Supplementary Material for "Provable ICA with Unknown Gaussian Noise, with Implications for Gaussian Mixtures and Autoencoders"

A Omitted proofs in Section 2

Lemma A.1 (Denoising Lemma). $P(u) = 2\sum_{i=1}^{n}(u^T A)_i^4$

Proof: The crucial observation is that $u^T y = u^T Ax + u^T \eta$ is the sum of two independent random variables, Ax and η and that $P(u) = -\kappa_4(u^T Ax + u^T\eta) = -\kappa_4(u^T Ax) - \kappa_4(u^T\eta) = -\kappa_4(u^T Ax)$. So in fact, the functional $P(u)$ is invariant under additive Gaussian noise **independent of the variance matrix** Σ. This vastly simplifies our computation:

$$\begin{aligned}\mathbf{E}[(u^T Ax)^4] &= \sum_{i=1}^{n}(u^T A)_i^4\,\mathbf{E}[x_i^4] + 6\sum_{i<j}(u^T A)_i^2(u^T A)_j^2\,\mathbf{E}[x_i^2]\,\mathbf{E}[x_j^2] \\ &= \sum_{i=1}^{n}(u^T A)_i^4 + 6\sum_{i<j}(u^T A)_i^2(u^T A)_j^2 = -2\sum_{i=1}^{n}(u^T A)_i^4 + 3(u^T AA^T u)^2\end{aligned}$$

Furthermore $\mathbf{E}[(u^T Ax)^2]^2 = (u^T AA^T u)^2$ and we conclude that

$$P(u) = -\kappa_4(u^T y) = -\mathbf{E}[(u^T Ax)^4] + 3\,\mathbf{E}[(u^T Ax)^2]^2 = 2\sum_{i=1}^{n}(u^T A)_i^4.$$

■

Claim A.2. *If u_0 is chosen uniformly at random then with high probability for all i,*

$$\min_{i=1}^{n}\|A_i\|_2^2 n^{-4} \le D_A(u_0))_{i,i} \le \max_{i=1}^{n}\|A_i\|_2^2\frac{\log n}{n}$$

Proof: We can bound $\max_{i=1}^{n}|A_i\cdot u|$ by $\max_{i=1}^{n}\|A_i\|_2\frac{\log n}{\sqrt{n}}$ thus the bound for $\max_{i=1}^{n}(D_A(u_0))_{i,i}$ follows. Note that with high probability the minimum absolute value of n Gaussian random variables is at least $1/n^2$, hence $\min_{i=1}^{n}(D_A(u_0))_{i,i} \ge \min_{i=1}^{n}\|A_i\|_2^2 n^{-4}$. ■

Lemma A.3. *If u_0 is chosen uniformly at random and furthermore we are given $2N = \mathrm{poly}(n, 1/\epsilon, 1/\lambda_{min}(A), \|A\|_2, \|\Sigma\|_2)$ samples of y, then with high probability we will have that $(1-\epsilon)AD_A(u_0)A^T \preceq \mathcal{H}(\widehat{P}(u_0)) \preceq (1+\epsilon)AD_A(u_0)A^T$.*

Proof: First we consider each entry of the matrix updates. For example, the variance of any entry in $\mathcal{H}((u^T y)^4) = 12(u^T y)^2 yy^T$ can be bounded by $\|y\|_2^8$, which we can bound by $\mathbf{E}[\|y\|_2^8] \le O(\mathbf{E}[\|Ax\|_2^8 + \|\eta\|_2^8])$. This can be bounded by $O(n^4(\|A\|_2^8 + \|\Sigma\|_2^4))$. This is also an upper bound for the variance (of any entry) of any of the other matrix updates when computing $\mathcal{H}(\widehat{P}(u_0))$.

