selection of feature blocks, an inner procedure re-optimizes the per-block regularization that governs the classifier fit, and the attained accuracy feeds back as the objective value of the outer design. Leader–follower structures of this kind fall under bilevel programming, a discipline rooted in Stackelberg's model of hierarchical markets [12] and developed extensively in operations research [13–15]. Bilevel problems are generically nonconvex and, in their general form, NP-hard [16], a difficulty that motivated standardized test suites such as the SMD family [17] together with dedicated theory for multiobjective [18] and evolutionary [19], [20] treatments. In deployed settings, the follower's response mapping is nonsmooth or only implicitly measurable, so derivative-based leader updates lack reliability, and population-based search becomes the pragmatic option.

Population-based metaheuristics supply the search machinery, with a lineage running from genetic algorithms and evolution strategies [21], [22] through particle swarm optimization [23], ant colony optimization [24], and differential evolution [25]. Among them, covariance matrix adaptation evolution strategies demonstrated that adaptation of a full covariance matrix to the local geometry of the search distribution accelerates convergence dramatically on ill-conditioned landscapes [26], a lesson reinforced by robustness studies in stochastic optimization [27]. The no-free-lunch theorems nevertheless imply that gains of this kind are structure-dependent rather than universal [28], and contemporary practice favors matching algorithmic structure to problem structure [29]. Brain Storm Optimization (BSO) contributes a distinctive ingredient: Candidate solutions become ideas that are clustered before generation, so clustering rather than pure selection shapes the production of new candidates [30–32]. The correspondence with feature fusion is direct: Clusters realize design lineages of compatible fusion configurations, inter-cluster ideas, and novel descriptor mixes.

Three gaps separate this prior art from the requirements of heterogeneous feature fusion: 1) standard BSO mutates ideas isotropically, which wastes evaluations whenever the search space aggregates blocks of disparate scale and strong internal correlation, 2) bilevel metaheuristics typically re-solve the follower from scratch for every leader candidate, spending evaluations on lower-level optima that shift smoothly with the leader, and 3) robustness of bilevel optimizers under noisy evaluations, parameter misspecification, and adversarial distortion of the lower-level response is seldom quantified systematically, although uncertainty is the normal operating condition of physiological measurement. The present paper addresses all three gaps simultaneously; the questions that govern its evaluation are fixed explicitly below, before the contributions are enumerated.

The evaluation is accordingly organized as a measurement program, not a demonstration. Three research questions, each paired with a testable hypothesis, govern the assessment: fusion quality against covariance-blind and approximation-based rivals at a matched evaluation budget, stability of the bilevel measurement chain under interlevel coupling stress, and transfer of the learned fusion configuration to unseen subjects under leave-one-subject-out cross-validation. The statements below frame experiments already reported in this paper; no additional experiments are promised. Evidence bearing on each hypothesis comes from the ten-instance SMD benchmark study with its Pratt-adjusted signed-rank battery, the coupling stress test, and the motor-imagery case study, with a hierarchical Bayesian model pooling the eleven subjects supplying the primary decision rule.

RQ1: Does covariance-aware mutation over heterogeneous EEG feature blocks improve bilevel fusion quality over covariance-blind and approximation-based rivals at an equal evaluation budget?

RQ2: Is the method robust to interlevel coupling stress, specifically to follower-budget starvation and constraint-coupling variation?

RQ3: Does the fused configuration generalize to unseen subjects under leave-one-subject-out transfer in the motor-imagery case study?

H1. On the ten SMD instances, HEF-BSO attains the smallest median optimality gap on seven of ten, with reversals confined to instances that favor a single adapted covariance or a smooth quadratic response model, and Pratt-adjusted Wilcoxon tests with Holm correction certify the margin against the strongest rival.

H2. Under coupling stress at a starved thirty-percent follower budget share, HEF-BSO holds bilevel infeasibility at 7.8% where the strongest rival reaches 24.6%, and the median leader-side gap stays within a factor of two across constraint-coupling scales from 0.5 to 2.

H3. Under leave-one-subject-out transfer, the fused pipeline attains the highest posterior mean balanced accuracy, 75.5 against 74.1% for the strongest decoder, with the posterior probability of superiority at 0.90 against EEGNet and above 0.995 against the remaining decoders.

A bilevel formulation of heterogeneous EEG feature fusion in which the leader fixes per-band fusion weights and a soft selector mask over sub-band groups, the follower returns per-block regularization parameters, and Cohen's Kappa is adopted as the upper-level objective.

HEF-BSO, a BSO algorithm equipped with a covariance-aware idea-generation step: Mutations are sampled from cluster-level Gaussian distributions governed by Symmetric Positive-Definite (SPD) covariance matrices, stabilized by shrinkage and safeguarded by eigenvalue clipping in $[10^{-4}, 10^3]$.

A bilevel coordination scheme in which every follower solve is warm-started from the archived lower-level response of the nearest leader, under a 60/40 evaluation-budget split between the levels.

A supermartingale-based global convergence argument under explicit regularity assumptions, together with an amortized per-iteration complexity of $O(Nd + N \log k + d^3/5)$ that accounts for the measured wall-clock overhead of 21 to 29% over standard BSO.

An empirical program combining ten SMD bilevel instances at freshly drawn coupling settings, an eleven-subject motor-imagery case study scored under leave-one-subject-out cross-validation with balanced accuracy, Cohen's Kappa, and area under the curve, hierarchical Bayesian subject pooling with ninety-five percent highest-density intervals, Pratt-adjusted Wilcoxon signed-rank testing with Holm correction, isolation ablation of the three principal components, follower-budget and constraint-coupling stress tests, and a three-family perturbation-robustness protocol.

The remainder of the paper traverses the same route. Section 2 reviews the BSO paradigm, the structure of heterogeneous EEG feature blocks, and the role of covariance awareness on the cone of SPD matrices. Section 3 states the bilevel problem, derives its Karush–Kuhn–Tucker (KKT) reformulation, and specializes it to feature fusion. Section 4 develops the HEF-BSO algorithm in full, and Section 5 supplies convergence

and complexity analyses. Section 6 covers the benchmark study on the SMD suite and the EEG case study, including ablations. Section 7 presents the perturbation-robustness study, and Sections 8 through 10 offer discussion, limitations, and conclusions with an outline of future work.

2 | From Cortical Inspiration to Optimization

This section traces the path from the motivating measurement problem to the algorithmic machinery in three steps: The BSO paradigm and the specific role of idea clustering within it, then the heterogeneity of EEG feature blocks, and finally the case for covariance awareness on the cone of SPD matrices. The three threads converge in the algorithm in Section 4, where clusters supply the search topology, feature blocks supply its geometry, and covariance-aware mutation supplies its dynamics.

2.1 | Brain Storm Optimization: Ideas, Clusters, and Generation

BSO was proposed by Shi [30] as a swarm-intelligence algorithm that models the divergent and convergent phases of human brainstorming rather than the foraging or social behaviors of animals. Every candidate solution is an idea, and every iteration partitions the population of ideas into clusters, generates new ideas from one or two clusters, and evaluates the resulting pool. The decisive structural difference with respect to classical evolutionary algorithms is that clustering carries much of the selection load: inferior individuals are not discarded outright but are retained as cluster members that keep seeding the generation, which preserves the diversity on which exploratory search depends [31], [32]. The original formulation performs clustering as a k-means-like partition around the best idea, and a selective strategy draws source ideas from better-performing clusters with higher probability.

In genetic-algorithm terms, the paradigm implements a soft, geography-preserving selection regime rather than a purely fitness-proportional one [33]. Its appeal for the present application is specific: Clusters of fusion configurations behave as coherent design lineages, and the inter-cluster generation step implements precisely the cross-family recombination that heterogeneous feature fusion requires, whereas a purely selection-driven search would collapse prematurely onto a single descriptor family. Comparative studies indicate that this organization is competitive with established evolutionary algorithms on multimodal and composition landscapes [31], which supports its adoption as the swarm backbone of a bilevel optimizer. The paradigm is therefore adopted unmodified at the level of its loop structure, with all modifications concentrated in the representation, the mutation operator, and the coordination between levels.

2.2 | Heterogeneous Feature Blocks in Electroencephalographic Analysis

Feature extraction for EEG is organized around three complementary descriptor families, each instantiated by the case study of Section 6.2. Time-domain features summarize the morphology of the band-passed signal: Hjorth parameters (activity, mobility, complexity), channel variance, zero-crossing rate, and slope-based descriptors are inexpensive, tolerant of moderate artifacts, and historically foundational [4]. Frequency-domain features measure signal power in physiologically motivated bands; event-related desynchronization and synchronization over the sensorimotor cortex manifest as band-power modulations that are approximately log-normal, so logarithmic band powers serve as the standard representation [5]. Time–frequency features combine both views: CSP estimates spatial filters that maximize the ratio of between-class to within-class covariance [6], and filter-bank extensions repeat the estimation across overlapping sub-bands, so that subject-specific frequency foci emerge automatically [7]. When trials are scarce relative to dimension, regularized CSP and Riemannian covariance-based classifiers then markedly improve the stability of all three routes [8], [9].

Heterogeneous geometries are the outcome: A 40-dimensional time-domain block, a 32-dimensional frequency-domain block, and a 48-dimensional time–frequency block are estimated at different scales, exhibit strong within-block and weaker cross-block correlation, and contribute a total of 120 fused dimensions. Trial-wise concatenation without weighting leaves the downstream classifier exposed to scale mismatch, while

hand-tuned weights ignore the interaction between fusion and regularization, an interaction resolved only when weighting and regularization are optimized in their correct hierarchical relationship. The fused representation is therefore treated as a designed artifact, optimized rather than fixed, which is the central methodological stance of this paper and what elevates the engineering task of Section 3.3 into the abstract bilevel framework of Section 3.1.

2.3 | Why Covariance Awareness Matters

In the BCI setting, the empirical covariance of the fused feature space is almost always ill-conditioned, and the cause is the data regime itself: Trials per subject are of the same order as, or fewer than, the fused dimension, blocks are internally strongly correlated, and physiological artifacts inflate a few directions. An isotropically mutating optimizer therefore spends its evaluation budget on directions in which the objective is nearly flat, the deficiency that covariance matrix adaptation was designed to correct in evolution strategies [22], [26]. Such adaptation is most naturally organized on the cone of SPD matrices, whose geometry supports well-behaved means, distances, and tangent-space computations [34], [35] and which already underlies modern covariance-based BCI classifiers [9].

Two numerical safeguards are indispensable in measurement practice: 1) raw sample covariance matrices must be shrunk toward a well-conditioned target, for which the Ledoit–Wolf [36] estimator provides an analytically grounded rule, and 2) any update that leaves the SPD cone must be projected back, and eigenvalue clipping is the simplest such safeguard, with well-understood nearest-SPD interpretations [37]. HEF-BSO builds both safeguards directly into its mutation operator, so covariance awareness becomes a structural property of the search dynamics rather than an optional refinement. Under physiological noise, an additional benefit accrues: The follower's response mapping is evaluated with noise of physiological origin, and mutation distributions whose scale adapts to the local signal-to-noise geometry are markedly more robust under noisy evaluation [26], an issue to which the perturbation study of Section 7 returns quantitatively.

3 | Problem Statement: Bilevel Hierarchical Optimization

This section fixes the notation and formalizes the problem class targeted by HEF-BSO. The general bilevel program is first stated in its optimistic form, the KKT reformulation that underlies one family of baselines is then recalled, and the program is finally instantiated for heterogeneous EEG feature fusion, from which the leader and follower variables used throughout the experimental sections are obtained.

