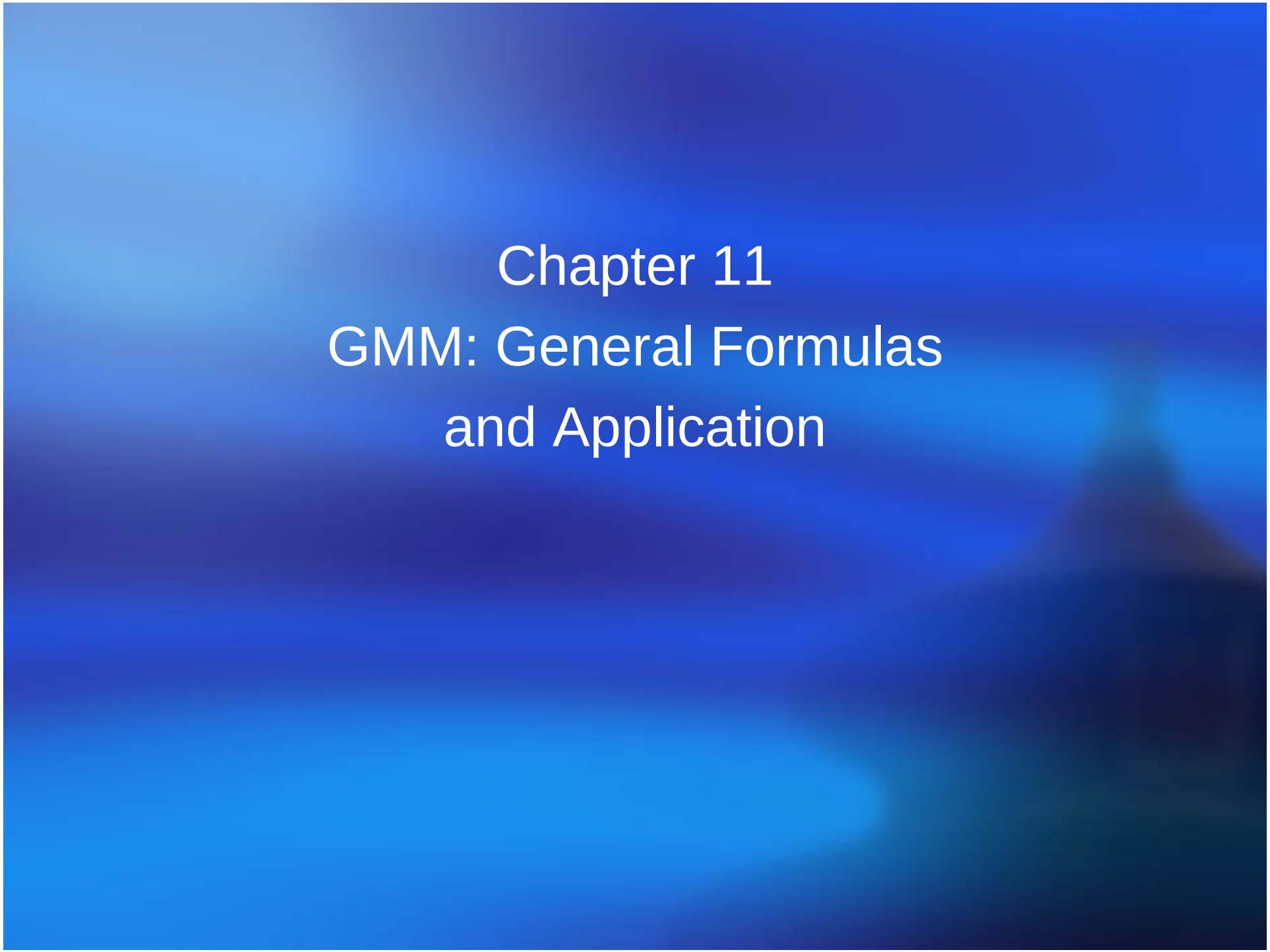

Chapter 11
GMM: General Formulas
and Application

Main Content

- General GMM Formulas
- Testing Moments
- Standard Errors of Anything by Delta Method
- Using GMM for Regressions
- Prespecified weighting Matrices and Moment Conditions
- Estimating on One Group of Moments,Testing on Another
- Estimating the Spectral Density Matrix

11.1 General GMM Formulas

- GMM procedures can be used to implement a host of estimation and testing exercises.
- To estimate the parameter, you just have to remember (or look up) a few very general formulas, and then map them into your case.
- Express a model as

$$E[f(x_t, b)]$$

- Everything is a vector: f can represent a vector of L sample moments, x_t can be M data series, b can be N parameters.