Applying standard concentration bounds, $\mathrm{poly}(n, 1/\epsilon', \|A\|_2, \|\Sigma\|_2)$ samples suffice to guarantee that all entries of $\mathcal{H}(\widehat{P}(u_0))$ are ϵ' close to $\mathcal{H}(P(u))$. The smallest eigenvalue of $\mathcal{H}(P(u)) = AD_A(u_0)A^T$ is at least $\lambda_{min}(A)^2\min_{i=1}^{n}\|A_i\|_2^2 n^{-4}$ where here we have used Claim 2.9. If we choose $\epsilon' = \mathrm{poly}(1/n, \lambda_{min}(A), \epsilon)$, then we are also guaranteed $(1-\epsilon)AD_A(u_0)A^T \preceq \mathcal{H}(\widehat{P}(u_0)) \preceq (1+\epsilon)AD_A(u_0)A^T$ holds. ■

Lemma A.4. *Suppose that $(1-\epsilon)AD_A(u_0)A^T \preceq \widehat{M} \preceq (1+\epsilon)AD_A(u_0)A^T$, and let $\widehat{M} = BB^T$. Then there is a rotation matrix R^* such that $\|B^{-1}AD_A(u_0)^{1/2} - R^*\|_F \le \sqrt{n}\epsilon$.*

Proof: Let $M = AD_A(u_0)A^T$ and let $C = AD_A(u_0)^{1/2}$, and so $M = CC^T$ and $\widehat{M} = BB^T$. The condition $(1-\epsilon)M \preceq \widehat{M} \preceq (1+\epsilon)M$ is well-known to be equivalent to the condition that for all vectors x, $(1-\epsilon)x^T Mx \le x^T\widehat{M}x \le (1+\epsilon)x^T Mx$.

Suppose for the sake of contradiction that $S = B^{-1}C$ has a singular value outside the range $[1-\epsilon, 1+\epsilon]$. Assume (without loss of generality) that S has a singular value strictly larger than $1+\epsilon$

(and the complementary case can be handled analogously). Hence there is a unit vector y such that $y^T SS^T y > 1 + \epsilon$. But since $BSS^T B^T = CC^T$, if we set $x^T = y^T B^{-1}$ then we have $x^T \widehat{M} x = x^T BB^T x = y^T y = 1$ but $x^T M x = x^T CC^T x = x^T BSS^T B^T x = y^T SS^T y > 1 + \epsilon$. This is a contradiction and so we conclude that all of the singular values of $B^{-1}C$ are in the range $[1-\epsilon, 1+\epsilon]$.

Let $U\Sigma V^T$ be the singular value decomposition of $B^{-1}C$. If we set all of the diagonal entries in Σ to 1 we obtain a rotation matrix $R^* = UV^T$. And since the singular values of $B^{-1}C$ are all in the range $[1-\epsilon, 1+\epsilon]$, we can bound the Froebenius norm of $B^{-1}C - R^*$: $\|B^{-1}C - R^*\|_F \leq \sqrt{n}\epsilon$, as desired. ■

B Omitted proofs in Section 3

Theorem B.1. *Suppose we are given samples of the form $y = Ax + \eta$ where x is uniform on $\{+1, -1\}^n$, A is an $n \times n$ matrix, η is an n-dimensional Gaussian random variable independent of x with unknown covariance matrix Σ. There is an algorithm that with high probability recovers $\|\widehat{A} - A\Pi diag(k_i)\|_F \leq \epsilon$ where Π is some permutation matrix and each $k_i \in \{+1, -1\}$ and also recovers $\|\widehat{\Sigma} - \Sigma\|_F \leq \epsilon$. Furthermore the running time and number of samples needed are* $\mathrm{poly}(n, 1/\epsilon, \|A\|_2, \|\Sigma\|_2, 1/\lambda_{min}(A))$

Proof: In Step 1, by Lemma 2.11 we know once we use $z = B^{-1}y$, the whitened function $P'(u)$ is inverse polynomially close to $P^*(u)$. Then by Lemma 5.3, the function $\widehat{P'}(u)$ we get in Step 2 is inverse polynomially close to $P'(u)$ and $P^*(u)$. Theorem 4.6 and Lemma 5.5 show that given $\widehat{P'}(u)$ inverse polynomially close to $P^*(u)$, Algorithm 2: : ALLOPT finds all local maxima with inverse polynomial precision. Finally by Theorem 5.6 we know A and W are recovered correctly up to additive ϵ error in Frobenius norm. The running time and sampling complexity of the algorithm is polynomial because all parameters in these Lemmas are polynomially related. ■

