Appendix

1 Notation

A : Tall matrix in $\mathbb{R}^{N\times M}, N > M$. We assume A has full rank.

A^+ : The pseudo-inverse matrix of A. $A^+ = (A^TA)^{-1}A^T$.

$\begin{bmatrix} U_\parallel & U_\perp \end{bmatrix}$: Left singular vectors of A. This is the "U" matrix from the SVD of A.

$U_\parallel$: The M leftmost left-singular vectors of A. Its columns span the image of A.

$U_\perp$: The $N - M$ rightmost left-singular vectors of A. Its columns span the nullspace of A^+.

$A^\perp$: Short hand for $U_\perp U_\perp^T$. This matrix projects onto Null(A^+).

t : "Tall" vector in $\mathbb{R}^N$.

s : "Short" vector in $\mathbb{R}^M, M < N$.

ϵ : Noise vector in $\mathbb{R}^{N-M}$.

γ : Noise vector in $\mathbb{R}^N$.

$\gamma_\perp$: Short hand for $A^\perp\gamma$. This is the projection of γ onto Null(A^+).

2 List of relevant identities

Below are some of the main identities used in the derivations:

$$U_\parallel = A(A^TA)^{\frac{-1}{2}} \tag{1}$$

$$U_\parallel U_\parallel^T = AA^+ \tag{2}$$

$$U_\perp U_\perp^T = I - AA^+ \tag{3}$$

$$A^+ = V\begin{bmatrix} S^{-1} & 0 \end{bmatrix}\begin{bmatrix} U_\parallel^T \\ U_\perp^T \end{bmatrix} \tag{4}$$

$$|S| = |A^TA|^{\frac{1}{2}} \tag{5}$$

$$\int f(z)\delta(z-a)dz = f(a) \tag{6}$$

$$\delta(Px) = \delta(x)|P|^{-1}, \quad P \in \mathbb{R}^{N\times N} \tag{7}$$

$$\int \delta(x - f(z))dx = 1, \quad \forall f(z) \tag{8}$$

$$\delta(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}) = \delta(x_1)\delta(x_2) \tag{9}$$

3 Derivations of main text

Here we present the derivations for the equations presented in the main text.

3.1 Tall MVP

Theorem 1. *The probability density function of a tall MVP is:*

$$t = As, \quad s \sim p_s(s) \tag{10}$$
$$p_t(t) = \delta(U_\perp^T t) p_s(A^+ t) |A^T A|^{\frac{-1}{2}} \tag{11}$$

Proof.

$$p_t(t) = \int p_s(s)\delta(t - As)ds \tag{12}$$
$$= \int p_s(s)\delta(\begin{bmatrix} U_\parallel^T \\ U_\perp^T \end{bmatrix}(t - As))ds \tag{13}$$
$$= \int p_s(s)\delta(U_\parallel^T(t - As))\delta(U_\perp^T(t - As))ds \tag{14}$$
$$= \int p_s(s)\delta(U_\parallel^T(t - As))\delta(U_\perp^T t - \underbrace{U_\perp^T A}_{0} s)ds \tag{15}$$
$$= \int p_s(s)\delta((A^T A)^{\frac{-1}{2}} A^T(t - As))ds\delta(U_\perp^T t) \tag{16}$$
$$= \int p_s(s)\delta(A^+(t - As))ds|A^T A|^{\frac{-1}{2}}\delta(U_\perp^T t) \tag{17}$$
$$= \int p_s(s)\delta(A^+ t - s)ds|A^T A|^{\frac{-1}{2}}\delta(U_\perp^T t) \tag{18}$$
$$= \delta(U_\perp^T t) p_s(A^+ t)|A^T A|^{\frac{-1}{2}} \tag{19}$$

□

3.2 Tall MVP with additive orthogonal noise

Theorem 2. *The probability density function of a tall MVP with additive orthogonal noise is:*

$$t = As + U_\perp \epsilon, \quad s \sim p_s(s), \quad \epsilon \sim p_\epsilon(\epsilon|s) \tag{20}$$
$$p_t(t) = p_s(A^+ t) p_\epsilon(U_\perp^T t|A^+ t)|A^T A|^{\frac{-1}{2}} \tag{21}$$

Proof.

