Identifying eigenfunctions of a Markov process using trajectory data

Robert J. Webber†

†Courant Institute of Mathematical Sciences

Motivation

- Eigenfunctions are used for reducing dimensionality and analyzing dynamics in biochemistry.
- Monte Carlo methods are needed to estimate eigenfunctions, but these methods have limited accuracy.
- What is driving the error?
- How can we optimize Monte Carlo's efficiency for the future?

Contributions

– In [4], I proved the convergence of the leading spectral estimation method in biochemistry, called the "variational approach to conformational dynamics" (VAC), and derived detailed error bounds.

– In our follow-up work [2], we extended VAC to make it more robust for applications with limited data and flexible neural network approximation spaces.

VAC algorithm

The variational approach to conformational dynamics (VAC) estimates eigenvalues and eigenfunctions of the transition operator $T_\tau f(x) = \mathrm{E}[f(\boldsymbol{X}_\tau) | \boldsymbol{X}_0 = x]$ for a Markov process $\boldsymbol{X}_\tau$ with a time-reversible distribution μ.

simulation: complete lots of short, independent simulations ($\sim 100{-}1000$) or a few longer simulations ($\sim 1{-}10$).

preparation: choose a set of basis functions ($\sim 10{-}1000$) and estimate expectations $\mathrm{E}_\mu[\phi_i(\boldsymbol{X}_0)\,\phi_j(\boldsymbol{X}_\tau)]$.

spectral estimation: apply VAC to estimate eigenvalues and eigenfunctions.

post-processing: look at top eigenfunctions ($\sim 1-10$) to find meaningful and interpretable patterns.

Specifically:

1. Form matrix $\hat{\boldsymbol{C}}(0)$ with entries $\hat{\boldsymbol{C}}_{ij}(0) \approx \mathrm{E}_\mu[\phi_i(\boldsymbol{X}_0)\,\phi_j(\boldsymbol{X}_0)]$.
2. Form matrix $\hat{\boldsymbol{C}}(\tau)$ with entries $\hat{\boldsymbol{C}}_{ij}(\tau) \approx \mathrm{E}_\mu[\phi_i(\boldsymbol{X}_0)\,\phi_j(\boldsymbol{X}_\tau)]$.
3. Solve eigenvalue problem $\hat{\lambda}_i^\tau \hat{\boldsymbol{C}}(0)\,\hat{\boldsymbol{v}}^i(\tau) = \hat{\boldsymbol{C}}(\tau)\,\hat{\boldsymbol{v}}^i(\tau)$.
4. Return estimated eigenvalues $\hat{\lambda}_i^\tau = e^{-\hat{\sigma}_i^\tau \tau}$ and eigenfunctions $\hat{\gamma}_i^\tau = \sum_j \hat{\boldsymbol{v}}_j^i(\tau)\,\phi_j$.

Problems with VAC

1. No one knows how to choose a lag time.
2. How do we know if VAC is **VACcurate**?

Establishing VAC convergence

I established the convergence of VAC eigenspace estimates under a spectral gap condition.

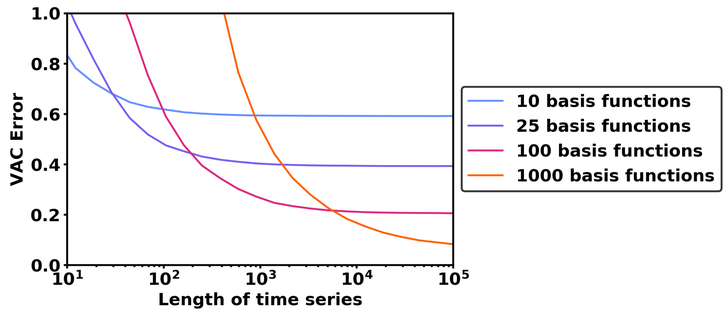

Fig. 1: VAC error vanishes as the number of basis functions increases and the length of the time series increases.

Error bounds for VAC

The analysis of VAC required proving original error bounds.

- Standard bounds for the approximation of eigenspaces (e.g., [3, 1]) depend on the inverse gap between eigenvalues.
- However, the gap between any two non-trivial eigenvalues of the transition operator vanishes exponentially fast with the lag time parameter τ. Therefore, the standard bounds increase exponentially as $\tau \to \infty$.
- In contrast to this asymptotic scaling, I contributed a sharp new perturbation bound that depends only on the inverse *relative* gap between eigenvalues. This new bound reaches its minimal value in the large τ limit, demonstrating the benefit of long lag times for reducing approximation error.