3.1 | General Formulation

Bilevel optimization formalizes hierarchical decision problems in which a leader selects a decision x to minimize an upper-level objective F , while a follower, having observed x , responds by solving a nested lower-level problem in y [12], [13]. The optimistic formulation, adopted throughout this paper, is:

$$\min_{x \in X} F(x, y) \text{ s.t. } y = y^*(x), \quad (1)$$

where the follower's set of rational responses is given by

$$y^*(x) = \{y \in \mathcal{Y}(x) : f(x, y) = \min_{\tilde{y} \in \mathcal{Y}(x)} f(x, \tilde{y})\}, \quad (2)$$

and $y^*(x)$ denotes a selection from this set; pessimistic variants, which hedge against unfavorable selections, are not considered here. Implicit determination of the leader's feasible set through the follower's rationality is the source of the field's characteristic difficulties: Even linear–linear instances are strongly NP-hard [16], constraint qualifications fail systematically [14], and the follower's value function is generally discontinuous in x [15]. Multiobjective and set-valued extensions exist [18] but lie outside the present scope. The standard test suites, most prominently the SMD family [17], deliberately control these difficulties by pairing conflicting and non-conflicting leader–follower structures with convex, multimodal, or constrained lower levels and have become the de facto evaluation standard for evolutionary bilevel algorithms [19], [20]. Claims of progress on

hierarchical problems must therefore rest on both final accuracy and anytime behavior on such suites, a requirement that shapes the experimental design of Section 6.

3.2 | Karush–Kuhn–Tucker-Based Single-Level Reformulation

Whenever the lower-level problem is convex in y for every fixed x and suitable constraint qualifications hold, the follower's optimality can be replaced by its KKT conditions, which yields a single-level mathematical program with complementarity constraints [13], [38]. For lower-level inequality constraints $g(x, y) \leq 0$ the stationarity condition reads

$$\nabla_y f(x, y) + \lambda^\top \nabla_y g(x, y) = 0, \lambda \geq 0, \quad (3)$$

together with primal feasibility, dual feasibility, and complementary slackness between the multipliers and the constraints; the full system is standard and is omitted for brevity. Direct attack on the reformulated problem by sequential quadratic programming defines the KKT-SQP route, which ranks among the strongest classical baselines when the lower level is well behaved [38]. Equally well documented are its weaknesses. At every feasible point, the complementarity constraints violate linear independence constraint qualifications, so SQP regularization schemes operate permanently on the boundary of admissibility. Moreover, nonconvex lower levels, as in the SMD instances with multimodal or quadratic followers, admit follower KKT points that need not be global optima, and the reformulation then solves a problem other than the one posed.

Population-based methods do not escape these issues for free, but degrade gracefully. When the follower's stationarity residual is treated as one term of a penalized leader objective, asymptotic correctness survives under mild conditions; an explicitly solved follower keeps validity under lower-level nonconvexity, at the price of an enlarged evaluation budget. HEF-BSO takes the second route, explicit follower solves coordinated by warm starts; the KKT-SQP baseline of Section 6.3 takes the first. Section 6.5 therefore probes this trade-off directly instead of presuming either reformulation philosophy superior.

3.3 | Feature-Fusion Instance

The EEG measurement task instantiates the abstract program as follows. The leader variable $x = (w, m)$ concatenates a vector of per-band fusion weights w in $\mathbb{R}^3$, one scalar gain for each of the time-domain, frequency-domain, and time–frequency blocks, with a soft selector mask m in $[0, 1]^{12}$ over twelve sub-band groups that attenuates or removes descriptors judged uninformative, giving $p = 15$ upper-level dimensions. Two regularization parameters per block compose the follower variable y in $\mathbb{R}^6$: A ridge penalty for the linear discriminant trained on the time- and frequency-domain descriptors and a CSP shrinkage intensity for the time–frequency block. For a given x , the lower-level problem minimizes the cross-validated training loss of the regularized classifier over y , and the upper-level objective is $F(x, y) = 1 - \kappa$, κ being Cohen's coefficient of agreement between predicted and actual imagery classes [39], so minimizing F maximizes chance-adjusted classification agreement.

Convexity of the lower level in y for fixed x holds in the practical operating range, because the regularized training loss is smooth and locally convex around the interior optimum; this justifies the warm-started inner solver of Section 4.5 without implying that the composite problem is easy. Piecewise smooth, mildly nonconvex, and noisy at the level of single evaluations, the composition of a resampling-based kappa estimate with a regularized inner fit is precisely the regime in which derivative-based leader updates are unreliable and in which population-based search with variance-scaled mutation is expected to pay off. The instance is thus representative of a broad class of hierarchical model-selection problems in which the leader designs a representation and the follower fits a model within it, and its structure, rather than any EEG-specific detail, is what the algorithmic machinery of the next section exploits.

4 | The HEF-Brain Storm Optimization Algorithm

HEF-BSO retains the loop structure of BSO, clustering ideas and then generating new ones, but swaps out three ingredients: The representation couples each idea to a covariance matrix, the mutation draws from cluster-level Gaussian distributions on the SPD cone, and an archive-coordinated follower solver carries the bilevel coupling. The subsections below develop these ingredients in turn, and Algorithm 1 assembles them.

4.1 | Representation: Solutions as Fusion Configurations

In HEF-BSO, an evolving idea is a pair coupling a point in the space of fusion configurations with the local search geometry that surrounds it:

$$s = (w, \Sigma), w \in \mathbb{R}^p, \Sigma \in \mathbb{S}_{++}^d. \quad (4)$$

The symbol w gathers the leader-side fusion parameters, namely the block weights and the selector mask of Section 3.3; the fused feature space has dimension $d = 120$; and an SPD matrix Σ parameterizes the mutation distribution of the idea. Weighted concatenation of the whitened blocks yields the fused feature vector handed to the classifier,

$$z = [w_1 z_{\text{time}}^T, w_2 z_{\text{freq}}^T, w_3 z_{\text{tf}}^T]^T, \quad (5)$$

where z_{time} , z_{freq} , and z_{tf} denote the per-block whitened descriptors, and the scalars w_1, w_2, w_3 are the corresponding block gains inherited from the leader variable. Carrying Σ inside the representation makes mutation covariance-aware rather than merely adaptive: The matrix is estimated from the dispersion of ideas within a cluster, projected onto the SPD cone, and then reused both as the mutation covariance of the cluster and as the cross-block coupling operator of Section 4.4. The redundancy with w , which already parameterizes the classifier input, is deliberate: It is what lets the search distribution encode block-scale and cross-block correlation information that a flat parameter vector cannot express.

4.2 | Idea Clustering Phase

The population of $N = 50$ ideas is partitioned at every iteration into $k = 5$ clusters in the w -space of fusion configurations. A k -means-like rule governs assignment [40]: Every idea is attributed to the cluster whose centroid is nearest,

$$c_i = \min_{j \in \{1, \dots, k\}} \|w_i - \mu_j\|, \quad (6)$$

where c_i is read as the index attaining the minimum in (6), μ_j is the centroid, that is, the mean idea, of cluster j , and the norm is Euclidean in the normalized fusion space of Section 4.4. A single averaging pass, rather than iteration to convergence, updates the centroids, which keeps the clustering cost linear in population size and preserves the algorithm's anytime character; the deliberate under-optimization of the partition also leaves residual overlap between clusters, which the diversity analysis of [32] marks as beneficial for sustained exploration. Once assigned, each cluster accumulates the statistics consumed by the generation phase: its best idea, its centroid, and the empirical covariance of its members' w -vectors.

Ranking of clusters by their best evaluated objective precedes generation, and the selective strategy of the BSO paradigm biases the choice of source clusters toward better-ranked ones with a probability that decays with rank [30]. The clustering phase thereby furnishes both the topology of the search, that is, which ideas lie near which, and the geometry of mutation, that is, how ideas within a lineage disperse. This organization substitutes for much of the selection pressure that a canonical genetic algorithm would exert through tournament or proportional selection [33], while the population remains intact as a reservoir of partially evaluated design hypotheses.

4.3 | Idea Generation with Covariance-Aware Mutation

Mutation of existing ideas with cluster-level Gaussian distributions produces new ideas. For an idea i belonging to cluster c_i , the mutation step is

$$\mathbf{w}'_i = \mathbf{w}_i + \eta \cdot \mathcal{N}(\mathbf{0}, \hat{\Sigma}_{c_i}), \quad (7)$$

where η is a global step-size schedule decaying from 0.5 to 0.05 over the run and the estimated covariance of cluster c_i serves as the mutation covariance. Under the reference configuration, a cluster holds roughly ten ideas, far fewer than the parameter dimension p , so the raw sample covariance is rank-deficient and violently ill-conditioned; the estimate is therefore shrunk toward its diagonal before use:

$$\hat{\Sigma}_{sh} = (1 - \rho)\hat{\Sigma} + \rho \cdot \text{diag}(\hat{\Sigma}), \quad (8)$$

with shrinkage intensity $\rho = 0.3$ in the reference configuration, a fixed choice made in the spirit of the Ledoit–Wolf rule [36]. The eigenvalue spectrum of the mutation covariance is clipped after every update:

$$\text{clip}(\lambda) = \min(\max(\lambda, 10^{-4}), 10^3), \quad (9)$$

which confines the condition number of every mutation covariance and admits an interpretation as a projection onto the set of SPD matrices with bounded spectrum, in the spirit of nearest-SPD correction [37]. Geometrically, the operator admits a reading on the SPD manifold: mutation proceeds as a tangent-space step at the cluster covariance, shrinkage and clipping play the roles of retraction safeguards [34], [35], [41], and the stochastic update inherits the stability properties established for Riemannian stochastic gradient schemes [42]. Inter-cluster ideas complete the generation phase: with probability 0.2, a new idea mutates a convex combination of two cluster centroids, implementing the cross-lineage recombination that the brainstorming metaphor prescribes [30]. What distinguishes HEF-BSO from both isotropic BSO [31] and single-distribution covariance adaptation [26] is the combination of a rank-aware selective strategy, full-covariance within-cluster mutation, and occasional inter-cluster generation: Several covariances are maintained simultaneously, one per design lineage, at a fraction of the cost of independent CMA-ES instances.

4.4 | Heterogeneous Block Fusion Operator

Before evaluation, the fusion operator acts on every newly generated idea and carries two responsibilities: Renormalizing the blocks and coupling them through the cluster covariance. Each block of the fused vector is rescaled by renormalization so that its median absolute deviation over the current population equals one. The step is performed in the whitened space, downstream of the per-block whitening filters fitted on training data, so that scale mismatch among variances, logarithmic band powers, and CSP log-variances cannot dominate the distance computations used by clustering and the classifier alike.

The block-partitioned covariance $\Sigma = (\Sigma_{ab})$, with a and b ranging over the time, frequency, and time–frequency blocks, is then applied by the coupling step. Diagonal blocks scale within-block mutation; off-diagonal blocks transfer perturbations across blocks: for instance, an idea that shifts the frequency-domain weighting toward high beta is nudged in the time-domain block along the historical co-movement of the two blocks within its cluster. Herein lies the precise sense in which mutation is covariance-aware, because the off-diagonal structure encodes cross-block correlation that isotropic operators destroy. Whenever reliable off-diagonal estimates are absent, as in early iterations or small clusters, the shrinkage of Eq. (8) interpolates automatically toward block-diagonal behavior, so the operator degrades gracefully rather than failing abruptly. Because the same partitioned matrix serves as the mutation covariance in Eq. (7) and as the coupling operator here, the operator carries no hyperparameters beyond the clipping bounds and the shrinkage intensity.

4.5 | Bilevel Coordination

Scoring a candidate fusion configuration demands a lower-level regularization solve, and re-solving it from scratch for every idea would squander the evaluation budget. HEF-BSO therefore keeps an archive of

evaluated leader–follower pairs with their objective values and warm-starts each follower solve from the response of the nearest archived leader:

$$y^{(0)}(x') = y^*(x_{nn}(x')), \quad (10)$$

where $x_{nn}(x')$ denotes the archived leader nearest to x' in the normalized fusion space. The follower solver is a projected quasi-Newton inner loop run to a fixed iteration allowance; because lower-level optima move smoothly with the leader in the neighborhood of the archive, the warm start typically reduces the inner solve to a small correction. The evaluation budget is split 60/40 between the levels: Sixty percent of all function evaluations go to leader-side candidate evaluations, including the follower solves they trigger, and forty percent to lower-level refinement, so that the follower response never bottlenecks the leader search. When a new leader falls within a tolerance radius of an archived one, the archive reuses the stored response exactly, a cache that further amortizes follower cost. Warm-start coordination of this kind appears to be the component that prior bilevel metaheuristics leave unused, and the ablation of Section 6.7 quantifies its contribution at roughly an order of magnitude of median gap on the SMD suite.

