Supplementary Materials

A. Derivation of CFAD log-likelihood

Here, we provide a derivation of the CFAD log-likelihood to the form presented in Eq. S5 of the main text. The generative structure of the CFAD model, as described in the main text, is as follows:

$$X \mid Y \sim \mathcal{N}(\alpha\mu_y, \alpha\Lambda_y\alpha^\top + \alpha_0\Lambda_0\alpha_0^\top + \Psi) \tag{S1}$$

Here, we assume the high-dimensional data $X \in \mathbb{R}^{N\times p}$ contains N samples and has dimensionality p. The output, $Y \in \mathbb{R}^{N\times 1}$, has h classes with π_y denoting the fraction of points belonging to a particular class y. Other variables are described in Sec. 3.1. Let $\Sigma_y = \alpha\Lambda_y\alpha^\top + \alpha_0\Lambda_0\alpha_0^\top + \sigma^2 I_p$, eq. S1 trivially leads to the following log-likelihood:

$$\begin{aligned}\mathcal{L}_{\text{CFAD}} = &-\frac{Np}{2}\log(2\pi) - \frac{N}{2}\sum_y \pi_y \log|\Sigma_y| \\ &- \frac{1}{2}\sum_y \sum_{i=1}^{N\pi_y}(X_y^i - \alpha\mu_y)^T\Sigma_y^{-1}(X_y^i - \alpha\mu_y)\end{aligned} \tag{S2}$$

Now, we note that the last term in the above expression is a scalar and so it is equal to its trace. Let us assume that $\widehat{\mu}_y$ is the sample per-class mean and $\widehat{\Sigma}_{X|y}$ is the per-class covariance. Then, we can re-write the scalar term by introducing the sample per-class means as follows:

$$\begin{aligned}&\mathrm{Tr}\left(\frac{1}{2}\sum_y \sum_{i=1}^{N\pi_y}(X_y^i - \alpha\mu_y)^T\Sigma_y^{-1}(X_y^i - \alpha\mu_y)\right) \\ &= \frac{N}{2}\sum_y \pi_y \,\mathrm{Tr}\left(\Sigma_y^{-1}\left(\widehat{\Sigma}_{X|y} + (\widehat{\mu}_y - \alpha\mu_y)(\widehat{\mu}_y - \alpha\mu_y)^\top\right)\right)\end{aligned} \tag{S3}$$

Further, it is easy to show that the maximum likelihood estimate of $\mu_y = \alpha^\top\widehat{\mu}_y$. Substituting this in eq. S3 renders it equivalent to:

$$\frac{N}{2}\sum_y \pi_y \,\mathrm{Tr}\left(\Sigma_y^{-1}\left(\widehat{\Sigma}_{X|y} + (I - \alpha\alpha^\top)\widehat{\mu}_y\widehat{\mu}_y^\top(I - \alpha\alpha^\top)\right)\right) \tag{S4}$$

Fixing $B_y = (I - \alpha\alpha^\top)\widehat{\mu_y}\widehat{\mu_y}^\top(I - \alpha\alpha^\top)$, we obtain the CFAD log-likelihood in Eq. S5:

$$\begin{aligned}\mathcal{L}_{\text{CFAD}} = &-\frac{Np}{2}\log(2\pi) - \frac{N}{2}\sum_{y=1}^{h}\pi_y \log|\Sigma_y| \\ &- \frac{N}{2}\sum_{y=1}^{h}\pi_y \,\mathrm{Tr}\left(\Sigma_y^{-1}\left(\widehat{\Sigma}_{X|y} + B_y\right)\right)\end{aligned} \tag{S5}$$

B. Derivation of Sufficient Dimension Reduction for CFAD

Here we show that CFAD is formally a sufficient dimension reduction (SDR) method. The CFAD model assumes $X \mid Y$ has the following distribution:

$$X \mid (Y = y) \sim \mathcal{N}(\alpha\nu_y, \alpha\Lambda_y\alpha^\top + \alpha_0\Lambda_0\alpha_0^\top + \Psi) \tag{S6}$$

