Bayesian Estimation and Comparison of Moment Condition Models

Dimitri Meunier
ENSAE & ENS Paris Saclay
dimitri.meunier@ensae.fr

Clement Guillo
ENSAE & ENS Paris Saclay
clement.guillo@ensae.fr

Abstract

This report is based on the article "Bayesian Estimation and Comparison of Moment Condition Model" by Siddharta Chib and Anna Simoni [3]. First we will detail the main ideas for simulation from a non parametric posterior density with only moment conditions on the model. Secondly, we will replicate the regression experiment of the paper. And finally, we apply the method for model selection. On the way, we will demonstrate that the asympotic normality is satisfied in all our experiments. The main challenge was the implementation that requires numerous computational components. Since no Python code was available we implemented our own functions and our codes can be found here: https://github.com/DimSum2k20/BETEL.

Introduction: Moment Condition Models

We say that a statistical model is a moment condition model if the model is specified only through moment restrictions: $\mathbb{E}^P[g(X,\theta)] = 0$. P is the unknown distribution of the $\mathbb{R}^d$-valued random vector X, $g : \mathbb{R}^{d_X} \times \Theta \mapsto \mathbb{R}^d$ is a nonlinear function with values in $\mathbb{R}^d$ and $\theta = (\theta_1, .., \theta_p) \in \Theta \subset \mathbb{R}^p$.

If $d > p$, i.e. if the setting is over-identified by the moment conditions, it is possible that no $\theta \in \Theta$ verifies these moment conditions. Thus, for some of these conditions we can relax the constraint by introducing a vector $V \in \mathbb{R}^d$ such as $V_k \neq 0$ if we accept that the moment condition k can be inactive. Thus, the moment condition becomes:

$$\mathbb{E}^P[g^A(X,\theta,V)] = 0 \qquad g^A(X,\theta,V) := g(X,\theta) - V$$

This type of situation is very often found when estimating the coefficients of linear regression models such as :

$$y_i = \alpha + \beta z_i + e_i \qquad i = 1 \dots n \tag{1}$$

where (z_i, e_i) are independently drawn from some distribution P. In this case the moment conditions can be:

$$\mathbb{E}^P[e_i(\theta)] = 0 \qquad \mathbb{E}^P[e_i(\theta) z_i] = 0 \qquad \mathbb{E}^P[(e_i(\theta))^3] = v \qquad e_i(\theta) := y_i - \alpha - \beta z_i$$

The first two moments are the standard orthogonality conditions and if v is set to 0 the third condition mean that the distribution of e_i is symmetric. Since the symmetry is not necessarely true, treating v as a free parameter allow the model to stay correctly specified under asymmetry.

We will see in the following section that in a Bayesian framework it is possible to obtain a semi-parametric posterior of Moment Condition Models.

1 Bayesian Exponentially Tilted Empirical Likelihood

There exists several ways to estimate a Moment Condition Model and we will focus on the ones based on the Empirical Likelihood (EL). The main advantage of empirical likelihood is that it enables inference that does not require distributional assumptions. The EL re-weight the samples so that is satisfies the moment conditions exactly, while maximising the likelihood function of a multinomial supported on the sample ($x_{1:n} = (x_1, \cdots, x_n)$):

$$\begin{aligned} EL(\theta) = \max_{p_1,\cdots,p_n} \quad & \sum_{i=1}^{n} p_i \\ \text{subject to} \quad & \sum_{i=1}^{n} p_i = 1 \qquad \sum_{i=1}^{n} p_i g(x_i, \theta) = 0 \end{aligned}$$

[4] has shown that it is possible to derive a posterior from the Empirical Likelihood. It is thus possible to perform semi-parametric Bayesian inference. They start with the basic setup of parametric Bayesian inference with a noisy parameter u that provides additional degrees of freedom in case of mis-specification. Then, they derive a non parametric version of this procedure by letting the dimension of u to go to infinity and using the fact that any distribution can be approximated by a mixture of uniform distributions (if the number of components can grow). The main result of [4] is the proof that the asymptotic formulation of the non parametric formulation can be re-written as a concave maximisation problem. Under a given prior π on (θ, v), the posterior distribution has the following form:

$$\pi(\theta, v|x_{1:n}) \propto \pi(\theta, v) p(x_{1:n}|\theta, v)$$

$p(x_{1:n}|\theta, v)$ is the Exponentially Tilted Empirical Likelihood (ETEL) function defined as:

$$p(x_{1:n}|\theta, v) = \prod_{i=1}^{n} p_i^*(\theta, v)$$

Where the p_i^* are solutions of the following concave optimisation program:

$$\begin{aligned} \underset{p_1,\cdots,p_n}{\text{maximize}} \quad & \sum_{i=1}^{n} -p_i \log np_i = f_0(p) \\ \text{subject to} \quad & \sum_{i=1}^{n} p_i = 1 \qquad \sum_{i=1}^{n} p_i g^A(x_i, \theta, v) = 0_{\mathbb{R}^d} \end{aligned}$$

We refer to the project done on the paper [4] for more details and proof as in this project we focus on [3]. We are lucky since this is a well defined entropy maximisation problem and using duality theory (see [1] for a comprehensive introduction), we can show that the p_i^* can be obtained by solving an unconstrained problem involving the dual variables. Thus, one can use any solver to quickly solve for p_i^*. We now show how to solve the primal from the dual.
We introduce the matrix $A \in \mathbb{R}^{d \times n}$ such that $(A)_{i,j} = g_i^A(x_j, \theta, v)$. The problem can be rewritten as,

$$\begin{aligned} \underset{p_1,\cdots,p_n}{\text{maximize}} \quad & f_0(p) \\ \text{subject to} \quad & p^T \mathbb{1}_n = 1 \qquad Ap = 0_{\mathbb{R}^d} \end{aligned}$$

The Lagrangian function is, $\mathcal{L}(p, \lambda, \alpha) = f_0(p) + \lambda^T(Ap) + \alpha(p^T \mathbb{1}_d - 1)$ and the dual function is therefore,

$$\begin{aligned} g(\lambda, \alpha) &= \sup_p \big(f_0(p) + \lambda^T(Ap) + \alpha(p^T \mathbb{1}_n - 1)\big) \\ &= -\alpha + \sup_p \big(f_0(p) + (A^T\lambda + \alpha \mathbb{1}_d)^T p\big) \\ &= -\alpha + f^*(A^T\lambda + \alpha \mathbb{1}_d) \end{aligned}$$

Where f^* is the Fenchel transformation of $-f_0$ (see [1], section 3.3). It a simple exercise to show that $f^*(x) = \frac{1}{n}\sum_{i=1}^{n} e^{x-1}$, thus,

$$g(\lambda, \alpha) = -\alpha + \frac{1}{n}\sum_{i=1}^{n} e^{A_{i:}^T \lambda} e^{\alpha - 1}$$

The dual problem is therefore $\inf g(\lambda, \alpha)$. Let's solve it to find the dual variables $(\hat{\lambda}, \hat{\alpha})$. For a fixed λ, g is convex in α, thus $\hat{\alpha}(\lambda)$ is solution of,

$$\nabla_\alpha g(\lambda, \alpha) = 0 \implies -1 + \frac{1}{n}\sum_{i=1}^{n} e^{A_{i:}^T\lambda} e^{\alpha - 1} = 0 \implies \hat{\alpha}(\lambda) = -\log\big(\frac{1}{n}\sum_{i=1}^{n} e^{A_{i:}^T\lambda}\big) + 1$$

By injecting $\hat{\alpha}(\lambda)$ in the dual problem, we find that $\hat{\lambda}$ is solution of,

$$\hat{\lambda} = \underset{\lambda \in \mathbb{R}^d}{\text{argmin}} \quad \frac{1}{n}\sum_{i=1}^{n} e^{A_{i:}^T\lambda}$$

Then, using KKT conditions, we know that the first order conditions must be satisfied: $\nabla_p L(p, \hat{\lambda}, \hat{\alpha}) = 0$,

$$\nabla_p L(p, \hat{\lambda}, \hat{\alpha})_i = 0 \implies -\log(np_i) - 1 + \hat{\lambda}_T A_i + \hat{\alpha} = 0 \implies p_i = \frac{1}{n} e^{\hat{\lambda}_T A_i + \hat{\alpha} - 1} = \frac{e^{\hat{\lambda}(\theta,v)^T g^A(x_i,\theta,v)}}{\sum_{j=1}^{n} e^{\hat{\lambda}(\theta,v)^T g^A(x_j,\theta,v)}}$$

Finally, the posterior distribution has the following form:

$$\pi(\theta, v | x_{1:n}) \propto \pi(\theta, v) \prod_{i=1}^{n} \frac{e^{\hat{\lambda}(\theta,v)^T g^A(x_i,\theta,v)}}{\sum_{j=1}^{n} e^{\hat{\lambda}(\theta,v)^T g^A(x_j,\theta,v)}}$$

When the specification of the considered model is correct and under certain assumptions of regularity, the posterior converges in total variation to a normal law. Let's note that any evaluation of the posterior requires to solve an optimisation problem.