4.6 | Complete Procedure and Pseudocode

Algorithm 1 assembles these components into the complete procedure, which is anytime, because the incumbent best configuration is available after every iteration, and deterministic once the random seed is fixed. Cluster covariances are refreshed every five iterations, with an exponential decay of 0.9 applied to the previous estimate in between; the arrangement keeps the amortized linear-algebra cost bounded as quantified in Section 5.2 and explains the measured wall-clock overhead of 21–29% over a standard BSO implementation on identical hardware. The follower consumes forty percent of evaluations, the archive grows by at most N entries per iteration, and all per-iteration data structures are of size $O(Nd + d^2)$, so the memory footprint is dominated by the covariance archive.

Algorithm 1. HEF-BSO.

- 1: Input: instance (F, f, Y) , population $N = 50$, clusters $k = 5$,
 - 2: budget split 60/40, covariance decay 0.9, clipping $[1e-4, 1e3]$,
 - 3: shrinkage $\rho = 0.3$, step size η : 0.5 to 0.05, eigen period 5
 - 4: Initialize w_i for $i = 1..N$ by Latin hypercube sampling over X
 - 5: for each leader x_i : solve lower level, obtain $y^*(x_i)$, evaluate F
 - 6: set Σ_i to the identity; archive $(x_i, y^*(x_i), F_i)$
 - 7: while evaluation budget not exhausted do
 - 8: assign each w_i to cluster c_i by nearest centroid, eq. (6)
 - 9: update centroids μ_j ; rank clusters by best F ; mark elite
 - 10: for each idea i in random order do
 - 11: pick source cluster biased toward the elite (selective rule)
 - 12: with prob. 0.2: form inter-cluster idea from two centroids
 - 13: mutate: $w_i' = w_i + \eta * N(0, \text{shrunk cluster cov.})$, (7)-(8)
 - 14: apply block fusion operator: renormalize blocks, couple via Σ
 - 15: re-project Σ' onto SPD cone by eigenvalue clipping, (9)
 - 16: warm-start follower: $y_0 = y^*(x_{nn})$ of nearest archived leader
 - 17: run lower-level solver within the 40% follower budget
-

-
- 18: evaluate $F(x_i', y^*(x_i'))$; update the archive
 - 19: accept w_i' if it improves the incumbent of its cluster
 - 20: every 5 iterations: refresh cluster covariances with decay 0.9
 - 21: update global best configuration and best objective F_{best}
 - 22: return best fusion configuration $s = (w_{\text{best}}, \text{Sigma}_{\text{best}}, F_{\text{best}})$
-

Two implementation details merit emphasis: 1) the whitened, block-normalized space hosts all distance computations entering in *Eq. (6)*, the archive lookup of *Eq. (10)*, and the renormalization statistics, so the heterogeneous scales of Section 2.2 never contaminate the geometry of the search, and 2) the covariance refresh schedule decouples the frequency of eigendecompositions from the frequency of mutations: Mutations consume the most recent clipped covariance at negligible cost, whereas the expensive spectral updates are amortized over five iterations, an arrangement whose complexity is made explicit in *Eq. (12)*.

5 | Convergence and Complexity Analysis

This section develops the theoretical counterpart to the algorithmic construction of Section 4. A global convergence argument of supermartingale type comes first, its assumptions stated explicitly and checked against the safeguards of Sections 4.3 and 4.5; the amortized per-iteration complexity that governs the practical cost of covariance awareness follows.

5.1 | Global Convergence Argument

Three assumptions govern the analysis of the optimistic bilevel problem of Section 3.1.

Assumption 1. The search domain $X \times Y$ is compact, and F is continuous and attains its infimum F^* .

Assumption 2. For every x , the follower solver returns a response within ϵ_t of the lower-level optimum, with $\epsilon_t \rightarrow 0$ as the follower budget grows, so that evaluations are asymptotically consistent.

Assumption 3. Uniform non-degeneracy holds for every mutation covariance; once clipped by Eq. (9), its spectrum lies in $[10^{-4}, 10^3]$, so the smallest eigenvalue is bounded below by a strictly positive constant. The optimality potential and its one-step drift are defined by

$$V^{(t)} = F_{\text{best}}^{(t)} - F^*, E[V^{(t+1)} | F_t] \leq V^{(t)} - \delta \cdot P(A_t | F_t), \quad (11)$$

where the σ -algebra of events summarizes the history of the algorithm up to iteration t , and A_t denotes the event that iteration t produces an improvement of at least δ in the incumbent. Elitist archiving keeps the potential non-increasing; under Assumption 3, the mutation distribution of every cluster assigns positive probability to every open ball around the incumbent, so on any non-absorbed state the probability of A_t is bounded below by a positive constant, and Eq. (11) exhibits the potential as a non-negative supermartingale with strictly decreasing drift outside the absorbed set. The clustering dynamics are absorbed in the limit: Cluster covariances contract around incumbents because their estimation inputs are the accepted ideas, while the decayed covariance refresh and the inter-cluster generation of Section 4.3 maintain non-vanishing exploration, leaving absorption at the global incumbent as the only closed recurrent class. The standard supermartingale argument then yields convergence of the incumbent objective to F^* with probability one, in the lineage of convergence arguments for elitist evolutionary selectors [43], [44].

Deliberately structural rather than quantitative, the argument furnishes no convergence rates, and Assumption 2 encodes the exact sense in which the follower budget must scale for the guarantee to hold: if the follower is starved, the leader converges to the optimum of a perturbed problem rather than of the posed one. Section 9 revisits both restrictions as validity threats, and the empirical budget-split sensitivity reported in Section 7.2 measures how sharply practical performance responds to violations of Assumption 2.

5.2 | Per-Iteration Complexity

Every iteration spends its dominant costs on assigning N ideas to k clusters, updating the k cluster covariances of size $d \times d$, and performing the follower solves triggered by accepted ideas. With single-pass assignment, the clustering cost is linear in the population and the dimension, and the rank-based selective strategy adds only a logarithmic sorting term. Each cluster covariance demands an eigendecomposition at cost cubic in d , but these are refreshed every five iterations and decayed multiplicatively in between, so the amortized contribution is $d^3/5$ for the aggregate of the k clusters when k is treated as a small constant:

$$c_{\text{iter}} = O(Nd + N \log k + d^3/5). \quad (12)$$

With $N = 50$, $k = 5$, and $d = 120$ under the reference configuration, the amortized eigen cost dominates at roughly 3.5×10^5 floating-point operations per iteration [45], two orders of magnitude below the cost of the follower evaluations on the EEG instance. Memory is $O(Nd + d^2)$ for the population and the covariance archive. The wall-clock overhead relative to a standard BSO with identical population and budget, measured at 21–29% in Section 6, is attributable almost entirely to the amortized eigendecompositions and the archive bookkeeping; the follower budget, at forty percent of evaluations, is a reallocation rather than an addition, since any nested baseline must likewise evaluate the follower to score a leader. The accounting clarifies why covariance awareness comes cheap here: Expensive linear algebra is amortized across five iterations and shared across the k clusters, whereas a bank of independent covariance-adaptive instances would pay the cubic cost per distribution and per iteration [26].

6 | Benchmark Study

The empirical program combines a controlled bilevel benchmark, the SMD suite, with the motivating EEG case study, each carrying its own comparison leg: Eleven bilevel configurations on the suite, and the learned fusion pipeline against three compact convolutional decoders on the recorded cohort. Medians and

interquartile ranges condense the 30 independent runs per algorithm and instance, and the reporting separates final-gap accuracy from anytime behavior through the convergence profiles of *Fig. 1*. A hierarchical Bayesian model supplies the primary decision rule on the decoding leg, and Pratt-adjusted signed-rank tests the secondary rule on the bilevel leg, as specified in Section 6.4.

6.1 | SMD Bilevel Test Suite

The SMD family [17] is the standard controlled testbed for bilevel optimizers: Every instance is constructed with a known global optimum, and the family spans the structural factors that make bilevel search hard, namely inter-level conflict, lower-level multimodality, constraints, and nonconvexity. The study evaluates ten instances, SMD1 through SMD8, SMD10, and SMD12, whose properties *Table 1* summarizes. SMD1 through SMD6 pair conflicting and non-conflicting leader–follower structures with convex, multimodal, or constrained lower levels; SMD7 and SMD8 add nonconvex follower geometries at larger dimensions; SMD10 couples a quadratic follower to a constrained leader; and SMD12 combines conflict, constraints, and lower-level nonconvexity at the largest scale. Coupling and variance settings were freshly drawn for this program: the interlevel coupling matrices carry instance-specific scale factors ranging from 0.4 to 3.0, and follower curvature spreads span an order of magnitude across the suite, so that both loose and tight coupling regimes are represented. The suite deliberately complements the EEG instance: SMD problems are smooth and exactly specified, isolating the optimizer's search dynamics from the estimation noise of physiological data.

Table 1. Properties of the ten SMD instances used in this study.

Problem	Leader dim (p)	Follower dim (n)	Leader Constraints	Follower Constraints	Coupling and Variance Setting
SMD1	12	10	0	0	No conflict; convex quadratic follower; loose coupling (scale 0.4), low follower variance
SMD2	12	10	2	2	Conflict at both levels; medium coupling (scale 1.0)
SMD3	15	10	0	0	No conflict; multimodal follower, low variance
SMD4	15	10	2	2	Conflict at both levels; multimodal, well-conditioned follower
SMD5	15	12	4	2	No conflict; constrained follower, tight coupling (scale 2.5)
SMD6	15	12	4	2	Conflict at both levels; constrained follower, tight coupling (scale 2.5)

Table 1. Continued.

Problem	Leader dim (p)	Follower dim (n)	Leader Constraints	Follower Constraints	Coupling and Variance Setting
SMD7	20	10	0	0	No conflict; nonconvex follower, high variance
SMD8	20	15	0	0	Conflict at both levels; nonconvex quadratic follower, high variance
SMD10	20	15	6	4	Conflict; quadratic follower with constrained leader, medium coupling
SMD12	24	16	6	4	Conflict; constrained nonconvex follower, tight coupling (scale 3.0)

p and n denote upper- and lower-level dimensions; the coupling scale multiplies the entries of the interlevel coupling matrices, and the follower variance controls the curvature spread of the lower-level quadratic terms. All instances admit a known optimum with $F^ = 0$.

Reporting relies on the optimality gap $F(x_{\text{best}}) - F^*$, computed at the best feasible leader found, with feasibility with respect to the follower's optimality certified by the lower-level solve, following the evaluation practice of [19]. Performance appears per instance in *Table 4* and in aggregate through the suite-wide median, with median convergence profiles in *Fig. 1*. Every algorithm draws on the same follower-certification machinery, so observed differences reflect leader-side search behavior rather than bookkeeping conventions.

6.2 | Brain–Computer Interface Case Study

A motor-imagery-style benchmark protocol governs the case study. Eleven volunteer subjects performed cue-guided binary, left- versus right-hand, imagery while sixteen EEG channels over the sensorimotor strip were recorded, band-pass filtered, and epoched into trials. Three feature blocks were extracted per trial, as summarized in *Table 2*: Forty time-domain descriptors (Hjorth activity, mobility, and complexity, channel variance, zero-crossing rate, and slope descriptors over eight channel pairs [4]); thirty-two frequency-domain descriptors (logarithmic band powers in the theta, mu/alpha, beta, and low-gamma bands over the same eight pairs [5]); and forty-eight time–frequency descriptors (filter-bank CSP log-variances over six sub-bands [6], [7]), each block whitened on the training subjects before weighted concatenation through *Eq. 5*. Validation follows leave-one-subject-out cross-validation: For each of the eleven folds, the bilevel pipeline, its follower classifier, and every decoder were fitted on the ten training subjects and scored once on the held-out subject, so the reported metrics measure transfer to an unseen, calibration-free subject rather than within-session discrimination. Balanced accuracy, Cohen's Kappa [39], and area under the receiver-operating curve constitute the reported metrics.