Using the CFAD-discovered projection $\alpha \in \mathbb{O}^{p\times d}$, and it's null space $\alpha_c \in \mathbb{O}^{p\times(p-d)}$, we can rotate eq. (S18) into the Gaussian joint distribution

$$\left[\begin{array}{c}\alpha^\top X \\ \alpha_c^\top X\end{array}\right]\Bigg| \sim \mathcal{N}\left(\left[\begin{array}{c}\nu_y \\ 0\end{array}\right], \left[\begin{array}{cc}\Lambda_y + \alpha^\top\Psi\alpha & \alpha^\top\Psi\alpha_c \\ \alpha_c^\top\Psi\alpha & \alpha_c^\top(\alpha_0\Lambda_0\alpha_0^\top + \Psi)\alpha_c\end{array}\right]\right) \tag{S7}$$

Note that α_0 spans only part of the nullspace of α, so in general we would have $\alpha_0 \subset \alpha_c$. We'll use the following property of Gaussian joint distributions to generate conditional distributions for $\alpha_c^\top X$:

$$\left[\begin{array}{c}x \\ y\end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c}a \\ b\end{array}\right], \left[\begin{array}{cc}A & C \\ C^T & B\end{array}\right]\right) \tag{S8}$$

$$y|x \sim \mathcal{N}\left(b + C^\top A^{-1}(x-a), B - C^\top A^{-1}C\right) \tag{S9}$$

It follows that

$$\alpha^\top X \mid (Y = y) \sim \mathcal{N}\left(\nu_y, \Lambda_y + \alpha^\top\Psi\alpha\right) \tag{S10}$$

$$\alpha_c^\top X \mid (\alpha^\top X, Y = y) \sim \mathcal{N}\left(\alpha_c^\top\Psi\alpha(\alpha^\top X - \nu_y), K\right) \tag{S11}$$

where K = $\alpha_c^\top(\alpha_0\Lambda_0\alpha_0^\top + \Psi)\alpha_c - \alpha_c^\top\Psi\alpha\left(\Lambda_y + \alpha^\top\Psi\alpha\right)^{-1}\alpha^\top\Psi\alpha_c$

At first glance, it appears that eq. (S11) depends on y. However, if Ψ is diagonal, as we assume in the CFAD model, the product $\alpha_c^\top\Psi\alpha = 0$ everywhere, and the y-dependent terms vanish. The resulting distribution for $\alpha_c^\top X \mid (\alpha^\top X, Y = y)$ is factored such that we can readily lift it to the distribution of $X \mid (\alpha^\top X, Y = y)$, from which we see

$$\alpha_c^\top X \mid (\alpha^\top X, Y = y) \sim \mathcal{N}\left(0, \alpha_c^\top[\alpha_0\Lambda_0\alpha_0^\top + \Psi]\alpha_c\right) \tag{S12}$$

$$\Longrightarrow X \mid (\alpha^\top X, Y = y) \sim \mathcal{N}\left(0, \alpha_0\Lambda_0\alpha_0^\top + \Psi\right) \tag{S13}$$

and $X \mid \alpha^\top X$ is independent of Y. We conclude under these conditions that CFAD is an SDR method.

C. fMRI Classification Results

In Sec.5, we demonstrated the application of CFAD (with smoothing prior) on fMRI data. Our choice of d relied on fixing $d + q$, which was chosen such that $d + q$ principal components explain 90% of variance in the data. We restricted the variance to 90%, in part to motivate the selection of small d since a dimensionality reduction method is useful only when a substantially low-dimensional space can be obtained. From Table 2, we know that sCFAD performed best in all subjects except subject 1 and subject 6, in which case DR outperformed sCFAD by using a much higher d. In this section, we show some more classification results on the visual object recognition fMRI dataset to establish that if we allow a higher range for d and q, sCFAD can outperform all existing methods. We also benchmark CFAD against voxel selection using ANOVA, which is classically used in fMRI anaysis.