2 Generation of parameters

To sample parameters θ and v from the posterior distribution, the article suggests to use a Markov Chain Monte Carlo method (MCMC). More precisely a version of the Metropolis Hasting algorithm is used as follow:

First we chose an initial value (θ_0, v_0) and we choose a proposition law, $q(\theta, v|x_{1:n})$, then at each step t ($t \in 1, ..T$) we perform the following:

1. generate a proposal (θ_*, v_*) from q
2. calculate $\alpha((\theta_*, v_*), (\theta_t, v_t)|x_{1:n}) = \min\left(1, \frac{\pi(\theta_*, v_*|x_{1:n})}{\pi(\theta_t, v_t|x_{1:n})} \frac{q(\theta_t, v_t|x_{1:n})}{q(\theta_*, v_*|x_{1:n})}\right)$
3. Set $(\theta_{t+1}, v_{t+1}) = (\theta_*, v_*)$ with probability $\alpha((\theta_*, v_*), (\theta_t, v_t)|x_{1:n})$

Since this algorithm creates a Markov Chain with a stationary distribution equal to $\pi(.,.)$, after several iterations, the generated parameters follow the posterior distribution. It is common with MCMC to burn the first iterations (burn-in phase) before keeping the samples.

We can make the following remarks. First, we see that it is not necessary to know the multiplicative constant of the posterior. This is common to all MCMC methods. Secondly, each iteration of the algorithm requires to solve the maximization program described in the previous part which can be very costly in computation time. Last but not least the choice of q is quite important in the BETEL framework as we will see in the next section.

3 Application I: Regression Model Parameter Estimation

3.1 Data

To replicate the results contained in the paper, we generated two samples of 250 and 2500 observations from the regression model (1) with $z_i \sim \mathcal{N}(0.5, 1)$, $\alpha = 0$, $\beta = 1$ and :

$$e_i \sim \begin{cases} \mathcal{N}(0.75, 0.75^2) & \text{with probability } 0.5 \\ \mathcal{N}(-0.75, 1.25^2) & \text{with probability } 0.5 \end{cases}$$

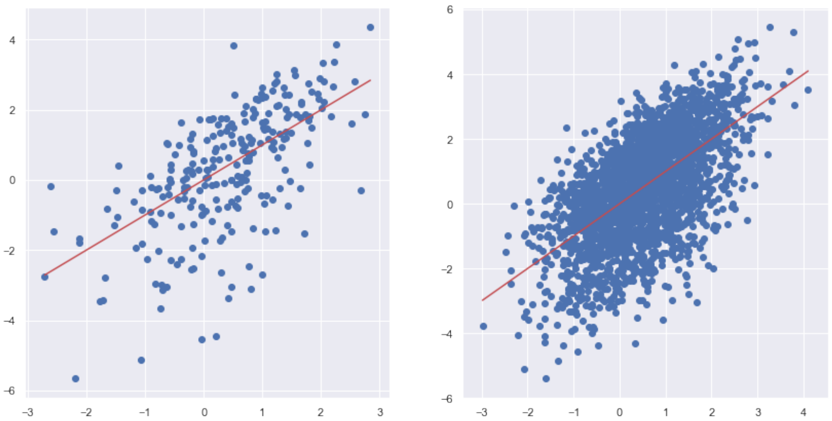

Figure 1: Observations and true regression line with n=250 (left) and n=2500 (right)