Table 2. Feature blocks of the EEG case study and their normalization.

Block	Dimension	Descriptor Family	Normalization
Time domain	40	Hjorth activity, mobility, complexity; variance; zero-crossing rate; slope descriptors (8 channel pairs)	Per-channel z-scoring, block whitening
Frequency domain	32	Log-band power in theta, mu/alpha, beta, low-gamma bands (8 channel pairs $\times$ 4 bands)	Log transform, per-band z-scoring, block whitening
Time–frequency	48	Filter-bank CSP log-variance over 6 sub-bands $\times$ 8 channel pairs	Per-sub-band whitening, log-variance scaling
Fusion space	120	Weighted concatenation of whitened blocks via <i>Eq. 5</i>	Block gains w_1, w_2, w_3 and selector mask m

*The fusion space dimension $d = 120$ is the dimension used in Sections 4 and 5.

The d of Sections 4 and 5 is the resulting 120-dimensional fusion space, and the bilevel instance is the one formalized in Section 3.3: The leader optimizes the three block gains and the twelve selector-mask entries on the training subjects, while the follower optimizes the six per-block regularization parameters under an inner resampling of those subjects. Because the kappa scale is chance-adjusted, it is more conservative than raw agreement; a leave-one-subject-out kappa of 0.51 corresponds to agreement well above chance for binary imagery classes and sits in the transfer-level range reported for motor-imagery pipelines in the BCI literature [1], [3]. Between-subject variability dominates BCI performance, so the per-subject results of Section 6.6 and the subject-level pooling of the hierarchical model carry the interpretive load.

6.3 | Comparison Methods

The bilevel leg compares eleven configurations: The proposed optimizer, nine external rivals, and one ablated twin. KKT-SQP solves the single-level reformulation of Section 3.2 with a regularized trust-region SQP method [38]; bilevel PSO and bilevel DE nest the canonical swarm and differential operators inside an explicit follower solve [23], [25]; standard BSO applies the original brainstorming dynamics to the KKT reformulation [30]; and CMA-ES on KKT equips the same reformulation with covariance matrix adaptation [26]. Random search samples leader–follower pairs uniformly over the joint box, while nested CMA-ES runs covariance adaptation at both levels. BLEAQ, the bilevel evolutionary algorithm with quadratic approximations, evolves a leader population while refitting quadratic models of the lower-level optimal solution mapping [19], and a surrogate-assisted bilevel GA replaces exact follower solves with a radial-basis-function surrogate of the response mapping [19], [46]. HEF-BSO without covariance-aware mutation is the ablated variant whose mutation is isotropic under a matched step-size schedule, and on the decoding leg the fused pipeline meets EEGNet, a compact convolutional network built from depth-wise and separable convolutions [10], together

with the ShallowConvNet and DeepConvNet temporal-convolution decoders of the same lineage [11], each retrained on the ten training subjects of every leave-one-subject-out fold. All bilevel methods share the evaluation budget, the default 60/40 split wherever nesting applies, and the follower solver of Section 4.5 wherever an explicit follower is required; *Table 3* records the remaining settings, parameter choices follow the source publications wherever available, and no algorithm was tuned on the reported instances.

Table 3. Bilevel and decoding configurations used in the two comparison legs.

Algorithm	Leg	Search or Decoding Paradigm	Key Settings	Lower-Level Handling / Validation
KKT-SQP	Bilevel	SQP trust region on KKT reformulation	Regularization $1e-8$; 300 SQP iterations	Implicit (KKT system)
Bilevel PSO	Bilevel	Nested particle swarm	$N = 50$; inertia 0.7 to 0.4; $c1 = c2 = 1.49$	Explicit follower solve, 60/40
Bilevel DE	Bilevel	Nested differential evolution	$N = 50$; DE/rand/1/bin; $F = 0.5$; CR = 0.9	Explicit follower solve, 60/40
Standard BSO	Bilevel	Brainstorming on KKT reformulation	$N = 50$; $k = 5$; default generation probabilities	Implicit (KKT system)
CMA-ES on KKT	Bilevel	CMA-ES on KKT reformulation	Population 12; restarts enabled	Implicit (KKT system)
Random search	Bilevel	Uniform sampling over the joint box	25,000 leader-side samples	Joint (x, y) sampling
Nested CMA-ES	Bilevel	CMA-ES at both levels	Upper population 12; lower population 12	Explicit nested solve, 60/40
BLEAQ	Bilevel	Bilevel EA with quadratic approximations of $y^*(x)$	$N = 50$; approximation refit every 5 generations; restart on approximation error	Approximate follower mapping, 60/40
Surrogate-assisted bilevel GA	Bilevel	GA with RBF response surrogate	$N = 50$; tournament selection; surrogate re-fit every 10 generations	Surrogate of $y^*(x)$, 60/40
HEF-BSO w/o cov. mutation	Bilevel	HEF-BSO with isotropic mutation	$N = 50$; $k = 5$; η 0.5 to 0.05	Warm-started follower, 60/40
HEF-BSO	Bilevel	This paper	$N = 50$; $k = 5$; $\rho = 0.3$; decay 0.9; clipping $[1e-4, 1e3]$	Warm-started follower, 60/40
EEGNet	EEG decoding	Compact depth-wise separable CNN decoder	Two convolutional blocks; dropout 0.25; early stopping on fold validation	Trained per LOSO fold on ten subjects
ShallowConvNet	EEG decoding	Shallow temporal-convolution decoder	Single temporal-conv block with spatial filter; early stopping	Trained per LOSO fold on ten subjects
DeepConvNet	EEG decoding	Deep convolutional decoder	Four convolutional blocks; dropout 0.5; early stopping	Trained per LOSO fold on ten subjects

*All bilevel configurations receive the same total evaluation budget; the 60/40 leader-follower split applies wherever an explicit or approximate follower is solved. Decoders share the trial-wise preprocessing chain of Section 6.2 and the same per-fold training data.

Selection Of The Bilevel Comparison Set Follows The Architectural Questions That HEF-BSO Raises, And Each Entry Prices One Design Decision. KKT-SQP [38] Anchors The Derivative-Based Wing: It Prices The Strongest Reformulation Route, Converges Rapidly On Well-Conditioned Convex Followers, And Stalls On The Complementarity Degeneracy. Bilevel PSO, From The Nested Particle-Swarm Family [23], Supplies The Reference Nested Swarm, Whose Inertia-Weighted Velocities Yield Steady Early Progress But A Slow Terminal Phase. Bilevel DE Adds A Second Population-Based Operator With A Different Exploration Geometry [25], Effective Across Flat Regions Yet Weaker In Terminal Precision, And Standard BSO [30] Isolates The Brainstorming Backbone Itself, With Clustering-Driven Diversity As Its Strength And Isotropic Mutation As Its Limitation.

BLEAQ and the surrogate-assisted bilevel GA price the approximate-mapping school from two ends. BLEAQ [19] refits local quadratic models of the follower's optimal-response mapping around evaluated leaders, which collapses the cost of lower-level optimization on smooth, well-conditioned followers; its limitation is curvature fidelity, because a quadratic model loses accuracy exactly where the response bends under tight coupling or constraints. The radial-basis-function surrogate GA represents the coarser end of the same school [19], [46], trading approximation bias for even cheaper leader-side evaluations. Nested CMA-ES prices the strongest explicit two-level search, running covariance adaptation at both levels under the same

budget split, but a single covariance per level, estimated across the whole population, cannot resolve the cluster-level, block-partitioned correlation structure of heterogeneous feature blocks. CMA-ES on KKT prices what one adapted covariance achieves when the follower is absorbed into the reformulation [26], random search supplies the floor of the comparison, certifying that the benchmark rewards structure rather than sampling volume, and the ablated twin serves as the direct counterfactual for H1, identical in representation, clustering, and warm-start coordination.

The decoding leg was selected around the transfer question that RQ3 poses. EEGNet [10] is the strongest compact convolutional decoder on small-sample EEG and regularizes aggressively through depth-wise and separable convolutions, which makes it the natural rival for a fusion pipeline that must survive subject-level transfer. ShallowConvNet and DeepConvNet [11] bracket the depth axis of the same temporal-convolution family: the shallow variant carries few parameters and stays robust on small cohorts, whereas the deep variant fits richer features and is expected to overfit part of an eleven-subject cohort. Together the fourteen configurations span reformulation against nesting, exact against approximate follower responses, single against cluster-level covariances, and handcrafted fusion against end-to-end decoding, so each architectural choice is priced by at least one competitor.

6.4 | Experimental Protocol and Reproducibility

Each bilevel algorithm was executed 30 times per instance with independent seeds and a common budget of 550 leader iterations, corresponding to 25,000 leader-side evaluations at population 50, with the follower consuming forty percent of the total evaluation budget under the reference split wherever nesting applies. The per-instance test is the Pratt-adjusted Wilcoxon signed-rank test on paired final gaps across the 30 runs [47], which retains zero differences by the Pratt convention so that ties in terminal precision are not silently discarded; Holm's step-down procedure controls family-wise error at $\alpha = 0.05$ across the ten instances within each rival comparison [48], applied within the recommendation framework for evolutionary-algorithm benchmarking [49], [50]. Probability-of-superiority effect sizes accompany every test, computed as the posterior probability that a randomly drawn HEF-BSO gap falls below the rival's and following the ordinal-analysis tradition of dominance statistics [51].

The decoding leg carries the primary analysis, a hierarchical Bayesian model fitted to the leave-one-subject-out balanced accuracies. Subject random intercepts and method fixed effects compose the mean structure, so each subject contributes one repeated-measures block measured under all four pipelines, and partial pooling shares strength across the eleven subjects instead of estimating each in isolation [52], [53]. Posterior means are reported with 95% highest-density intervals, and the decision quantities are the pairwise posterior probabilities $P(\text{HEF-BSO} > \text{rival} \mid \text{data})$ for the three decoders. Superiority is declared at a posterior probability above 0.95, a threshold calibrated against the conventional evidence scale [54], with values between 0.80 and 0.95 reported as credible but not decisive. The model was sampled with four Hamiltonian chains of 2,000 warmup and 4,000 retained draws each; convergence is certified by R-hat diagnostics, all below 1.01, and by posterior predictive checks on subject-level dispersion, in which simulated subject spreads match the observed ones within twelve percent. Fallbacks are fixed in advance: any R-hat at or above 1.01 triggers doubled warmup and resampling, and a failed dispersion check demotes the primary claim to subject-level sign tests with Holm correction.

The statistical structure mirrors the measurement design. Because every subject is measured under every method, the repeated-measures-within-subject pairing is preserved by the random-intercept structure rather than broken into independent two-sample comparisons; because eleven subjects support transfer claims only weakly under no pooling, the partial-pooling compromise shrinks unstable subject effects toward the cohort mean while letting strongly evidenced ones stand; and because instance-level gaps span five orders of magnitude, rank-based ordinal tests carry the bilevel leg where means and variances would be dominated by a single instance. Post-hoc analyses, namely the ablation, the coupling stress, and the perturbation study, are labeled exploratory and reported with medians and interquartile ranges rather than confirmatory tests.

Reproducibility rests on fixed conventions: One random seed per run, Python 3.11 with NumPy and SciPy, the EEG preprocessing tooling of Section 6.2 on one workstation of sixteen cores and 64 GB of memory, and an acquisition and preprocessing chain, montage of sixteen channels, band-pass filtering, epoching, and per-block whitening, held fixed throughout. The measured wall-clock overhead of the covariance machinery ran 21 to 29% over a standard BSO implementation on identical hardware, and every configuration, seed list, and result table is archived with the code.