C Omitted proofs in Section 4

Lemma C.1. *Given v_1, v_2, ..., v_k, each γ-close respectively to local maxima $v_1^*, v_2^*, ..., v_k^*$ (this is without loss of generality because we can permute the index of local maxima), then there is an orthonormal basis $v_{k+1}, v_{k+2}, ..., v_n$ for the orthogonal space of span$\{v_1, v_2, ..., v_k\}$ such that for any unit vector $w \in \mathbb{R}^{n-k}$, $\sum_{i=1}^{n-k} w_k v_{k+i}$ is $3\sqrt{n}\gamma$ close to $\sum_{i=1}^{n-k} w_k v_{k+i}^*$.*

Proof: Let S_1 be *span*$\{v_1, v_2, ..., v_k\}$, S_2 be *span*$\{v_1^*, v_2^*, ..., v_k^*\}$ and $S_1^\perp$, $S_2^\perp$ be their orthogonal subspaces respectively. We first prove that for any unit vector $v \in S_1^\perp$, there is another unit vector $v' \in S_2^\perp$ so that $v^T v' \geq 1 - 4n\gamma^2$. In fact, we can take v' to be the unit vector along the projection of v in $S_2^\perp$. To bound the length of the projection, we instead bound the length of projection to S_2. Since we know $v_i^T v' = 0$ for $i \leq k$ and $\|v_i - v_i^*\| \leq \gamma$, it must be that $(v_i^*)^T v' \leq 2\gamma$ when $\gamma < 0.01$. So the projection of v' in S_2 has length at most $2\sqrt{n}\gamma$ and hence the projection of v' in $S_2^\perp$ has length at least $1 - 4n\gamma^2$.

Next, we prove that there is a pair of orthornormal basis $\{\tilde{v}_{k+1}, \tilde{v}_{k+2}, ..., \tilde{v}_n\}$ for $S_1^\perp$ and $\{\tilde{v^*}_{k+1}, \tilde{v^*}_{k+2}, ..., \tilde{v^*}_n\}$ for $S_2^\perp$ such that $\sum_{i=1}^{n-k} w_k \tilde{v}_{k+i}$ is close to $\sum_{i=1}^{n-k} w_k \tilde{v^*}_{k+i}$. Once we have such a pair, we can simultaneously rotate the two basis so that the latter becomes $v_{k+1}^*, ..., v_n^*$.

To get this set of basis we consider the projection operator to $S_2^\perp$ for vectors in $S_1^\perp$. The squared length of the projection is a quadratic form over the vectors in $S_1^\perp$. So there is a symmetric PSD matrix M such that $\left\|\mathrm{Proj}_{S_2^\perp}(v)\right\|_2^2 = v^T M v$ for $v \in S_1^\perp$. Let $\{\tilde{v}_{k+1}, \tilde{v}_{k+2}, ..., \tilde{v}_n\}$ be the eigenvectors of this matrix M. As we showed the eigenvalues must be at least $1 - 8n\gamma^2$. The basis for $S_2^\perp$ will just be unit vectors along directions of projections of $\tilde{v}_i$ to $S_2^\perp$. They must also be orthogonal because the projection operator is linear and

$$\left\|\mathrm{Proj}_{S_2^\perp}(\sum_{i=1}^{n-k} w_i \tilde{v}_{k+i})\right\|_2^2 = \left\|\sum_{i=1}^{n-k} w_i \mathrm{Proj}_{S_2^\perp}(\tilde{v}_{k+i})\right\|_2^2 = \sum_{i=1}^{n-k} \lambda_i w_i^2$$

The second equality cannot hold if these vectors are not orthogonal. And for any w,

$$\left(\sum_{i=1}^{n-k} w_k \tilde{v}_{k+i}\right)^T \left(\sum_{i=1}^{n-k} w_k \tilde{v^*}_{k+i}\right) = \sum_{i=1}^{n-k} w_k^2 (\tilde{v}_{k+i})^T \tilde{v^*}_{k+i} \geq 1 - 8n\gamma^2$$

So we conclude that the distance between these two vectors is at most $3\sqrt{n}\gamma$. ■