Definition of the GMM Estimate

- We estimate parameters $\hat{b}$ to set some linear combination of sample means of f to zero

$$\hat{b} : set \quad a_T g_T(\hat{b}) = 0$$

where

$$g_T(b) = \frac{1}{T}\sum_{t=1}^{T} f(x_t, b)$$

- a_t is a matrix that defines which linear combination of $g_T(b)$ will be set to zero.
- If you estimate b by $\min g_T'(b) W g_T(b)$,the first-order condition are $\frac{\partial g_T'}{\partial b} W g_T(b) = 0$
- This is mean $a_T = \frac{\partial g_T'}{\partial b} W$

Standard Error of Estimate

- Hansen (1982), Theorem 3.1 tells us that the asymptotic distribution of the GMM estimate is

$$\sqrt{T}(\hat{b} - b) \rightarrow N[0, (ad)^{-1} aSa'(ad)^{-1'}]$$

where

$$d \equiv E[\frac{\partial f}{\partial b'}(x_t, b)] = \frac{\partial g_T(b)}{\partial b'} \qquad a = p\lim a_T$$

$$S \equiv \sum_{j=-\infty}^{\infty} E[f(x_t, b), f(x_{t-j}, b)']$$

- In practical terms,this means to use

$$\text{var}(\hat{b}) = \frac{1}{T}(ad)^{-1} aSa'(ad)^{-1'}$$

Distribution of the moments

- Hansen's Lemma 4.1 gives the sampling distribution of the moments $g_T(b)$:

$$\sqrt{T}\, g_T(\hat{b}) \to N[0, (I - d(ad)^{-1}a)S(I - d(ad)^{-1}a)']$$

- The $I - d(ad)^{-1}a$ terms account for the fact that in each sample some linear combinationsof g_T are set to zero. Then S is singular.

χ^2 Test

- A sum of squared standard normals is distributed χ^2, so we have

$$Tg_T(\hat{b})'[(I - d(ad)^{-1}a)S(I - d(ad)^{-1}a)']^{-1}g_T(\hat{b})$$

is distributed χ^2 which has degrees of freedom given by number of nonzero linear combinations of g_T, the number of moments less the number of estimated parameters

- It does, but with a hitch: The variance-covariance matrix is singular, so you have to pseudo-invert it.
- For example, you can perform an eigenvalue decomposition $\sum = Q \Lambda Q'$ and then invert only the non-zero eigenvalues.

Efficient Estimates

- Hansen shows that one particular choice is statistically optimal, $a = d'S^{-1}$
- This choice is the first order condition to $\min_{\{b\}} g_T(b)'S^{-1}g_T(b)$ that we studied in the last Chapter.
- With this weighting matrix, the standard error of b reduces to

$$\sqrt{T}(\hat{b} - b) \to N[0, (d'S^{-1}d)^{-1}]$$

- With the optimal weights S^{-1} the variance of the moments simplifies to

$$\operatorname{cov}(g_T) = \frac{1}{T}(S - d(d'S^{-1}d)^{-1}d')$$

- Proof:

$$\operatorname{var}(g_T(\hat{b})) = \frac{1}{T}(I - d(ad)^{-1}a)S(I - d(ad)^{-1}a)'$$

$$a = d'S^{-1}$$

$$(I - d(ad)^{-1}a)S(I - d(ad)^{-1}a)')$$

$$= (I - d(d'S^{-1}d)d'S^{-1})S(I - d(d'S^{-1}d)d'S^{-1})$$

$$= (S - d(d'S^{-1}d)d')(I' - S^{-1}d(d'S^{-1}d)^{-1'}d')$$

$$= S - d(d'S^{-1}d)^{-1'}d' - d(d'S^{-1}d)d' + d(d'S^{-1}d)d'$$

$$= S - d(d'S^{-1}d)^{-1'}d'$$

- Using this matrix in a test, there is an equivalent and simpler way to construct this test

$$Tg_T(\hat{b})S^{-1}g_T(\hat{b}) \rightarrow \chi^2(\#moments - \#parameters)$$

- Alternatively, note that S^{-1} is a pseudo-inverse of the second stage $\text{cov}(g_T)$
- Proof: A pseudo inverse times $\text{cov}(g_T)$ should result in an idempotent matrix of the same rank as $\text{cov}(g_T)$

$$S^{-1}\text{cov}(g_T) = S^{-1}(S - d(d'S^{-1}d)d') = I - S^{-1}d(d'S^{-1}d)d'$$

- Then, check that the result is idempotent

$$(I - S^{-1}d(d'S^{-1}d)^{-1}d')(I - S^{-1}d(d'S^{-1}d)^{-1}d') = I - S^{-1}d(d'S^{-1}d)^{-1}d')$$

- This derivation not only verifies that J_T has the same distribution as $g_T'\text{cov}(g_T)g_T$, but that they are numerically the same in every sample.

Model Comparisons

- You often want to compare one model to another. If one model can be expressed as a special or restricted case of the other or unrestricted model we can perform a statistical comparison that looks very much like a likelihood ratio test.

$$TJ_T(restricted) - TJ_T(unrestricted) \sim \chi^2(\#restriction)$$

- If the restricted model is really true, it should nor rise “much.
- This is a “ χ^2 difference” test due to Newey and West(1987), who call it the “D-test”

11.2 Test Moments

- How to test one or a group of pricing error.
  (1)Use the formula for $\text{var}(g_T)$
  (2)A χ^2 difference test
- We can use the sampling distribution of g_T , to evaluate the significance of individual pricing errors, to construct a t-test(for a single moment) or a χ^2 test(for groups of moments)