Table S1 shows the 8-classification on all subjects at the best d for sCFAD. This optimal d is chosen by varying d in increments of 10 such that $d + q$ is set to the number of components required to explain 95% variance in the data. We find that sCFAD performs better than all other methods at this d.

We also compare sCFAD with all other methods at their respective best d in Table S2 (Note that the results for all methods, except sCFAD, are the same as Table 2). We see that sCFAD (with $d + q$ set to the number of principal components needed to achieve 95% variance) outperforms the other methods for all subjects, hence establishing the utility of our method for high-dimensional small-sample size datasets.

D. Relationship of CFAD to other generative methods

Under the CFAD model,

$$X \mid (Y = y) \sim \mathcal{N}(\alpha\nu_y, \alpha\Lambda_y\alpha^\top + \alpha_0\Lambda_0\alpha_0^\top + \Psi) \quad \text{(S14)}$$

If the the class means or (ν_y) are 0 and $q = 0$, i.e., there is no distinct latent subspace containing $Y-$independent correlations in X, CFAD reduces to:

$$X \mid (Y = y) \sim \mathcal{N}(0, \alpha\Lambda_y\alpha^\top + \Psi) \quad \text{(S15)}$$

Let $L_y \triangleq \alpha\Lambda_y^{1/2}$,

$$X \mid (Y = y) \sim \mathcal{N}(0, L_y L_y^\top + \Psi) \quad \text{(S16)}$$

Hence, CFAD reduces to **Factor Analysis** for each class with L as the loading matrix which is constrained to be spanned by α for each class.

Along with the above conditions, if $\Psi = \sigma^2 I_p$ CFAD reduces to **Probabilistic PCA**:

$$X \mid (Y = y) \sim \mathcal{N}(0, L_y L_y^\top + \sigma^2 I_p) \quad \text{(S17)}$$

Further, if all classes are constrained to have the same covariance $\Lambda \triangleq \Lambda_y$, then CFAD reduces to Factor Analysis or Probabilistic PCA (depending on Ψ) on the whole dataset X. Let $L \triangleq \alpha\Lambda^{1/2}$, hence:

$$X \mid (Y = y) \sim \mathcal{N}(0, LL^\top + \Psi) \quad \text{(S18)}$$

$$X \sim \mathcal{N}(0, LL^\top + \Psi) \quad \text{(S19)}$$

E. Comparison to GLLiM

Deleforge et al. (2015) developed a probabilistic regression method for mapping high-dimensional data to low-dimensional targets. However, unlike SDR methods, their approach does not provide an estimate of the "central subspace" (capturing the statistical dependencies of X on Y). The inverse-regression structure of the GLLiM model connecting low-dimensional target $Y \in \mathbb{R}^L$ to high-dimensional input $X \in \mathbb{R}^D$ is as follows:

$$X = \sum_{k=1}^{K} \mathbb{I}(Z = k)\,(A_k Y + b_k + E_k) \quad \text{(S20)}$$

Here, matrix $A_k \in \mathbb{R}^{D\times L}$ and $b_k \in \mathbb{R}^D$ define the transformation to the input variable and E_k is an error term set to a zero mean Gaussian. The discrete variable Z defines which of the K mappings to choose from for a particular input-output pair, hence the name Gaussian "locally-linear" mapping. They also develop a hybrid extension to their model which includes an additional unobserved output variable.

We downloaded the GLLiM package and applied it to the example DR problems used by Cook (2007) (which we also show in Fig. 2 of our paper). Fig. S1 shows that both CFAD and LAD outperform GLLiM on all but the first example (which happens to be the only linear case):

References

Cook, R. D. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

Deleforge, A., Forbes, F., and Horaud, R. High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25:893–911, 2015.