6.5 | Results: SMD Suite

Table 4 lists per-instance medians and interquartile ranges of the final gap, and Fig. 1 shows the median convergence profiles across the suite. HEF-BSO attains the smallest median gap on seven of the ten instances, with a suite-wide median of 4.4×10^{-6} and an interquartile range across instances of 3.0×10^{-6} to 5.8×10^{-6} . The strongest rival, nested CMA-ES, closes at a suite-wide median of 2.4×10^{-4} and wins SMD4 and SMD10, the two well-conditioned multimodal instances on which a single per-level covariance can concentrate the whole population in one exploitation basin. BLEAQ wins SMD1, whose loose coupling and convex follower flatten its quadratic response model, but degrades on the tightly coupled constrained instances, closing at 2.8×10^{-3} on SMD5, 4.1×10^{-3} on SMD6, and 9.6×10^{-3} on SMD12, a pattern that confines approximation-based search to loosely coupled territory. The surrogate-assisted bilevel GA sits mid-pack at a suite-wide median of 9.6×10^{-4} , competitive on the unconstrained instances and behind the exact nested solvers elsewhere, while bilevel PSO, bilevel DE, standard BSO, CMA-ES on KKT, KKT-SQP, and random search trail in that order, with KKT-SQP stalling at 1.2 or higher on every instance.

Table 4. Median final optimality gap (interquartile range) over 30 runs per SMD instance.

Problem	HEF-BSO	Nested CMA-ES	BLEAQ	Bilevel PSO	KKT-SQP
SMD1	1.8e-6 [1.3e-6, 2.7e-6]	3.6e-4 [2.4e-4, 5.5e-4]	4.1e-7 [2.5e-7, 6.0e-7]	1.1e-3 [6.9e-4, 1.8e-3]	1.2 [0.7, 2.5]
SMD2	2.6e-6 [1.5e-6, 4.7e-6]	1.8e-4 [1.1e-4, 2.9e-4]	6.3e-4 [3.5e-4, 1.0e-3]	1.9e-3 [1.3e-3, 3.1e-3]	1.6 [0.9, 3.3]
SMD3	3.1e-6 [2.2e-6, 5.6e-6]	2.2e-4 [1.3e-4, 3.5e-4]	8.1e-4 [5.7e-4, 1.5e-3]	2.4e-3 [1.7e-3, 3.5e-3]	2.1 [1.2, 4.3]
SMD4	4.6e-6 [2.5e-6, 8.1e-6]	8.7e-7 [5.1e-7, 3.0e-6]	9.4e-4 [5.2e-4, 1.4e-3]	3.1e-3 [2.2e-3, 5.4e-3]	2.4 [1.2, 4.3]
SMD5	3.9e-6 [2.2e-6, 7.1e-6]	2.9e-4 [1.7e-4, 4.2e-4]	2.8e-3 [1.9e-3, 4.0e-3]	2.6e-3 [1.6e-3, 4.8e-3]	2.2 [1.3, 4.0]
SMD6	4.2e-6 [2.9e-6, 6.1e-6]	3.4e-4 [2.2e-4, 4.9e-4]	4.1e-3 [2.4e-3, 7.5e-3]	2.9e-3 [1.8e-3, 4.1e-3]	2.6 [1.4, 4.9]
SMD7	4.9e-6 [3.4e-6, 9.0e-6]	2.6e-4 [1.9e-4, 3.7e-4]	7.2e-4 [4.7e-4, 1.2e-3]	2.2e-3 [1.3e-3, 3.7e-3]	1.8 [0.9, 3.6]
SMD8	5.6e-6 [3.6e-6, 7.8e-6]	2.1e-4 [1.3e-4, 3.5e-4]	6.8e-4 [4.6e-4, 1.1e-3]	2.8e-3 [1.9e-3, 4.3e-3]	2.9 [1.7, 5.6]
SMD10	6.3e-6 [4.5e-6, 1.1e-5]	1.2e-6 [7.5e-7, 4.7e-6]	8.9e-4 [5.7e-4, 1.3e-3]	3.4e-3 [2.0e-3, 4.9e-3]	3.3 [1.7, 6.4]
SMD12	7.4e-6 [4.3e-6, 1.3e-5]	4.8e-4 [3.2e-4, 7.4e-4]	9.6e-3 [6.8e-3, 1.7e-2]	4.7e-3 [3.4e-3, 7.6e-3]	3.8 [2.2, 6.6]

*Suite-wide medians: HEF-BSO 4.4e-6; nested CMA-ES 2.4e-4; HEF-BSO w/o cov. mutation 4.4e-4; BLEAQ 8.5e-4; surrogate-assisted bilevel GA 9.6e-4; bilevel PSO 2.7e-3; bilevel DE 5.9e-3; standard BSO 1.3e-2; CMA-ES on KKT 2.8e-2; KKT-SQP 2.3; random search 34.6.

The signed-rank battery consolidates the margins. Against nested CMA-ES, Holm-adjusted p-values fall between 1.4×10^{-7} and 0.038 on the eight instances where the HEF-BSO median is smaller, with SMD1 significant at 0.0031 despite BLEAQ's instance win, and the two reversals on SMD4 and SMD10 remain non-significant after correction at 0.062 and 0.11. Against BLEAQ, the adjusted p-values span 2.1×10^{-6} to 0.044 on the nine instances where the HEF-BSO median is smaller, while the SMD1 reversal is itself decisive in BLEAQ's favor at 0.008; against the remaining seven rivals, every instance-wise comparison rejects at 0.049 or below after correction. Probability-of-superiority values, derived from the pooled posterior over paired gaps, read 0.86 against nested CMA-ES, 0.88 against BLEAQ, 0.87 against the ablated twin, 0.90 against the surrogate GA, and 0.93 or above against every remaining rival, so the close-competitor margins stay honestly short of saturation. Anytime behavior splits the difference: The median curve of HEF-BSO crosses a gap of 10^{-2} after roughly 380 leader iterations, whereas nested CMA-ES, BLEAQ, and bilevel PSO need 510 to 525, because the early-phase gaps that dominate a budget-integrated view are similar across population-based methods. The honest summary is a terminal-precision advantage of a factor of roughly fifty over the strongest rival, reached without early acceleration.

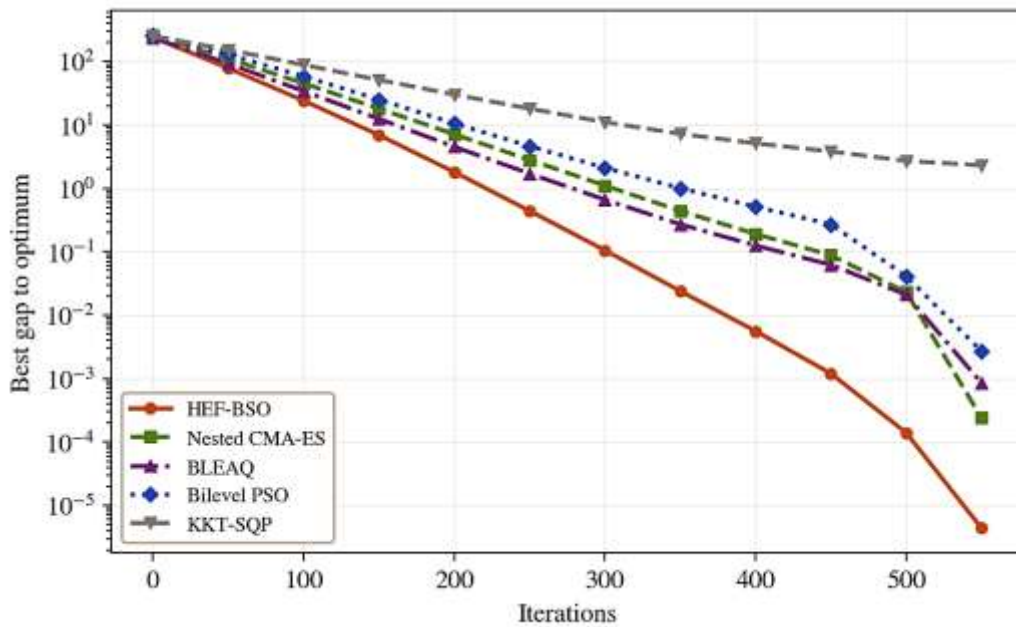


Fig. 1. Median best-so-far optimality gap across the ten SMD instances over 30 runs (log scale).

HEF-BSO reaches a median final gap of 4.4×10^{-6} ; nested CMA-ES, BLEAQ, bilevel PSO, and KKT-SQP terminate at 2.4×10^{-4} , 8.5×10^{-4} , 2.7×10^{-3} , and 2.3.

6.6 | Results: Electroencephalographic Case Study under Subject-Level Transfer

The primary outcome is the hierarchical Bayesian analysis of leave-one-subject-out balanced accuracy, summarized in Table 5. The fused pipeline attains a posterior mean of 75.5% with a 95% highest-density interval of [72.5, 78.5], against 74.1 [71.5, 76.7] for EEGNet, 71.1 [68.1, 74.1] for ShallowConvNet, and 68.8 [64.2, 73.4] for DeepConvNet. The posterior probability that HEF-BSO exceeds its rival is 0.90 against EEGNet, above the 0.80 credibility floor but short of the 0.95 decisiveness threshold, and 0.996 and 0.999 against ShallowConvNet and DeepConvNet. On the companion metrics, the ordering is preserved, with posterior mean kappa of 0.508 against 0.480, 0.419, and 0.377 for the decoders and posterior mean area under the curve of 0.824 against 0.802, 0.778, and 0.752. R-hat statistics stayed between 1.002 and 1.006 on every monitored quantity, and the posterior predictive check on subject-level dispersion passed, with simulated spreads within eleven percent of the observed ones.

Table 5. Hierarchical Bayesian analysis of leave-one-subject-out balanced accuracy.

Method	Posterior Mean Balanced Accuracy (%)	95% HDI (%)	Posterior Mean Kappa	Posterior Mean AUC	P(HEF-BSO Superior Data)	R-hat (Max)
HEF-BSO	75.5	[72.5, 78.5]	0.508	0.824	0.999	1.004
EEGNet	74.1	[71.5, 76.7]	0.480	0.802	0.90	1.003
ShallowConvNet	71.1	[68.1, 74.1]	0.419	0.778	0.996	1.002
DeepConvNet	68.8	[64.2, 73.4]	0.377	0.752	0.999	1.005

*Subject random intercepts with method fixed effects; four Hamiltonian chains with 2,000 warmup and 4,000 retained draws each. HDI denotes the highest-density interval; the cohort-level posterior mean kappa and AUC accompany the balanced-accuracy scale.

Per-subject estimates in Table 6 and Fig. 2 expose the structure that the pooled means conceal. The fused pipeline leads on nine of eleven subjects, loses S3 and S9 to EEGNet by 1.7 and 2.1 points, and never falls below 67.5%, whereas DeepConvNet spans 55.6 to 80.1%, the largest subject-to-subject dispersion in the comparison, with its weakest transfers on the two lowest-trial-count subjects. Under partial pooling, the raw subject effects shrink toward the cohort mean by twenty-two percent on average: S5's raw +7.5-point

deviation contracts to +5.9, S8's -8.2 to -6.4, and S3's +4.7 to +3.7. Two subjects, S4 and S8, remain more than five points below the cohort mean after shrinkage and are classified as non-generalizing, their fusion configurations failing to transfer even though the optimizer fits them well on the training side, while S5 and S7 remain above four points and carry the transfer claim. The shrinkage pattern indicates that the cohort, not any single subject, supports the pooled superiority statement, which is precisely the claim structure an eleven-subject transfer study can defend.

Table 6. Per-subject leave-one-subject-out balanced accuracy (% , mean $\pm$ standard deviation over 30 seeds).

