A review of parameter learning methods based on approximate versions of the method of moments

Richard E. Turner
Computational and Biological Learning Lab
Engineering Department, Trumpington Street, Cambridge

September 25, 2010

Abstract

The method of moments is an old statistical technique that fits parameters of a model by matching the average value of moments under the model with their average value over the data. Recently, approximate versions of this basic technique have been proposed in several diverse literatures including statistics, econometrics, population genetics, systems biology, ecology, epidemiology, and image texture modelling. This review connects these different approaches, addresses some key themes, and suggests several directions for future work.

1 The method of moments

In this section we provide a brief introduction to the parameter learning scheme called the method of moments. Many of the subsequent methods considered in this review can be seen as approximate versions of this scheme.

The goal of the method of moments is to estimate the parameters, $\boldsymbol{\theta}$, of a probabilistic model, $p(\mathbf{y}|\boldsymbol{\theta})$, from training data, $\mathrm{Y} = \{\mathbf{y}_n\}_{n=1}^N$. The basic idea is compute some empirical moments on the training data, $\Phi_i(\mathrm{Y}) = \frac{1}{N}\sum_{n=1}^N f_i(\mathbf{y}_n)$, and then to alter the parameters so that the expected moments under the model,

$$\langle f_i(\mathbf{y})\rangle_{p(\mathbf{y}|\boldsymbol{\theta})} = \int d\mathbf{y}\; f_i(\mathbf{y})p(\mathbf{y}|\boldsymbol{\theta}), \tag{1}$$

are identical to the empirical moments, that is $\Phi_i(\mathrm{Y}) = \langle f_i(\mathbf{y})\rangle_{p(\mathbf{y}|\boldsymbol{\theta}^{\mathrm{MM}})}$.

Typically the functions $f_i(\mathbf{y})$ are chosen so that their expectation under the model can be computed analytically, in which case learning reduces to solving a set of coupled and possibly non-linear equations. For example, consider learning the parameters of an Inverse Gamma distribution,

$$p(\mathrm{y}|\boldsymbol{\theta}) = \mathrm{InvGamma}(\mathrm{y};\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\mathrm{y}^{-\alpha+1}\exp(-\beta/\mathrm{y}), \tag{2}$$

by matching the first and second moments of the data,

$$\Phi_1(Y) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{y}_n = \langle \mathrm{y} \rangle_{p(\mathrm{y}|\boldsymbol{\theta}^{\mathrm{MM}})} = \frac{\beta^{\mathrm{MM}}}{\alpha^{\mathrm{MM}} - 1}, \tag{3}$$

$$\Phi_2(Y) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{y}_n^2 = \langle \mathrm{y}^2 \rangle_{p(\mathrm{y}|\boldsymbol{\theta}^{\mathrm{MM}})} = \frac{\left(\beta^{\mathrm{MM}}\right)^2}{(\alpha^{\mathrm{MM}} - 1)(\alpha^{\mathrm{MM}} - 2)}. \tag{4}$$

These equations are solved when, $\alpha^{\mathrm{MM}} = (2\Phi_2(Y) - \Phi_1^2(Y))/(\Phi_2(Y) - \Phi_1^2(Y))$ and $\beta^{\mathrm{MM}} = \Phi_1(Y)\Phi_2(Y)/(\Phi_2(Y) - \Phi_1^2(Y))$.

In the special case when the model is in the exponential family, *and* the empirical moments are the associated sufficient statistics,

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{i=1}^{I} \theta_i \Phi_i(\mathbf{y})\right), \tag{5}$$

the method of moments is identical to maximum-likelihood learning,

$$\boldsymbol{\theta}^{\mathrm{MM}} = \boldsymbol{\theta}^{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log p(\mathbf{y}_n|\boldsymbol{\theta}). \tag{6}$$

However, if these conditions are not met, then moment matching is not equivalent to maximum likelihood. In such cases the estimators derived from the method of moments are not guaranteed to be either consistent or efficient. Worse still, the parameter estimates may lie outside the domain of the parameter space. Returning to the example above, although the Inverse Gamma distribution is in the exponential family, the second moment is not a sufficient statistic. So, whilst the estimators derived above are consistent, they are not efficient. The method does have one advantage over maximum-likelihood learning in this particular case, which is that the estimators have a closed form solution for the inverse gamma distribution.

