VARIATIONAL METHODS IN DIMENSIONALITY REDUCTION

Václav Šmídl[1,2]**, Anthony Quinn**[2]

[1] *UTIA, Praha 8, Czech Republic*
[2] *Trinity College Dublin, Ireland*

Email: smidl@utia.cas.cz

Abstract: Reliable tools for reduction of dimensionality are needed for data processing in many areas of science. Well known algorithms like PCA are usually taken as golden standard and used as black box for any kind of problem. However, classical algorithms (like PCA) usually does not provide information about uncertainty of their results thus preventing further investigation of model structure. Proper Bayesian treatment is not feasible. Variational Bayes is a new and interesting approximative method of Bayesian estimation. In this text we summarize approximative Bayesian solution of PCA problem and identify its week spots. A new model is constructed and computationally advantageous method of variational PCA is presented.

Keywords: Dimensionality reduction, Bayesian estimation, Variational Bayes, Principal Component Analysis, Singular Value Decomposition, Von Mises-Fisher distribution

1 INTRODUCTION

Dimensionality reduction is considered as an inevitable subtask of many data processing applications when dealing with multivariate systems. It is usually defined as mapping

$$\mathcal{R}: \mathbf{y} \rightarrow \mathbf{x} \tag{1}$$

where $\mathbf{y}$ denotes space of p-dimensional vectors and x denotes space of r-dimensional vectors. The term *reduction* is used if $r < p$. The term *dimensionality reduction* is used across many areas of science with slightly different definitions and understanding.

In this paper we distinguish two basic reasons why we need to reduce dimensionality of measured variables

Necessity total amount of measured data is too high for further processing. Typical example might be data transfer with pre-specified bitrate. In this case we usually know what amount of data we can afford to process. The task is then reduced to selection of the best transformation function.

Physical nature of the process does not allow us to measure inner variables responsible for system behaviour. We are able to observe only external variables which are believed to be transformed inner variables corrupted by noise. The task is to verify whether measured data contain enough information to determine dimension of inner variables, r, together with optimal transformation.

Notice, that for the first case the question of optimality of chosen transformation can be easily overlooked. In many application the problem specific transformation is used without questioning its relevancy. However, once the issue of optimal transformation is raised the very next question is what kind of information is important and what can be neglected. This naturally leads to signal and noise separation and consequently to probabilistic interpretation of the whole problem

$$\mathbf{y} = \{y, \mathcal{R} : f\left(y|x\right), x \in \mathbf{x}\}$$

where transformation from one space to another is given by probability density function. Probabilistic calculus can be used to answer all above mentioned questions. Even though the solution is formally straightforward it is often infeasible from algorithmic and computational point of view.

That is why many practically used algorithms does not provide fully probabilistic solution but usually point estimates of mean value of involved parameters. This approach is acceptable in the first described case when we are interested in limited number of values. However, in the second case this approach reach its limits as it is not able to provide estimates of number of lower dimensions. This task falls into category of model structure estimation in which Bayesian methods has proven the most successful one.

However, exact Bayesian solution is not feasible. There are various methods that provide approximative solution, namely Laplace method (R. E. Kass, 1994), Monte Carlo Markov Chain methods and Variational approach. In this paper we will discuss application of Variational methods.

2 PROBLEM FORMULATION

Probabilistic interpretation of dimensionality reduction problem was studied in context of latent variable models (Everit, 1984), where unknown low dimensional variable x is the latent one. The data generating model is defined

$$y = g\left(x, \theta\right) + e \tag{2}$$

where y denotes observed variable, x latent variable, g transformation function parameterised by θ and e denotes additive noise. The model is fully determined is we specify transformation function g, and probability distribution of noise $f\left(e\right)$. Data generation model can be rewritten

$$f\left(y|\theta, x\right). \tag{3}$$

In order to obtain distribution of latent variable x conditioned by measured data we can use Bayes rule

$$f\left(x|y\right) \propto \int_{\theta^*} f\left(y|x,\theta\right) f_0\left(x|\theta\right) f_0\left(\theta\right) d\theta \tag{4}$$

where f_0 denotes prior distribution. As we will see later analytical solution of this integral is far from trivial even for very simple linear Gaussian model.