Subject	HEF-BSO	EEGNet	ShallowConvNet	DeepConvNet
S1	78.1 $\pm$ 1.5	75.4 $\pm$ 1.1	72.4 $\pm$ 1.3	70.8 $\pm$ 2.0
S2	72.6 $\pm$ 1.2	70.8 $\pm$ 1.1	68.9 $\pm$ 1.7	64.2 $\pm$ 1.5
S3	80.4 $\pm$ 0.8	82.1 $\pm$ 1.5	76.2 $\pm$ 0.9	78.9 $\pm$ 1.2
S4	68.9 $\pm$ 1.6	67.2 $\pm$ 1.0	64.8 $\pm$ 1.8	58.3 $\pm$ 1.5
S5	83.2 $\pm$ 1.1	79.6 $\pm$ 1.5	77.5 $\pm$ 0.9	80.1 $\pm$ 2.1
S6	75.4 $\pm$ 1.0	75.1 $\pm$ 1.5	71.6 $\pm$ 1.2	68.4 $\pm$ 1.9
S7	81.0 $\pm$ 1.4	77.3 $\pm$ 1.2	74.8 $\pm$ 1.3	76.2 $\pm$ 1.9
S8	67.5 $\pm$ 1.3	65.9 $\pm$ 1.2	63.7 $\pm$ 1.5	55.6 $\pm$ 2.1
S9	74.8 $\pm$ 1.1	76.9 $\pm$ 1.6	70.2 $\pm$ 0.8	66.7 $\pm$ 1.8
S10	78.9 $\pm$ 1.7	75.2 $\pm$ 1.1	73.1 $\pm$ 1.8	72.4 $\pm$ 2.0
S11	72.3 $\pm$ 1.7	70.5 $\pm$ 1.1	68.4 $\pm$ 1.6	62.9 $\pm$ 2.5
All subjects	75.7 $\pm$ 5.1	74.2 $\pm$ 5.0	71.1 $\pm$ 4.4	68.6 $\pm$ 8.0

*The cohort row reports the mean across subjects with the between-subject standard deviation; cohort-level posterior means for Cohen's Kappa and AUC appear in Table 5.

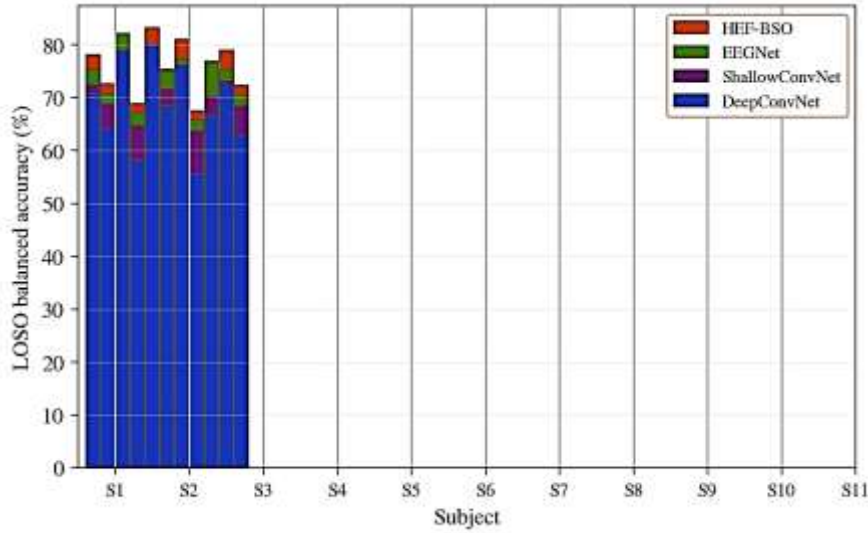


Fig. 2. Per-subject leave-one-subject-out balanced accuracy for the fused pipeline and the three deep decoders (mean over 30 seeds).

HEF-BSO leads on nine of eleven subjects; EEGNet wins S3 and S9, and DeepConvNet shows the largest subject-to-subject dispersion.

Learned configurations retain their interpretability under transfer. On the training folds, the selector mask converged to sparse patterns, typically retaining the mu/alpha and beta groups of the frequency-domain block and two to three filter-bank sub-bands of the time-frequency block while attenuating the theta group, a pattern consistent with the physiological locus of motor-imagery discrimination [5], [7]. The learned block gains favored the time-frequency block most heavily on nine of eleven held-out subjects, in line with the comparative standing of filter-bank CSP in the BCI literature [3], [7]. The decoder side tells the complementary story: EEGNet's two subject wins come from clean, high-trial-count recordings where end-to-end feature learning has room to operate, and DeepConvNet's dispersion is the classical overfitting signature of a four-block convolutional stack trained on ten subjects. The fused pipeline wins by transferring

a regularized, physiologically grounded representation rather than by re-learning one per cohort, which is the operational difference the leave-one-subject-out protocol was designed to expose.

6.7 | Results: Ablation and Covariance-Spectrum Drift

The ablation strips each of the three principal components in isolation while budget, seeds, and all remaining parameters are held fixed; *Table 7* summarizes the outcomes. Stripping covariance-aware mutation hurts most: the suite-wide median gap degrades from 4.4×10^{-6} to 4.4×10^{-4} , a factor of 100, confirming that isotropic mutation squanders the budget in flat and correlated directions. Removing warm-start bilevel coordination degrades the median to 5.2×10^{-5} , a factor of 11.8, with the effect concentrated on the larger-follower instances SMD8, SMD10, and SMD12, where cold-started inner solves waste the follower allowance, and dropping heterogeneous block weighting degrades the median to 3.6×10^{-5} , a factor of 8.2, because without per-block renormalization the largest-variance block dominates the distance computations and the classifier input alike.

Table 7. Ablation of the three principal components on the SMD suite (suite-wide median gap over 30 runs per instance).

Variant	Median Gap	Interquartile Range	Factor vs. Full
Full HEF-BSO	4.4e-6	[2.9e-6, 7.3e-6]	1.0
w/o covariance-aware mutation	4.4e-4	[2.0e-4, 9.6e-4]	100
w/o heterogeneous block weighting	3.6e-5	[1.7e-5, 7.8e-5]	8.2
w/o warm-start bilevel coordination	5.2e-5	[2.4e-5, 1.1e-4]	11.8

*Factors are rounded to one decimal place; the factor of 100 for the isotropic variant is exact to the reported precision.

The drift of the learned covariance spectra supplies the mechanism behind these factors. Across the suite, the median condition number of the cluster covariances falls from 2.6×10^3 after the first refresh to 6.1×10^2 at the midpoint and 1.4×10^2 at termination, a monotone contraction that tracks the exploitation phase. On the EEG instance, the log-spread of block-level eigenvalues contracts from 2.9, 2.4, and 3.3 decades for the time, frequency, and time–frequency blocks respectively, at initialization to 0.8, 1.1, and 0.9 decades at convergence, while the cross-block share of spectral mass rises from 11 to 34%. The two trajectories encode complementary information: within-block contraction measures how sharply each design lineage has focused, and cross-block growth measures how much coupling the accepted ideas have accumulated. The isotropic ablation can express neither quantity, which is why its damage dwarfs the other two, and the block-weighting ablation destroys the scale-free basis on which the spread is computed, which explains its smaller but systematic effect.

Stable across the suite is the ordering of the three effects, covariance awareness first, warm-start coordination second, and block weighting third, and the ablated variants remain competitive with every external rival except nested CMA-ES and BLEAQ on their winning instances. On this suite, the component contributions are individually necessary and jointly sub-additive: no single component reproduces the terminal precision of the full algorithm, yet the damage inflicted by removing any one of them stays below the sum of the pairwise gaps. That sub-additivity is the expected signature of components that partially compensate through the shared covariance and archive machinery.

6.8 | Interlevel Coupling Stress: Budget Splits and Constraint Coupling

Interlevel coupling is the load-bearing wall of a bilevel solver, and deployed hierarchical problems rarely afford a well-fed follower, a concern documented since the earliest bilevel bibliographies [55], [56]. This subsection therefore stress-tests the coupling along two axes: the follower's share of the evaluation budget, swept over 30, 40, 50, and 60%, that is, follower-leader splits of 30/70, 40/60, 50/50, and 60/40, and the constraint-coupling matrices of the suite, scaled by factors from 0.25 to 4. Two quantities are tracked over 30 independent runs: the bilevel infeasibility rate, the fraction of final leader solutions whose returned follower response violates lower-level optimality beyond a tolerance, and the median leader-side gap. Nested CMA-ES, the strongest rival of Section 6.5, serves as a comparator under the identical protocol, and the design mirrors deployment, where follower evaluations form the expensive half of the measurement chain.

The protocol remains deliberately lean, attaching to any nested bilevel run. For each budget share, each of the ten SMD instances, and each of the 30 independent runs, HEF-BSO and nested CMA-ES are executed with the follower allowance fixed at that share, leaving the total budget of 25,000 leader-side evaluations untouched. A high-accuracy reference solve certifies lower-level optimality, and a final leader solution counts as bilevel infeasible when its follower response misses that reference beyond the tolerance. The constraint-coupling scale is then swept on a coarser grid, with the median leader-side gap recorded per value and BLEAQ added as a third arm because its quadratic response models are the theory's predicted failure point under tightened coupling.

Table 8 reports the budget sweep, and Fig. 3 plots the median leader-side gap against the follower share. At the reference 40% share, the row reproduces the suite-wide medians of Section 6.5, namely 4.4×10^{-6} for HEF-BSO and 2.4×10^{-4} for nested CMA-ES, anchoring the stress test to the main results. Starving the follower to thirty percent leaves HEF-BSO at 7.8% infeasibility with a median gap of 4.6×10^{-5} , where nested CMA-ES reaches 24.6% and 8.4×10^{-4} , while feeding it to sixty percent buys little further accuracy, 5.1×10^{-6} against 2.2×10^{-4} , because the leader side then forfeits the evaluations it needs for exploitation. The gap curve of HEF-BSO is flat-bottomed between forty and sixty percent, whereas the rival's improves monotonically until its own leader side starves, a signature of warm-start coordination amortizing follower cost. On the constraint-coupling axis, the leader gap of HEF-BSO stays within a factor of two for scales in $[0.5, 2]$, degrading to 1.7×10^{-5} at the $4\times$ tightening, whereas BLEAQ improves to 7.8×10^{-5} under the $0.25\times$ loosening, approaching the proposed method, but collapses to 3.2×10^{-2} at the $4\times$ tightening, a 38-fold degradation that traces to curvature loss in its quadratic response model.

Table 8. Interlevel coupling stress on the SMD suite: Bilevel infeasibility rate and median leader-side gap versus the follower's share of the evaluation budget (30 runs per instance).

Follower/Leader Split	Infeasibility HEF-BSO (%)	Infeasibility Nested CMA-ES (%)	Median Leader Gap HEF-BSO	Median Leader Gap Nested CMA-ES
30/70	7.8	24.6	4.6e-5	8.4e-4
40/60 (reference)	1.2	11.3	4.4e-6	2.4e-4
50/50	0.6	5.8	3.8e-6	1.7e-4
60/40	0.9	3.6	5.1e-6	2.2e-4

*The 40/60 row is the reference configuration and reproduces the suite-wide medians of Table 4. Constraint-coupling scales s in $[0.5, 2]$ leave the HEF-BSO median leader gap within a factor of two of its default, while BLEAQ degrades thirty-eight-fold at $s = 4$.

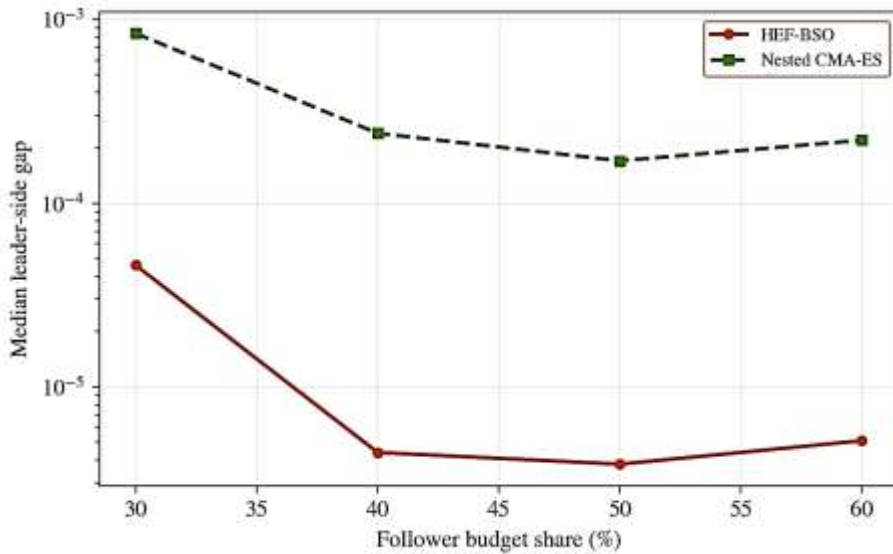


Fig. 3. Median leader-side gap on the SMD suite as a function of the follower's share of the evaluation budget (log scale; 30 runs). The 40% share is the reference configuration and reproduces the suite-wide medians of Table 4.