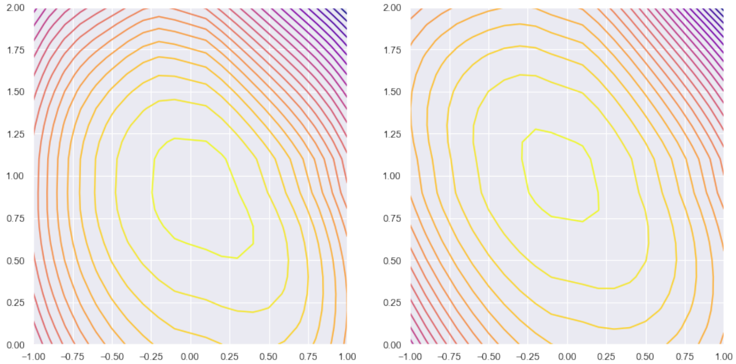

Figure 2: Contour plots of the logETEL function for n=250 and n=2500 in the coordinates (α, β) after performing a grid evaluation

Even if the model is simple, as mentioned in the previous section, the application of the method is not straight forward and requires many computational components such as convex optimisation (with the simplex method) and MCMC. In the next section we detail the choice for q, it will require the use of gradient ascent, gradient approximations and hessian approximations.

3.2 Proposal function

The proposition function for Metropolis Hasting is important as a poor choice will make the Metropolis Hasting algorithm rejects all the samples. The article mentions that an appropriate choice for q is a Student distribution with the mode of the logETEL function $(\log \pi(\theta, v|x_{1:n}))$ for the location parameter and the negative inverse of the logETEL at the mode for the scaling parameter.

Finding the mode is not easy as we can only evaluate $\pi(\theta, v|x_{1:n})$ point by point by solving a convex problem. A first idea is to use a grid search to find the mode. This approach takes a really long take and scale badly with the dimension of the parameters. It is also not useful in practice as we need to use our knowledge of the true parameters to center the grid, otherwise it is almost impossible to find the mode.

The only advantage of the grid search is that it allows to obtain a contour plot of the logETEL function. Asymptotically the ETEL function is a Gaussian Distribution, thus, the logETEL function should have level curves that looks like ellipsoïds. Figure 2 shows the contour plot for 250 and 2500 observations in the coordinates (α, β). The ellipsoïds are more precise with $n = 2500$.

The fact that $\pi(\theta, v|x_{1:n})$ is asymptotically a Gaussian distribution indicates that it should be unimodal and easy to maximise with gradient ascent. We start form an initialisation parameter x_0 and

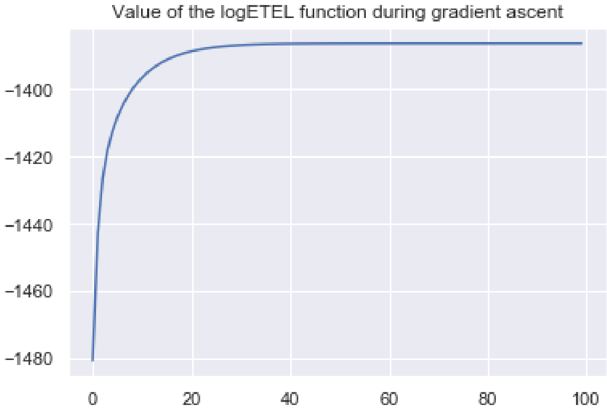

Figure 3: Approximated Gradient ascent in the regression model to find the mode of the LogETEL function

we perform the following gradient updates,

$$x_{t+1} = x_t + \lambda \nabla f(x_t)$$

where f is $\pi(\theta, v|x_{1:n})$ and λ is the step. Since we do not have an analytic expression for ∇f we need to approximate it with differences,

$$\nabla f(x)_i \sim \frac{f(x + he_i) - f(x)}{h}$$

This approach is more robust as it converges even if we start from an initial point far from the true value of the parameters. We therefore don't need any information about the true parameters of the model. The normality of $\pi(\theta, v|x_{1:n})$ makes the gradient ascent quite simple, Figure 3 shows that the convergence towards the mode necessitates less than 50 iterations which is way faster than a grid search and takes less than 30 seconds on a standard CPU.

Once we have the mode, we need to obtain the Hessian of the logETEL at the mode. Since we do not have access to the Gradient nor the Hessian we make the following second order approximation:

$$\tilde{H}_{i,j}(x) = \frac{f(x + h_1 e_i + h_2 e_j) - f(x + h_1 e_i) - f(x + h_2 e_j) + f(x)}{h_1 \times h_2}$$

As this approximation might not be symmetric, we symmetries it with

$$\hat{H} = \frac{\tilde{H} + \tilde{H}^T}{2}$$

We then take the negative inverse of this quantity for the scaling parameter of q. We chose 3 for the degrees of freedom.

