Review of Estimation Methods

MLE and GMM in Applied Settings

Erica Moszkowski[a]
emoszkowski@g.harvard.edu

August 30, 2022

Harvard University, Econ 2610

[a]Based on previous notes by David Hao Zhang, Ashvin Gandhi, Daniel Pollman, Tom Wollman, and Michael Sinkinson

Outline

Identification of Structural Parameters

- A *structrural parameter* is one that is invariant to a particular set of counterfacturals the researcher is interested in.
- *Identification* can refer to a lot of things (see Lewbel, JEL forthcoming), but at its most basic, it means that parameters or features of a model are uniquely determined from the observable population that generates the data.
    - Identification is a statement about *a model in the population.* It is not about the data you have!
    - Identification is obtained by imposing exclusions or other kinds of restrictions (read: assumptions) on your data.

Extremum Estimators

“Extremum estimators are a wide class of estimators for ***parametric*** *models that are calculated through maximization (or minimization) of a certain* ***objective function****, which depends on the* ***data****.” – Wikipedia*

Setup of an extremum estimator:

- Parameter space: $\Theta \subset \mathbb{R}^K$. K must be finite and independent of sample size.
- Data: $W_i = (Y_i, X_i)$ for observations $i = \{1, \ldots, n\}$, iid
- Objective function: $Q_n(\theta_0)$

The *estimate*, $\hat{\theta}$, is the value that minimizes the objective function:

$$\hat{\theta} = \arg\min_t Q_n(t)$$

Identification and Consistency of an Extremum Estimator

Identificaton:

- Condititions for identification:
  - Θ is compact
  - $Q(\theta)$ is continuous in θ
  - θ_0 *uniquely* minimizes $Q(\theta)$

Consistency:

- An estimator is *consistent* if

$$\hat{\theta} \xrightarrow{p} \theta_0$$

  or, more formally,

$$\lim_{n\to\infty} \Pr\left[\left\|\hat{\theta} - \theta_0\right\| > \epsilon\right] = 0, \forall \epsilon > 0.$$

Notation

- For distributions, upper case letters denote c.d.f.s and lower case letters denote p.d.f.s.
- For random variables, upper case letters denote the random variable itself and lower case letters denote the realization of the random variable (e.g., data before it is observed is W_i; once it's observed, it's w_i)
- For parameters, the subscript 0 denotes the true value of the parameter and a $\hat{\cdot}$ denotes an estimate. No annotation denotes a generic argument to a function.

Outline

Maximum Likelihood (MLE)

1. Let $(W_1, \dots, W_n)$ be iid random variables, where W_i has distribution F. The realizations $(w_1, \dots, w_n)$ correspond to the observed data.
2. Researcher picks a family of distributions, F_θ, indexed by a parameter $\theta \in \Theta$
   - For each observation, pdf $f(w_i|\theta)$ is the probability that you draw observation $W_i = w_i$ from a population distributed according to F_θ
3. The *likelihood* of observing your exact dataset is:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(w_i|\theta)$$

4. The maximum likelihood estimate, $\hat{\theta}^{MLE}$, of θ is the value that makes the observed data the "most probable" according to your model:

$$\hat{\theta}^{MLE} = \arg\max_t \mathcal{L}(\theta)$$

Identification of MLE

As we know, MLE is identified if the likelihood is **uniquely** maximized at the true value; that is:

$$\arg\max_{\theta} \mathcal{L}(\theta) = \theta' \iff \theta' = \theta_0$$

Computation of MLE

- We usually maximize the log likelihood, because summation is faster than multiplication:

$$\hat{\theta}^{MLE} = \arg\max_{t} \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(w_i|\theta)$$

- Most programming languages' optimization functions do *minimization* rather than *maximization*: don't forget to multiply the log-likelihood by -1!
- I recommend writing a function, `likelihood`, that computes the log-likelihood for you. This makes your code much cleaner.

```
% In its own file:
function likelihood(theta,w) = sum(log(normpdf(w,theta)))
% When you estimate:
theta_hat = fminsearch(@(t) likelihood(t,w), theta_start)
```

