Probabilistic Expert Knowledge Elicitation of Feature Relevances in Sparse Linear Regression

Pedram Daee* **Tomi Peltola*** **Marta Soare*** **Samuel Kaski**
Helsinki Institute for Information Technology HIIT,
Department of Computer Science, Aalto University
firstname.lastname@aalto.fi
*Authors contributed equally.

1 Introduction

In this extended abstract, we consider the "small n, large p" prediction problem, where the number of available samples n is much smaller compared to the number of covariates p. This challenging setting is common for multiple applications, such as precision medicine, where obtaining additional samples can be extremely costly or even impossible. Extensive research effort has recently been dedicated to finding principled solutions for accurate prediction. However, a valuable source of additional information, domain experts, has not yet been efficiently exploited.

We propose to integrate expert knowledge as an additional source of information in high-dimensional sparse linear regression. We assume that the expert has knowledge on the relevance of the features in the regression and formulate the knowledge elicitation as a sequential probabilistic inference process with the aim of improving predictions. We introduce a strategy that uses Bayesian experimental design [2] to sequentially identify the most informative features on which to query the expert knowledge. The evaluation of our method in simulation experiments shows improved prediction accuracy already with a small effort from the expert.

By interactively eliciting and incorporating expert knowledge, our approach fits into the interactive learning literature [1, 10]. The ultimate goal is to make the interaction as effortless as possible for the expert. This is achieved by identifying the most informative features on which to query expert feedback and asking about them first, similarly to active learning strategies [12], where the most informative additional samples are identified.

2 Method

We introduce a probabilistic model that subsumes both a sparse regression model which predicts external targets, and a model for encoding expert knowledge. We then present a method to query expert knowledge sequentially (one feature at a time), with the aim of getting fast improvement in the predictive accuracy of the regression with a small number of queries.

For the regression, a Gaussian observation model with a spike-and-slab sparsity-inducing prior [4] on the regression coefficients is used:

$$\begin{aligned} \boldsymbol{y} &\sim \mathrm{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \mathbf{I}), && (1)\\ w_j &\sim \gamma_j \,\mathrm{N}(0, \psi^2) + (1-\gamma_j)\delta_0, && j = 1, \ldots, p,\\ \gamma_j &\sim \mathrm{Bernoulli}(\rho), && j = 1, \ldots, p, \end{aligned}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ are the output values and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the matrix of covariate values. The regression coefficients are denoted by $w_1, \ldots, w_p$, and σ^2 is the residual variance. The γ_j indicate inclusion ($\gamma_j = 1$) or exclusion ($\gamma_j = 0$) of the covariates in the regression (δ_0 is a point mass at zero). The prior expected sparsity is controlled by ρ.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

The expert knowledge on the relevance of the features for the regression is encoded by a feedback model:

$$f_j \sim \gamma_j \,\mathrm{Bernoulli}(\pi) + (1-\gamma_j)\,\mathrm{Bernoulli}(1-\pi), \tag{2}$$

where $f_j = 1$ indicates that feature j is *relevant* and $f_j = 0$ *not-relevant*, and π is the probability that the expert feedback is correct relative to the state of the covariate inclusion indicator γ_j.

As the number of covariates p can be large, we assume that it is infeasible, or at least unnecessarily burdensome, to ask the expert about each feature. Instead, we aim to ask first about the features that are estimated to be the most informative given the (small) training data, and frame this problem as a Bayesian experimental design task [2, 11].

We prioritize the features based on their expected information gain for the predictive distribution of the regression. As the expert is queried for the feedbacks sequentially, the posterior distribution of the model and the prioritization is recomputed after each feedback in order to use the latest knowledge. At iteration t for feature j, the expected information gain is

$$\mathrm{E}_{p(\tilde{f}_j|\mathcal{D}_t)}\left[\sum_i \mathrm{KL}[p(\tilde{y}|\mathcal{D}_t, \boldsymbol{x}_i, \tilde{f}_j) \parallel p(\tilde{y}|\mathcal{D}_t, \boldsymbol{x}_i)]\right],$$

where $\mathcal{D}_t = \{(y_i, x_i) : i = 1, \ldots, n\} \cup \{f_{j_1}, \ldots, f_{j_{t-1}}\}$ denotes the training data together with the feedback that has been given at previous iterations and $p(\tilde{f}_j|\mathcal{D}_t)$ is the posterior predictive distribution of the feedback for the jth feature. The summation over i goes over the training dataset. We assume that each feature will be queried about only once (or not at all if the iterations are terminated before reaching p).

This query scheme goes beyond pure prior elicitation [3, 7, 9] as the training data is used to facilitate an efficient expert knowledge elicitation. This is a crucial aspect that enables the elicitation in high-dimensional regression. The Bayesian probabilistic framework provides a natural way to sequentially update the inferences.

The probabilistic model does not have a closed form posterior distribution or solution to the information gain maximization problem. To achieve fast computation, important for possible real-time interactive knowledge elicitation, we use expectation propagation [8] to deterministically approximate the posterior distribution and the required quantities for the expected information gain [5, 6, 11].

3 Experiments

We evaluate the performance of the proposed method in a "small n, large p" regression problem with synthetic data.