The method of moments was extensively applied to simple probabilistic models, but it was then superseded by maximum-likelihood parameter estimators because of their superior theoretical properties. However, recently there has been a resurgence of interest in this area especially for estimating the parameters of models for which maximum-likelihood is intractable.

2 The Generalised Method of Moments

The method of moments provides a recipe for estimating the parameters of simple models. For more complex models several complications arise. For instance, there might not be an analytical solution to the moment equations, or the model might not be able to contort itself to match the observed moments (as is often the case when there are more moments to be matched than there are parameters in the model). The Generalised Method of Moments (GMM) [1, 2] addresses these difficulties.

The approach is to specify a cost-function, $\Delta(\mathrm{Y}, \boldsymbol{\theta})$, which penalises large differences between the data-statistics and model-statistics, $g_i(\mathrm{Y}, \boldsymbol{\theta}) = \frac{1}{N}\sum_n f_i(\mathbf{y}_n) - \langle f_i(\mathbf{y}) \rangle_{p(\mathbf{y}|\boldsymbol{\theta})}$. The generalised method of moments uses a cost function which is

a quadratic form,

$$\boldsymbol{\theta}^{\text{GMM}} = \arg\min_{\boldsymbol{\theta}} \Delta(\mathrm{Y}, \boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \sum_{i,j} g_i(\mathrm{Y}, \boldsymbol{\theta}) W_{i,j} g_j(\mathrm{Y}, \boldsymbol{\theta}), \qquad (7)$$

where W is positive definite. In general, this is a non-linear optimisation problem. However, if a unique solution exists then the estimator $\boldsymbol{\theta}^{\text{GMM}}$ can be shown to be consistent and asymptotically normal. Moreover, if the matrix,

$$W^{-1} = \langle \boldsymbol{f}(\mathbf{y}) \boldsymbol{f}(\mathbf{y})^T \rangle_{p(\mathbf{y}|\theta^{\text{GMM}})} - \langle \boldsymbol{f}(\mathbf{y}) \rangle_{p(\mathbf{y}|\theta^{\text{GMM}})} \langle \boldsymbol{f}(\mathbf{y})^T \rangle_{p(\mathbf{y}|\theta^{\text{GMM}})}, \qquad (8)$$

then the estimator can also shown to be efficient. The intuition is that moments with less variance are more informative and should therefore be given more weight. However, this result does not immediately lead to a practical algorithm as it requires *a priori* knowledge of the solution. An alternative is use,

$$W^{-1} = \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{f}(\mathbf{y}_n) - \langle \boldsymbol{f}(\mathbf{y}) \rangle_{p(\mathbf{y}|\boldsymbol{\theta})} \right) \left(\boldsymbol{f}(\mathbf{y}_n) - \langle \boldsymbol{f}(\mathbf{y}) \rangle_{p(\mathbf{y}|\boldsymbol{\theta})} \right)^T, \qquad (9)$$

either updating W to this value at each iteration (the 2-step feasible GMM), or substituting this form into the cost-function and minimising with respect to the model parameters (the Continuously Updated GMM). However, for models of high dimensional data which have a large number of parameters it is possible that the additional computational cost associated with these modifications (and the possibly more difficult optimisation problem) outweigh any potential improvement in efficiency. One direction for future research is in low-rank versions of these schemes which may retain the advantages, but decrease the computational complexity.