$$p_t(t) = \int\int p_s(s)p_\epsilon(\epsilon|s)\delta(t - As - U_\perp\epsilon)d\epsilon ds \tag{22}$$
$$= \int p_s(A^+(t - U_\perp\epsilon))p_\epsilon(\epsilon|A^+ t)\delta(U_\perp^T(t - U_\perp\epsilon))d\epsilon|A^T A|^{\frac{-1}{2}} \tag{23}$$
$$= p_s(A^+ t)\int p_\epsilon(\epsilon|A^+ t)\delta(U_\perp^T t - \epsilon)d\epsilon|A^T A|^{\frac{-1}{2}} \tag{24}$$
$$= p_s(A^+ t)p_\epsilon(U_\perp^T t|A^+ t)|A^T A|^{\frac{-1}{2}} \tag{25}$$

□

3.3 Wide MVP

Theorem 3. *The probability density function of a wide MVP is*

$$s = A^+ t, \quad t \sim p_t(t) \tag{26}$$

$$p_s(s) = \int p_t(As + U_\perp \epsilon) d\epsilon |A^T A|^{\frac{1}{2}} \tag{27}$$

Let $R = \begin{bmatrix} A^+ \\ U_\perp^T \end{bmatrix}$. Then $R^{-1} = \begin{bmatrix} A & U_\perp \end{bmatrix}$ and $|R|^{-1} = |A^T A|^{\frac{1}{2}}$:

Proof.

$$R^{-1}R = \begin{bmatrix} A & U_\perp \end{bmatrix} \begin{bmatrix} A^+ \\ U_\perp^T \end{bmatrix} \tag{28}$$

$$= AA^+ + (I - AA^+) \tag{29}$$

$$= I \tag{30}$$

$$|R|^{-1} = | \begin{bmatrix} A^+ \\ U_\perp^T \end{bmatrix} |^{-1} \tag{31}$$

$$= | \begin{bmatrix} VS^{-1}U_\parallel^T \\ U_\perp^T \end{bmatrix} |^{-1} \tag{32}$$

$$= | \begin{bmatrix} VS^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} U_\parallel^T \\ U_\perp^T \end{bmatrix} |^{-1} \tag{33}$$

$$= |S| \tag{34}$$

$$= |A^T A|^{\frac{1}{2}} \tag{35}$$

□

With these identities, we can proceed with the proof.

Proof.

$$p_s(s) = \int p_t(t) \delta(s - A^+ t) dt \tag{36}$$

$$= \int p_t(t) \delta(s - A^+ t) \underbrace{\int \delta(\epsilon - U_\perp^T t) d\epsilon}_{1} dt \tag{37}$$

$$= \int \int p_t(t) \delta(\begin{bmatrix} s \\ \epsilon \end{bmatrix} - \underbrace{\begin{bmatrix} A^+ \\ U_\perp^T \end{bmatrix}}_{R} t) d\epsilon dt \tag{38}$$

$$= \int \int p_t(t) \delta(\underbrace{\begin{bmatrix} A & U_\perp \end{bmatrix}}_{R^{-1}} \begin{bmatrix} s \\ \epsilon \end{bmatrix} - t) d\epsilon dt \underbrace{|A^T A|^{\frac{1}{2}}}_{R^{-1}} \tag{39}$$

$$= \int p_t(As + U_\perp \epsilon) d\epsilon |A^T A|^{\frac{1}{2}} \tag{40}$$

□

3.4 Definition of $Z(x|Az, \Sigma)$

Definition 4. *Let $x \in \mathbb{R}^N, z \in \mathbb{R}^M$, $A \in \mathbb{R}^{N \times M}$ and let $N(x|\mu, \Sigma)$ denote the probability density function of a Gaussian centered at μ with covariance Σ. We construct the function $Z(x|A, \Sigma)$ to be equal to the following:*

$$\begin{aligned} Z(x|A,\Sigma) :&= \int N(x|Az,\Sigma)dz \\ &= \frac{N(x|0,\Sigma)}{N(h|0,J)}|J|^{-1} \\ J =&A^T\Sigma^{-1}A, \quad h = A^T\Sigma^{-1}x \end{aligned} \tag{41}$$

3.4.1 Partition Function of Gaussian

$$\int \exp\{-\frac{1}{2}z^TJz + z^Th\}dz = N(h|0,J)^{-1}|J|^{-1} \tag{42}$$

Proof.