Outline

Generalized Method of Moments (GMM): Setup

- The researcher specifies a model that implies the *moment condition*

$$\mathbb{E}\left[\psi\left(W_i, \theta_0\right)\right] = 0, \tag{1}$$

  where ψ is known and has dimension L.
- For us to be able (with infinite data) to tell a false θ apart from the true value, we need

$$\mathbb{E}\left[\psi(W_i, \theta)\right] \neq 0 \text{ for all } \theta \neq \theta_0 \tag{2}$$

- Identification also requires $L \geq K$ (i.e., more moment conditions than parameters $(\theta_1, \ldots, \theta_K)$).
- The sample analog of the moment $\mathbb{E}[\psi\left(W_i, \theta\right)]$ is:

$$m_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi(w_i, \theta) \in \mathbb{R}^L$$

Defining the Estimator

- Since θ_0 is the only θ that satisfies $\mathbb{E}[\psi(W_i, \theta)] = 0$, we might want to look for an estimator $\hat{\theta}^{GMM}$ that satisfies the system of L equations in K unknowns:

$$m_n(\hat{\theta}^{GMM}) = 0$$

- If $L = K$ then this system typically has a solution
- But if $L > K$ (the *over-identified* case), a solution may not exist.

  $\rightsquigarrow$ pick $\hat{\theta}^{GMM}$ to satisfy $m_n(\hat{\theta}^{GMM}) \approx 0$ as closely as possible.
- The GMM objective function is:

$$Q^{GMM}(\theta) = \underset{\theta}{\operatorname{argmin}} \, \mathbb{E}\left[\psi\left(w;\theta\right)\right]' C \mathbb{E}\left[\psi\left(w;\theta\right)\right] \quad (3)$$

  - This is like the “sum of squared residuals”, but C lets us be more flexible
- The GMM estimate, $\hat{\theta}^{GMM}$, is the value that minimizes the sample analog of eqn 3:

$$\hat{\theta}^{GMM} = \underset{\theta}{\operatorname{argmin}} \, Q^{GMM}_{C,n}(\theta) = m_n(\theta)'\hat{C} m_n(\theta) \quad (4)$$

Familiar Examples of Moment Conditions

- **Regression**: $Y_i = X_i'\theta + \varepsilon_i$

$$\mathbb{E}\left[X_i \varepsilon_i\right] = 0.$$

- **Instrumental Variables**: $Y_i = X_i'\theta + \varepsilon_i$, $\mathbb{E}[X_i'\varepsilon_i] \neq 0$

$$\mathbb{E}\left[Z_i \varepsilon_i\right] = 0.$$

- **Maximum likelihood**: $\max_\theta \mathcal{L}(Y_i|X_i, \theta)$

$$\mathbb{E}\left[\frac{\partial \log \mathcal{L}(Y_i|X_i, \theta)}{\partial \theta}\right] = 0.$$

Typical moment conditions in IO applications

1. **Orthogonality conditions**:
   - Examine your model for zero-correlation conditions.
   - Example: Any variable Z that is uncorrelated with unobserved heterogeneity in product characteristics, ξ, can be used as an instrument. The moment condition is

$$\mathbb{E}[Z'\xi] = 0$$

2. **First order conditions**:
   - **Best-Response/Nash conditions**: Equilibrium conditions which we assume to hold on the supply side, such as Differentiated Products Bertrand Equilibrium:

$$\max \underbrace{s(p)}_{\text{share}} \underbrace{[p-c]}_{\text{markup}} \implies \text{FOC: } \underbrace{s'(p)[p-c] + s(p) = 0}_{\text{moment condition}}$$

   - **Consumer optimality**: Consumer may optimally stockpile, for example, based on sales frequencies and amounts.