3 Approximating the model moments using sampling

The method of moments and its generalisation both require the user to choose statistics to match whose expectation can be analytically computed under the model, $\langle f(\mathbf{y}) \rangle_{p(\mathbf{y}|\boldsymbol{\theta})} = \int d\boldsymbol{y} f(\boldsymbol{y}) p(\mathbf{y}|\boldsymbol{\theta})$. This can be a restrictive assumption, for example, if either the model or the statistics are non-linear. However, for models from which it is efficient to sample data $\mathbf{y}_m \sim p(\mathbf{y}|\boldsymbol{\theta})$, like causal generative models, the moments can be simply approximated by averaging, $\langle f(\mathbf{y}) \rangle_{p(\mathbf{y}|\boldsymbol{\theta})} \approx \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{y}_m)$.

A variant on the generalised method of moments that uses this general idea is described by Wood for a time-series model from the ecology literature [3]. Wood shows that the log-likelihood for this model is rough, meaning that it is plagued by local optima, and so an alternative parameter learning scheme is proposed. The new method works as follows: First, as with the generalised method of moments, the starting point is a set of statistics which are computed from the data $\Phi(\mathrm{Y}) = \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{y}_n)$. These statistics are then treated like 'new data'. A new Gaussian likelihood function is defined for the model parameters, by sampling N synthetic data-sets from the model, $\mathrm{Y}_n \sim p(\mathrm{Y}|\theta)$, computing the statistics on these data-sets $\Phi_n = \Phi(\mathrm{Y}_n)$, and computing the mean and covariance of these statistics, $\mu(\theta) = \frac{1}{N} \sum \Phi(\mathrm{Y}_n)$, $\Sigma(\theta) = \frac{1}{N} \sum_{n=1}^{N} \Phi(\mathrm{Y}_n) \Phi^T(\mathrm{Y}_n) - \mu(\theta) \mu(\theta)^T$. Sampling

is required because the expectations cannot be computed analytically. The new log-likelihood is then given by,

$$\log p(\Phi(\mathrm{Y})|\theta) = -\frac{1}{2}\log\det\Sigma(\theta) - \frac{1}{2}(\Phi(\mathrm{Y}) - \mu(\theta))^T\Sigma^{-1}(\theta)(\Phi(\mathrm{Y}) - \mu(\theta)). \tag{10}$$

This likelihood is demonstrated to be better behaved (less rough and less local optima) than the normal likelihood for their application. However, local optima mean that a simple Metropolis-Hastings method was used to optimise the above expression, rather than simple gradient ascent.

The above objective function is clearly similar to that used in the Continuously Updated Generalised Method of Moments, the difference being the extra term arising from the determinant of the covariance of the statistics. In general, the additional computational complexity arising from a) estimating the covariance of the statistics under the model, and b) the more complicated objective function, probably makes this method less preferable than a vanilla GGM approach in which the model moments are computed by sampling.

One possible extension to the model for situations applicable where there is a large amount of training data, would be to estimate the mean and covariance of the statistics on both the synthetic *and* the real data. The KL-divergence between these two distributions could then be used as an objective function. This will not affect the computational complexity greatly.

4 Approximate Bayesian Computation

The aim of the moment matching approaches described above was to return a point estimate of the model parameters. However, the goal of the Bayesian approach to parameter learning is to return the posterior distribution over the parameters of the model, given the data,

$$p(\theta|\mathrm{Y}) = \frac{1}{p(\mathrm{Y})}p(\mathrm{Y}|\theta)p(\theta). \tag{11}$$

Approximate Bayesian Computation (ABC) [4, 5, 6, 7] is the name given to a family of sampling methods that use method of moment ideas to *approximately* sample from the posterior distribution.

ABC was originally designed for models where computation of the likelihood is intractable, but for which it is possible to sample data given the model parameters, $\mathrm{Y} \sim p(\mathrm{Y}|\theta)$.

ABC starts, like the other moment matching methods, by choosing a vector of statistics to compute from the data $\Phi(\mathrm{Y})$ and specifying a distance measure, or cost function, between two vectors of statistics, $\Delta(\Phi(\mathrm{Y}), \Phi(\mathrm{Y}'))$. The ABC algorithm then proceeds as follows,

1. Sample the parameter from the prior $\theta' \sim p(\theta)$
2. Sample a data-set from the model $\mathrm{Y}' \sim p(\mathrm{Y}|\theta')$
3. Measure the distance between statistics computed from the observed and generated data-sets, and accept θ' if $\Delta(\Phi(\mathrm{Y}), \Phi(\mathrm{Y}')) \leq \epsilon$
4. Return to step 1

The stationary distribution of this scheme is $p(\theta|\Delta(\Phi,\Phi') \leq \epsilon)$. In general it is conjectured that ABC gives back less compact distributions than the true posterior distribution, but little theoretical analysis has been possible.