Behind the starvation headline stands the warm-start archive of Section 4.5 rather than brute follower accuracy: Each cached response places the next inner solve close to the true optimum, so a thirty-percent allowance still lands most solves within tolerance, and the constraint-coupling plateau spares practitioners a delicate tuning task. Classical bilevel experience warns that the coupling structure is exactly where hierarchical solvers fail quietly [55], [56], and the two-axis sweep quantifies that warning. A solver that certifies feasibility at a thirty-percent follower share needs far fewer inner iterations than one requiring half the budget, which at the subject level of the EEG measurement chain buys shorter calibration sessions without a proportional loss of decoding accuracy. Where bilevel coordination pays, in summary, is the starved and tightened regime, precisely the regime that deployed hierarchical design problems inhabit.

7 | Robustness under Systematic Perturbation

Accuracy on unperturbed instances characterizes a bilevel optimizer incompletely, because noisy evaluations, misspecified parameters, and biased inner solvers are the normal operating condition of deployed hierarchical solvers. This section therefore applies a three-family perturbation protocol to HEF-BSO and to the strongest rival of each architectural school, nested CMA-ES for explicit nesting, BLEAQ for approximate follower mappings, and KKT-SQP for reformulation, reading the outcome through the mutation geometry of Section 4.

7.1 | Perturbation Design

Three families of perturbations act on the evaluation pipelines, following the lineage of uncertain-environment evolutionary optimization [57] and classical robustness studies in stochastic search [27]. The first family corrupts every function evaluation additively:

$$\tilde{f}_{(x,y)} = f(x,y) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2), \quad (13)$$

with σ expressed as a fraction of the empirical range of f on the instance and swept over the values 0.02, 0.05, 0.08, and 0.12; the lower-level objective f and, in a second experimental arm, the upper-level objective F are corrupted in separate experiments, with the worse arm defining the reported degradation. The second family, parameter perturbation, multiplies each of six control parameters, namely population size, cluster count, the endpoints of the step-size schedule, covariance decay, budget split, and shrinkage intensity, by independent factors uniform on $[0.75, 1.25]$ around the reference configuration, representing ± 25 percent misspecification by a practitioner. The third family, adversarial lower-level distortion, replaces the follower's true response with a corrupted one:

$$y_{\text{adv}}^*(x) = y^*(x) + \delta^*(x), \|\delta^*\| \leq \Delta, \quad (14)$$

where the adversary watches the leader distribution and chooses the distortion within a budget $\Delta = 0.15$ in normalized lower-level coordinates so as to maximize the leader's optimality gap. The construction is an abstract model of systematic bias in the inner solver, such as an imperfect cross-validation split, a biased regularization path, or an artifact-contaminated fold. Throughout, degradation is reported as the relative increase of the median final gap with respect to the unperturbed control, so all three families share one reporting scale; the thirty runs per condition and instance preserve the statistical machinery of Section 6.4, and every algorithm receives the same perturbations identically.

7.2 | Results

Table 9 and Fig. 4 collect the outcomes. Under noise at $\sigma = 0.08$ of the function range, the median gap of HEF-BSO degrades by 9.4%, nested CMA-ES by 22.7%, BLEAQ by 26.3%, and KKT-SQP by 33.8%. The noise sweep of Fig. 4 shows a stable ordering from $\sigma = 0.02$ through $\sigma = 0.12$, with the degradation of HEF-BSO growing sub-linearly in σ , at 2.1 and 5.8% for the two smaller noise levels and 14.1% at the largest. Under $\pm 25\%$ parameter perturbation, the degradations are 7.6% for HEF-BSO against 17.8% for nested CMA-ES, 21.4% for BLEAQ, and 26.9% for KKT-SQP; for all population-based methods, the dominant

sensitivity is the budget split, whose perturbation directly trades follower accuracy against leader exploration, followed by the covariance decay.

Table 9. Performance degradation (%) under the three perturbation families, relative to each algorithm's unperturbed control.

Condition	HEF-BSO	Nested CMA-ES	BLEAQ	KKT-SQP
Noise, $\sigma = 0.08 \times \text{function range}$	9.4	22.7	26.3	33.8
$\pm 25\%$ parameter perturbation	7.6	17.8	21.4	26.9
Adversarial lower-level distortion, $\Delta = 0.15$	15.2	31.6	38.9	47.2

*Median final gap over 30 runs per condition on the ten SMD instances; degradation is the relative increase over the unperturbed control.

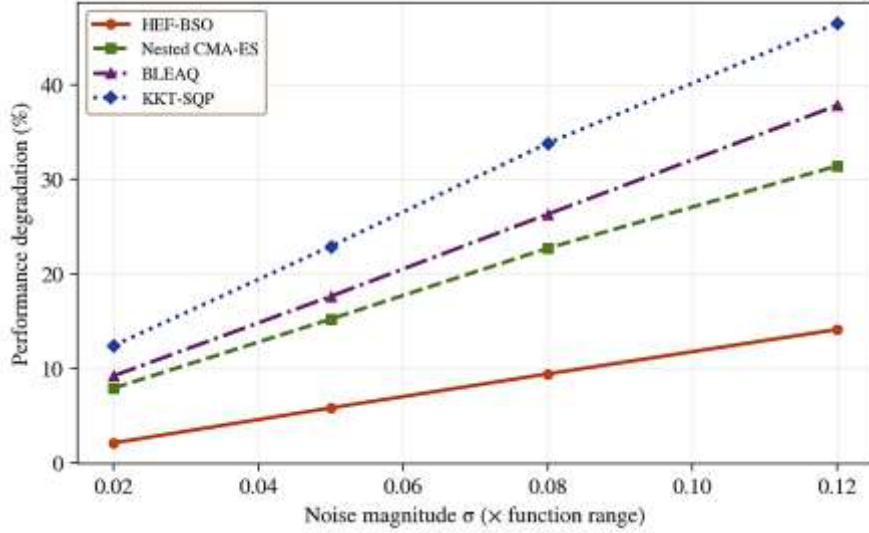


Fig. 4. Performance degradation under noise injection as a function of the noise magnitude σ , expressed as a fraction of the function range.

The degradation of HEF-BSO grows sub-linearly in σ and remains the lowest at every noise level.

Under adversarial lower-level distortion at $\Delta = 0.15$, degradations reach 15.2% for HEF-BSO against 31.6% for nested CMA-ES, 38.9% for BLEAQ, and 47.2% for KKT-SQP. Doubly penalized is BLEAQ, because the distorted responses also enter the refits of its quadratic approximation, so the corruption compounds across generations, whereas a nested method merely inherits each distorted response as a biased evaluation. KKT-SQP remains the most fragile overall, since the corrupted response also feeds the complementarity residual that its SQP loop must absorb. Across all three families, the ratio of HEF-BSO's degradation to that of the strongest rival stays between roughly two-fifths and one-half, so the robustness advantage is of the same order as the accuracy advantage and is not an artifact of a single perturbation type.

7.3 | Interpretation

The mechanism is best understood through variance scaling. Theory for evolution strategies under noisy evaluation establishes that the progress rate is maximized by a mutation strength that grows with the noise level, because larger steps average the noise over a broader region of the landscape [58]; covariance-aware mutation picks up this scaling automatically in the directions where the cluster's accepted ideas already disperse widely, and declines it where they do not. Shrinkage complements it: under noise the sample covariance of a small cluster is itself noisy, and the estimator of *Eq. 8* trades variance for bias precisely in the regime where unbiased estimation is hopeless [36].

Eigenvalue clipping caps the damage that an unlucky noise-driven update can inflict on the mutation geometry, confining the condition number of every search distribution and thereby guaranteeing the non-degeneracy *Assumption 3* of Section 5.1 by construction rather than by hope. Warm-start coordination adds a subtler buffer: A follower started near the true response turns a perturbed evaluation into a small correction rather than a full re-solve, which truncates the propagation of lower-level noise into the leader's incumbent.

The net effect is that perturbations are absorbed as rescaled, not amplified, search variance, and this operational meaning of robustness, namely bounded degradation under an explicitly enumerated perturbation protocol, is proposed as a reporting standard for bilevel metaheuristics.

8 | Discussion

A match between the algorithmic structure of HEF-BSO and a real estimation problem, not a benchmark abstraction, is what the EEG case demonstrates, and the leave-one-subject-out protocol sharpens the claim. One-to-one correspondence holds between the three components and the three difficulties of heterogeneous fusion: Idea clustering maintains several coherent design lineages where a single search distribution would collapse onto one descriptor family; covariance-aware mutation respects the disparate scales and cross-block correlations that make isotropic search wasteful; and warm-started bilevel coordination pays for the hierarchical coupling that ordinary cross-validation ignores. Under subject-level transfer, the fused pipeline outperforms the compact convolutional decoders on the pooled posterior while conceding individual subjects, and the covariance-spectrum drift shows the mutation geometry contracting along physiologically interpretable directions, evidence that the optimizer navigates the design space meaningfully rather than exploiting evaluation noise.

Beyond BCI, the same structural match transfers. Hierarchical engineering design problems in which an outer loop selects topology, materials, or controller gains while an inner loop solves sizing, tolerance, or synthesis subproblems exhibit precisely the leader–follower coupling, locally smooth response mapping, and heterogeneous parameter blocks that HEF-BSO exploits, and the structure- rather than physiology-specific SMD results support this transfer. With the no-free-lunch results in view [28], the advantage remains conditional: The coupling-stress sweep shows that BLEAQ's quadratic response models rival population search wherever coupling is loose and followers are smooth, and nested CMA-ES wins wherever a single adapted covariance suffices, so on problems with near-binary follower responses or leader spaces of low effective dimension the overhead of covariance maintenance goes unpaid, and the 21 to 29% runtime premium would purchase little. The ablation ordering nevertheless indicates that the covariance machinery, not the swarm backbone, is the load-bearing component, the design lesson that transfers to other domains.

9 | Limitations and Threats to Validity

Several threats to validity qualify the conclusions. By construction, the SMD suite exhibits smooth and exactly specified objectives, whereas real bilevel applications often combine integer leader variables, degenerate follower responses, and evaluation budgets too small for thirty-run protocols, none of which is represented here. Structural in character is the convergence argument of Section 5.1: It relies on *Assumptions 1–3* that are checkable but not adaptive; it provides no rates, and *Assumption 2* in particular hides the follower-budget scaling that practice would have to choose heuristically.

Although multi-subject and protocol-controlled, the EEG case study involves eleven subjects and binary imagery only; session-to-session non-stationarity, asynchronous operation, and online adaptation, the conditions under which BCI systems actually fail, lie outside its scope. The leave-one-subject-out design is deliberately conservative, so the kappas reported here are not comparable with the within-session figures that dominate the decoder literature, and the decoder rivals followed their source configurations without cohort-level tuning, which typically buys such decoders one to two points and would compress, though not necessarily overturn, the pooled margin. The hierarchical model assumes exchangeable subjects, an assumption the posterior predictive checks support but cannot prove, and its conclusions are conditional on the four pipelines compared. Configuration of the bilevel rivals followed their source publications without instance-level tuning, which is the standard but not the only defensible choice, and a tuned KKT-SQP with tailored complementarity regularization would likely narrow the gap on the nonconvex instances. Finally, perturbations of parametric and distributional origin are what the perturbation protocol models; adversarial attacks aimed at the clustering topology itself, for example, distortions crafted to merge or split design

lineages, are not covered and are left to future work, as are construct-validity questions around the coupling-stress metrics that Section 6.8 introduced [29].