- Alternatively, you can use the χ^2 difference approach.
- Start with a general model that includes all the moments, and form an estimate of the *spectral density matrix* $\mathcal{S}$.
- *Set* to zero the moments you want to test, and denote $g_{sT}(b)$ the vector of moments, including the zeros(s for “smaller)

  $Tg_T(\hat{b})'S^{-1}g_T(\hat{b}) - Tg_{sT}(\hat{b}_s)'S^{-1}g_{sT}(\hat{b}_s) \sim \chi^2$ (#eliminated moments)
- If moments we want to test truly are zero,the criterion should not be that much lower

11.3 Standard Errors of Anything by Delta Method

- we want to estimate a quantity that is a nonlinear function of sample means

$$b = \phi[E(x_t)] = \phi(u)$$

- In this case, we have

$$\text{var}(b_T) = \frac{1}{T}[\frac{d\phi}{du}]' \sum_{-\infty}^{\infty} \text{cov}(x_t, x_{t-j}')[\frac{d\phi}{du}]$$

- For example, a correlation coefficient can be written as a function of sample means as

$$corr(x_t, y_t) = \frac{E(x_t y_t) - E(x_t)E(y_t)}{\sqrt{E(x_t^2) - E^2(x_t)}\sqrt{E(y_t^2) - E^2(y_t)}}$$

- Thus, take $u = [E(x_t), E(x_t^2), E(y_t), E(y_t^2), E(x_t y_t)]'$

11.4 Using GMM for Regression

- Mapping any statistical procedure into GMM makes it easy to develop an asymptotic distribution that corrects for statistical problems such as *non-normality*, *serial correlation* and *conditional heteroskedasticity*.
- For example, I map OLS regressions into GMM.
- When errors do not obey the OLS assumptions, OLS is consistent, and often more robust than GLS, but its standard errors need to be corrected.

- OLS picks parameters β to minimize the variance of the residual:

$$\min_{\{\beta\}} E_T[(y_t - \beta' x_t)^2]$$

- We find $\widehat{\beta}$ from the first order condition, which states that the residual is orthogonal to the right hand variable:

$$g_T(\widehat{\beta}) = E[x_t(y_t - x_t'\widehat{\beta})] = 0$$

- It is exactly identified. We set the sample moments exactly to zero and there is no weighting matrix ($a = I$). We can solve for the estimate analytically,

$$\widehat{\beta} = [E_T(x_t x_t')]^{-1} E_T(x_t y_t)$$

- This is the familiar OLS formula. But its standard error need to be corrected.

- We can use GMM to obtain the standard errors through $\sqrt{T}(\widehat{b}-b) \to N[0,(d'S^{-1}d)^{-1}]$, so that

$$d = E(x_t x_t')$$

$$f(x_t, \beta) = x_t(y_t - x_t\beta) = x_t\varepsilon_t$$

$$\text{var}(\widehat{\beta}) = \frac{1}{T}E(x_t x_t')^{-1}[\sum_{j=-\infty}^{\infty} E(\varepsilon_t x_t x_{t-j}' \varepsilon_{t-j}')]E(x_t x_t')^{-1}$$

Serially Uncorrelated, Homoskedastic Errors

- Formally,the OLS assumptions are

$$E(\varepsilon_t \mid x_t, x_{t-1}, \cdots \varepsilon_{t-1}, \varepsilon_{t-2} \cdots) = 0$$

$$E(\varepsilon_t^{\,2} \mid x_t, x_{t-1}, \cdots \varepsilon_{t-1}, \varepsilon_{t-2} \cdots) = cons\tan t = \sigma_E^2$$

- The first assumption means that only the j=0 term enter the sum

$$\sum_{j=-\infty}^{\infty} E(\varepsilon_t x_t x'_{t-j} \varepsilon'_{t-j}) = E(\varepsilon_t^2 x_t x_t')$$

- The second assumption means that

$$E(\varepsilon_t^2 x_t x_t') = E(\varepsilon_t^2) E(x_t x_t')$$

- Hence the standard errors reduces to our old form

$$\mathrm{var}(\hat{\beta}) = \frac{1}{T} \sigma_\varepsilon^2 (X'X)^{-1}$$

Heteroskedastic Errors

- If we delete the condition homoskedasticity assumption

$$E(\varepsilon_t^{\,2} \mid x_t, x_{t-1}, \cdots \varepsilon_{t-1}, \varepsilon_{t-2} \cdots) = cons\tan t = \sigma_\varepsilon^2$$

- The standard errors are

$$\mathrm{var}(\widehat{\beta}) = \frac{1}{T} E(x_t x_t')^{-1} E(\varepsilon_t^2 x_t x_t') E(x_t x_t')$$

- These are known as “heteroskedasticity consistent standard errors” or “white standard errors” after White (1980)

Hansen-Hodrick Errors

- When the regression notation is

$$y_{t+k} = \beta_k' x_t + \varepsilon_{t+k}$$

under the null that one-period returns are unforecastable, we still see correlation in the e_t due to the overlapping data. Unforecastable returns imply

$$E(\varepsilon_t \varepsilon_{t-j}) = 0 \qquad \text{for} \quad |j| \geq K$$

- Under this condition, the standard errors are

$$\text{var}(\beta_k) = \frac{1}{T} E(x_t x_t')^{-1} [\sum_{j=-k+1}^{k} E(\varepsilon_t x_t x_{t-j}' \varepsilon_{t-j}')] E(x_t x_t')^{-1}$$

11.5 Prespecified Weighting Matrices and Moment Conditions

- In the last chapter, our final estimates were based on the "efficinet" S^{-1} weighting matrix.
- A prespecified weighting matrix lets you *specify which moments or linear combination of moments* GMM will value in the minimization.
- So you can also go one step further and impose which linear combinations a_T of moment conditions will be set to zero in estimation rather than use the choice resulting from a minimization.