For models of modest complexity, several improvements are necessary to make ABC a practical algorithm. For example, sampling from the prior will be disastrous because the volume of parameter space that is likely to give rise to statistics close to the observed statistics is typically small, resulting in many rejections. One solution is to use information about the accepted parameter values, for instance, by sampling in a neighbourhood around these values. A Metropolis Hastings scheme is one possibility,

1. Start at an initial parameter value θ
2. Propose a sample θ' from $q(\theta'|\theta)$
3. Sample a data-set from the model $\mathrm{Y}' \sim p(\mathrm{Y}|\theta')$
4. If $\Delta(\Phi(\mathrm{Y}),\Phi(\mathrm{Y}')) \leq \epsilon$ then go to step 4, otherwise return to step 1
5. Calculate $h = h(\theta,\theta') = \min\left(1, \frac{p(\theta')q(\theta|\theta')}{p(\theta)q(\theta'|\theta)}\right)$
6. Move to θ' with probability h, else stay at θ.
7. Return to step 1

Additionally, it is usual for to anneal ϵ in ABC.

There are versions of ABC where the hard rejection of parameter values which do not give rise to statistics close enough to the data statistics, is softened. One way to do this is to represent posterior samples by a weighted average of samples, where higher weights represent better satisfaction of the statistics. This reduces to the hard thresholding case when the weights are 0/1.

5 Sampling

The aim of the methods described previously was to estimate the parameters of a model. Data can then be generated from the models by sampling. An alternative to this two step scheme for generating data, which does not require an explicit parametric model, is to iteratively shape white noise until its statistics match the statistics of the observed data. This approach has been successfully applied to generate naturalistic image and sound textures [8, 9].

In more detail, the approach first involves computing a selection of statistics on the training data, $\Phi_i(\mathrm{Y}) = \frac{1}{N}\sum_{n=1}^{N} f_i(\mathbf{y}_n)$. The synthetic data is initialised as white noise $\mathrm{Y}' \sim \mathrm{Norm}(0,1)$ and it is then shaped by minimising a cost function that penalises differences between the statistics of the training data and the statistics on the white noise. For example, a quadratic cost is often employed,

$$\arg\min_{Y'} \sum_{i=1}^{I} \gamma_i (\Phi_i(\mathrm{Y}) - \Phi_i(\mathrm{Y}'))^2. \tag{12}$$

6 Conclusions

The key issues that arise in all of the approaches above are,

- How to choose the moments (analytic form, minimum number, maximally informative, prune uninformative moments).
- How to scale the sensitivity to each moment in the cost function (e.g. down-weight statistics in proportion to their variability/informativeness)
- How to optimise the cost-function (which method, methods to handle weights which depend on model parameters)
- How to approximate complex moments of complex models (sampling versus deterministic approximations)
- How to extend the method of moments beyond point estimation of the parameters (ABC is one method, but is there a more principled approach?)

References

[1] Lars P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[2] R. Hall, Alastair. *Generalized Method of Moments*. Oxford University Press, 2004.

[3] Simon N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, August 2010.

[4] W. Zhang M. A. Beaumont and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.

[5] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15324–15328, 2003.

[6] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1760–1765, 2007.

[7] Mark A. Beaumont, Jean michel Marin, and Jean marie Cornuet. Adaptivity for abc algorithms: the abc-pmc. arXiv, 2008.

[8] J Portilla and E P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40:49–71, 2000.

[9] J.H. McDermott, A.J. Oxenham, and E. Simoncelli. Sound texture synthesis via filter statistics. In *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk NY*, 2009.