$$\int N(z|J^{-1}h, J^{-1})dz = 1 \tag{43}$$

$$\int \exp\{-\frac{1}{2}(z - J^{-1}h)^TJ(z - J^{-1}h) - \frac{1}{2}\log|J^{-1}| - \frac{\dim(z)}{2}\log(2\pi)\}dz = 1 \tag{44}$$

$$\int \exp\{-\frac{1}{2}z^TJz + z^Th - \frac{1}{2}h^TJ^{-1}h + \frac{1}{2}\log|J| - \frac{\dim(z)}{2}\log(2\pi)\}dz = 1 \tag{45}$$

$$\int \exp\{-\frac{1}{2}z^TJz + z^Th\}dz = \exp\{\frac{1}{2}h^TJ^{-1}h - \frac{1}{2}\log|J| + \frac{\dim(z)}{2}\log(2\pi)\} \tag{46}$$

Finally,

$$\exp\{\frac{1}{2}h^TJ^{-1}h - \frac{1}{2}\log|J| + \frac{\dim(z)}{2}\log(2\pi)\} \tag{47}$$

$$= \exp\{\frac{1}{2}h^TJ^{-1}h + \frac{1}{2}\log|J| + \frac{\dim(z)}{2}\log(2\pi)\}|J|^{-1} \tag{48}$$

$$= N(h|0,J)^{-1}|J|^{-1} \tag{49}$$

□

3.4.2 Regression Marginal

$$\int N(x|Az,\Sigma)dz = \frac{N(x|0,\Sigma)}{N(h|0,J)}|J|^{-1} \tag{50}$$

$$:= Z(x|A,\Sigma) \tag{51}$$

$$\text{where} \tag{52}$$

$$J =A^T\Sigma^{-1}A, \quad h = A^T\Sigma^{-1}x \tag{53}$$

Proof.

$$\int N(x|Az,\Sigma)dz = \int \exp\{-\frac{1}{2}(x - Az)^T\Sigma^{-1}(x - Az) - \frac{1}{2}\log|\Sigma| - \frac{\dim(x)}{2}\log(2\pi)\}dz \tag{54}$$

$$= \exp\{-\frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\log|\Sigma| - \frac{\dim(x)}{2}\log(2\pi)\} \tag{55}$$

$$\int \exp\{-\frac{1}{2}z^T\underbrace{A^T\Sigma A}_{J}z + z^T\underbrace{A^T\Sigma^{-1}x}_{h}\}dz \tag{56}$$

$$= N(x|0,\Sigma)N(h|0,J)^{-1}|J|^{-1} \tag{57}$$

$$= \frac{N(x|0,\Sigma)}{N(h|0,J)}|J|^{-1} \tag{58}$$

□

3.5 Tall RealMVP

Theorem 5. *A probability density function for a tall MVP with additive orthogonal noise that scales to high dimensions is*

$$t = As + A^{\perp}\gamma, \quad s \sim p_s(s), \quad \gamma \sim N(\gamma|\mu(s), \Sigma(s)) \tag{59}$$

$$p_t(t) = p_s(A^+t)Z(\mu(A^+t) - A^{\perp}t|A, \Sigma(A^+t)) \tag{60}$$

Proof.

$$p_t(t) = p_s(A^+t)p_{U_{\perp}^T\gamma}(U_{\perp}^T t|A^+t)|A^TA|^{\frac{-1}{2}} \tag{61}$$

$$= p_s(A^+t)\int p_{\gamma}(U_{\perp}U_{\perp}^T t + U_{\parallel}r|A^+t)dr|A^TA|^{\frac{-1}{2}} \tag{62}$$

$$= p_s(A^+t)\int p_{\gamma}(A^{\perp}t + A(A^TA)^{\frac{-1}{2}}r|A^+t)|A^TA|^{\frac{-1}{2}}dr \tag{63}$$

$$= p_s(A^+t)\int p_{\gamma}(A^{\perp}t + As|A^+t)ds \tag{64}$$

$$= p_s(A^+t)\int N(A^{\perp}t + As|\mu(A^+t), \Sigma(A^+t))ds \tag{65}$$

$$= p_s(A^+t)\int N(\mu(A^+t) - A^{\perp}t|As, \Sigma(A^+t))ds \tag{66}$$

$$= p_s(A^+t)Z(\mu(A^+t) - A^{\perp}t, A, \Sigma) \tag{67}$$

□

The step from Eq.61 to Eq.62 is due to the fact that $p_{U_{\perp}^T\gamma}$ is the probability density function of a wide MVP, so we can apply theorem 3.