3.3 Results

we generate 10000 values of our parameters via the metropolis hasting algorithm (the first 1000 observations are burnt). We do this for a sample size of $n = 250$ and $n = 2500$.

The graph 4 shows the confidence ellipse described by our alpha-beta generations for a fixed value of v. This empirically confirms the normality of the previously computed posterior distribution. Moreover, we can notice that this ellipse becomes more refined from 250 to 2500 observations.

The table shows the statistics obtained on our generations. Here again, the distribution observed for $n = 2500$ is tighter around the parameters.

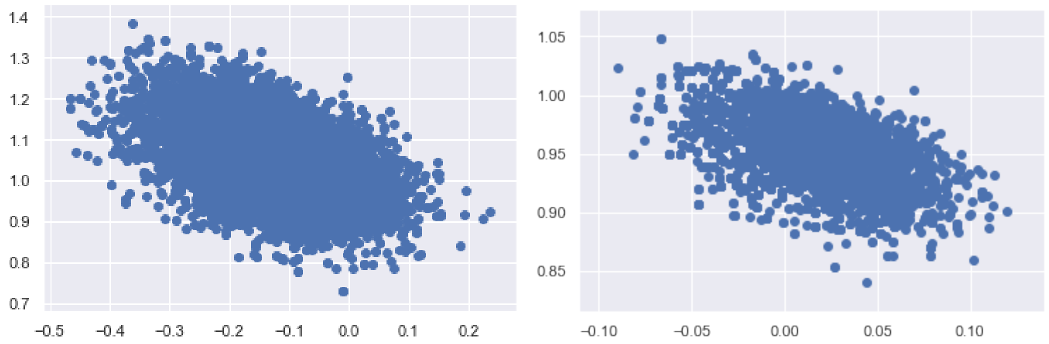

Figure 4: generations of α, β obtained for a fixed v

	mean	sd	median	lower	upper
			$n = 250$		
α	-0.13	0.09	-0.13	-0.32	0.05
β	1.06	0.08	1.05	0.89	1.22
v	-1.93	0.44	-1.9	-2.87	-1.17
			$n = 2500$		
α	0.01	0.03	0.01	-0.05	0.07
β	0.95	0.03	0.95	0.9	1
v	-1.07	0.12	-1.06	-1.3	-0.86

Table 1: Statistics obtained on Metropolis-Hasting generations

4 Application II: Model Selection

Another application of the BETEL framework is model selection. It can select the best model from a countable set of L models. This is made possible by comparing the marginal densities associated with each model.

$$m(x_{1:n}; M_l) = \int_{\Psi} p(x_{1:n}|\psi, M_l)\pi(\psi, M_l)d\psi$$

with $\psi = (\theta, v)$

For any point $\tilde{\psi}$ in the support of the posterior we can write :

$$\pi(\tilde{\psi}|x_{1:n}, M_l) = \frac{p(x_{1:n}|\tilde{\psi}, M_l)\pi(\tilde{\psi}, M_l)}{m(x_{1:n}; M_l)}$$

and thus

$$\log(m(x_{1:n}; M_l)) = \log(p(x_{1:n}|\tilde{\psi}, M_l)) + \log(\pi(\tilde{\psi}, M_l)) - \log(\pi(\tilde{\psi}|x_{1:n}, M_l))$$

The two first terms of the right hand side of this last inequalities are already known, and the last term can be calculate as [2]:

$$\pi(\tilde{\psi}|x_{1:n}, M_l) = \frac{E_1[\alpha(\psi, \tilde{\psi}|x_{1:n}, M_l)q(\tilde{\psi}|x_{1:n}, M_l)]}{E_2[\alpha(\tilde{\psi}, \psi|x_{1:n}, M_l)]}$$

with E_1 the expectation w.r.t $\pi(\psi|x_{1:n})$ and E_2 the expectation w.r.t $q(\psi|x_{1:n}, M_l)$. Indeed, the M-H is built such as the Markov chain created is reversible, *i.e*

$$\pi(\psi|x_{1:n}, M_l)\alpha(\psi, \tilde{\psi}|x_{1:n}, M_l)q(\tilde{\psi}|x_{1:n}, M_l) = \pi(\tilde{\psi}|x_{1:n}, M_l)\alpha(\tilde{\psi}, \psi|x_{1:n}, M_l)q(\psi|x_{1:n}, M_l)$$