Typical moment conditions in IO applications, ctd

3. **Moment matching**:

   Lots of papers do the following thing:

   1. Write up a model with some parameter vector θ.
   2. Determine how the model implies moments of the data should depend on θ.
      - You can derive this relationship analytically or by simulation.
      - Example: The probability that an individual chooses an insurance plan is a known function of their risk aversion (same for everyone) and health risk (drawn from some parameterized distribution). Simulate a lot of individuals' choice probabilities, then aggregate.
   3. Pick $\hat{\theta}$ such that the model-implied moments, match the empirical moments as closely as possible (using whatever metric you like, often sum-of-squares)

Identification of GMM

Recall that an extremum estimator is identified iff

$$\theta = \arg\min_t Q^{GMM}(t) \iff \theta = \theta_0$$

Therefore, GMM is identified if:

1. $\mathbb{E}\left[\psi\left(W_i;\theta\right)\right] = 0 \iff \theta = \theta_0$
    - It's generally hard to show $\mathbb{E}[\psi(W_i,\theta_0)] = 0$. Many papers don't actually prove their model is identified.
2. the weight matrix C is nonsingular, and
3. $L \geq K$

Consistency of GMM

- For GMM, pointwise convergence of the sample objective function to the population objective function is easy:
  - $m_n(\theta) \xrightarrow{p} \mathbb{E}[\psi(W_i, \theta)]$ by the law of large numbers
  - $\hat{C} \rightarrow C$ by assumption
  - So for all $\theta \in \Theta$, $m_n(\theta)'\hat{C}m_n(\theta) \xrightarrow{p} \mathbb{E}[\psi(W_i, \theta)]C\mathbb{E}[\psi(W_i, \theta)]$
- There are stronger conditions under which the sample GMM objective function, eqn 4, becomes *uniformly* close to the limiting objective function, eqn 3
  - Loosely, "uniform convergence" means there's an ε s.t. as $n \rightarrow \infty$the distance at *any* θ between expressions 4 and 4 is less than ε
- In applications, make sure you know which n is being referred to: are you taking the number of firms to ∞? The number of markets?

Asymptotic Normality of GMM

- Under some (standard) assumptions,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V\right),$$

where

$$V = \left(\Gamma' C \Gamma\right)^{-1} \Gamma' C \Delta C \Gamma \left(\Gamma' C \Gamma\right)^{-1}$$

- $\Gamma = \mathbb{E}\left[\frac{\partial \psi}{\partial \theta}\left(x, \theta_0\right)\right]$: gradient of the moment condition w.r.t. to the parameters/Hessian of the unweighted objective function (size = $L \times K$)
- $\Delta = \mathbb{E}\left[\psi\left(x, \theta_0\right)\psi\left(x, \theta_0\right)'\right]$: covariance of the moment conditions at θ_0 (size = $L \times L$)
- Note that this is only one component of error. There is also:
    - sampling error (if your data is a sample of the population).
    - simulation error (if you compute the moments via simulation).

  These will enter into the Δ term.

Standard Errors

As you'd expect, the sample analog is a consistent estimator of V:

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi(x,\hat{\theta})}{\partial\theta'}$$

$$\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\psi(x,\hat{\theta})\psi(x,\hat{\theta})'$$

$$\hat{V} = (\hat{\Gamma}'\hat{C}\hat{\Gamma})^{-1}\hat{\Gamma}'\hat{C}\hat{\Delta}\hat{C}\hat{\Gamma}(\hat{\Gamma}'\hat{C}\hat{\Gamma})^{-1}$$

Therefore standard errors are:

$$SE = \sqrt{\frac{\text{diag}(\hat{V})}{n}}$$

Ok, it's time to talk about $\hat{C}$.

The Optimal Weighting Matrix

- In the over-identified case, the weight matrix C assigns "importance" to satisfying the different moment conditions. We can choose whatever C we like, as long as it is positive definite.
- So, choose C to make our estimate as precise as possible – that is, "minimize" V

Just-identified case, $C = I$	Over-identified case, $C = \Delta^{-1}$
$V = (\Gamma' C \Gamma)^{-1} \Gamma' C \Delta C \Gamma (\Gamma' C \Gamma)^{-1}$	$V = (\Gamma' C \Gamma)^{-1} \Gamma' C \Delta C \Gamma (\Gamma' C \Gamma)^{-1}$
$= \Gamma^{-1} C^{-1} \Gamma'^{-1} \Gamma' C \Delta C \Gamma \Gamma^{-1} C^{-1} \Gamma'^{-1}$	$= (\Gamma' \Delta^{-1} \Gamma)^{-1} \Gamma' \Delta^{-1} \Delta \Delta^{-1} \Gamma (\Gamma' \Delta^{-1} \Gamma)^{-1}$
$= \Gamma^{-1} \Delta \Gamma'^{-1} = (\Gamma' \Delta^{-1} \Gamma)^{-1}$	$= (\Gamma' \Delta^{-1} \Gamma)^{-1}$

- Find the proof that $(\Gamma' \Delta^{-1} \Gamma)^{-1}$ is positive semi-definite in any econometrics text.
- **Intuition**: we want to more heavily weight the moments that are "precisely measured" (ie the least variable sample moments).