3.6 Lemma 6

Lemma 6. *Let* $B = \begin{bmatrix} A^+ \\ A^{\perp} \end{bmatrix}$. *Then* $B^+ = \begin{bmatrix} A \\ A^{\perp} \end{bmatrix}$ *and* $|B^TB| = |A^TA|^{-1}$.

Proof.

$$B^+B = \begin{bmatrix} A & A^{\perp} \end{bmatrix}\begin{bmatrix} A^+ \\ A^{\perp} \end{bmatrix} \tag{68}$$

$$= AA^+ + A^{\perp}A^{\perp} \tag{69}$$

$$= AA^+ + A^{\perp} \tag{70}$$

$$= I \tag{71}$$

□

Proof.

$$|B^TB| = |\begin{bmatrix} {A^+}^T & U_{\perp}U_{\perp}^T \end{bmatrix}\begin{bmatrix} A^+ \\ U_{\perp}U_{\perp}^T \end{bmatrix}| \tag{72}$$

$$= |{A^+}^T A^+ + U_{\perp}U_{\perp}^T| \tag{73}$$

$$= |{A^+}^T A^+ + I - AA^+| \tag{74}$$

$$= |I + ({A^+}^T - A)A^+| \tag{75}$$

$$= |I + A^+({A^+}^T - A)| \tag{76}$$

$$= |A^+{A^+}^T| \tag{77}$$

$$= |(A^TA)^{-1}A^TA(A^TA)^{-1}| \tag{78}$$

$$= |A^TA|^{-1} \tag{79}$$

□

3.7 Lemma 7

Lemma 7. *Let* $B = \begin{bmatrix} A^+ \\ A^\perp \end{bmatrix}$. *Consider the left singular vectors of* B *that are orthogonal to the image of* B, $U_\perp(B)$. *Then there exists an orthogonal matrix,* Q, *such that* $U_\perp(B) = \begin{bmatrix} 0 \\ U_\parallel \end{bmatrix} Q$.

Proof.

$$U_\perp(B)U_\perp(B)^T = I - BB^+ \tag{80}$$

$$= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} A^+ \\ A^\perp \end{bmatrix} \begin{bmatrix} A & A^\perp \end{bmatrix} \tag{81}$$

$$= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} I & 0 \\ 0 & A^\perp A^\perp \end{bmatrix} \tag{82}$$

$$= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} I & 0 \\ 0 & A^\perp \end{bmatrix} \tag{83}$$

$$= \begin{bmatrix} 0 & 0 \\ 0 & AA^+ \end{bmatrix} \tag{84}$$

$$= \begin{bmatrix} 0 & 0 \\ 0 & U_\parallel U_\parallel^T \end{bmatrix} \tag{85}$$

The only way this is true is if the claim is true. □

3.8 Wide MVP - 1

Lemma 8. *An expression for the probability density function of a wide MVP is*

$$s = A^+ t, \quad t \sim p_t(t) \tag{86}$$

$$p_s(s) = \int \delta(U_\parallel^T \gamma_\perp) p_t(Ax + A^\perp \gamma_\perp) d\gamma_\perp |A^T A|^{\frac{1}{2}} \tag{87}$$

Proof.

$$p_s(s) = \int p_t(t)\delta(s - A^+ t)dt \tag{88}$$

$$= \int p_t(t)\delta(s - A^+ t) \underbrace{\int \delta(\gamma_\perp - A^\perp t)d\gamma_\perp}_{1} dt \tag{89}$$

$$= \int\int p_t(t)\delta(\begin{bmatrix} s \\ \gamma_\perp \end{bmatrix} - \underbrace{\begin{bmatrix} A^+ \\ A^\perp \end{bmatrix}}_{B} t)d\gamma_\perp dt \tag{90}$$

B is a tall matrix, so apply the tall change of variables equation

$$= \int \delta(U_\perp^T(B) \begin{bmatrix} s \\ \gamma_\perp \end{bmatrix}) p_t(B^+ \begin{bmatrix} s \\ \gamma_\perp \end{bmatrix}) d\gamma_\perp |B^T B|^{\frac{-1}{2}} \tag{91}$$

$$= \int \delta(Q^T \begin{bmatrix} 0 & U_\parallel^T \end{bmatrix} \begin{bmatrix} s \\ \gamma_\perp \end{bmatrix}) p_t(As + A^\perp \gamma_\perp) d\gamma_\perp |A^T A|^{\frac{1}{2}} \tag{92}$$