Setting. The covariates of n training data points are generated from $\boldsymbol{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$. Out of the p regression coefficients $w_1, \ldots, w_p \in \mathbb{R}$, p^* are generated from $w_j \sim \mathrm{N}(0, \psi^2)$ and the rest are set to zero. The output values are generated from $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2\,\mathbf{I})$. The expert has noisy knowledge about non-relevant/relevant features (Eq. 2 with $\gamma_j = 1$ if w_j is non-zero, and $\gamma_j = 0$ otherwise, and $\pi = 0.95$). For a generated set of training data, the expert feedback is queried one feature at a time. Mean squared error (MSE) is used as the performance measure to evaluate query strategies. We use the known data-generating values for the fixed hyperparameters: $\sigma^2 = 1$, $\psi^2 = 1$, and $\rho = p^*/p$.

We compare four query strategies:

- random feature suggestions *(green line, triangle up)*,
- an "oracle" strategy that knows the relevant features beforehand and queries the expert about them first, and then chooses at random from the features not already selected *(red line, triangle down)*[1],
- our sequential experimental design strategy (Sect. 2) *(blue line, squares)*,
- a non-sequential version of our strategy, which chooses the sequence of features to be queried before observing any expert feedback *(magenta line, circles)*.

Results. In Fig. 1, we consider a "small n, large p" scenario, with $n = 10,\ p = 100,\ p^* = 10$ and report the average MSE value over 500 runs (repetitions of the data generation). The results in the

[1] Although unrealistic, this "oracle" strategy allows to see the performance gain obtainable by an intuitively good strategy which first queries experts about the relevant features.

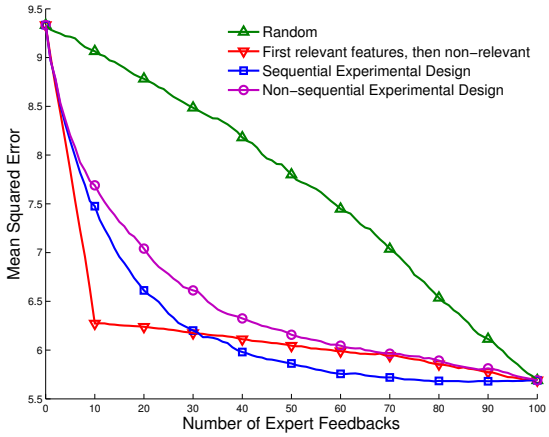

Figure 1: Mean squared errors for the four query strategies. The number of relevant features $p^* = 10$ out of $p = 100$ features and the number of training data points $n = 10$. Note that the red strategy is not available in practice.

plot are shown for an increasing number of feedbacks, up to the number of dimensions where all strategies converge. For the "oracle" strategy, the reduction in the MSE is fast in the first feedback iterations where it queries the expert on the 10 relevant features only. This implies that querying about these features is important in this setting. The experimental design strategies are able to identify these features reasonably well and their performances are very close to just asking about relevant features in the first interactions. After 30 feedback, our method reaches the same MSE level as the oracle, and performs better thereon until the last step, indicating that even the order of asking feedback on the non-relevant features affects the speed of MSE reduction.

With regard to the realistic scenario of a limited number of feedbacks, both experimental design strategies have a faster increase in the prediction accuracy in the first iterations compared to the random strategy. This implies that both experimental design strategies are able to identify and ask with priority about more informative features. Compared to the non-sequential selection strategy that does not take into account the observed expert feedback, the more carefully selected sequence of queries done by the sequential experimental design strategy reduces the prediction error faster, indicating that the accumulated expert feedback affects the next query.

4 Conclusions

We presented an expert knowledge elicitation approach for high-dimensional sparse linear regression. The results for a "small n, large p" problem in simulated data, with expert knowledge on the relevance of features, showed improved prediction accuracy already with a small number of expert feedbacks. Compared to pure prior elicitation, the approach can be used in knowledge elicitation for high-dimensional parameters without overwhelming the expert.

Acknowledgements

This work was financially supported by the Academy of Finland (Finnish Center of Excellence in Computational Inference Research COIN; grants 295503, 294238, 292334, and 284642), Re:Know funded by TEKES, and MindSee (FP7–ICT; Grant Agreement no 611570).

References

[1] Amershi, S. (2012). *Designing for Effective End-User Interaction with Machine Learning*. PhD thesis, University of Washington.

[2] Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304.

[3] Garthwaite, P. H. and Dickey, J. M. (1988). Quantifying expert opinion in linear regression problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 462–474.

[4] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

[5] Hernández-Lobato, D., Hernández-Lobato, J. M., and Dupont, P. (2013). Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14(1):1891–1945.

[6] Hernández-Lobato, J. M., Hernández-Lobato, D., and Suárez, A. (2015). Expectation propagation in linear regression models with spike-and-slab priors. *Machine Learning*, 99(3):437–487.

[7] Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., and Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, 75(372):845–854.

[8] Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 362–369.

[9] O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements. Eliciting Experts' Probabilisties*. Wiley, Chichester, England.

[10] Porter, R., Theiler, J., and Hush, D. (2013). Interactive machine learning in data exploitation. *Computing in Science & Engineering*, 15(5):12–20.

[11] Seeger, M. W. (2008). Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813.

[12] Settles, B. (2010). Active learning literature survey. Computer sciences technical report 1648, University of Wisconsin, Madison.