Theorem 1 (Traditional error bound)**.** *Let $\mathcal{H} \succeq 0$ be a Hermitian operator with top eigenvalues $\lambda_0(\mathcal{H}) \geq \cdots \geq \lambda_k(\mathcal{H})$, and let $\mathcal{U}_k$ be the span of the top k eigenfunctions of $\mathcal{H}$. Let $\mathcal{P}_\Phi$ be the orthoprojector onto a subspace Φ. Set $\mathcal{H}' = \mathcal{P}_\Phi \mathcal{H}|_\Phi$, and let $\mathcal{U}_k'$ be the span of the top k eigenfunctions of $\mathcal{H}'$. Then the distance between $\mathcal{U}_k$ and $\mathcal{U}_k'$ is bounded by*

$$1 \leq \frac{d(\mathcal{U}_k, \mathcal{U}_k')^2}{d(\mathcal{U}_k, \Phi)^2} \leq 1 + \frac{1}{4} \frac{\|\mathcal{H}\|_2^2}{|\lambda_k(\mathcal{H}) - \lambda_{k+1}(\mathcal{H}')|^2} \tag{1}$$

Theorem 2 (Improved error bound)**.** *In the setting of Theorem 1, $d(\mathcal{U}_k, \mathcal{U}_k')$ also satisfies*

$$1 \leq \frac{d(\mathcal{U}_k, \mathcal{U}_k')^2}{d(\mathcal{U}_k, \Phi)^2} \leq 1 + \frac{1}{4} \frac{1}{|\lambda_k(\mathcal{H}')/\lambda_{k+1}(\mathcal{H}) - 1|^2}. \tag{2}$$

Lastly, my asymptotic expressions for the estimation error do depend on the inverse spectral gap and grow in the large τ limit. Therefore, it is best to select an intermediate lag time.

New IVAC algorithm

- The VAC algorithm is sensitive to lag time, so we developed a new *integrated VAC* (IVAC) algorithm that incorporates multiple lag times, thereby increasing the robustness.
- IVAC searches for orthonormal eigenfunctions $\gamma_1^\tau, \ldots, \gamma_k^\tau$ that maximize

$$\sum_{i=1}^k \int_{\tau=\tau_{\min}}^{\tau_{\max}} \mathrm{E}[\gamma_i(X_0)\,\gamma_i(X_\tau)]\,d\tau\,.$$

- We can do this either using a linear combination of basis functions (linear IVAC) or a nonlinear combination of basis functions (nonlinear IVAC), such as the output of a neural network.

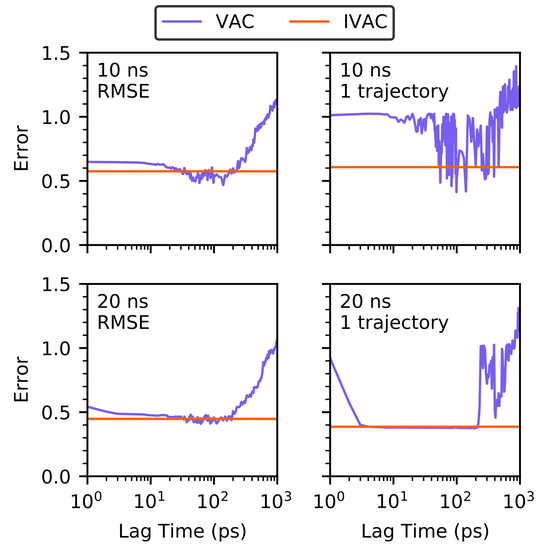

Fig. 2: IVAC performs about as well as VAC with the perfect lag time and much better than VAC with a bad lag time.

References

[1] Andrew Knyazev. "New estimates for Ritz vectors". In: *Mathematics of computation* 66.219 (1997), pp. 985–995.

[2] Chatipat Lorpaiboon et al. "Integrated Variational Approach to Conformational Dynamics: A Robust Strategy for Identifying Eigenfunctions of Dynamical Operators". In: *The Journal of Physical Chemistry B* 124.42 (2020), pp. 9354–9364.

[3] Yousef Saad. *Numerical methods for large eigenvalue problems: Revised edition*. Vol. 66. Siam, 2011.

[4] Robert J Webber et al. "Error bounds for dynamical spectral estimation". In: *SIAM Journal on the Mathematics of Data Science* (2021, to appear).