- For example, if $g_T = [g_T^1, g_T^2]$, $W = I$, but $\partial g_T / \partial b = [1, 10]$ so that the second moment is 10 times more sensitive to the parameter value than the first moment, then GMM with fixed weighting matrix set

$$1 * g_T^1 + 10 * g_T^2 = 0$$

  If we want GMM to pay equal attention to the two moment, we can fix the a_T matrix directly.
- Using a prespecified weighting matrix is not the same thing as ignoring correlation of the error u_t in the distribution theory.

How to Use Prespecified Weighting Matrices

- If we use weighting matrix W , the first-order conditions to $\min_{\{b\}} g_T'(b) W g_T(b)$ are

$$\frac{\partial g_T(b)'}{\partial b} W g_T(b) = d' W g_T(b) = 0$$

- So the variance-covariance matrix of the estimated coefficients is

$$\text{var}(\hat{b}) = \frac{1}{T}(d'Wd)^{-1} d'WSWd(d'Wd)^{-1}$$

- The variance-covariance matrix of the moments g_T

$$\text{var}(g_T) = \frac{1}{T}(I - d(d'Wd)^{-1}d'W)S(I - Wd(d'Wd)^{-1}d')$$

- The above equation can be the basis of χ^2 test for the overidentifying restrictions.

- If we interpret $(\bullet)^{-1}$ to be a generalized inverse, then

$$g_T' \operatorname{var}(g_T)^{-1} g_T \sim \chi^2(\#moment - \#parameters)$$

- If $\operatorname{var}(g_T)$ is singular, you can inverting only the nonzero eigenvalues.

Motivations for Prespecified Weighting Matrices

Robustness, as with OLS vs. GLS

- When errors are *autocorrelated or heteroskedastic* and we *correctly model the error covariance* matrix and *the regression is perfectly specified* , the GLS procedure can improve efficiency.
- If the error covariance matrix is incorrectly, the GLS estimates can be much worse than OLS.
- In these cases, it is often a good idea to use OLS estimates. But we need to correct the *standard error* of the OLS estimates

- For GMM, first-stage or other fixed weighting matrix estimates may give up something in asymptotic efficiency, standard errors and model fit tests.
- They are still consistent and more robust to statistical and economic problems.But we use the *S* matrix in computing standard error.
- When the parameter estimates have a great different between the first stage and the second stage, we should decide what cause this.
- It is truly due to efficiency gain or a model misspecification.

Near-Singular S

- The spectral density matrix is often nearly singular, since asset returns are highly correlated with each other.
- As a result, second stage GMM tries to minimize differences and differences of differences of asset returns in order to extract statistically orthogonal components with lowest variance.
- This feature leads GMM to place a lot of weight on poorly estimated, economically uninteresting, or otherwise non-robust aspects of the data.

- For example, suppose that S is given by

$$S = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- So

$$S^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

- We can write

$$C'C = S^{-1}$$

where

$$C = \begin{bmatrix} \frac{1}{\sqrt{1-\rho^2}} & \frac{-\rho}{\sqrt{1-\rho^2}} \\ 0 & 1 \end{bmatrix}$$

- Then, the GMM criterion is $\min g_T' S^{-1} g_T$ is equivalent to $\min (g_T' C')(C g_T)$

- Cg_T gives the linear combination of moments that efficient GMM is trying to minimize.
- As $\rho \to 1$, for the matrix C, the (2,2) element stay at 1, but the (1,1) and (1,2) elements get very large.
- If $\rho = 0.95$ then

$$C = \begin{bmatrix} 3.20 & -3.04 \\ 0 & 1 \end{bmatrix}$$

- This mean that GMM pay a little attention to the second moment, and play three times as much weight on the difference between first and second moment.
- Through the decomposition of *S*, we can see what moments GMM is prizing.

- GMM wants to focus on well-measured moments.
- In asset pricing applications, the errors are close to uncorrelated over time, so GMM is *looking for portfolios with small values of* $\mathrm{var}(m_{t+1}R^e_{t+1})$. Those will be assets with small return variance.
- Then, *GMM will pay most attention to correctly pricing the sample minimum-variance portfolio*.
- This cause that *sample minimum-variance portfolio may have little to do with the true minimum-variance portfolio*.
- Like any portfolio on the sample frontier, its composition largely reflects luck.