Lemma C.2. *Let g^* be the projection of f^* into the space spanned by the rest of local maxima, then $|g^*(w) - g(w)| \leq \delta/8 + \delta'/20 \leq \delta'/8$.*

Proof: The proof is straight forward because $|g^*(w) - g(w)| \leq |f^*(u) - f(u)| + |f^*(u) - f^*(u')|$ for some $\|u - u'\|_2 \leq 3\sqrt{n}\gamma$, we know the first one is at most $\delta/8$ and the second one is at most $\delta'/20$ by Lipschitz Condition. ■

Theorem C.3. *Suppose function $f^*(u) : \mathbb{R}^n \to \mathbb{R}$ satisfies the following properties*

1. *Orthogonal Local Maxima: The function has n local maxima v_i^* and they are orthogonal to each other.*
2. *Locally Improvable: f^* is (γ, β, δ) Locally Improvable.*
3. *Improvable Projection: The projection of the function to any subspace spanned by a subset of local maxima is $(\gamma', \beta', \delta')$ Locally Improvable. The step size $\delta' \geq 10\delta$.*
4. *Lipschitz: If two points $\|u - u'\|_2 \leq 3\sqrt{n}\gamma$, then the function value $|f^*(u) - f^*(u')| \leq \delta'/20$.*
5. *Attraction Radius: Let $Rad \geq 3\sqrt{n}\gamma + \gamma'$, for any local maximum v_i^*, let T be $\min f^*(u)$ for $\|u - v_i^*\|_2 \leq Rad$, then there exist a set U containing $\|u - v_i^*\|_2 \leq 3\sqrt{n}\gamma + \gamma'$ and does not contain any other local optima, such that for every u that is not in U but is β close to U, $f^*(u) < T$.*

If we are given function f such that $|f(u) - f^*(u)| \leq \delta/8$ and f is both (β, δ) and (β', δ') Locally Approximable, then Algorithm 2 can find all local optima of f^* within distance γ.

Proof: By Theorem 4.4 the first column is indeed γ close to a local maximum. We then prove by induction that if v_1, v_2, ..., v_k are γ close to different local maxima, then v_{k+1} must be close to a new local maximum.

By Lemma 4.8 we know g_{k+1} is $(\gamma', \beta', \delta')$ Locally Improvable, and because it is a projection of f its derivatives are also bounded so it is (β', δ') Locally Approximable. By Theorem 4.4 u' must be γ' close to local maximum for the projected function. Then since the projected space is close to the space spanned by the rest of local maxima, u' is in fact $\gamma' + 3\sqrt{n}\gamma$ close to v_{k+1}^* (here again we are reindexing the local maxima wlog.).

Now we use the Attraction Radius property, since u is currently in U, $f^*(u) \geq T$, and each step we go to a point u' such that $\|u' - u\| \leq \beta$ and $f^*(u') > f^*(u) \geq T$. The local search in Algorithm 1 can never go outside U, therefore it must find the local maximum v_{k+1}^*. ■

D Omitted proofs in Section 5

Theorem D.1 ([5]). *When $\beta < d_{min}/10d_{max}n^2$, the function $P^*(u)$ is $(3\sqrt{n}\beta, \beta, P^*(u)\beta^2/100)$ Locally Improvable and $(\beta, d_{min}\beta^2/100n)$ Locally Approximable. Moreover, the local maxima of the function is exactly $\{\pm R_i^*\}$.*

Proof: The proof appears in [5]. Here for completeness we show the proof using our notations.

First we establish that $P^*(u)$ is Locally Improvable.Observe that this desirada is invariant under rotation, so we need only prove the theorem for $P^*(v) = \sum_{i=1}^n d_i v_i^4$. The gradient of the function is $\nabla P^*(v) = 4(d_1 v_1^3, d_2 v_2^3, ..., d_n v_n^3)$. The inner product of $\nabla P^*(v)$ and v is exactly $4\sum_{i=1}^n d_i v_i^4 = 4P^*(v)$. Therefore the projected gradient $\phi = \text{Proj}_{\perp v} \nabla P^*(v)$ has coordinate $\phi_i = 4v_i(d_i v_i^2 - P^*(v))$. Furthermore, the Hessian $H = \mathcal{H}(P^*(v))$ is a diagonal matrix whose $(i,i)^{th}$ entry is $12 d_i v_i^2$.