2.1 Linear Gaussian Model

Linear Gaussian dimensionality reduction is probably the first model ever studied. Its roots goes back to 19th century. It represents a special case of the model (2) where transformation function $g(x, A)$ is linear combination and additive noise is chosen homogenous Gaussian with zero mean and variance σ. The data generating models is then

$$\begin{aligned} f\left(y|x, A\right) &= \mathcal{N}\left(Ax, \sigma I_p\right) \text{ for one realization} \\ f\left(Y|X, A\right) &= \mathcal{N}\left(AX, \sigma I_p \otimes I_n\right) \text{ for } n \text{ realizations} \end{aligned} \tag{5}$$

where the compact notation of multiple realizations $Y = [y_1, y_2, \ldots, y_n]$, $X = [x_1, x_2, \ldots, x_n]$ is possible due to properties of Gaussian noise. Matrix notation simplifies readability of many following results.

This model is widely used as it is possible to show (Bishop, 1998), that its maximum likelihood estimates correspond with well known Principal Component Analysis (PCA) (Jolliffe, 1986). As it was mentioned in the introduction, maximum likelihood estimates does not provide estimation of structure, e.g. number of dimensions r.

Exact Bayesian inference (4) for this model was derived (Press and Shigemasu, 1989). The resulting distribution is a complicated one with surprising moments. Mean values for both A and X are zero. This finding is (probably) closely related to principal indeterminacy of basic model

$$D = AX + E$$

which yields the same result for any invertible matrix T of appropriate dimensions

$$AX = ATT^{-1}X. \tag{6}$$

If you realize that AX is reached with the same probability as $(-A)(-X)$ the zero mean of estimates is not so surprising. More practical approach would require solution of the the inference with restrictions on positivity of involved matrices which would be far too complicated.

2.2 *Other models*

There are various models published in the literature that follows the basic scenario (2) in the literature. They can be classified by transformation function, distribution of the noise and prior information on parameters. These models are widely used in neural computation, artificial inteligence, biology etc. However, Bayesian analysis of these models is not available.

Another famous dimensionality reduction algorithm is Factor analysis. It is the whole branch of science by itself, however its basic algorithm can be interpreted as maximum likelihood estimate of the linear model

$$f(y|x, A) = \mathcal{N}(Ax, R) \tag{7}$$

with diagonal covariance matrix R. It looks like a small difference compared to identity matrix in PCA model however it has serious consequences on estimation algorithm.

3 VARIATIONAL BAYES

Variational approximation is based on application of Jensen's inequality to the logarithm of the likelihood function, where D denotes measured realizations of random variable y

$$\ln f(D) = \ln \int_{\theta^*, x^*} f(D, x, \theta) d\theta dx \tag{8}$$

$$= \ln \int_{\theta^* x^*} \frac{q(\theta, x)}{q(\theta, x)} f(D, \theta, x) d\theta dx \geq \int_{\theta^* x^*} q(\theta, x) \ln \frac{f(D, \theta, x)}{q(\theta, x)} d\theta dx = \mathcal{L}(D). \tag{9}$$

We see that the function $\mathcal{L}(D)$ forms a rigorous lower bound on the true logarithm of the likelihood function. If $q(\theta)$ is pdf the error of approximation is given by

$$\mathrm{KL}\left(q(\theta, x) \,||\, f(\theta, x|D)\right) = -\int_{\theta^*} q(\theta, x) \ln \frac{f(\theta, x|D)}{q(\theta, x)} d\theta \tag{10}$$

which is the Kullback-Leibler distance (Kullback and Leibler, 1951).

Note that equation (8) is valid for any pdf $q\left(\theta, x\right)$. The choice of approximation pdf q is then subject of design aim. For a choice $q\left(\theta, x\right) = f\left(\theta, x\right)$ KL distance is equal to zero, however it does not solve intractability of integral (4). If we want to find a tractable solution we have to restrict $q\left(\theta, x\right)$ to a space of tractable functions. This can be done basically in two ways:

1. direct specification of parametric form of $q\left(\theta, x\right) = Q\left(\theta, x|\phi\right)$ and minimisation of KL distance with respect to parameters ϕ. Hence, parameters
$$\hat{\phi} = \arg\max_{\phi} \mathrm{KL}\left(Q\left(\theta, x|\phi\right)||f\left(\theta, x|D\right)\right)$$
determine the best approximation $Q\left(\theta, x|\hat{\phi}\right)$of $f\left(\theta, x|D\right)$ in the space of Q.
2. non parametric restriction of functional $q\left(\theta, x\right)$. The solution has to be found by functional minimisation of KL distance. The solution is more complicated, however it provides even the form of pdf.