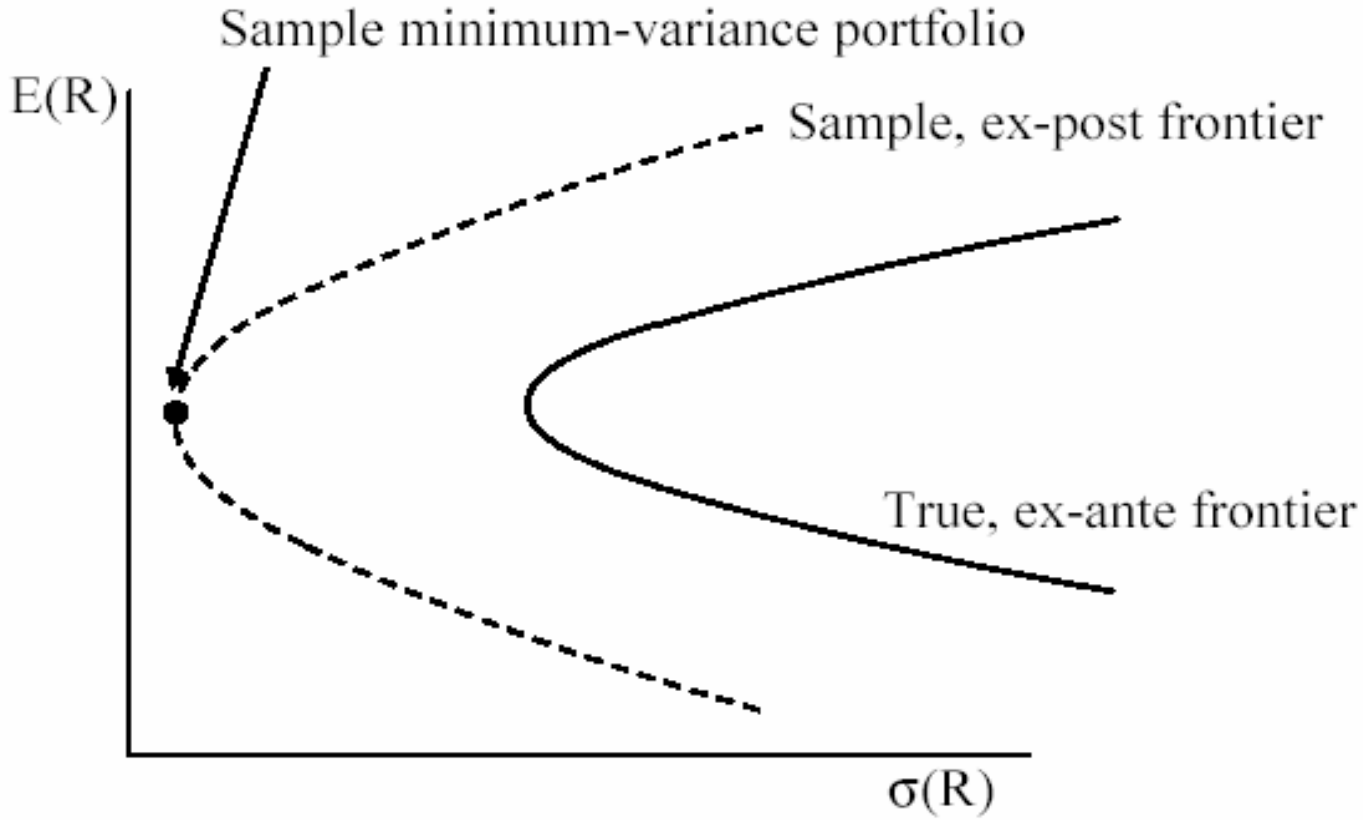


Figure 24. True or ex ante and sample or ex-post mean-variance frontier. The sample often shows a spurious minimum-variance portfolio.

Economically Interesting Moment

- The optimal weighting matrix makes GMM pay close attention to linear combinations of moments with *small sampling error* in both estimation and evaluation.
- We want to force the estimation and evaluation to pay attention to *economically* interesting moments instead.

Level Playing Field

- The S matrix changes as the model and as its parameters change.
- As the S matrix changes, which assets the GMM estimate tries hard to price well changes as well.
- For example we take a model m_t and create a new model by simply adding noise, unrelated to asset returns (in sample), $m_t' = m_t + \varepsilon_t$ then the moment condition $g_T = \mathrm{E}_\mathrm{T}(m_t' R_t^e)$ is unchanged. However, the spectral density matrix $S = E((m_t + \varepsilon_t)^2 R_t^e R_t^e)$ rise dramatically. This can reduce the J_T , leading to a false sense of “improvement”.

Level Playing Field

- If the sample contains a nearly riskfree portfolio of the test assets, or a portfolio with apprently small variances of $m_{t+1}R^e_{t+1,}$ then J_T test will focus on the pricing of this portfolio and will lead to a false rejection, since there is an eigenvalue of S that is too small.
- Some stylized facts, such as the RMSE, pricing errors, are as interesting as the statistical tests.

Some Prespecified Weighting Matrices

- When the second-moment matrix of payoffs $W = E(xx')^{-1}$ in place of *S* *(Hansen and Jagannathan (1997)).* The minimum distance (second moment) between a candidate discount factor y and the space of true discount factors is the same as the minimum value of the GMM criterion with $W = E(xx')^{-1}$ as weighting matrix.
- Why is this true?