Consider the case in which $\|\phi\| \geq P^*(v)\beta/4$. We can obtain an improvement to $P^*(v)\beta^2/100$ because we can take ξ in the direction of ϕ and with $\|\xi\|_2 = \beta/20$. The contribution of the Hessian term is nonnegative and the third term $-2P^*(u)\,\|\xi\|_2^2$ is small in comparison.

Hence, we can assume $\|\phi\| \leq P^*(v)\beta/4$. Now let us write out the expression of $\|\phi\|^2$

$$\sum_{i=1}^{n} v_i^2(d_i v_i^2 - P^*(v))^2 \leq \beta^2(P^*(v))^2/16.$$

In particular every term $v_i^2(d_i v_i^2 - P^*(v))^2$ must be at most $\beta^2(P^*(v))^2/16$.. Thus for any i, either $v_i^2 \leq \beta^2$ or $(d_i v_i^2 - P^*(v))^2 \leq (P^*(v))^2/16$.

If there are at least 2 coordinates k and l such that $(d_i v_i^2 - P^*(v))^2 \leq (P^*(v))^2/16$, then we know for these two coordinates $v_i^2 \in [0.75P^*(v)/d_i, 1.25P^*(v)/d_i]$. We choose the vector ξ so that $\xi_k = \tau v_l$ and $\xi_l = -\tau v_k$. Wlog assume $\xi \cdot \phi \geq 0$ otherwise we use $-\xi$. Take τ so that $\tau^2(v_l^2 + v_k^2) = \beta^2$. Clearly $\|\xi\| = \beta$ and $\xi \cdot v = 0$ so ξ is a valid solution. Also τ^2 is lower bounded by $\beta^2/(v_l^2 + v_k^2) \geq \frac{4}{5}\frac{\beta^2}{P^*(u)(1/d_l + 1/d_k)}$.

Consider the function we are optimizing:

$$\begin{aligned}\phi \cdot \xi + 1/2\xi^T \mathcal{H}\xi - 2P^*(u)\,\|\xi\|_2 &\geq 1/2\xi^T H\xi - 2P^*(u)\beta^2 = 6\tau^2 v_k^2 v_l^2(d_k + d_l) - 2P^*(u)\beta^2 \\ &\geq \frac{27}{8}\tau^2 P^*(u)^2\frac{d_k + d_l}{d_k d_l} - 2P^*(u)\beta^2 \geq \frac{7}{10}P^*(u)\beta^2.\end{aligned}$$

In the remaining case, all of the coordinates except for at most one satisfy $v_i^2 \leq \beta^2$. Since we assumed $\beta^2 < \frac{1}{n}$, there must be one of the coordinate v_k that is large, and it is at least $1 - n\beta^2$. Thus the distance of this vector to the local maxima e_k is at most $3\sqrt{n}\beta$. ■

Claim D.2. $Z = O(d_{min}^2 \lambda_{min}(A)^8 \|\Sigma\|_2^4 + d_{min}^2)$.

Proof: We will start by bounding $\mathbf{E}[(z_i z_j z_k z_l)^2] \leq \mathbf{E}[(z_i^8 + z_j^8 + z_k^8 + z_l^8)]$. Furthermore $\mathbf{E}[z_i^8] \leq O(\mathbf{E}[(B^{-1}Ax)_i^8 + (B^{-1}\eta)_i^8])$. Next we bound $\mathbf{E}[(B^{-1}\eta)_i^8]$, which is just the eighth moment of a Gaussian with variance at most $\|B^{-1}\Sigma B^{-T}\|_2 \leq \|B^{-1}\|_2^2 \|\Sigma\|_2 \leq d_{min}^{1/2}\lambda_{min}(A)^{-2}\|\Sigma\|_2$. Hence we can bound this term by$O(\|B^{-1}\Sigma B^{-T}\|_2^4) = O(d_{min}^2 \lambda_{min}(A)^8 \|\Sigma\|_2^4)$. Finally the remaining term $\mathbf{E}[(B^{-1}Ax)_i^8]$ can be bounded by $O(d_{min}^2)$ because the variance of this random variable is only larger if we instead replace x by an n-dimensional standard Gaussian. ■