2-Step GMM

- From the last slide, we would like $C \propto \Delta^{-1} = \mathbb{E}[Cov(\psi(W_i, \theta_0))]^{-1}$.
- **Problem:** we don't know θ_0.
- **Solution:** Form a consistent estimate $\hat{\Delta}$ using a consistent though inefficient estimate of θ_0. This is good enough to achieve the optimal asymptotic variance.

2-step GMM:

- **Step 1:** Estimate $\hat{\theta}^{GMM1}$ by minimizing $Q_{C,n}(\theta)$ with an arbitrary choice of (positive semi-definite) C (usually the identity matrix)
- **Step 2:** Estimate the optimal weighting matrix as:

$$\hat{\Delta}^{-1} = \left\{ \mathbb{E}_n \left[\psi\left(w_i, \hat{\theta}^{GMM1}\right) \psi\left(w, \hat{\theta}^{GMM1}\right)' \right] \right\}^{-1}$$

and use this to then solve for $\hat{\theta}_{GMM2} = \arg\min_\theta Q_{\hat{\Delta}^{-1}}(\theta)$.

Two-Step GMM: OLS example

Consider the following OLS model:

$$Y_i = X_i'\theta_0 + \varepsilon_i, \quad \mathbb{E}\left(\varepsilon_i | X_i\right) = 0$$

Orthogonality of X_i and ε_i implies:

$$\mathbb{E}\left(Y_i - X_i'\theta_0 | X_i\right) = 0 \Rightarrow \mathbb{E}\left[\left(Y_i - X_i'\theta_0\right) h\left(X_i\right)\right] = 0$$

for any function $h(\cdot)$, in particular $h(X) = X$. So choose

$$\psi\left(W_i; \theta\right) = \left(Y_i - X_i'\theta\right) X_i$$

and we get a moment conditon: $\mathbb{E}\left[\psi\left(W_i; \theta_0\right)\right] = 0$.

In a more general problem, using "optimal instruments" means optimal choice of $h(\cdot)$, an approximation to which we will discuss later.

Two-Step GMM: Linear IV example

Consider the following linear IV model, where X_i is a $K \times 1$ vector:

$$Y_i = X_i'\theta_0 + \varepsilon_i, \quad i = 1, \ldots, n$$

You know how to do this with 2SLS, but let's set it up in the GMM framework.

- Suppose $\mathbb{E}[X_{ik}\varepsilon_i] \neq 0$ for some $k \in 1, \ldots, K$
- You have an $L \times 1$ vector Z_i of instruments such that $\mathbb{E}[Z_i\varepsilon_i] = 0$ and $Cov(Z_i, X_i) > 0$ (exclusion restriction + relevance hold)
- Let $W_i = (Y_i, X_i, Z_i)$
- Define $\psi(W_i, \theta_0) = Z_i\varepsilon_i = Z_i(Y_i - X_i'\theta_0)$ so we can use GMM
- If only some elements of X_i are endogenous, Z_i will also include the remaining subset.
    - Notice: if $\dim(z_i) = \dim(x_i)$, the model is just-identified; for $\dim(Z_i) > \dim(X_i)$, it is over-identified.