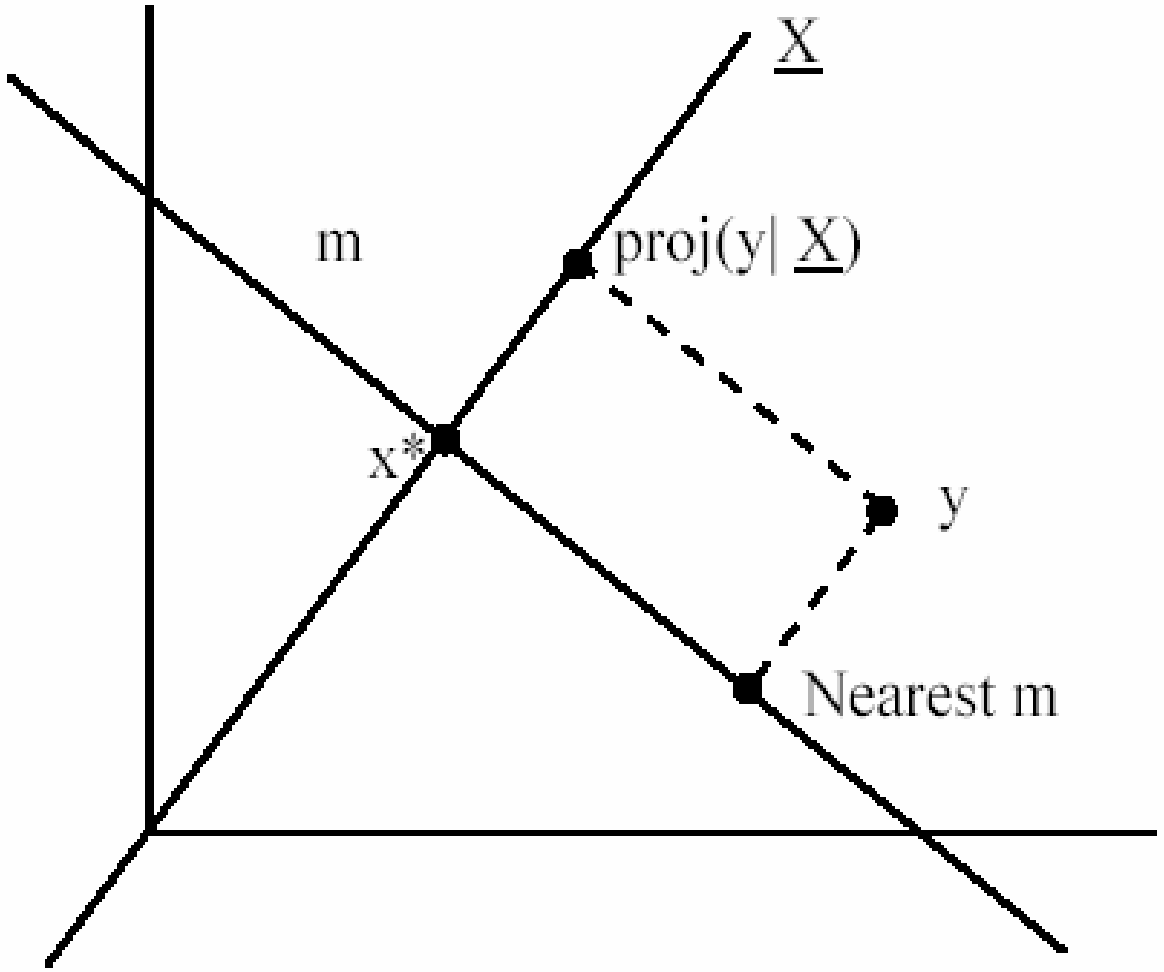


Figure 25. Distance between y and nearest m = distance between $proj(y|X)$ and x^*.

- Proof :The distance between y and the nearest valid m is the same as the distance between is the same as the distance between $proj(y \mid X)$ and x^*.

From the OLS formula,

$$proj(y \mid X) = E(yx')E(xx')^{-1}x$$

x^* is the portfolio of x that price x

$$x^* = p'E(xx')^{-1}x$$

Then, the distance between y and x^* is

$$\begin{aligned}
\left\| y - x^* \right\| &= \left\| proj(y \mid X) - x^* \right\| \\
&= \left\| E(yx')E(xx')^{-1}x - p'E(xx')^{-1}x \right\| \\
&= \left\| (E(yx') - p')E(xx')^{-1}x \right\| \\
&= [E(xy') - p]'E(xx')^{-1}[E(xy') - p] \\
&= g_T'E(xx')^{-1}g_T
\end{aligned}$$

Identity Matrix

- Using the identity matrix weights has a particular advantage with large systems in which S is nearly singular.
- It avoids most of the problems associated with inverting a near-singular S matrix.