Lemma D.3. *Given $2N$ samples $y_1, y_2, ..., y_N, y'_1, y'_2, ..., y'_N$, suppose columns of $R' = B^{-1}AD_A(u_0)^{1/2}$ are ϵ close to the corresponding columns of R^*, with high probability the function $\widehat{P}'(u)$ is $O(d_{max}n^{1/2}\epsilon + n^2(N/Z\log n)^{-1/2})$ close to the true function $P^*(u)$.*

Proof: $\widehat{P}'(u)$ is the empirical mean of $F(u, y, y') = -(u^T B^{-1}y)^4 + 3(u^T B^{-1}y)^2(u^T B^{-1}y')^2$. In Section 2 we proved that $P'(u) = \mathbf{E}_{y,y'}\,F(u, y, y') = \sum_{i=1}^{n} 2D_{i,i}^{-1/2}(u^T R_i)^4 = \sum_{i=1}^{n}\lambda_i(u^T R_i)^4$. First, we demonstrate that $P'(u)$ is close to $P^*(u)$, and then using concentration bounds we show that $\widehat{P}'(u)$ is close to $P'(u)$ (with high probability) over all u.

The first part is a simple application of Cauchy-Schwartz:

$$\begin{aligned}|P'(u) - P^*(u)| &= \sum_{i=1}^{n} d_i\left[(u^T R'_i) - (u^T R^*_i)\right] \cdot \left[(u^T R'_i + u^T R^*_i)((u^T R'_i)^2 + (u^T R^*_i)^2)\right] \\ &\leq d_{max}\sqrt{\sum_{i=1}^{n}(u^T(R'_i - R^*_i))^2} \cdot (3\left\|u^T R' + u^T R^*\right\|_2) \leq 6d_{max}n^{1/2}\epsilon.\end{aligned}$$

The first inequality uses the fact that $((u^T R'_i)^2 + (u^T R^*_i)^2) \leq 3$, the second inequality uses the fact that when ϵ is small enough, $\left\|u^T R'\right\|_2 \leq 2$.

Next we prove that the empirical mean $\widehat{P}'(u)$ is close to $P'(u)$. The key point here is we need to prove this for all points u since a priori we have no control over which directions local search will choose to explore. We accomplish this by considering $\widehat{P}'(u)$ as a degree-4 polynomial over u and prove that the coefficient of each monomial in $\widehat{P}'(u)$ is close to the corresponding coefficient in $P'(u)$. This is easy: the expectation of each coefficient of $F(u,y,y')$ is equal to the correct coefficient, and the variance is bounded by $O(Z)$. The coefficients are also sub-Gaussian so by Bernstein's inequality the probability that any coefficient of $\widehat{P}'(u)$ deviates by more than ϵ' (from its expectation) is at most $e^{-\Omega(\epsilon'^2 N/Z)}$. Hence when $N \geq O(Z\log n/\epsilon'^2)$ with high probability all the coefficients of $\widehat{P}'(u)$ and $P'(u)$ are ϵ' close. So for any u:

$$|P'(u) - \widehat{P}'(u)| \leq \epsilon'(\sum_{i=1}^{n} |u_i|)^4 \leq \epsilon' n^2.$$

Therefore $\widehat{P}'(u)$ and $P^*(u)$ are $O(d_{max} n^{1/2}\epsilon + n^2(N/Z\log n)^{-1/2})$ close. ■

This proof can also be used to show that the derivatives of the function $\widehat{P}'(u)$ is concentrated to the derivatives of the true function $P^*(u)$ because the derivatives are only related to coefficients, therefore $\widehat{P}'(u)$ is also $(\beta, d_{min}\beta^2/100n)$ Locally Approximable.