$$= \int \delta(U_\parallel^T \gamma_\perp) p_t(As + A^\perp \gamma_\perp) d\gamma_\perp |A^T A|^{\frac{1}{2}} \tag{93}$$

$$\tag{94}$$

□

3.9 Orthogonally projected noise

Lemma 9. *Let* $\gamma_\perp = A^\perp \gamma, \quad \gamma \sim N(\gamma|\mu, \Sigma)$. *Then*

$$p_{\gamma_\perp}(\gamma_\perp) = \delta(U_\parallel^T \gamma_\perp) Z(\mu - \gamma_\perp | A, \Sigma) |A^T A|^{\frac{1}{2}} \tag{95}$$

Proof.

$$\begin{aligned}
p_{\gamma_\perp}(\gamma_\perp) &= \delta(U_\parallel^T \gamma_\perp) p_{U_\perp^T \gamma}(U_\perp^T \gamma_\perp) && (96)\\
&= \delta(U_\parallel^T \gamma_\perp) \int p_\gamma(A^\perp \gamma_\perp + U_\parallel r) dr && (97)\\
&\quad \text{The } A^\perp \text{ drops because } A^\perp \gamma_\perp = A^\perp A^\perp \gamma = A^\perp \gamma = \gamma_\perp \\
&= \delta(U_\parallel^T \gamma_\perp) \int p_\gamma(\gamma_\perp + As) |A^T A|^{\frac{1}{2}} ds && (98)\\
&= \delta(U_\parallel^T \gamma_\perp) \int N(\gamma_\perp + As | \mu, \Sigma) ds |A^T A|^{\frac{1}{2}} && (99)\\
&= \delta(U_\parallel^T \gamma_\perp) \int N(\mu - \gamma_\perp | As, \Sigma) ds |A^T A|^{\frac{1}{2}} && (100)\\
&= \delta(U_\parallel^T \gamma_\perp) Z(\mu - \gamma_\perp | A, \Sigma) |A^T A|^{\frac{1}{2}} && (101)
\end{aligned}$$

□

3.10 Wide RealMVP

Theorem 10. *Let* $q(\gamma_\perp|s)$ *be the pdf of* $\gamma_\perp$ *where* $\gamma_\perp = A^\perp \gamma, \quad \gamma \sim N(\gamma|\mu(s), \Sigma(s))$. *Then a probability density function for a wide MVP that scales to high dimensions is*

$$s = A^+ t, \quad t \sim p_t(t) \tag{102}$$

$$p_s(s) = \int q(\gamma_\perp|s) \frac{p_t(As + \gamma_\perp)}{Z(\mu(s) - \gamma_\perp | A, \Sigma(s))} d\gamma_\perp \tag{103}$$

Proof.

$$\begin{aligned}
p_s(s) &= \int \delta(U_\parallel^T \gamma_\perp) p_t(As + \gamma_\perp) d\gamma_\perp |A^T A|^{\frac{1}{2}} && (104)\\
&= \int \frac{q(\gamma_\perp|s)}{q(\gamma_\perp|s)} \delta(U_\parallel^T \gamma_\perp) p_t(As + \gamma_\perp) d\gamma_\perp |A^T A|^{\frac{1}{2}} && (105)\\
&\quad \text{Let } q(\gamma_\perp|s) \text{ be the distribution from Lemma 9} \\
&= \int q(\gamma_\perp|s) \frac{\delta(U_\parallel^T \gamma_\perp) p_t(As + U_\perp U_\perp^T \gamma_\perp)}{\delta(U_\parallel^T \gamma_\perp) Z(\mu(s) - \gamma_\perp | A, \Sigma(s)) |A^T A|^{\frac{1}{2}}} d\gamma_\perp |A^T A|^{\frac{1}{2}} && (106)\\
&= \int q(\gamma_\perp|s) \frac{p_t(As + \gamma_\perp)}{Z(\mu(s) - \gamma_\perp | A, \Sigma(s))} d\gamma_\perp && (107)
\end{aligned}$$

□

3.11 ELBO for wide MVP

Corollary 11. *The ELBO of* $\log p_s(s)$ *is*

$$\log p_s(s) \geq \mathbb{E}_{q(\gamma_\perp|s)}[\log \frac{p_t(As + \gamma_\perp)}{Z(\mu(s) - \gamma_\perp | A, \Sigma(s))}] \tag{108}$$