Comparing the Second-Moment and Identity Matrices

- The second moment matrix and the optimal weighting matrix *S* give an objective that is invariant to the initial choice of assets or portfolios.

$$[E(yAx) - Ap]'E(Axx'A')^{-1}[E(yAx) - Ap]$$
$$= [E(yx) - p]'E(xx')^{-1}[E(yx) - p]$$

- It is not true of the identity or other fixed matrices. The results depend on the initial choice of portfolios.
- The second-moment matrix is often even more nearly singular than the spectral density matrix. It is no help on overcoming the near singularity of *S*.

Estimating on One Group of Moment, Testing on Another

- We can use one set of moment for estimate and another for testing
- We can also using one set of asset returns and then see how the model does “out of sample” on another set of asset
- We can do all this very simply by using an appropriate weighting matrix or a prespecified moment matrix a_T , for example

$$a_T = [I_N, 0_{N+M}]$$

11.7 Estimating the Spectral Density Matrix

- The optimal weighting matrix S depend on population moments, and depend on the parameters b.

$$S = \sum_{j=-\infty}^{\infty} E(u_t u'_{t-j}), u \equiv (m_t(b)x_t - p_{t-1})$$

- There are a lot of parameters.
- How do we estimate this matrix in practice?

Use a sensible first-stage W, or transform the data

- In the first-stage b estimates, we should use a sensible weighting matrix.
- Sometimes, some moments will have different unit than other moment.
- For example the moment formed by $R_{t+1} * \frac{d}{p_t}$ and the moment formed by $R_{t+1} * 1$.
- It is also useful to start with moments that are not horrendously correlated with each other.

- For example, you might consider R^a and $R^a - R^b$ rather than R^a and R^b .

$$W = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

Remove means

- Under the null, $E(u_t) = 0$ whether we estimate the covariance matrix by removing mean.

$$\frac{1}{T}\sum_{t=1}^{T}[(u_t - \bar{u})(u_t - \bar{u})'], \bar{u} = \frac{1}{T}\sum_{t=1}^{T} u_t$$

- Hansen and Singleton (1982) advocate removing the means in sample.
- But this mothed also make that estimate S matrices are often nearly singular. Since

$$E(uu') = \mathrm{cov}(u, u') + E(u)E(u')$$

- $E(u)E(u')$ is a singular matrix

Downweight highter-order correlations

- When we estimate S, we want to construct consistent estimates that are automatically positive definite in every sample.
- For example the Newey and West estimate, it is

$$\hat{S} = \sum_{j=-k}^{k} \left(\frac{k-|j|}{k}\right) \frac{1}{T} \sum_{t=1}^{T} (u_t u'_{t-j})$$

- The Newey-West estimator is base on the variance of kth sums. So it is positive definite

$$\mathrm{var}\left(\sum_{j=1}^{k} u_{t-j}\right) = kE(u_t u'_t) + (k-1)[E(u_t u'_{t-1}) + E(u_{t-1} u'_t)] + \cdots$$

$$+[E(u_t u'_{t-k}) + E(u_{t-k} u'_t)] = k \sum_{j=-k}^{k} \frac{k-|j|}{k} E(u_t u'_{t-j})$$

- What value of *k*, or how wide a window if of another shape, should you use?
- The rate at which *k* should increase with sample size, but not as quickly as the sample size increases.

Consider parametric structures for autocorrelation and heteroskedasticity

- GMM is not inherently tied to "nonparametric" covariance matrix estimates.
- We can impose a parametric structure on the S matrix.
- For example, if we model a scalar u as an AR(1) with parameter σ_u^2 and ρ ‹ then

$$S = \sum_{j=-\infty}^{\infty} E(u_t u_{t-j}) = \sigma_u^2 \sum_{j=-\infty}^{\infty} \rho^{|j|} = \sigma_u^2 \frac{1+\rho}{1-\rho}$$

- So we only need to estimate two parameter

Use the null limit correlations

- In the asset pricing setup, the null hypothesis specifies that $E_t(u_{t+1}) = E_t(m_{t+1}R_{t+1} - 1) = 0$ as well as $E(u_{t+1}) = 0$
- In this situation, you can get

$$\hat{S} = \frac{1}{T}\sum_{t=1}^{T} u_t u_t'.$$

- However, the null might not be correct, if the null is not correct, you have a inconsistent estimate.
- If the null is not correct ,estimating extra lags that should be zero under the null only loses a little bit of power.

- Monte Carlo evidence suggest that adding null hypothesis can help with the *power* of test statistics.
- Small-sample performance of the nonparametric estimators with many lags is not very good.
- We can test the autocorrelated of u_t to decide whether the model is right.
- If there is a lot of correlation, this is an indication that something is wrong with the estimate.

$$\hat{S} = \frac{1}{T}\sum_{t=1}^{T} u_t u_t'.$$

Size problems; consider a factor or other parametric cross-sectional structure

- when the number of moments is more than around 1/10 the number of data points, S estimates tend to become unstable and near-singular.
- It might be better to estimate an S imposing a factor structure on all the primitive assets.
- One might also use a highly structured estimate of S as weighting matrix, while using a less constrained estimate for the standard errors.

Alternatives to the two-stage procedure: iteration and one-step.