Then:

$$\begin{aligned} E_1[\alpha(\psi, \tilde{\psi}|x_{1:n}, M_l)q(\tilde{\psi}|x_{1:n}, M_l)] &= \int \alpha(\psi, \tilde{\psi}|x_{1:n}, M_l)q(\tilde{\psi}|x_{1:n}, M_l)\pi(\psi|x_{1:n})d\psi \\ &= \int \alpha(\tilde{\psi}, \psi|x_{1:n}, M_l)q(\psi|x_{1:n}, M_l)\pi(\tilde{\psi}|x_{1:n})d\psi \\ &= \pi(\tilde{\psi}|x_{1:n})E_2[\alpha(\tilde{\psi}, \psi|x_{1:n}, M_l)] \end{aligned}$$

In order to test the model selection we want to choose between this two models :

$$\begin{aligned}&\text{M1 : } \mathbb{E}^P[e_i(\theta)] = 0 \qquad \mathbb{E}^P[e_i(\theta)z_i] = 0\\ &\text{M2 : } \mathbb{E}^P[e_i(\theta)] = 0 \qquad \mathbb{E}^P[e_i(\theta)z_i] = 0 \qquad \mathbb{E}^P[(e_i(\theta))^3] = 0\end{aligned}$$

However, it is not possible to systematically compare two models. In order to be able to compare two models, the convex Hulls associated with the two models must have the same dimensions. In other words, the g functions must take their values in spaces of the same dimension. The solution proposed in the article is to define a large model that contains all the models. Thus we rewrite the two models as follow :

$$\begin{aligned}&\text{M1 : } \mathbb{E}^P[e_i(\theta)] = 0 \qquad \mathbb{E}^P[e_i(\theta)z_i] = 0 \qquad \mathbb{E}^P[(e_i(\theta))^3] = v\\ &\text{M2 : } \mathbb{E}^P[e_i(\theta)] = 0 \qquad \mathbb{E}^P[e_i(\theta)z_i] = 0 \qquad \mathbb{E}^P[(e_i(\theta))^3] = 0\end{aligned}$$

Now, the condition on the v parameter allows us to compare the two models. The parameter v is free in model 1 which means that there is no restriction for $\mathbb{E}^P[(e_i(\theta))^3]$ in model 1.

We then generate $n = 250$ observations from the regression model (1), calculate the marginals of the two models and select the model with the highest. Model 2 should be rejected insofar as the condition $\mathbb{E}^P[(e_i(\theta))^3] = 0$ is not verified by our generated datas. We repeat the process one hundred time, each time the model M1 is the one selected.
It would have been possible to compare more than two models and to test several other different conditions. Nevertheless, each new moment condition implies to enlarge the big model accordingly, which increases the size of the problem and the computation time. Indeed, the computation time of the marginal for each model is relatively long, in particular for $\pi(\tilde{\psi}|x_{1:n}, M_l)$ which requires several generations.

Conclusion

We demonstrated the use of the BETEL framework in the case of a simple regression model. We had some troubles to implement the method as it requires advanced optimisation tools that we created from scratch in Python and some of the computation steps were not mentioned in the paper. The next step would have been to improve the codes to run it faster and to apply the method to more complicated models.

We have shown that the Metropolis Hastings combined with a well chosen proposal function can generate samples from a non parametric posterior and that those samples are asymptotically normal. We then confirmed that the framework can be used for relevant model selection with comparing simple regression models.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. USA: Cambridge University Press, 2004. ISBN: 0521833787.

[2] Jeliazkov Chib. “Marginal Likelihood from the Metropolis-Hastings Output”. In: *Journal of the American Statistical Association* 55.1 (2001), pp. 1657–1663.

[3] Siddhartha Chib, Minchul Shin, and Anna Simoni. “Bayesian estimation and comparison of moment condition models”. In: *Journal of the American Statistical Association* 113.524 (2018), pp. 1656–1668.

[4] Susanne M Schennach. “Bayesian exponentially tilted empirical likelihood”. In: *Biometrika* 92.1 (2005), pp. 31–46.