Proof. Consider the optimal importance sampler distribution for Eq.104, $p^*(\gamma_\perp|s) = \frac{\delta(U_\parallel^T \gamma_\perp) p_t(As + \gamma_\perp) |A^T A|^{\frac{1}{2}}}{p_s(s)}$. If

we let $q(\gamma_\perp|s)$ take the form of Lemma 9, we have that:

$$KL[q(\gamma_\perp|s)||p^*(\gamma_\perp|s)] = \int q(\gamma_\perp|s) \log \frac{q(\gamma_\perp|s)}{p^*(\gamma_\perp|s)} d\gamma_\perp \tag{109}$$

$$= \int q(\gamma_\perp|s) \log \frac{\delta(U_\parallel^T \gamma_\perp) Z(\mu - \gamma_\perp|A, \Sigma)|A^T A|^{\frac{1}{2}} p_s(s)}{\delta(U_\parallel^T \gamma_\perp) p_t(As + \gamma_\perp)|A^T A|^{\frac{1}{2}}} d\gamma_\perp \tag{110}$$

$$= \int q(\gamma_\perp|s) \log \frac{Z(\mu - \gamma_\perp|A, \Sigma) p_s(s)}{p_t(As + \gamma_\perp)} d\gamma_\perp \tag{111}$$

$$= \int q(\gamma_\perp|s) \log \frac{Z(\mu - \gamma_\perp|A, \Sigma)}{p_t(As + \gamma_\perp)} d\gamma_\perp + \log p_s(s) \tag{112}$$

$KL[q(\gamma_\perp|s)||p^*(\gamma_\perp|s)]$ is a positive value, so the corollary follows directly from Eq.112. □

4 Square MVP sanity check

Consider a square matrix that is written as the product of a wide matrix and a tall matrix $C = \begin{bmatrix} B \end{bmatrix} \begin{bmatrix} A \end{bmatrix}$.

Let the singular value decomposition of A and B be $A = \begin{bmatrix} U_\parallel & U_\perp \end{bmatrix} \begin{bmatrix} S_A \\ 0 \end{bmatrix} V^T$ and $B = U \begin{bmatrix} S_B & 0 \end{bmatrix} \begin{bmatrix} V_\parallel^T \\ V_\perp^T \end{bmatrix}$. We can compute the change of probability density function of a matrix-vector product with C using the change of variables formula. If we let $x = Cz$ for $z \sim p_z(z)$,

$$p_x(x) = p_z(C^{-1}x)|C|^{-1} \tag{113}$$

However if we apply the methods presented in the main text to $u = Az$ and then $x = Bu$, we end up with a different expression:

$$p_x(x) = \int p_{Az}(B^+ x + V_\perp \epsilon) d\epsilon |BB^T|^{\frac{-1}{2}} \tag{114}$$

$$= \int \delta(U_\perp^T B^+ x + U_\perp^T V_\perp \epsilon) p_z(A^+ B^+ x + A^+ V_\perp \epsilon) d\epsilon |BB^T|^{\frac{-1}{2}} |AA^T|^{\frac{-1}{2}} \tag{115}$$

Although this second expression seems very different, it turns out to be identical to the first expression. We will show that the delta term in equation 115 accounts for the different inverse and determinant terms between equations 113 and 115.

First we write BA, $(BA)^{-1}$ and A^+B^+ out in terms of the SVD components of A and B

$$BA = US_B(V_\parallel^T U_\parallel)S_A V^T \tag{116}$$

$$(BA)^{-1} = VS_A^{-1}(V_\parallel^T U_\parallel)^{-1} S_B^{-1} U^T \tag{117}$$

$$A^+B^+ = VS_A^{-1} U_\parallel V_\parallel^T S_B^{-1} U^T \tag{118}$$

$$\tag{119}$$

It immediately follows that

$$|BA| = |A^T A|^{\frac{1}{2}} |BB^T|^{\frac{1}{2}} |V_\parallel^T U_\parallel| \tag{120}$$

Before we can show that Eq.115 is the same as Eq.113, we need two more identities:

Lemma 12.