The later approach is adopted as a basis for method called Variational Bayes.

Theorem 3.1 (Variational Bayes) *Let $f\left(\theta, x, D\right)$ be a joint pdf of data matrix D and variables θ, x and $q\left(\theta, x\right)$ be approximative pdf restricted to independent distribution of variables θ, x*

$$q\left(\theta, x\right) = q_\theta\left(\theta\right) q_x\left(x\right).$$

Then, minimum of KL distance $KL\left(q\left(\theta, x\right)||f\left(\theta, x, D\right)\right)$ is reached for

$$q_\theta\left(\theta\right) \propto \exp\left(\mathcal{E}_{q_x}\left\{\ln\left(f\left(\theta, x, D\right)\right)\right\}\right), \; q_x\left(x\right) \propto \exp\left(\mathcal{E}_{q_\theta}\left\{\ln\left(f\left(\theta, x, D\right)\right)\right\}\right)$$

where $\mathcal{E}_q$ denotes expected value with respect to distribution q.

The above mentioned theorem is quite a powerful tool for estimation of latent variables. Notice that it provide even the form of resulting pdf. However, moments of the resulting distribution are mutually dependent. Consequences are illustrated on the following special case.

3.1 Variational PCA

Variational approach was applied to the PCA model (5) by Bishop (Bishop, 1999). The model has to be extended by selection of prior distribution on parameters and latent variables

$$f_0\left(X\right) = \mathcal{N}\left(0, I_p \otimes I_n\right), \;\; f_0\left(A\right) = \mathcal{N}\left(0, \mathrm{diag}\left(\alpha\right) \otimes I_n\right), \;\; f_0\left(\sigma\right) = \Gamma\left(c_0, d_0\right)$$

where α denotes vector of hyperparameters $\alpha_i, i = 1 \ldots p-1$ with priors $f\left(\alpha_i\right) = \Gamma\left(a_0, b_0\right)$.

Application of theorem 3.1 gives following results

$$f\left(X\right) = \mathcal{N}\left(M_X, \Sigma_X \otimes I_{(N\times N)}\right) \qquad f\left(A\right) = \mathcal{N}\left(M_A, I_{(p\times p)} \otimes \Sigma_A\right) \tag{11}$$
$$f\left(\sigma\right) = \Gamma\left(c, d\right) \qquad f\left(\alpha\right) = \prod_{i=1}^{R} \Gamma\left(a, b_i\right)$$

with moments given by set of implicit equations

$$\begin{aligned} M_X &= \frac{c}{d}\Sigma_X' M_A' D & \Sigma_X &= \left(I_r + \frac{c}{d}\left(\Sigma_A + M_A' M_A\right)\right)^{-1} \\ M_A &= \frac{c}{d}\Sigma_A M_X D' & \Sigma_A &= \left(\operatorname{diag}\left(\frac{a_i}{b_i}\right) + \frac{c}{d}\left(\Sigma_X + M_X' M_X\right)\right)^{-1} \\ a &= a_0 + 0.5p & b &= \operatorname{diag}\left(b + \left(\Sigma_A + M_A' M_A\right)\right) \\ c &= c_0 + 0.5np & d &= d_0 + \frac{1}{2}\operatorname{trace}\left(DD'\right) - 2\operatorname{trace}\left(D' M_A M_X\right) + \\ & & &+\operatorname{trace}\left(\left(\Sigma_A + M_A' M_A\right)\left(\Sigma_X + M_X' M_X\right)\right). \end{aligned} \tag{12}$$

Number of latent dimensions is determined by number of α_i that remains higher than zero after all iterations. Technically, as there is a prior on α we need to set up some threshold above prior value a_0/b_0.

The set of implicit equation has to be solved iteratively which implies usual issues of starting point, stopping rule and convergency. Starting point in the original implementation was chosen randomly, stopping rule by some small threshold on increment of parameter estimates. Convergency of algorithm is quite stable but it is quite sensitive to initial guess of variance σ. If it is chosen too high algorithm does not detect all underlying dimensions, if too low convergency of algorithm is very slow.