Table S1. 8-class classification accuracy on fMRI data after dimensionality reduction. (d is optimal for sCFAD such that $d + q$ contain 95% variance; 12.5% is chance performance)

SUB.	d	sCFAD	LDA	SIR	SAVE	DR	LAD	PCA	LOL	RRR	ANOVA
1	20	**89.3**	59.3	59.4	6.8	54.1	50.4	40.1	42.9	23.0	66.6
2	30	**74.7**	58.9	59.2	11.3	62.3	38.9	42.3	38.2	18.7	57.7
3	50	**66.6**	60.3	60.5	8.9	61.7	49.2	51.9	47.9	16.0	54.6
4	30	**65.4**	21.4	21.1	11.1	32.3	27.8	25.3	29.0	19.6	49.8
5	30	**78.5**	60.2	61.2	11.8	65.1	47.8	48.5	41.2	18.1	55.4
6	50	**78.4**	71.3	71.0	9.5	74.0	58.6	61.4	53.0	21.2	63.1

Table S2. 8-class classification accuracy on fMRI data after dimensionality reduction (at optimal d for the respective method)

SUB.	sCFAD		LDA		SIR		SAVE		DR		LAD		PCA		LOL		RRR	ANOVA	
	d	%	d	%	d	%	d	%	d	%	d	%	d	%	d	%	%	d	%
1	20	**89.3**	10	59.3	10	59.5	180	12.6	350	75.8	50	56.1	90	57.3	70	46.5	23.0	550	73.1
2	30	**74.7**	10	58.9	10	59.9	460	20.1	10	62.3	40	40.2	460	46.1	360	41.2	18.7	450	67.4
3	50	**66.6**	10	60.3	20	62.9	300	17.0	10	61.7	50	49.2	250	55.1	260	51.5	16.0	300	62.2
4	30	**65.4**	10	22.0	50	21.2	80	12.5	50	58.1	10	29.3	310	32.6	530	30.3	19.6	450	61.2
5	30	**78.5**	10	60.2	10	61.8	420	35.9	180	69.4	10	50.8	360	54.3	80	51.4	18.1	400	67.1
6	50	**78.4**	10	71.5	30	71.2	240	14.4	50	74.0	10	65.0	230	67.5	240	63.5	21.2	300	72.2

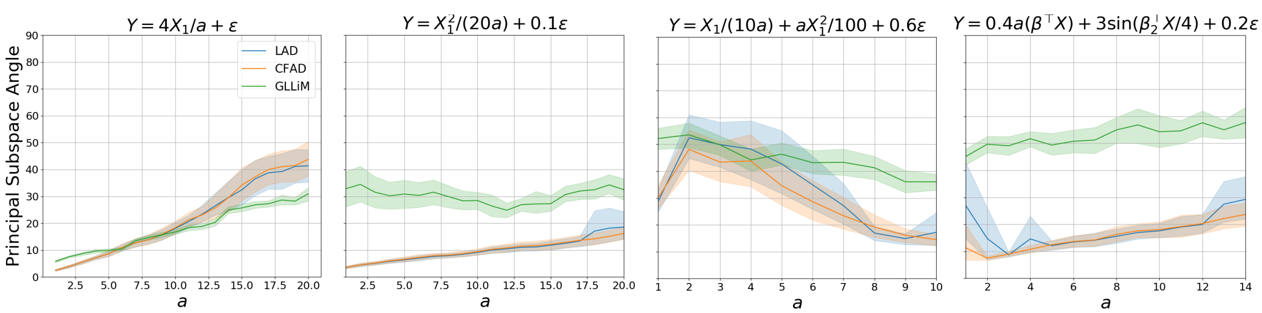

Figure S1. Principal subspace angles between the true and estimated DR subspaces for LAD, CFAD and GLLiM under varying input-output relationships. For GLLiM, we varied the number of mixtures $K \in \{1, \ldots, 20\}$ and reported the best results. We also tested the hybrid GLLiM by varying the latent dimensionality $L_w \in \{0, \ldots, 8\}$ and found the best results with L_w=0. (Note that GLLiM does not natively produce a subspace estimate; to obtain it, we took the top d-singular vectors of the inferred $\{A_k\}$, the same approach used in the SIR estimator).