- ***Iterate***: we can use this formula

$$\hat{b}_2 = \min_{\{b\}} g_T'(b) S^{-1}(b_1) g_T(b)$$

Where b_1 is a first-stage estimate, held fixed in the minimization over b_2 , then use $\hat{b}_2$ to find $S(\hat{b}_2)$, find

$$\hat{b}_3 = \min_{\{b\}} [g_T(b)' S(\hat{b}_2)^{-1} g_T(b)$$

and so on. We can find this estimate serial converge to one value.

- This procedure is also likely to produce estimates that do not depend on the initial weighting matrix.
- ***Pick b and S simultaneously.***

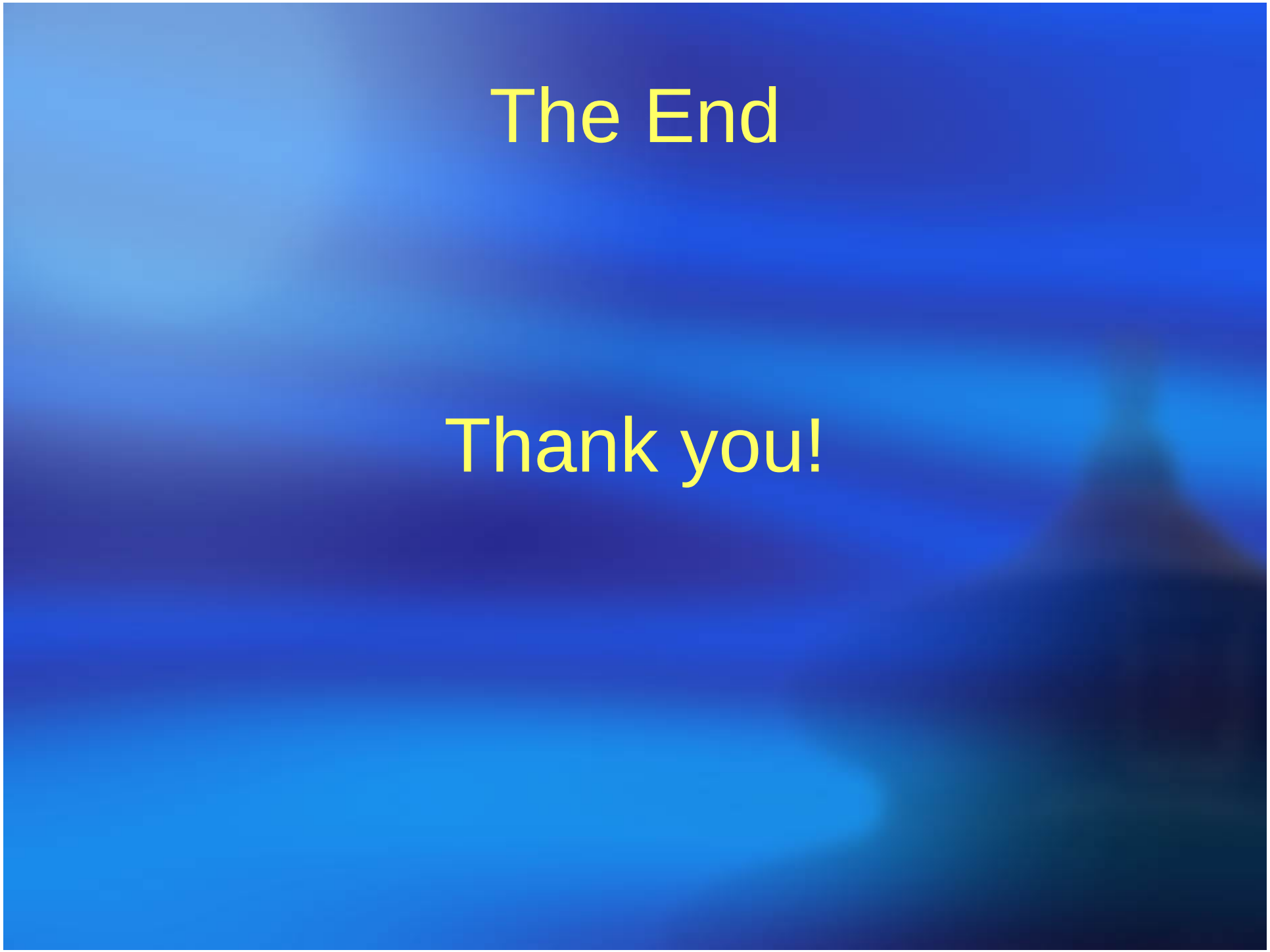
When search for b, the S also change. Then the object become into

$$\min_{\{b\}}[g_T(b)'S^{-1}(b)g_T(b)]$$

The first-order conditions are

$$2*(\frac{\partial g_T}{\partial b})'S^{-1}(b)g_T(b)+g_T(b)'\frac{\partial S^{-1}(b)}{\partial b}g_T(b)=0$$

- In the iteration method, each step involves a numerical search over $g_T(b)'Sg_T(b)$, may be much quicker to minimize once over $g_T(b)'S(b)g_T(b)$.
- On the other hand, the latter is not a locally quadratic form, so the search may run into greater numerical difficulties.


The End
Thank you!