Overall, it is possible to say that with proper tuning of priors and thresholds algorithm is usable. However, random starts produce different results. It does not provide guidelines for selection of initial starting point and various tuning knobs. If we apply this algorithm on more complicated models (see section 3.3) problems with selection of starting point and tuning knobs are growing. Moreover, speed of convergency is significantly decreasing.

We believe that these problems are closely related to principal indeterminacy of the model as described earlier. Thus we decided to address these problems by choice of unique data generating model.

3.2 *Orthogonal Variational PCA*

This modification of Variational PCA is based on reformulation of data generating model in a unique way. It is common to model lower rank matrices by means of singular value decomposition (svd) algorithm:

$$D = ALX + E \tag{13}$$

which is unique if $A'A = I_r$, $XX' = I_r$ and L is a diagonal matrix $L = \operatorname{diag}\left(l_1, \ldots, l_r\right)$ of positive numbers. Notice that some indeterminacy is still present (signs of matrices A and X), however there is an unwritten agreement that numerical implementations of svd algorithm returns matrices with first column positive, which makes decomposition truly unique.

Solution follows methodology of variational estimates described in section 3. However, we have to keep in mind restrictions of orthogonality of matrices A, X. These live on Stiefel manifold with finite volume and thus we can choose prior distribution to be uniform. Normalising coefficient is known (Khatri and Mardia, 1977), but it does not influence further evaluations and thus we do not state it explicitely.

Prior information for remaining parameters is chosen as

$$l_i \sim \mathcal{N}\left(0, \lambda^{-1}\right) \qquad \sigma \sim \Gamma\left(a, b\right)$$

where parameters λ, a, b can be chosen as very small values or totally neglected. Non-informative prior on those variables provides almost identical results.

Direct application of theorem 3.1 gives us following results

$$\begin{array}{ll} A \sim \mathcal{M}\left(\overline{LX}D'\overline{\sigma}\right) & l \sim \mathcal{N}\left(\frac{\overline{\sigma}}{\overline{\sigma}+\alpha}\mathrm{diag}\left(\overline{X}D'\overline{A}\right), \left(\overline{\sigma}+\alpha\right)^{-1} I_r\right) \\ X \sim \mathcal{M}\left(D'\overline{AL}\overline{\sigma}\right) & \sigma \sim \Gamma\left(\frac{pN}{2}, \frac{1}{2}\mathrm{tr}\left[D'D - \overline{X}'\overline{LA}'D - D'\overline{ALX} + \left(\overline{LL} + \Sigma_L\right)\right]\right). \end{array} \tag{14}$$

We can see that minimisation of KL distance results in a bit exotic distribution, $\mathcal{M}$, known as von Mises-Fisher.

Moments for variables l and σ are quite easy to evaluate, $\overline{\sigma}$ denotes mean value of variable σ (same for other variables) and Σ_L denotes second moment of L

$$\begin{aligned} \overline{L} &= \frac{\overline{\sigma}}{\overline{\sigma}+\alpha}\mathrm{diag}\left(\overline{X}D'\overline{A}\right) \\ \Sigma_L &= \left(\overline{\sigma}+\alpha\right)^{-1} I_r \\ \overline{\sigma} &= \frac{pN}{\mathrm{tr}\left[D'D - \overline{X}'\overline{LA}'D - D'\overline{ALX} + \left(\overline{LL} + \Sigma_L\right)\right]}. \end{aligned} \tag{15}$$

However, moments of von Mises-Fisher distribution are more complicated. Fortunately, we need only the fist one, mean value, which is known (Downs, 1972; Khatri and Mardia, 1977) but hard to evaluate.

Theorem 3.2 (Moments of von Mises-Fisher distribution) *Let $X \sim \mathcal{M}\left(F\right)$ be distributed as von Mises-Fisher with given parameter F, and $F = USV$ be* svd *decomposition of matrix F. Then mean value of X,*

$$\overline{X} = U\mathbf{D}V \tag{16}$$

where $\mathbf{D}$ is diagonal matrix with elements

$$\mathbf{d}_i = \frac{\partial}{\partial d_i}\log\left({}_0F_1\left(\frac{1}{2}p, \frac{1}{4}S^2\right)\right) \tag{17}$$

where ${}_0F_1$ denotes generalised hypergeometric function (James, 1964).