$$U_\parallel V_\parallel^T - (V_\parallel^T U_\parallel)^{-1} = U_\parallel^T V_\perp (U_\perp^T V_\perp)^{-1} U_\perp^T V_\parallel \tag{121}$$

$$|V_\parallel^T U_\parallel| = |U_\perp^T V_\perp| \tag{122}$$

Proof. Consider $W = \begin{bmatrix} U_\parallel^T \\ U_\perp^T \end{bmatrix} \begin{bmatrix} V_\parallel & V_\perp \end{bmatrix} = \begin{bmatrix} U_\parallel^T V_\parallel & U_\parallel^T V_\perp \\ U_\perp^T V_\parallel & U_\perp^T V_\perp \end{bmatrix}$. Its inverse is trivially known because $\begin{bmatrix} U_\parallel^T \\ U_\perp^T \end{bmatrix}$ and $\begin{bmatrix} V_\parallel & V_\perp \end{bmatrix}$ are orthogonal: $W^{-1} = \begin{bmatrix} V_\parallel^T \\ V_\perp^T \end{bmatrix} \begin{bmatrix} U_\parallel & U_\perp \end{bmatrix}$. We can immediately see that the top left block of W^{-1} is $V_\parallel^T U_\parallel$. However, we can also use the block matrix inversion lemma on W to get that the top left block of W^{-1} is $(U_\parallel^T V_\parallel - U_\parallel^T V_\perp (U_\perp^T V_\perp)^{-1} U_\perp^T V_\parallel)^{-1}$. By setting the two equal to each other, we get

$$V_\parallel^T U_\parallel = (U_\parallel^T V_\parallel - U_\parallel^T V_\perp (U_\perp^T V_\perp)^{-1} U_\perp^T V_\parallel)^{-1} \tag{123}$$

The first identity follows almost directly from Eq.123 and the second identity follows using Schur's determinant identity:

$$\begin{aligned}
\underbrace{|W|}_{1} &= |\begin{bmatrix} U_\parallel^T V_\parallel & U_\parallel^T V_\perp \\ U_\perp^T V_\parallel & U_\perp^T V_\perp \end{bmatrix}| && (124)\\
&= |U_\perp^T V_\perp||U_\parallel^T V_\parallel - U_\parallel^T V_\perp (U_\perp^T V_\perp)^{-1} U_\perp^T V_\parallel| && (125)\\
&= |U_\perp^T V_\perp||(V_\parallel^T U_\parallel)^{-1}| && (126)\\
&\text{therefore} \\
|U_\perp^T V_\perp| &= |V_\parallel^T U_\parallel| && (127)
\end{aligned}$$

□

We are now ready to show that Eq.115 is the same as Eq.113.

$$\begin{aligned}
p_x(x) &= \int \delta(U_\perp^T B^+ x + U_\perp^T V_\perp \epsilon) p_z(A^+ B^+ x + A^+ V_\perp \epsilon) d\epsilon |BB^T|^{\frac{-1}{2}} |AA^T|^{\frac{-1}{2}} && (128)\\
&= \int \delta((U_\perp^T V_\perp)^{-1} U_\perp^T B^+ x + \epsilon) p_z(A^+ B^+ x + A^+ V_\perp \epsilon) d\epsilon |BB^T|^{\frac{-1}{2}} |AA^T|^{\frac{-1}{2}} |U_\perp^T V_\perp|^{-1} && (129)\\
&= p_z(A^+ B^+ x - A^+ V_\perp (U_\perp^T V_\perp)^{-1} U_\perp^T B^+ x) |BB^T|^{\frac{-1}{2}} |AA^T|^{\frac{-1}{2}} |V_\parallel^T U_\parallel|^{-1} && (130)\\
&= p_z(A^+ B^+ x - V S_A^{-1} U_\parallel V_\perp (U_\perp^T V_\perp)^{-1} U_\perp^T V_\parallel^T S_B^{-1} U^T x) |BA|^{-1} && (131)\\
&= p_z(A^+ B^+ x - V S_A^{-1} (U_\parallel V_\parallel^T - (V_\parallel^T U_\parallel)^{-1}) S_B^{-1} U^T x) |BA|^{-1} && (132)\\
&= p_z(A^+ B^+ x - (A^+ B^+ - (BA)^{-1}) x) |BA|^{-1} && (133)\\
&= p_z((BA)^{-1} x) |BA|^{-1} && (134)\\
&= p_z(C^{-1} x) |C|^{-1} && (135)
\end{aligned}$$

4.1 The importance of the dirac delta term

In the related work section of the main text, we conjecture that the dirac delta term accounts for the difference between the our modular method and the standard manifold change of variables formula. We point to the above example as a piece of evidence that the dirac delta term plays a role in connecting the algorithms described in our paper with a change of variable formula that cannot be evaluated in component-wise. To evaluate the change of variables formula of $x = BAz$, one must perform computations on B and A at once instead of first on B and then on A. Our presented method affords this component-wise property and we hope to explore a more general relationship in the future.