Analytical solution to linear GMM

The sample moment is:

$$m_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i\theta) = (S_{zy} - S_{zx})\theta; \quad S_{zy} = \frac{1}{N}\sum_{t=1}^{N} z_t y_t, \quad S_{zx} = \frac{1}{N}\sum_{t=1}^{N} z_t x_t'$$

You can set $m_n = 0$ and solve:

$$\hat{\theta} = (S_{zx}' C S_{zx})^{-1} S_{zx}' C S_{zy}$$

The asymptotic variance is:

$$Cov(\hat{\theta}) = (S_{zx}' C S_{zx})^{-1} S_{zx}' C \hat{S} C S_{zx} (S_{zx}' C S_{zx})^{-1}, \quad \hat{S} = \frac{1}{N}\sum_{t=1}^{N} z_t z_t' \hat{\varepsilon}_t^2$$

and C is the weight matrix. The weight matrix can be estimated after the first-step via $\hat{C} = \hat{S}^{-1}$. $Cov(\hat{\theta})$ should be estimated in the second step with the second step $\hat{S}$.

- This is equivalent to 2SLS if errors are homoscedastic (but they may not be!)

Logit Example: Another Linear GMM!

Suppose instead, we have:

- Just market shares and characteristics of J goods
- Endogeneity of certain characteristics (need to instrument)

$\rightsquigarrow$ hard to construct a likelihood.

Let $\delta_j = \beta X_j + \xi_j$, so that market shares (aggregated across all consumers i) are:

$$s_j = \frac{\exp(\delta_j)}{1 + \sum_k \exp(\delta_k)} \tag{5}$$

Given the share of an outside good s_0, **?** says we can recover δ_j:

$$\delta_j = \log(s_j) - \log(s_0) \tag{6}$$

Given instruments Z_j we can form a familiar linear GMM moment condition:

$$\mathbb{E}[\xi_j Z_j] = \mathbb{E}[(\delta_j - \beta X_j)Z_j] = \mathbb{E}[(\log(s_j) - \log(s_0) - \beta X_j)Z_j] = 0 \tag{7}$$

An Approximation to Optimal Instruments

More generally, suppose we want to estimate α, β using the following moment condition:

$$\mathbb{E}(\xi_j h_j(Z)) = \mathbb{E}[(\delta_j - \beta X_j - \alpha p_j)h_j(Z)] = 0 \tag{8}$$

Let $T(z)'T(z) = \Delta^{-1}$ (so $T(z)$ normalizes the error matrix). **?** tells us that the optimal set of instruments is:

$$h_j(z) = \mathbb{E}\left[\frac{\partial \xi_j(\theta_0)}{\partial \theta} | Z\right] T(z_j) \tag{9}$$

Intuition:

- Give larger weights to observations that generate ξs whose computed values are very sensitive to the choice of θ

Approximating Optimal Instruments

Problem: it's hard to compute $\mathbb{E}\left[\frac{\partial \xi_j(\theta_0)}{\partial \theta}|Z\right]$

- We would need to compute the pricing equilibrium for different sequences of ξ_j, compute $\frac{\partial \xi_j}{\partial \theta}$ at that price, and integrate over all such sequences.

The approximation of **?**:

1. Obtain an initial estimate of $\hat{\alpha}, \hat{\beta}$ using any instruments.
2. Use the initial estimate to construct $\hat{\delta}_j = \hat{\beta} X_j + \hat{\alpha} p_j$, $(\xi = 0)$.
3. Solve the FOC of the model to find $\hat{p}, \hat{s}$ as a function of $\alpha, \beta, \hat{\delta}, X$.
4. Get $\hat{\xi}_j(\alpha, \beta) = \hat{\delta}_j(\alpha, \beta) - \beta X_j - \alpha \hat{p}_j(\alpha, \beta)$, and take the derivatives $\frac{\partial \hat{\xi}_j}{\partial \alpha}, \frac{\partial \hat{\xi}_j}{\partial \beta}|\hat{\alpha}, \hat{\beta}$ as an approximation to $E\left[\frac{\partial \xi_j(\theta_0)}{\partial \theta}|Z\right]$.

Is GMM a "black box"?

- Under a researcher's maintained assuptions a_0, $\hat{\theta}^{GMM}$ is consistent and asymptotically normal
- But what if readers want to assess the bias in $\hat{\theta}^{GMM}$ if some other alternative $a \neq a_0$ were the case?
- **?** provide an expression for the direction and magnitude of bias: For any local perturbation to the true model leading to the moments converging asymptotically to $\tilde{\psi}$ instead of 0, the first-order asymptotic bias to the estimates $\tilde{\theta}$ is:

$$\mathbb{E}[\tilde{\theta}] = \Lambda \mathbb{E}[\tilde{\psi}]$$

  where $\Lambda = -(\Gamma' C \Gamma)^{-1} \Gamma' C$ is the sensitivity of estimated parameters to the model.
    - For OLS, $\Lambda = -\Gamma^{-1} = -E[XX']$. Intuition = omitted variables: the bias from not including an endogenous variable is related to its covariance with included variables.