This theorem has very useful consequences for the whole iteration algorithm. Natural starting point of the iterative algorithm is svd decomposition of data matrix

$$D = U_0 S_0 V_0 \tag{18}$$

which determines $\overline{A}_0 = U_0$, $\overline{L}_0 = S_0$, $\overline{X}_0 = V_0$, reasonable guess for σ_0 is the lowest (non-zero) singular value. First iteration yields

$$\begin{aligned} \overline{A}_1 &= D\overline{X}_0'\overline{L}_0\overline{\sigma}_0 = U_0S_0V_0V_0'S_0\overline{\sigma}_0 = U_0\wr\left(S_0^2\overline{\sigma}_0\right) \\ \overline{X}_1 &= \overline{\sigma}_0\overline{L}_0\overline{A}_0'D = \overline{\sigma}_0S_0U_0U_0S_0V_0 = \wr\left(S_0^2\overline{\sigma}_0\right)V_0 \end{aligned}$$

where $o\left(S\right)$ is operation defined by (17). Following the iteration steps we can see that mean value of matrices preserves the form $\overline{A}_n = U_0S_A$, $\overline{X}_n = S_XV_0$ where S_A, S_X are diagonal matrices of size $p \times p$. Thus all we have to do in each iteration step is to perform operation (17) for matrices S_A and S_X and operations (15). Which brings significant speedup in comparison with multiplication of matrices A and X in each step which is required for original version.

The most problematic point remains operation (17) because definition of hypergeometric function of matric argument is quite complicated and exact evaluation is far from trivial (James, 1964). Fortunately, Khatri and Mardia (Khatri and Mardia, 1977) provides approximative formulas which are suitable for preliminary evaluation. Further study on this issue is very needed.

3.3 Other models

The application of variational theorem to more complex models is straightforward and that is why it has already been published for mixtures of factor analysers (Ghahramani and Beal, 1999) and independent component analysis (ICA) (Miskin, 2000). Those works are pioneering however still a lot of work remains to be done, especially with respect to speed of convergence and guidelines for selection of initial conditions of iterative algorithm.

Above described orthogonal reformulation of model is surely possible for model known as factor analysis (7). However, application of variational theorem gives A being distributed as generalised Bingham distribution (Khatri and Mardia, 1977) which moments are not published in available literature. Further research in this area may bring very interesting and widely useful algorithms.

4 RESULTS

Performance of both methods was tested on both synthetic and real life data. Synthetic data were generated by the model (5) with values $n = 4000$, $r = 3$, $p = 60$. With various levels of noise variance. Comparison of typical run of both methods is listed in the following table

	Simulated data		Real life data	
	VPCA	OVPCA	VPCA	OVPCA
estimated dimensionality, r	3	3	123	113
CPU time consumed	290s	3s	4800s	18s

As we can see both methods estimated correctly underlying dimensionality of latent variables. In general our experience based on many experiments is that OVPCA has tendency to underestimate number of latent variables while VPCA even though it sometimes provides different results (random initialisation) is usually correct. The behaviour of OVPCA can be explained by used approximation which is valid for high values of singular values thus eliminating dimensions with lower evidence in data.

Significant differences in CPU time consumption can be partially explained by improper choice of initial conditions of VPCA. It is possible to tune initial guess of σ to achieve better performance. However, it would be necessary to do for each realization of the noise. In this respect deterministic initial conditions of OVPCA are of high interest. The main computational burden of OVPCA lies in svd decomposition of data matrix. The iterative algorithm is very fast because it uses very simple approximation of hypergeometric function. It gives us hope that even more sophisticated approximation would be still computationally feasible.

Real life data were taken from Prague traffic light control system. We have collected 2300 data records from 173 sensors. Correct dimensionality of data is naturally not known. Unbiased comparison of both methods is again complicated by the fact that VPCA is started randomly and provides slightly different results for each start. Moreover, number of selected components is influenced by selected stopping rule as well.

Figure 1 shows typical evolution of parameters determining latent dimensionality. It is parameter L for OVPCA and α for VPCA. As we can see evolution of L is very fast but evolution of α is singificantly slower and it seems to slide to higher values. This behaviour is, however, allowed by the original model (5) increasing values of A are compensated by decreasing values of X and their mutual product remains constant. Chosen prior distributions are designed to deal with this problem but they are not strong enough to keep it steady.