5 Pseudo-code for wide MVP Experiments

The tall MVP architectures are made of two linear layers and two non-linearities. The wide MVP used a multi-layer perceptron with 7 layers and 256 hidden units each to create the network that learns $\mu(s), \Sigma(s)$. The logistic CDF mixture layer used a 4 layer multi-layer perceptron with 128 hidden units each as the conditioner network.

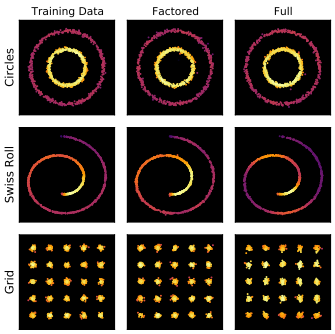
```
def make_flow(factored):

    if(factored):
        dim_change = sequential(WideMVP(input_dim=2,
                                        output_dim=4,
                                        factored=True),
                                AffineDense())
    else:
        dim_change = RectangularDense(input_dim=2,
                                      output_dim=4,
                                      factored=False)

    model = sequential(dim_change,
                       CouplingLogitsticCDFMixture(n_components=8),
                       Logit(),
                       AffineDense(),
                       CouplingLogitsticCDFMixture(n_components=8),
                       Logit(),
                       UnitGaussianPrior())

    return model
```

6 Pseudo-code for tall MVP Experiments

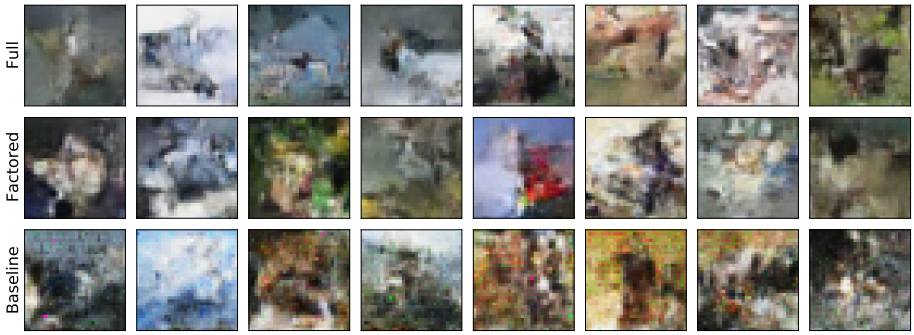
Below is the pseudo-code for the tall MVP experiments. We used a similar architecture to the Flow++ paper [Ho et al., 2019]. All of our conditioner networks for our coupling layers (logitstic cdf mixture and tall mvp) are residual networks with 4 residual blocks. Each residual block used a sequence of a 3x3 conv with 64 channels, relu non-linearity, a 1x1 conv with 64 channels, relu non-linearity and then a 3x3 conv to the original number of channels. The tall MVP is implemented as a one-by-one convolution that halves the number of channels in an input image. The multiscale architecture baseline was set by fixing the matrix in the tall MVP to be $\begin{bmatrix} I \\ 0 \end{bmatrix}$ and setting $\mu(s), \Sigma(s) = (0, I)$.

```
def FlowPP():
    layers = []
    for i in range(3):
        layers.append(ActNorm())
        layers.append(OneByOneConv())
        layers.append(CouplingLogitsticCDFMixture(n_components=32))
        layers.append(Logit())

    return sequential(*layers)

def multiscale(flow, factored):
    return sequential(Squeeze(),
                      ActNorm(),
                      TallMVP(out_channels=in_channels//2,
                              factored=factored),
                      flow)

def build_network(flow, factored):
    layers = [FlowPP()]
    if factored:
        layers.append(OneByOneConv())
    layers.append(multiscale(flow, factored))

    return sequential(*layers)

def make_network(factored):
    flow_layers = FlowPP()
    for i in range(3):
        flow_layers = build_network(flow_layers, factored)

    return sequential(UniformDequantization(),
                      Logit(),
                      flow_layers,
                      UnitGaussianPrior())
```

7 Samples from models used in experiments



(a) Samples from data, factored model and full model.



(b) Samples from full, factored and baseline models.

Figure 1: Samples from models trained for the wide and tall MVP experiments.