Lemma D.4. *For any $\|u-u'\|_2 \leq r$, $|P^*(u) - P^*(u')| \leq 5d_{max}n^{1/2}r$. All local maxima of P^* has attraction radius $Rad \geq d_{min}/100d_{max}$.*

Proof: The Lipschitz condition follows from the same Cauchy-Schwartz as appeared above. When two points u and u' are of distance r, $|P^*(u) - P^*(u')| \leq 5d_{max}n^{1/2}r$. Finally for the Attraction Radius, we know when $3\sqrt{n}\gamma + \gamma' \leq d_{min}/100d_{max}$, we can just take the set U to be $u^T R_i^* \geq 1 - d_{min}/50d_{max}$. For all u such that $u^T R_i^* \in [1 - d_{min}/25d_{max}, 1 - d_{min}/50d_{max}]$ (which contains the β neighborhood of U), we know the value of $P^*(u) \leq T$. ■

Theorem D.5. *Given a matrix $\widehat{R}$ such that there is permutation matrix Π and $k_i \in \{\pm 1\}$ with $\|\widehat{R}_i - k_i(R^*\Pi)_i\|_2 \leq \gamma$ for all i, Algorithm 3 returns matrix $\widehat{A}$ such that $\|\widehat{A} - A\Pi Diag(k_i)\|_F \leq O(\gamma \|A\|_2^2 n^{3/2}/\lambda_{min}(A))$. If $\gamma \leq O(\epsilon/\|A\|_2^2 n^{3/2}\lambda_{min}(A)) \times \min\{1/\|A\|_2, 1\}$, we also have $\|\widehat{\Sigma} - \Sigma\|_F \leq \epsilon$.*

Proof: By Lemma 2.11 we know the columns of R' is close the the columns of R (the parameters will be set so that the error is much smaller than γ), thus $\|\widehat{R}_i - k_i(R'\Pi)_i\|_2 \leq \gamma$. Applying Lemma 5.3 we obtain: $|\widehat{P}'(\widehat{R}_i) - P^*(\widehat{R}_i)| \ll \gamma$. Furthermore, when $\|\widehat{R}_i - k_i R^*_{\Pi^{-1}(i)}\|_2 \leq \gamma$ we know that $P^*(\widehat{R}_i)/d_{\Pi^{-1}(i)} \in [1-3\gamma, 1+3\gamma]$ (here we are abusing notation and use the permutation matrix as a permutation). Hence $\widehat{D}_A(u)_{i,i}/\,(D_A(u))_{\Pi^{-1}(i),\Pi^{-1}(i)} \in [1-3\gamma, 1+3\gamma]$. We have:

$$\widehat{A}_i = B\widehat{R}_i\widehat{D}_A(u)_{i,i}^{-1/2} \text{ and } (A\Pi\mathrm{Diag}(k_i))_i = BR'_{\Pi^{-1}(i)}\,(D_A(u))^{-1/2}_{\Pi^{-1}(i),\Pi^{-1}(i)}$$

and their difference is at most $O(\gamma\|B\|_2\,(D_A(u))^{-1/2}_{\Pi^{-1}(i),\Pi^{-1}(i)})$. Hence we can bound the total error by $O(\gamma\|B\|_2\left\|D_A(u)^{-1/2}\right\|_F)$. We also know $\|B\|_2 \leq \|A\|_2\|D_A(u)^{1/2}\|_2$ because $BB^T \approx AD_A(u)A^T$, so this can be bounded by $O(\gamma\|A\|_2\|D_A(u)\|_2^{1/2}\|D_A(u)^{-1/2}\|_F)$. Applying Claim 2.9, we conclude that (with high probability) the ratio of the largest to smallest diagonal entry of $D_A(u)$ is at most $n^2\|A\|_2^2/\lambda_{min}(A)^2$. So we can bound the error by $O(\gamma\|A\|_2^2 n^{3/2}/\lambda_{min}(A))$.

Consider the error for Σ: Using concentration bounds similar but much simpler than those used in Lemma 5.3, we obtain that $\|\widehat{C} - C\|_F \leq 1/2\epsilon$, so $\|\widehat{\Sigma} - \Sigma\|_F \leq \|\widehat{C} - C\|_F - \|\widehat{A}\widehat{A}^T - AA^T\|_F \leq \epsilon/2 + 2\|A\|_2\|A\Pi\mathrm{Diag}(k_i) - \widehat{A}\|_F + \|A\Pi\mathrm{Diag}(k_i) - \widehat{A}\|_F^2 \leq \epsilon$. ■