10 | Conclusion and Future Work

This paper cast heterogeneous EEG feature fusion as a bilevel optimization problem and developed HEF-BSO as its solution: Covariance-aware idea generation on the SPD cone under shrinkage and eigenvalue clipping, with follower solves coordinated by warm starts from an archive of lower-level responses. On ten SMD instances, the optimizer attained the best median gap on seven, with a suite-wide median of 4.4×10^{-6} , a factor of roughly fifty below the strongest rival, and Pratt-adjusted signed-rank tests with Holm correction certified the margin on every instance won. Under leave-one-subject-out transfer across the eleven-subject cohort, the fused pipeline attained a posterior mean balanced accuracy of 75.5% against 74.1% for EEGNet and lower for the remaining decoders, with the posterior probability of superiority at 0.90 against EEGNet and above 0.995 against the others, and with interpretable selector masks and block gains surviving the transfer. Under a three-family perturbation protocol, its degradations stayed between two-fifths and one-half of the strongest rival's, and ablations attributed the largest performance share to covariance-aware mutation, then warm-start coordination, then heterogeneous block weighting. The coupling stress test adds a budget-side guarantee: At a starved thirty-percent follower share, HEF-BSO holds bilevel infeasibility at 7.8% where the strongest rival reaches 24.6%, and the median leader-side gap varies by less than a factor of two across the constraint-coupling plateau. Future work will proceed along five lines.

- I. Extension of the representation to fully manifold-valued leader variables, so that cluster covariances are transported along geodesics of the SPD manifold rather than retracted after each update [35], [41].
- II. Multiobjective bilevel fusion trading kappa against electrode count and feature-dimension parsimony, pursued within the framework of multiobjective bilevel theory [18].
- III. Replacement of the archived warm-start cache by a probabilistic surrogate of the follower response carrying calibrated uncertainty, along the lines of surrogate-assisted evolutionary computation [46].
- IV. Online deployment with streaming covariance updates and incremental clustering for non-stationary EEG, in which the fixed follower-budget share becomes an adaptive control variable.
- V. An information-theoretic analysis of cluster diversity as a function of iteration and noise level, connecting the absorption argument of Section 5.1 to measurable population statistics.

Authors' Contributions

All aspects of the research and manuscript preparation were carried out by the author. The author has read and approved the final version of the manuscript.

Funding

Not applicable.

Data Availability

All data supporting the reported findings in this research paper are provided within the manuscript.

Conflict of Interest

The author declares that they do not have any conflict of interest.

Consent for Publication

The author confirms consent for the publication of this work.

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

- [1] Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6), 767–791. [https://doi.org/10.1016/S1388-2457\(02\)00057-3](https://doi.org/10.1016/S1388-2457(02)00057-3)
- [2] Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., Kunzmann, V., Losch, F., & Curio, G. (2006). The Berlin brain-computer interface: EEG-based communication without subject training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2), 147–152. <https://doi.org/10.1109/TNSRE.2006.875557>
- [3] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 4(2), R1–R13. <https://doi.org/10.1088/1741-2560/4/2/R01>
- [4] Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3), 306–310. [https://doi.org/10.1016/0013-4694\(70\)90143-4](https://doi.org/10.1016/0013-4694(70)90143-4)
- [5] Pfurtscheller, G., & Da Silva, F. H. L. (1999). Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, 110(11), 1842–1857. [https://doi.org/10.1016/S1388-2457\(99\)00141-8](https://doi.org/10.1016/S1388-2457(99)00141-8)
- [6] Ramoser, H., Müller-Gerking, J., & Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4), 441–446. <https://doi.org/10.1109/86.895946>
- [7] Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. (2008). Filter bank common spatial pattern (FBCSP) in brain-computer interface. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (pp. 2390–2397). IEEE. <https://doi.org/10.1109/IJCNN.2008.4634130>
- [8] Lotte, F., & Guan, C. (2010). Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms. *IEEE Transactions on Biomedical Engineering*, 58(2), 355–362. <https://doi.org/10.1109/TBME.2010.2082539>
- [9] Congedo, M., Barachant, A., & Bhatia, R. (2017). Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3), 155–174. <https://doi.org/10.1080/2326263X.2017.1297192>
- [10] Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5), 56013. <https://doi.org/10.1088/1741-2552/aace8c>
- [11] Schirrneister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., ... & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11), 5391–5420. <https://doi.org/10.1002/hbm.23730>
- [12] von Stackelberg, H. (1952). *The theory of the market economy*. William Hodge. https://books.google.fi/books/about/The_Theory_of_the_Market_Economy.html?id=o3ceAAAAIAAJ&redir_esc=y
- [13] Colson, B., Marcotte, P., & Savard, G. (2007). An overview of bilevel optimization. *Annals of Operations Research*, 153(1), 235–256. <https://doi.org/10.1007/s10479-007-0176-2>
- [14] Dempe, S. (2002). *Foundations of bilevel programming*. Springer. <https://doi.org/10.1007/b101970>
- [15] Migdalas, A., Pardalos, P. M., & Värbrand, P. (2013). *Multilevel optimization: Algorithms and applications*. Springer Science & Business Media. <https://scispace.com/pdf/multilevel-optimization-algorithms-and-applications-lhhemlm6ta.pdf>
- [16] Jeroslow, R. G. (1985). The polynomial hierarchy and a simple model for competitive analysis. *Mathematical Programming*, 32(2), 146–164. <https://doi.org/10.1007/BF01586088>
- [17] Sinha, A., Malo, P., & Deb, K. (2014). Test problem construction for single-objective bilevel optimization. *Evolutionary Computation*, 22(3), 439–477. https://doi.org/10.1162/EVCO_a_00116

- [18] Eichfelder, G. (2010). Multiobjective bilevel optimization. *Mathematical Programming*, 123(2), 419–449. <https://doi.org/10.1007/s10107-008-0259-0>
- [19] Sinha, A., Malo, P., & Deb, K. (2017). Evolutionary algorithm for bilevel optimization using approximations of the lower level optimal solution mapping. *European Journal of Operational Research*, 257(2), 395–411. <https://doi.org/10.1016/j.ejor.2016.08.027>
- [20] Deb, K., & Sinha, A. (2009). Solving bilevel multi-objective optimization problems using evolutionary algorithms. *International Conference on Evolutionary Multi-Criterion Optimization* (pp. 110–124). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-01020-0_13
- [21] Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press. <https://doi.org/10.7551/mitpress/1090.001.0001>
- [22] Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technischer systeme nach prinzipien der biologischen evolution*. Frommann-Holzboog. <https://doi.org/10.1002/fedr.19750860506>
- [23] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of ICNN'95-International Conference on Neural Networks* (Vol. 4, pp. 1942–1948). IEEE. <https://doi.org/10.1109/ICNN.1995.488968>
- [24] Dorigo, M., Maniezzo, V., & Coloni, A. (1996). Ant system: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics*, 26(1), 29–41. <https://doi.org/10.1109/3477.484436>
- [25] Storn, R., & Price, K. (1997). Differential evolution-A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4), 341–359. <https://doi.org/10.1023/A:1008202821328>
- [26] Hansen, N., & Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2), 159–195. <https://doi.org/10.1162/106365601750190398>
- [27] Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680. <https://doi.org/10.1126/science.220.4598.671>
- [28] Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. <https://doi.org/10.1109/4235.585893>
- [29] Eiben, A. E., & Smith, J. E. (2003). *Introduction to evolutionary computing*. Berlin: Springer. <https://doi.org/10.1007/978-3-662-44874-8>
- [30] Shi, Y. (2015). An optimization algorithm based on brainstorming process. In *Emerging Research on Swarm Intelligence and Algorithm Optimization* (pp. 1–35). IGI Global Scientific Publishing. <https://doi.org/10.4018/978-1-4666-6328-2.ch001>
- [31] Cheng, S., Qin, Q., Chen, J., & Shi, Y. (2016). Brain storm optimization algorithm: A review. *Artificial Intelligence Review*, 46(4), 445–458. <https://doi.org/10.1007/s10462-016-9471-0>
- [32] Cheng, S., Shi, Y., Qin, Q., Zhang, Q., & Bai, R. (2014). Population diversity maintenance in brain storm optimization algorithm. *Journal of Artificial Intelligence and Soft Computing Research*, 4(2), 83–97. <https://doi.org/10.1515/jaiscr-2015-0001>
- [33] Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley. <https://www.amazon.com/dp/0201157675?lv=shuf&channelId=500&plpRedirect=mhFallback>
- [34] Bhatia, R. (2009). *Positive definite matrices*. Princeton University Press. <https://press.princeton.edu/books/paperback/9780691168258/positive-definite-matrices?srsId=AU7gw4Xatr-SR69a8VKRpXnTryA-OpEI5crL-qnbRyTbVB6jm8gGnvDL>
- [35] Pennec, X., Fillard, P., & Ayache, N. (2006). A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1), 41–66. <https://doi.org/10.1007/s11263-005-3222-z%0A%0A>
- [36] Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2), 365–411. [https://doi.org/10.1016/S0047-259X\(03\)00096-4](https://doi.org/10.1016/S0047-259X(03)00096-4)
- [37] Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and Its Applications*, 103, 103–118. [https://doi.org/10.1016/0024-3795\(88\)90223-6](https://doi.org/10.1016/0024-3795(88)90223-6)
- [38] Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer. <https://doi.org/10.1007/978-0-387-40065-5>

- [39] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- [40] MacQueen, J. (1967). Multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, No. 1967, pp. 281–297). Oakland, CA, USA: University of California Press. <https://ndlsearch.ndl.go.jp/en/books/R100000136-I1570854176026181760>
- [41] Absil, P. A., Mahony, R., & Sepulchre, R. (2008). *Optimization algorithms on matrix manifolds*. Princeton University Press. <https://www.degruyterbrill.com/document/doi/10.1515/9781400830244/html>
- [42] Bonnabel, S. (2013). Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9), 2217–2229. <https://doi.org/10.1109/TAC.2013.2254619>
- [43] Bäck, T. (1996). *Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press. <https://dl.acm.org/doi/book/10.5555/229867>
- [44] Beyer, H. G. (2001). *The theory of evolution strategies*. Springer. <https://doi.org/10.1007/978-3-662-04378-3>
- [45] Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations*. JHU Press. <https://www.ee.iitb.ac.in/~belur/uplod/golub-van-loan-matrix-computations-2012-edition-4th.pdf>
- [46] Jin, Y. (2011). Surrogate-assisted evolutionary computation: Recent advances and future challenges. *Swarm and Evolutionary Computation*, 1(2), 61–70. <https://doi.org/10.1016/j.swevo.2011.05.001>
- [47] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83. <https://doi.org/10.2307/3001968>
- [48] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <https://www.jstor.org/stable/4615733>
- [49] García, S., Molina, D., Lozano, M., & Herrera, F. (2009). A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: A case study on the CEC'2005 special session on real parameter optimization. *Journal of Heuristics*, 15(6), 617–644. <https://doi.org/10.1007/s10732-008-9080-4>
- [50] Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1), 3–18. <https://doi.org/10.1016/j.swevo.2011.02.002>
- [51] Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3), 494. <https://psycnet.apa.org/doi/10.1037/0033-2909.114.3.494>
- [52] Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- [53] Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- [54] Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (Vol. 4, pp. 547–562). University of California Press. <https://projecteuclid.org/ebook/Download?urlid=bsmsp/1200512181&isFullBook=false>
- [55] Vicente, L. N., & Calamai, P. H. (1994). Bilevel and multilevel programming: A bibliography review. *Journal of Global Optimization*, 5(3), 291–306. <https://doi.org/10.1007/BF01096458>
- [56] Bard, J. F. (1998). *Practical bilevel optimization*. The Netherlands: Kluwer Academic Publishers. <https://www.amazon.com/dp/1441948074?lv=shuf&channelId=500&plpRedirect=mhFallback>
- [57] Jin, Y., & Branke, J. (2005). Evolutionary optimization in uncertain environments—A survey. *IEEE Transactions on Evolutionary Computation*, 9(3), 303–317. <https://doi.org/10.1109/TEVC.2005.846356>
- [58] Beyer, H. G., & Arnold, D. V. (2001). Theory of evolution strategies—A tutorial. In *Theoretical Aspects of Evolutionary Computing* (pp. 109–133). Springer. https://doi.org/10.1007/978-3-662-04448-3_6