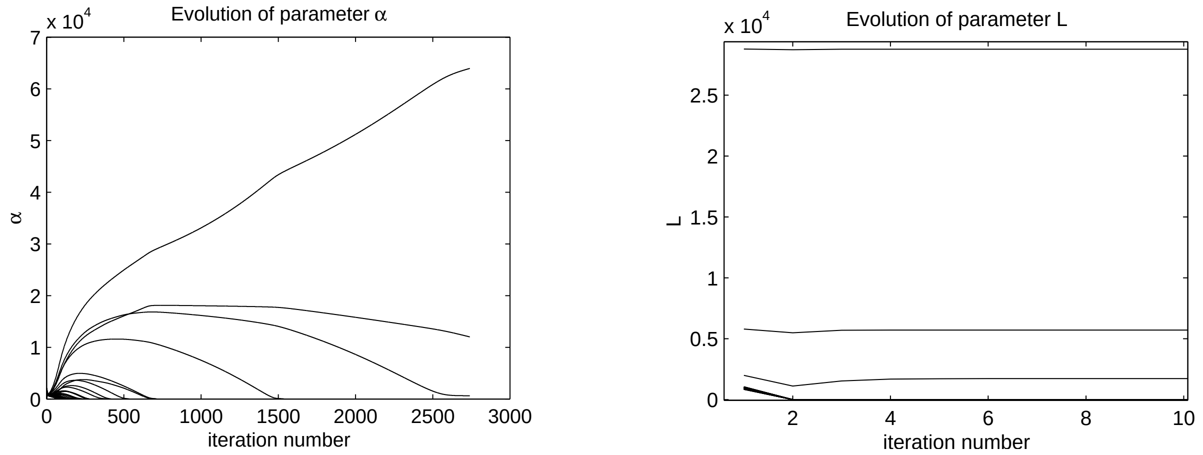

Figure 1: Comparison of evolution of hyperparameters α of VPCA (left) and parameters L of OVPCA (right). These parameters determines estimated dimensionality of latent variable in each method and thus allow as to compare its efficiency.

5 CONCLUSION

In this paper we have shown that variational methods represent very interesting tool for estimation of latent variable models. Drawbacks that prevented use of these method for use in large dimensions were identified as missing guidelines for selection of initial conditions and tuning knobs and speed of convergence.

We have shown that these illnesses can be overcome by reformulation of the model at least for the simplest case, known as PCA model. Reformulation of the model for more complex models is possible. However, numerical evaluation of moments of resulting distributions is far form trivial and further research is needed for successful realization.

REFERENCES

Bishop, C. M. (1998), ‘Bayesian PCA’, *Neural Information Processing Systems* **11**, 382–388.

Bishop, C. M. (1999), Variational principal components, *in* ‘Proceedings of Ninth International Conference on Artificial Neural Networks’.

Downs, T. D. (1972), ‘Orientational statistics’, *Biometrica* **59**, 665–676.

Everit, B. V. (1984), *An Introduction to Latent Variable Models*, Monographs on Statistics and Applied Probability, Chapman & Hall, London, New York.

Ghahramani, Z. and Beal, M. J. (1999), ‘Variational inference for bayesian mixtures of factor analyzers’, *Neural Information Processing Systems* **12**.

James, A. T. (1964), ‘Distribution of matrix variates and latent roots derived from normal samples’, *Annals of Mathematical Statistics* **35**, 475–501.

Jolliffe, I. (1986), *Principal Component Analysis*, Springer Verlag.

Khatri, C. G. and Mardia, K. V. (1977), ‘The von Mises-Fisher distribution in orientation statistics’, *Journal of Royal Statistical Society B* **39**, 95–106.

Kullback, S. and Leibler, R. (1951), ‘On information and sufficiency’, *Annals of Mathematical Statistics* **22**, 79–87.

Miskin, J. W. (2000), Ensemble Learning for Independent Component Analysis, PhD thesis, University of Cambridge.

Press, S. J. and Shigemasu, K. (1989), *Contributions to Probability and Statistics*, Springer Verlag, New York, chapter Bayesian Inference in Factor Analysis, chapter 15.

R. E. Kass, A. E. R. (1994), Bayes factors and model uncertainty, Technical report, University of Washington.