Outline

Efficiency of Different Estimators

- We want to know whether our estimates are as precise as possible.
- MLE achieves the Cramer-Rao lower bound on variance among *all* unbiased estimators in the parametric setting:

$$Var\left(\hat{\theta}(X)\right) \geq \underbrace{\Im\left(\theta_0\right)^{-1}}_{\text{Cramer-Rao bound}} \quad \text{where} \quad \underbrace{\Im\left(\theta\right)}_{\text{Fisher Information matrix}} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \ln p\left(W|\theta\right)\right]$$

- GMM attains the *semi*-parametric efficiency bound (**?**), which is the lower bound on variance for an estimator using only the information contained in the moment restrictions.

How restrictive is the estimator?

- MLE assumes the distribution of the data is known, up to a parameter. This is very restrictive!
- GMM makes assumptions about the *moments* of the distributions, which is less restrictive.
- For reference, non-parametric estimators (not discussed here) make almost no assumptions about the underlying distribution of the population.

When would you want to use each estimator?

- Use MLE when you have a fully specified distributional model and aren't worried about unmodeled endogeneity.
  - Remember, MLE assumes any variable *not* in your model is exogenous.
- Use GMM if your model is "partially specified" in the sense that you are making assumptions about orthogonality of residuals or optimality of behavior.
  - If you care about endogeneity of unobservables, you probably want to use GMM.

Trade off between strength of assumptions/amount of structure placed on the data and efficiency.

Outline

Implementing Extremum Estimators in Julia

```julia
using Optim
function linear_gmm(X,y; beta_start = 0.0, method = NelderMead())
  gmm_obj(b) = (X'*(y-X*b))'*(X*(y-X*b))
  return optimize!(gmm_obj, beta_start, method)
end
bhat = linear_gmm(X,y, BFGS())
```

- `y` is a column vector and `x` is a matrix (rows = observations)
- `method` is a keyword argument for choosing your opimization routine (default = Nelder-Mead)
- `optimize!` takes a single-argument function and minimizes it, starting at `beta_start`, using the method you specify.
- `gmm_obj` is a *closure*: it's a function defined on 1 variable, baking in the values of `X` and `y`. Using closures carefully can make your code much cleaner.

Implementing Extremum Estimators in MATLAB

```
bet = fminsearch(@(b) (X'*(y-X*b))'*(X*(y-X*b)), beta_start, myopts)
```

- `y` is a column vector and `x` is a matrix (rows = observations)
- `myopts` is a struct containing lots of options
- The answer will be stored in a variable `bet`
- `@(b)` means the routine will attempt to minimize the expression `(X'*(y-X*b'))'*(X'*(y-X*b'))` with respect to `b`. This is called an *anonymous function*, but you can also use a named function as in the likelihood example
- The starting guess for `b` will be the value held in the vector `beta_start`
- The routine will follow the specifications in the options set “myopts”, which is set before this using a command like

  ```
  myopts = optimset('TolFun',10e-12, 'MaxFunEvals',1000000,'MaxIter',1000)
  ```

- Also see `fmincon` and `fminunc`

Computing Gradients

- Necessary for Γ in asymptotic variance.
- Exact differentiation (analytic derivatives) is always preferred to numerical differentiation due to approximation error. This is also runs *much* faster.
  - Logit models (including BLP) do allow one to compute exact gradients – just differentiate the logit!
- If not practical, approximate the gradient using finite differences with h:
  - Forward difference formula:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

  - Symmetric difference formula (more accurate):

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}$$

  - See Judd (1998, Ch. 7) for details.

Where to go for more info

- I will post (with permission) Mikkel Plagborg-Møller's notes on GMM, which you may find useful.
- Also see **?** for a more formal overview

References i